COMPLEX NETWORKS 2018

THE 7TH INTERNATIONAL CONFERENCE ON COMPLEX NETWORKS AND THEIR APPLICATIONS

December 11 - 13, 2018
Cambridge, United Kingdom

Book Of Abstracts
COMPLEX NETWORKS 2018

The 7th International Conference on Complex Networks & Their Applications
December 11 - 13, 2018 Cambridge, United Kingdom
Published by the International Conference on Complex Networks & Their Applications.

Editors:
Hocine Cherifi
University of Burgundy, France

Luca Maria Aiello
Nokia-Bell Labs, UK

Chantal Cherifi
University of Lyon, France

Pietro Lio'
University of Cambridge, UK

Renaud Lambiotte
Oxford University, UK

Luis Mateus Rocha Indiana
University, USA

Copyright Notice COMPLEX NETWORKS 2018 and the Authors

This publication contributes to the Open Access movement by offering free access to its articles and permitting any users to read, download, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software. The copyright is shared by authors and the 7th International Conference on Complex networks & Their Applications (COMPLEX NETWORKS 2018) to control over the integrity of their work and the right to be properly acknowledged and cited.

To view a copy of this license, visit http://www.creativecommons.org/licenses/by/4.0/

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use. While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained her.

Preface

We are proud to present the Book of Abstracts for the 7th International Conference on Complex Networks & their Applications: COMPLEX NETWORKS 2018. Since 2012 the event has been held around the world on a yearly Basis. After Sorrento (Italy), Kyoto (Japan), Marrakech (Morocco), Bangkok (Thailand), Milan (Italy), Lyon (France), the seventh edition is hosted by the Department of Computer Science and Technology of the University of Cambridge from December 11 to December 13, 2018. The originality of the conference lies in the strongly interdisciplinary nature of the topics covered. Indeed, complexity and network science are multidisciplinary fields that mobilize intellectual resources in virtually all-scientific communities. Nowadays, all disciplines (physics, biology, social sciences, economics, computer science, meteorology, etc.) are faced with a massive influx of data and an explosion of information to manage. Through the data and their interactions, network science aims at understanding these complex systems increasingly large. COMPLEX NETWORKS is very focused at being an interdisciplinary event. However, this is linked with willingness to the requirements that the quality of the contributions must be among the best work in each of the scientific fields covered. In order to guarantee the excellence and reputation of this event, for its seventh edition COMPLEX NETWORKS has brought together in its scientific committee many leading international experts from all over the world. Year after year the event has increased its international influence. The 421 contributions that we received this year, from more than 50 countries around the world have been peer reviewed by at least 3 independent reviewers. This publication gathers the 163 extended abstracts accepted for presentation together with abstracts of eight keynote speeches and two invited tutorials.

Each edition of the conference represents a challenge that cannot be successfully achieved without the deep involvement of plenty of people, institutions and sponsors. We would like to thank all of them. We record our thanks to our fellow members of the Organizing committee for their huge efforts for the success of the conference. The program committee members for their engagement in promoting the event and refereeing submissions as well as the local committee members for their great commitment over the past months. We are also indebted to our sponsors, in particular Tribe Communication for designing the visual identity of the Conference. We are equally grateful to all the institutions that have helped us, in particular, the University of Cambridge for hosting this event. We also wish to express our appreciation to all participants and presenters. On a final note, we would like to express our deep sense of appreciation to our keynote and tutorial speakers.

Cambridge, United Kingdom,
December 2018

Luca Maria Aiello
Chantal Cherifi
Hocine Cherifi
Renaud Lambiotte
Pietro Liò
Luis Mateus Rocha

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
## Table of Contents

### Tutorials

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Epidemiology: From simple to data-driven models</td>
<td>2</td>
</tr>
<tr>
<td>Jesús Gómez-Gardeñes</td>
<td></td>
</tr>
<tr>
<td>From micro to macro: ego-network analysis and its applications</td>
<td>3</td>
</tr>
<tr>
<td>Silvio Lattanzi</td>
<td></td>
</tr>
</tbody>
</table>

### Invited Speakers

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximizing diversity in social networks</td>
<td>5</td>
</tr>
<tr>
<td>Aristides Gionis</td>
<td></td>
</tr>
<tr>
<td>Vulnerability of networked host populations to epidemics</td>
<td>6</td>
</tr>
<tr>
<td>Vittoria Colizza</td>
<td></td>
</tr>
<tr>
<td>Effects of Social Influence on Collective Motion</td>
<td>7</td>
</tr>
<tr>
<td>Romualdo Pastor-Satorras</td>
<td></td>
</tr>
<tr>
<td>“Topological data analysis for investigation of dynamics and biological networks</td>
<td>8</td>
</tr>
<tr>
<td>Heather Harrington</td>
<td></td>
</tr>
<tr>
<td>Essential nodes and keystone species in the brain, ecosystems and social systems</td>
<td>9</td>
</tr>
<tr>
<td>Hernan Makse</td>
<td></td>
</tr>
<tr>
<td>Modeling minorities in social networks</td>
<td>10</td>
</tr>
<tr>
<td>Markus Strohmaier</td>
<td></td>
</tr>
<tr>
<td>Motifs in Social Networks</td>
<td>11</td>
</tr>
<tr>
<td>Donald Towsley</td>
<td></td>
</tr>
<tr>
<td>Measuring Social Networks with High Resolution: What have we learned?</td>
<td>12</td>
</tr>
<tr>
<td>Sune Lehmann</td>
<td></td>
</tr>
</tbody>
</table>

### Biological Networks

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic identification of protein structure domains using Infomap and GLOSIM</td>
<td>14</td>
</tr>
<tr>
<td>William Grant and Sebastian Ahnert</td>
<td></td>
</tr>
<tr>
<td>Expected Number of Fixed Points in Boolean Networks with Arbitrary Topology</td>
<td>16</td>
</tr>
<tr>
<td>Fumito Mori</td>
<td></td>
</tr>
</tbody>
</table>
Cardiac Alternans: Understanding Subcellular Calcium Patterns with the Master Stability Function
Yi Ming Lai, Stephen Coombes and Ruediger Thul

Threshold-activated transport stabilizes chaotic populations to steady states
Chandrakala Meena, Pranay Deep Rungta and Sudeshna Sinha

Reconstructing Boolean Functions and Networks from Noisy Observational Data
Jie Sun and Erik Bollt

Gene-gene interaction module identification insingle-cell RNA sequencing
Simone Marini, Danila Vella, Nelson Nazzicari and Riccardo Bellazzi

Chromatin Assortativity approaches to study phenotypic variability
Vera Pancaldi

Identifying Dominant Connection Patterns in Protein Structure Networks
Nora S. Martin and Sebastian E. Ahnert

Highlighting the Complex Network Structure of Epigenetic Regulation using Message-Passing
Kimberly Glass

Topological variables of habitat networks as predictors of species occurrence
Damian Omar Ortiz-Rodríguez, Maarten J. van Strien, Antoine Guisan and Rolf Holderegger

Evaluating the leximin method for community detection
Arya McCarthy and David Matula

Community detection in networks with unobserved edges
Till Hoffmann, Leto Peel, Renaud Lambiotte and Nick Jones

Detectability of Macroscopic Structures in Directed Networks: a Stochastic Block Model Approach
Mateusz Wilinski, Piero Mazzarisi, Daniele Tantari and Fabrizio Lillo

Comparing Graph Clusterings: Set partition measures vs. Graph-aware measures
Valérie Poulin and François Théberge

A Short Introduction to Local Graph Clustering Methods and Software
Kimon Fountoulakis, David Gleich and Michael Mahoney

Community detection by modular decomposition of random walk
Hiroshi Okamoto and Xule Qiu
Community Structure Based on Circular Flow in a Large-Scale Transaction Network

Yuichi Kichikawa, Hiroshi Iyetomi, Takashi Iino and Hiroyasu Inoue

III Diffusion and Epidemics

Network localization is unalterable by infections in bursts

Qiang Liu and Piet Van Mieghem

The mixed assortativity property of finite Barabasi-Albert networks and its influence on diffusion times

Maria Letizia Bertotti and Giovanni Modanese

Disease Risk Assessment of the German Cattle Trade Network: A Static and Temporal Network Analysis

Jason Bassett, Inia Steinbach, H. K. Hartmut Lentz and Philipp Hoeval

Effective epidemic containment using link importance

Joan T. Matamalas, Sergio Gómez and Alex Arenas

Reframing the Gossip Problem as Information Spreading over Complex, Dynamic Networks

Daniel Goldsmith, Yonah Shmalo and Yitzchak Shmalo

ViralRank: A new metric for influencers identification

Flavio Iannelli, Manuel Sebastian Mariani and Igor M Sokolov

Modelling epidemic spreading in urban areas with large-scale agent-based transport simulations

Jürgen Hackl and Thibaut Dubernet

Maxmin-omega: A New Threshold Model on Networks

Ebrahim Patel

Network Reconstruction from NIMFA Viral State Observations of Multiple Epidemic Outbreaks

Bastian Prasse and Piet Van Mieghem

Epidemic spreading with awareness and different time scales in multiplex networks

Francisco Rodrigues, Yamir Moreno, Paulo Cesar Ventura, Colm Connaughton, Federico Vazquez and Fatima Velasquez-Rojas

Interplay of hubs and cooperation in social contagion

Yongjoo Baek, Kihong Chung, Daniel Kim, Hawoong Jeong and Meesoon Ha

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Spreading and influence of misinformation and traditional fact-based news in Twitter .......................................................... 100
Alexandre Bovet and Hernan Makse

The effects of trust and influence on the spreading of low and high quality information .................................................. 103
Diego Fregolente Mendes de Oliveira and Kevin S. Chan

Inferring properties of the underlying network based on the epidemic prevalence ......................................................... 106
Long Ma, Qiang Liu and Piet Van Mieghem

A three-time-scales random walk on temporal networks .................. 109
Julien Petit, Martin Gueuning, Timoteo Carletti, Ben Lauwens and Renaud Lambiotte

Impacts of mutual selection in temporal networks on random walk process ...... 112
Mei Yang, Bing Wang and Yuexing Han

Cascading Technology in Networks ........................................... 116
Côme Billard

Reachability analysis of discrete state epidemiological models ............ 119
Gergely Szlobodnyik

Cost-efficient sentinel surveillance strategies for preventing epidemics on networks ............................................................ 122
Ewan Colman, Petter Holme, Hiroki Sayama and Carlos Gershenson

An Extension of Autologistic Actor Attribute Models for Multi-Attribute Influence Processes ........................................... 125
Brennan Antone, Alina Lungeanu and Noshir Contractor

IV Dynamics on/of Networks

Synchronization invariance under network structural transformations ...... 129
Lluís Arola-Fernandez, Albert Diaz-Guilera and Alex Arenas

Communication and Coordination for DEEP Agents ....................... 131
Carlos Merlos

Preference in the restructuring mechanisms of time evolving hierarchies ...... 134
Gergely Palla, Dániel Zagyva, Sámuel G. Balogh and Péter Pollner

V Ecological Networks

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
High-order nested subsets and the structure of sites-species-species hypernetworks ............................................ 138
Albert Solé-Ribalta, Claudio Juan Tessone, Carlo Ferrari and Javier Borge-Holthofer

Trophic Structure of Directed Networks: Effects on Topology and Dynamics ............................... 141
Samuel Johnson

Extinction Curves and Error Propagation in Mutualistic Ecological Networks: The Case of Frugivores in a Tropical Rainforest ................. 144
Gitanjali Yadav and Suresh Babu

The structure and dynamics of non-normal networks ......................................................... 146
Malbor Asllani, Renaud Lambiotte and Timoteo Carletti

Inferring social relations from presence data. Manta Rays case study ......................... 149
Juan Fernandez-Gracia, Jorge P. Rodriguez, Lauren Peel, Konstantin Klemm, Mark Meekan and Victor M Eguiluz

VI Link Analysis and Ranking

A ‘dramatic truth’ in link prediction: SBM inference fails to effectively predict even the structure of synthetic networks generated with the SBM model ........................................... 153
Alessandro Muscoloni and Carlo Vittorio Cannistraci

A parametric approach to information filtering in complex networks: The Pólya filter ........................................ 156
Riccardo Marcaccioli and Giacomo Livan

Link Prediction via Indirect Interaction Duration ......................................................... 159
Laxmi Amulya Gundala and Francesca Spezzano

Long distance connections for ranking and robustness enhancement in networks ............. 162
Vincenza Carchiolo, Marco Grassia, Alessandro Longheu, Michele Malgeri and Giuseppe Mangioni

VII Machine Learning and Networks

An Exact No Free Lunch Theorem for Community Detection ................................. 166
Arya McCarthy

Graphlet Count Estimation via Convolutional Neural Networks ............................ 169
Xutong Liu, Yu-Zhen Janice Chen, John C.S. Lui and Konstantin Avrachenkov

Towards quantitative methods to assess network generative models .......................... 172
Vahid Mirzaebrahim Mostofi and Sadegh Aliakbary

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
VII

Network Embeddings For Graph Classification and Visualization ........................................ 175
Leonardo Gutierrez Gomez and Jean-Charles Delvenne

Node Embeddings in Dynamic Graphs .......................................................... 178
Ferenc Beres, Robert Palovics, Domokos Miklos Kelen, David Szabo and
Andras A. Benczur

Network properties captured by graph embeddings ........................................... 181
Rémi Vaudaine, Christine Largeron and Rémy Cazabet

Detecting Suspicious Entities in the Panama Papers ........................................ 184
Mikel Joaristi, Edoardo Serra and Francesca Spezzano

Temporal Alignment of Reddit Network Embeddings ....................................... 187
Siobhán Grayson and Derek Greene

Prediction of the disease controllability in a complex network using machine
learning algorithms ....................................................................................... 190
Richa Tripathi, Amit Reza and Dinesh Garg

Topological Data Analysis of Attribute Networks using Diffusion Frechet
Function with Ego-Networks ........................................................................ 194
Warren Keil and Mehmet Aktas

VIII Modeling Human Behavior

Learning strategic behavior in social and economic networks ....................... 198
Nicolo Pagan and Florian Dorfler

Exploring Information Dissemination Strategies to Foster Collective Intelligence .... 201
Dirk-Jan van Veen and Hans Rudolf Heinimann

Mobility helps problem-solving systems to avoid Groupthink ............................ 204
Paulo Freitas Gomes, Sandro Martinelli Reia, Francisco Aparecido
Rodrigues and José Fernando Fontanari

Quantifying the Role of Inactive Links in Social Networks ............................... 207
Andres M. Belaza and Jan Ryckebusch

Network-Based Mapping of Narrative Structures ....................................... 210
Juyong Park and Semi Min

Towards understanding the multi-level complexity of human health: from
drug interactions to human sexual behavior ............................................... 213
Rion Brattig Correia, Ian Wood and Luis M. Rocha

Characterization of Ego-network Circles in Mobile Phone Graphs .................. 215
Christian Quadri, Matteo Zignani and Sabrina Gaito and Gian Paolo Rossi

The 7th International Conference on Complex Networks and
Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fairness in multiplayer ultimatum games through degree-based role assignment  
Andreia Sofia Teixeira, Francisco C. Santos, Alexandre P Francisco and Fernando P. Santos  

DyNAM-i: a statistical model for the analysis of face-to-face interactions  
Marion Hoffman, Per Block, Timon Elmer and Christoph Stadtfeld  

Opinion spreading on complex networks with local and global interactions  
Arkadiusz Jędrzejewski  

Learning Quadratic Games on Networks  
Yan Leng, Xiaowen Dong and Alex Pentland  

IX  Multilayer Networks  

Multiplex decomposition of non-Markovian dynamics and the hidden layer  
reconstruction problem
Lucas Lacasa  

Majority-vote dynamics on multiplex networks
Kwang-Il Goh  

Structural Invulnerability Evaluation of Complex Multi-layer Emergency  
Logistics System Based on Interdependent Network Theory
Chen Chang, Yong Jie Niu and Guo Ji Kun  

Multiplex Complex Contagion Co-Diffusion
Ho-Chun Herbert Chang and Feng Fu  

Extracting Multi-Layer Networks from Sentinel-2 Satellite Image Time Series
Roberto Interdonato, Raffaele Gaetano, Danny Lo Seen and Mathieu Roche  

The Development of the Organic Dairy Supply Chain in the U.S.A  
Juan Carlos S. Herrera and Carolyn Dimitri  

Coordination in Open-Source Software Engineering - A Multiplex Network  
Analysis
Angelika Schmid, Thomas Bock, Janet Siegmund and Sven Apel  

Extensivity in infinitely large multiplex networks
Maria Angelica Araujo and Murilo S. Baptista  

X  Network Analysis  

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Multiscale core-periphery structure in a global liner shipping network

Sadamori Kojaku, Mengqiao Xu, Haoxiong Xia and Naoki Masuda

Characterising vorticity interactions during the intermittency route to thermoacoustic instability

Abin Krishnan and Raman Sujith

Unpacking maritime trafficking from the complex network approach

Yi-Chun Chang, Wei Weng, Seng-Cho Chou, Yen-Sheng Chiang and Chia-Te Chiang

Time series analysis in earthquake complex networks

Denisse Pasten, Zbigniew Czechowski and Benjamin Toledo

Assortative mixing in weighted directed networks

Uta Pigorsch and Marc Sabek

Network property and learning model of weblike equation system in a scientific category

Tetsuya Kojima, Jun Ishida, Shigeyuki Miyagi and Osamu Sakai

XI Network Models

Random Network Models Based on Density

Theresa Migler-Vondollen, Glencora Borradaile and Gordon Wilfong

A Bayesian approach to reference Exponential Random Graph Models (ERGMs) for comparing networks of different size and composition: Models and empirical tests

Alessandro Lomi, Alberto Caimo and Francesca Pallotti

A geometrically equivalent transformation between time series and complex networks

Xiaoyi Peng and Yi Zhao

A Monte Carlo method for Large Deviations applied to Erdos-Rényi random graphs

Cristian Giardinà, Claudio Giberti and Elena Magnanini

Network construction: A learning framework through localizing principal eigenvector

Priodyuti Pradhan and Sarika Jalan

Local-community network automata modelling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more

Alessandro Muscoloni and Carlo Vittorio Cannistraci

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Minimum curvilinear automata with similarity attachment for network embedding and link prediction in the hyperbolic space .......................... 291
Alessandro Muscoloni and Carlo Vittorio Cannistraci

Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction .......... 294
Alessandro Muscoloni and Carlo Vittorio Cannistraci

The microstructure of the giant component in configuration model networks . . 297
Ofer Biham, Eytan Katzav and Reimer Kuhn

Reconstructing networks with unknown and heterogeneous errors ............. 300
Tiago Peixoto

Modified box dimension of trees and hierarchical scale-free graphs ............ 303
Júlia Komjáthy, Roland Molontay and Károly Simon

A generalized class of thresholding functions for the hidden variable model of scale-free networks ................................................................. 307
Gáspár Sámal Balogh, Péter Pollner and Gergely Palla

Characterizing Probabilistic Networks ...................................................... 310
Amin Kaveh, Matteo Magnani and Christian Rohner

XII Network Neuroscience

Internal connectivity and topological roles of nodes in functional brain networks 314
Onerva Korhonen, Elisa Ryyppö and Jari Saramäki

Can MRI brain networks reproducibly distinguish psychosis patients from control subjects? ................................................................. 317
Sarah Morgan, Jonathan Young, Ameera Patel, Kirstie Whitaker, Cristina Scarpazza, Therese van Amelsvoort, Machtedl Marcelis, Jim van Os, Gary Donohoe, David Mothersill, Aiden Corvin, Martijn van den Heuvel, Edward Bullmore and Michael Brammer

Effects of spatial smoothing on group-level differences in functional brain networks ................................................................. 320
Ana Triana, Enrico Glerean, Jari Saramäki and Onerva Korhonen

Identifying the Coupling Structure in Complex Systems through the Optimal Causation Entropy Principle, Information Flow and Information Fragility ...... 323
Erik Bollt and Jie Sun

Robust Navigable Cores in the Human Brain Networks .......................... 324
Zalan Heszberger, Andras Majdan, Andras Gulyas, Tibor Gyimothy, Vilmos Bilicki and Jozsef Biro

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Communicalibity indicates structural and functional connectivity changes in children with autism spectrum disorder.           
Roma Siugzdaitė, Hannelore Aerts and Daniele Marinazzo

XIII  Networks in Finance and Economics

Transition from plasticity to instability in the structure of economic networks.     
Celian Colon and Jean-Philippe Bouchaud

Optimising Systemic Risk in Interbank Networks.         
Christian Diem, Anton Pichler and Stefan Thurner

Stability in Core-Periphery Production Networks.    
Manish Sarkhel and Nagarajan Krishnamurthy

Hierarchical Identification of Key Firms of International Tax Avoidance in Global Ownership Network. 
Tembo Nakamoto, Abhijit Chakraborty and Yuichi Ikeda

Hierarchical structure of the cryptocurrency market.   
Uta Pigorsch and Marc Sabek

Point-process network of firms bankruptcies.          
Yoshi Fujiwara and Shinsuke Koyama

Anomaly detection in networks using spectral methods and network comparison approaches. 
Andrew Elliott, Paul Reidy, Milton Martinez Luaces, Mihai Cucuringu and Gesine Reinert

Triadic and quadratic closure as mechanisms for inter-firm transactional network dynamics.  
Tomomi Kito and Petter Holme

Effects of industrial and geographical proximities of firms on the Japanese inter-firm network.  
Akihiro Hamamoto and Tomomi Kito

Struggle for existence in the world economic ecosystem.  
Viviana Viña Cervantes, Michele Coscia and Renaud Lambiotte

Influence of network correlation structure in multivariate time series forecasting. 
Luis Ospina-Forero, Mihai Cucuringu and Gesine Reinert

Finding the most central countries in the international exchange multiplex using the multirank.    
Giovanni Bonaccorsi, Massimo Riccaboni, Giorgio Fagiolo and Gianluca Santoni
How far does a shock event spreads on a network? Detecting causality on the salmon trade network 
Juan Carlos Rocha Gordo and Jessica Gephart

XIV Quantifying Success

Using Network Science to Discover the Grand Masters of the Florentine Renaissance
Cameron DeChurch and Noshir Contractor

Lifelong Career Success: Joseon Dynasty’s Bureaucrats
Donghyeok Choi and Juyong Park

Reciprocity and success in academic careers
Weihua Li, Tomaso Aste, Fabio Caccioli and Giacomo Livan

Ties and Event Flow Network in Basketball
Gyuhyeon Jeon and Juyong Park

XV Resilience and Control

Trade-off between time and energy cost for controlling complex networks
Yongzheng Sun, Siyang Leng, Yingcheng Lai, Celso Grebogi and Wei Lin

Chimera states in complex networks: from nonlocal to fractal modular and multilayer topologies
Iryna Omelchenko

Robustness of hierarchical network organization in models of functional connectivity
Ali Safari and Paolo Moretti

Resilience of Hierarchical Networks and Interdependent Hierarchical Networks
Louis Shekhtman and Shlomo Havlin

Temporal network-based analysis of turbulent mixing
Giovanni Iacobello, Stefania Scarsoglio, Hans Kuerten and Luca Ridolfi

Using non-normality for control energy reduction in network controllability problems
Gustav Lindmark and Claudio Altafini

Alternative routing maps with smaller search information
Hengda Yin, Raul Mondragon and Richard Clegg

Colour-dependant percolation in complex networks
Ivan Kryven

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Rigidity percolation in random rod packings ................................................. 409
Samuel Heroy, Dane Taylor, Feng Shi, Greg Forest and Peter Mucha

Redundancy in the Structure and Dynamics of Complex Networks ............. 412
Rion Brattig Correia, Alexander Gates, Nathan D. Ratkiewicz, Alain Barrat and Luis M. Rocha

Percolation transition on scale-free networks with assortative degree correlation .... 414
Toshihiro Tanizawa

Heterogeneity in Percolation on Complex Networks ................................. 417
Reimer Kuehn and Tim Rogers

Dynamic competitive opinion control: theory, simulations, and experiments ...... 420
Valerio Restocchi, Markus Brede, Sebastian Stein, Lewis Hill and Soheil Eshghi

Pairwise Approximations of Non-Markovian Network Epidemics ................ 423
Gergely Röst

Co-occurrence plasticity increases modularity and stability in bipartite networks 426
Somaye Sheykhi, Juan Fernandez-Gracia and Victor M Eguiluz

Optimal attacks on random graphs ......................................................... 429
Nicole Balashov, Reuven Cohen, Simi Haber and Michael Krivelevich

Controlling synchronization in bow-tie architectures ............................... 432
Edgar Wright, Sooyeon Yoon, José Mendes and Alexander Goltsev

European hinterland container transport as a complex network - How is robustness affected by the multi-mode structure? ......................... 435
Camill Harter, Rob Zuidwijk and Otto Koppius

Precursor Theory of Adaptive Networks ............................................... 439
Leonhard Horstmeyer, Christian Kuehn, Jan Korbel, Tuan Minh Pham and Stefan Thurner

Social Networks

Social structure in towns and corruption risk at city hall .......................... 443
Johannes Wachs, Balazs Lengyel, Taha Yasseri and Janos Kertesz

The Contagion Effects of Repeated Activation in Social Networks .............. 446
Pablo Piedrahita, Javier Borge-Holthoefer, Yamir Moreno and Sandra Gonzalez-Bailon

Homophily explains perception biases in social networks ......................... 449
Fariba Karimi
Recruiting Mechanisms on Online Labour Markets: Agent-Based Models of Bipartite Social Networks ................................................. Martin Lukac

On the clustered structure of a decentralized social network and its dynamics . . Matteo Zignani, Christian Quadri and Sabrina Gaito and Gian Paolo Rossi

Temporal evolution of mentions and retweets networks in the Spanish General Elections .......................................................... Samuel Martin-Gutierrez, Juan Carlos Losada and Rosa M. Benito

The homophily shows the Dunbar’s number: modeling and data .................. Janos Torok, Yohsuke Murase, Hang-Hyun Jo, Janos Kertesz and Kimmo Kaski

Topological study of the convergence in the Voter Model ......................... Ines Caridi and Pablo Balenzuela

XVII Structural Network Measures

Dimension as an Invariant of Street Networks ................................Franz-Benjamin Mocnik

Interpreting and comparing the length of complex networks .................. Gorka Zamora-Lopez and Romain Brasselet

Coexisting Mesoscale Patterns in Complex Networks .......................... Maria Palazzi, Javier Borge-Holthoefer, Claudio Juan Tessone and Albert Sol-Ribalta

Mixing patterns and individual differences in networks ....................... George T. Cantwell and M. E. J. Newman

Centrality measures for graphons: Accounting for uncertainty in networks . Michael Schaub, Marco Avella-Medina, Francesca Parise and Santiago Segarra

The distribution of shortest path lengths in random networks .................. Eyuan Katzav, Ofer Biham and Reimer Kuh

Logistic core-periphery detection in networks .................................... Francesco Tudisco and Desmond Higham

Network Constraints on the Mixing Patterns of Binary Node Metadata ...... Matteo Cinelli, Leto Peel, Antonio Iovanella and Jean-Charles Delvenne
# XVIII  Urban Systems and Networks

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insights from self-organizing maps for characterizing accessibility to healthcare networks</td>
<td>493</td>
</tr>
<tr>
<td>Jerome Mayaud, Sam Anderson, Martino Tran and Valentina Radic</td>
<td></td>
</tr>
<tr>
<td>Simple Spatial Scaling Rules behind Complex Cities</td>
<td>496</td>
</tr>
<tr>
<td>Ruiqi Li and Gene Stanley</td>
<td></td>
</tr>
<tr>
<td>On the Statistics of Urban Street Networks</td>
<td>498</td>
</tr>
<tr>
<td>Jerome Benoit and Saif Eddin Jabari</td>
<td></td>
</tr>
<tr>
<td>Decentralized Optimal Route Navigation</td>
<td>501</td>
</tr>
<tr>
<td>Trivik Verma, Marc Folini, Yi Wang and Nuno Araujo</td>
<td></td>
</tr>
<tr>
<td>Application of Complex Network Analysis for Urban Drainage Systems</td>
<td>505</td>
</tr>
<tr>
<td>Julian David Reyes-Silva, Jonatan Zischg, Robert Sitzenfrei and Peter Krebs</td>
<td></td>
</tr>
<tr>
<td>Mapping the discourse on Smart Cities by means of NLP (Natural Language Processing) and complex network analysis</td>
<td>509</td>
</tr>
<tr>
<td>Efrat Blumenfeld-Lieberthal, Nimrod Serok, Orr Levy, Elya Lucy Milner and Shlomo Havlin</td>
<td></td>
</tr>
<tr>
<td>Decongestion of urban areas with hotspot-pricing</td>
<td>512</td>
</tr>
<tr>
<td>Albert Sole, Sergio Gómez and Alex Arenas</td>
<td></td>
</tr>
</tbody>
</table>
Tutorials
Network Epidemiology: From simple to data-driven models

Jesús Gómez-Gardeñes

University of Zaragoza, Spain

In this tutorial we will deal with a topic that has advanced enormously in the recent decades thanks to contribution of network science: the modeling of epidemics. We will begin by reviewing the building blocks of the broad field of theoretical epidemiology: the compartmental models. From this point, we will progressively add ingredients aimed at capturing the real patterns of connectivity (networks) and mobility (metapopulations) observed in real societies. Finally, after analyzing the behavior of these models from the theoretical point of view, we will address the current challenges of epidemics prediction and the design of containment strategies.

Jesus Gomez-Gardeñes is Associate Professor and head of the Group of Theoretical and Applied Modeling (GOTHAM) at the Institute of Biocomputation and Physics of Complex Systems (BIFI) of the University of Zaragoza (Spain). His main fields of research are statistical physics, nonlinear dynamics and the theory of complex networks. Within these disciplines he has mainly focused in the study of the emergence of collective phenomena out of nonlinearity and the structure of interactions in complex systems. Along these lines he has studied some paradigmatic problems such as energy localization, synchronization, random walks, traffic congestion, disease propagation and evolutionary dynamics. He has authored more than 100 scientific articles in international journals, including Nature Physics, PNAS, Physical Review Letters, Physics Reports, Science Advances, Nature Human Behavior among others. In the recent years he has focused on the study of multilayer networks and network epidemiology.
From micro to macro: ego-network analysis and its applications

Silvio Lattanzi
Google Research Europe, Switzerland

Detecting the clustering structure of real-world networks has emerged as an important primitive in a wide range of data analysis tasks such as community detection, event detection, spam detection, computational biology, link prediction and many others. As a result, the study of the topology of real world networks and of their clustering (or community) structure is central in modern network analysis. In particular, in recent years, several models have been introduced to capture the community structure of social networks and numerous empirical studies analyzed the community structures at a macroscopic and microscopic levels. One of the main observations in this line of work is the lack of a clear macroscopic community structure in real world networks. In sharp contrast with these findings, it has been observed that while the community detection problem is hard at a macroscopic level, it becomes simple at a microscopic level. This is especially true when we restrict our attention to local structures know as ego-nets (a.k.a. ego-networks) which consist of the subgraph induced over the neighborhood of a single node in the graph. Intuitively, this happens because, even if a node is part of many communities, if we restrict our attention to a node and one of her neighbors, there is only one or a limited number of communities in which the two nodes interact, which present a clearer structure at the level of the neighborhood. In this talk we will first present this phenomenon then we will discuss how to analyze ego-networks at scale and finally describe few applications of ego-network analysis.

Silvio Lattanzi is a Research Scientist at Google Research Europe since April 2017. Before he was in the NY Algorithm group at Google New York from January 2011 to March 2017. He received my PhD from Sapienza University of Rome under the supervision of Alessandro Panconesi. During his PhD he interned twice at Google and once at Yahoo! Research. His research interests are in the areas of algorithms, machine learning and information retrieval.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Invited Speakers
Maximizing diversity in social networks

Aristides Gionis
Aalto University, Finland

Online social media are a major venue of public discourse today, hosting the opinions of hundreds of millions of individuals. Social media are often credited for providing a technological means to break information barriers and promote diversity and democracy. In practice, however, the opposite effect is often observed: users tend to favor content that agrees with their existing world-view, get less exposure to conflicting viewpoints, and eventually create "echo chambers" and increased polarization. Arguably, without any kind of moderation, current social-media platforms gravitate towards a state in which net-citizens are constantly reinforcing their existing opinions. In this talk we present our ongoing line of work on analyzing and moderating online social discussions. We first consider the questions of detecting controversy using network structure and content. We then address the problem of designing algorithms to break filter bubbles, reduce polarization, and increase diversity. We discuss a number of different strategies such as user and content recommendation, as well as approaches based on information cascades.

Aristides Gionis is a professor in the department of Computer Science in Aalto University. His previous appointments include being a visiting professor in the University of Rome and a senior research scientist in Yahoo! Research. He is currently serving as an action editor in the Data Management and Knowledge Discovery journal (DMKD), an associate editor in the ACM Transactions on Knowledge Discovery from Data (TKDD), and an associate editor in the ACM Transactions on the Web (TWEB). He has contributed in several areas of data science, such as algorithmic data analysis, web mining, social-media analysis, data clustering, and privacy-preserving data mining. His current research is funded by the Academy of Finland (projects Nestor, Agra, AIDA) and the European Commission (project SoBigData).

Elsevier Online Social Networks and Media Lecture
Vulnerability of networked host populations to epidemics

Vittoria Colizza
INSERM, France

Our understanding of communicable diseases prevention and control is rooted in the theory of host population transmission dynamics. The network of host-to-host contacts along which transmission can occur drives the epidemiology of communicable diseases, determining how quickly they spread and who gets infected. A large body of epidemiological, mathematical and computational studies has provided a number of insights into the understanding of the process and the identification of efficient control strategies. The explosion of time resolved contact data has however opened the stage to new challenges. What are the structural and temporal aspects, and possibly their non-trivial interplay, that are critical for disease spread? To answer this question, I will introduce the infection propagator approach, a theoretical analytical framework for the assessment of the degree of vulnerability of a host population to disease epidemics, once we account for the time variation of its contact pattern. By reinterpreting the tensor formalism of multilayer networks, this approach allows the analytical computation of the epidemic threshold for an arbitrary time-varying network of host contacts, i.e. the critical pathogen transmissibility above which large-scale propagation occurs. I will apply this framework to a set of empirical time-varying contact networks and show how it can be used to test different intervention strategies for infection prevention and control in realistic settings.

Dr. Vittoria Colizza is Director of Research at Inserm (French National Institute of Health and Medical Research) & Sorbonne University, Faculté de Médecine, Paris, France. She leads the EPIcx lab within the Pierre Louis Institute of Epidemiology and Public Health. Her work focuses on real episodes of human and animal epidemics to gather context epidemic awareness and provide risk assessment analyses for preparedness, mitigation, and control. Her research also includes more theoretical approaches for the modeling of small- to large-scale diffusion events where contacts between hosts and their mobility are explicitly considered from data. Colizza received several awards, including the Erdős–Rényi Prize by the Network Science Society in 2017. She also served as Young Advisor to the Vice President of the European Commission Mrs. Neelie Kroes for the Digital Agenda for Europe.

Springer Applied Network Science Lecture
Effects of Social Influence on Collective Motion

Romualdo Pastor-Satorras
Universitat Politècnica de Catalunya, Spain

Collective motion in animals is able to produce such stunning patterns as flocks of birds turning in unison or shoals of fish splitting and reforming while outmaneuvering a predator. The study of these phenomena is mainly based in simple models, a paradigmatic example being the one proposed by Vicsek and collaborators in the 90s. The main assumption of this and similar models is that individuals tend to orient their velocity parallel to the average velocity of their local neighborhood. The Vicsek model predicts a phase transition between an ordered phase, with individuals travelling in a common direction, and a disordered one, that has been recently the subject of a large interest in the statistical mechanics community. Here we will consider variations of the Vicsek model incorporating social interactions between individuals, represented in terms of a complex social network. The main result of the numerical study of this model is the observation that the heterogeneity of the social network can increase the resilience of the ordered state, making it immune to external perturbations. A related scalar version of the Vicsek model in networks allows for a mathematical treatment that lends support to the numerical observations, and allows for further generalizations.

Romualdo Pastor-Satorras (Barcelona, Spain, 1967) received a PhD in Condensed Matter Physics from the Universitat de Barcelona in 1995. He spent four years as postdoctoral researcher at the Massachusetts Institute of Technology (1996-1998) and The Abdus Salam International Centre for Theoretical Physics, ICTP (1998-2000). At present, he is Associate Professor of Applied Physics at the Universitat Politècnica de Catalunya since 2006. He has been visiting scientist at, among others, Yale University (USA), the University of Notre Dame (USA), the Kavli Institute for Theoretical Physics (USA), the Helsinki University of Technology TKK (Finland), Indiana University (USA) and the Institute for Scientific Interchange (ISI) Foundation (Italy). He has been awarded twice with the national “ICREA Academia Prize” by the Government of Catalonia. He has published more than 100 publications in peer-reviewed journals in the field of statistical physics, and is author of the book “Evolution and Structure of the Internet” (Cambridge University Press, 2007), together with Professor Alessandro Vespignani.
‘Topological data analysis for investigation of dynamics and biological networks

Heather Harrington
Oxford University, UK

Persistent homology (PH) is a technique in topological data analysis that allows one to examine features in data across multiple scales in a robust and mathematically principled manner, and it is being applied to an increasingly diverse set of applications. We investigate applications of PH to dynamic biological networks with concrete examples from contagions, neuroscience, and blood vessels.

Prof. Heather Harrington is a Royal Society University Research Fellow and Associate Professor in the Mathematical Institute at the University of Oxford. She is Co-Director of the Centre for Topological Data Analysis. Her research focuses on the problem of reconciling models and data by extracting information about the structure of models and the shape of data. To develop these methods, Prof Harrington integrates techniques from a variety of disciplines such as computational algebraic geometry and topology, statistics, optimisation, network theory, linear algebra, and dynamical systems. Based on this research, she was recently awarded a London Mathematical Society Whitehead Prize.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Essential nodes and keystone species in the brain, ecosystems and social systems

Hernan Makse
City College of New York, USA

Identifying essential nodes in complex networks is a central problem for biological systems to social systems. We treat this problem in three paradigmatic cases: the brain, ecosystems and social networks. Mathematically, we find the set of influential nodes by optimizing the damage to the giant connected component with systematic inactivation of nodes. We then apply network theory and pharmacogenetic interventions in a rat brain to predict and target essential nodes responsible for global integration in a model of learning and memory. We find that the integration of the brain network is mediated by a set of weak nodes through optimization of influence in optimal percolation. Pharmacogenetic inhibitions confirm the theoretical predictions. We discuss the relevance of these influencers to ecological systems dominated by abrupt first order tipping points as well as connectomes with regularities.

Prof. Hernan Makse leads the Complex Networks and Data Science Lab of at the Levich Institute and Department of Physics of City College of New York in New York City. Hernan research focuses on the theoretical understanding of Complex Systems from a Statistical Physics viewpoint. He is working towards the development of new emergent laws for complex systems, ranging from brain networks to biological networks and social systems. Treating these complex systems from a unified theoretical approach, he uses concepts from statistical mechanics, network and optimization theory, machine learning, and big-data science to advance new views on complex systems and networks.

PLOS Lecture
Modeling minorities in social networks

Markus Strohmaier
RWTH Aachen University, Germany

Homophily can put minority groups in social networks at a disadvantage by restricting their ability to establish links with people from a majority group. This can limit the overall visibility of minorities in the network, and create biases. In this talk, I will show how the visibility of minority groups in social networks is a function of (i) their relative group size and (ii) the presence or absence of homophilic behavior. In addition, the results show that perception biases can emerge in social networks with high homophily or high heterophily and unequal group sizes, and that these effects are highly related to the asymmetric nature of homophily in networks. This work presents a foundation for assessing the visibility of minority groups and corresponding perception biases in social networks in which homophilic or heterophilic behaviour is present.

Markus Strohmaier is the Professor for Methods and Theories of Computational Social Sciences and Humanities at RWTH Aachen University (Germany), and the Scientific Coordinator for Digital Behavioral Data at GESIS - Leibniz Institute for the Social Sciences. Previously, he was a Post-Doc at the University of Toronto (Canada), an Assistant Professor at Graz University of Technology (Austria), a visiting scientist at (XEROX) Parc (USA), a Visiting Assistant Professor at Stanford University (USA) and the founder and scientific director of the department for Computational Social Science at GESIS (Germany). He is interested in applying and developing computational techniques to research challenges on the intersection between computer science and the social sciences / humanities.
Motifs in Social Networks

Donald Towsley
UMass Amherst, USA

Complex networks that occur in nature and engineering, often exhibit simple, network structural properties, or “motifs.” Network motifs refer to recurring, significant patterns of interaction between sets of nodes and represent basic building blocks of graphs. Motifs in social networks exhibit spatial patterns and temporal patterns that vary according to the type of network. This talk reports on these variations across several network types and identify several common substructures. Reciprocity of directed ties occurs much more frequently than expected by chance in all networks. Similarly, we find that completely connected triads and tetrads (i.e., four-node sub-graphs) occur more often than expected, highlighting the tendency of actors to form clusters of ties. We also identify motifs that suggest patterns of hierarchy. Motifs are also useful for the purpose of sub-graph classification. We demonstrate their value in identifying the type of network that a sub-graph belongs to. We also consider the challenge of characterizing motifs in large graphs, and show how carefully designed sampling algorithms can accurately characterize them using a small number of samples. Last, we close with open problems regarding motifs whose solution can lead to better understanding of social networks and analytical tools for characterizing them.

Professor Towsley’s research spans a wide range of activities from stochastic analyses of queueing models of computer and telecommunications to the design and conduct of measurement studies. He has performed some of the pioneering work on the exact and approximate analyses of parallel/distributed applications and architectures. More recently, he pioneered the area of network tomography and the use of fluid models for large networks. He has published extensively, with over 150 articles in leading journals. Towsley is the recipient of one of the IEEE’s most prestigious honors, the 2007 IEEE Koji Kobayashi Computers and Communications Award.

MDPI Future Internet Lecture

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Measuring Social Networks with High Resolution: What have we learned?

Sune Lehmann

Technical University of Denmark, Denmark

In other to understand the multi-layered and dynamic social interactions within a large social system, I equipped 1000 freshmen students at the Technical University of Denmark with top-of-the-line smartphones running custom software designed to collect interactions mediated through face-to-face meetings (proximity estimated via Bluetooth), telecommunication (phone-calls, text messages), and online social networks (Facebook friendships and interactions). The phones also collected geo-locations, wifi-signals, and a number of other data channels; participants also answered paneled questionnaires regarding personality, study habits, and health-related behavior. The data collection lasted 2.5 years. Through this rich dataset, we have learned about much more than social networks. In my talk, I will discuss key findings from this study, with an emphasis on communities in dynamic networks and recent results on human mobility.

Sune Lehmann is an associate professor at the Technical University of Denmark, an adjunct (full) professor at University of Copenhagen’s Department of Sociology, and an adjunct associate professor at the Niels Bohr Institute (Department of Physics, University of Copenhagen). Sune is the associate director of the Center for Social Data Science at University of Copenhagen. Sune’s work focuses on the dynamics of complex networks as well as processes unfolding on such evolving networks. He is the author of multiple highly cited papers and his work has received world-wide press coverage.
Part I

Biological Networks
Systematic identification of protein structure domains using Infomap and GLOSIM

William P. Grant1 and Sebastian E. Ahnert1,2

1 Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, UK.
wpg23@cam.ac.uk
2 Sainsbury Laboratory, University of Cambridge, UK.

Protein structure domains are compact, globular subunits which are represented across multiple proteins. Often these domains encode a particular function, and identifying domains may therefore provide insight into the underlying processes mapping structure to function. The process of extracting these domains in an automated fashion remains computationally challenging. One possible definition of a protein domain is that of a community on a spatial network generated from the protein’s structure. In this work, the feasibility and utility of this definition is explored.

Following previous work [4], experimentally determined protein structures are abstracted into networks in which the amino acids are the vertices, and edges are generated according to spatial proximity. The Infomap community detection algorithm [5] is then used, and the communities mapped back onto the protein structure. In this way, proteins can be systematically decomposed into globular, spatially compact subregions.

This modular decomposition may have applications in molecular dynamics, in de novo protein design, and in normal mode analysis for fitting electron microscopy maps. However, here we investigate the properties of the network-derived protein modules themselves, rather than as part of the whole protein. By calculating the structural similarity of the modules, we can identify modules which are repeated in multiple proteins.

An initial study was performed using 40 multi-domain proteins, all of which share a common SCOPe domain [3], but with differing structure outside this region. The proteins were fragmented as described above, and their structural similarities compared using the GLOSIM kernel [1]. This method, which has been successfully used to compare protein ligands and crystal structures, uses the SOAP atomic descriptors [2] to characterise the atoms of the molecule using the positions of their neighbours. This is then used to compare the molecules globally, resulting in a score ∈ [0, 1].

Using hierarchical clustering, we can group the communities/modules by structural similarity (Fig. 1). By choosing a cutoff threshold on the dendrogram, we can extract modules that occur across multiple proteins; these are then candidate protein domains.

By varying the level of community structure, and the threshold for clustering, the structure can be compared on multiple length scales. By comparing the overlap between these network-based domains and existing protein structure classifications, we can identify novel domains, which may shed insight on the organisational and evolutionary underpinnings of protein structure.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. The matrix of GLOSIM similarity scores [1] for a set of structural domains found using community detection, clustered by similarity. Here a score of 1 (coloured white) indicates a perfect structural match between the two modules. The modules are indexed using a "PDBReference ChainReference ModuleID" notation; not all rows are labelled for space reasons. The modules were generated using the Infomap community detection algorithm [5] on residue contact networks. The networks were created from a set of 40 multi-domain proteins (as characterised by SCOPe [3]).

References
Abstract. Dynamics on complex networks are generally dependent on the network topologies. Thus, the relationships between dynamical characteristics and topological properties of network systems have attracted a considerable amount of research interest. Boolean networks, which serve as models for social and biological network systems, still have the potential to provide novel analytical methods to express such relationships due to their simplicity. An important quantity of a Boolean network is the number of fixed points, which are typically considered to correspond to cell types in differentiation process, and solutions without frustrations in opinion formation process. We have investigated the expected number of fixed points in a Boolean network, with Boolean functions drawn from probability distributions that are not required to be uniform or identical. We have proven that it is one, and is independent of its network topology if only a feedback arc set satisfies a stochastic neutrality condition. Therefore, controlling only network topology does not contribute to increasing the expected number of fixed points. It is increased by the predominance of positive feedback in a cycle. Our theory does not require calculating the actual number of fixed points in a single realization. Thus, it can be applied to systems with any network topology and size.

1 Introduction

A state of a gene, a neuron, and sometimes a person, is updated depending on the states of the surrounding elements. For instance, a gene expression level is changed by other genes. A person may determine whether he or she runs in an election depending on whether other persons are going to run or not. Such update rules can be expressed by Boolean functions. Thus, many social and biological phenomena, such as in gene regulatory networks, neural networks, and opinion formation networks, can be described by Boolean networks (BNs). Even though five decades have elapsed since their development, challenges to establish analytical methods to describe characteristics of BNs have continued. It has been proved that in a randomly constructed BN, the average number of attractors grows faster than any power law with network size [1, 2]. Because real systems are not random, interest in BNs with special topologies, including small-world and scale-free networks, has increased.

One of the most fundamental characteristics of a BN is the number of fixed points. In cell differentiation process, a gene expression pattern shown in a gene regulatory network determines a cell type. This expression pattern is assumed to be a fixed point.
It has been pointed out that real gene regulatory networks should have special structure [3] because average number of fixed points in a randomly constructed BN is only one. In opinion formation process, no one wants to change their opinion at a fixed point, i.e., the fixed point is a solution without frustration. Therefore, it is important whether a fixed point exists or not, and how many fixed points exist in a BN.

The upper bound of the number of fixed points in a BN under the restriction of a (signed) network topology is given by the minimum (positive) feedback vertex set [4, 5]. However, this must differ from the expected number of fixed points. We should obtain the expected number of fixed points under both the topological restriction and the presence of biases in Boolean functions, because in many situations, it is difficult to determine forms of regulatory functions whereas biases in the functions are observed in addition to network topology. In the case of no bias, the expected number is easily calculated and it is independent of network topology [6].

![An example set of a Boolean network and probability distributions of Boolean functions. Here, f2 is drawn from P2(f2), which is biased in favor of high frequency in on-state outputs. P1(f1) and P3(f3) are uniform and more complicated distributions, respectively.](image)

In this work, we have investigated the expected number of fixed points in a BN under the restriction of network topology, where Boolean functions assigned to every vertex are drawn from identical or nonidentical probability distributions [7]. In a BN, a Boolean variable \( x_i \) of the vertex \( i \) is updated by a Boolean function \( f_i \) as follows:

\[
x_i(t+1) = f_i[x_j(t); j \in J_i].
\]

The variable \( x_i \) takes \( x_i = 0 \) for the off state and \( x_i = 1 \) for the on state at every discrete time instant. A set of input vertices to vertex \( i \) is denoted as \( J_i \). We assume that network topology is fixed, and \( f_i \) is drawn from a probability distribution of Boolean functions \( P_i(f_i) \) (Fig. 1). The expected number of fixed points is denoted by \( \langle n \rangle \).
2 Results

Theorem.— If all links in a feedback arc set (FAS) are stochastically neutral, \( \langle n \rangle = 1 \) [7].

An FAS is defined as a subset of links containing at least one link of every cycle in a directed network (Fig. 2). We also define that a link \( j \to i \) is stochastically neutral if \( P_i(f_i) = P_i(g_i) \) holds for any pair of \( f_i(x_j, \{ x_k : k \in J_i - j \}) \) and \( g_i(x_j, \{ x_k : k \in J_i - j \}) \equiv f_i(\bar{x}_j, \{ x_k : k \in J_i - j \}) \), where \( \bar{x} \) denotes the negation of \( x \). This means that the probability that the link is activation interaction is equal to the probability that it is inhibition interaction. As shown in Fig. 2 (b), \( \langle n \rangle = 1 \) can be obtained from partial information due to the theorem. Moreover, the predominance of positive feedbacks increases \( \langle n \rangle \).

Fig. 2. (a) A set of four colored links is an FAS. (b) Another set of three colored links to vertex A is also an FAS. If \( P_A(f_A) \) is uniform, then these links in the FAS are stochastically neutral, and we immediately obtain \( \langle n \rangle = 1 \), independently of other information, such as function form of \( P_B(f_B) \).

References
Cardiac Alternans: Understanding Subcellular Calcium Patterns with the Master Stability Function

Yi Ming Lai, Joshua Veasy, Stephen Coombes, and Rüdiger Thul

Centre for Mathematical Medicine and Biology,
School of Mathematical Sciences,
University of Nottingham,
Nottingham NG7 2RD, UK

1 Introduction

What we perceive as a single heartbeat is the coordinated contraction of millions of cardiac muscle cells called myocytes. In the healthy heart, each myocyte performs a tightly orchestrated sequence of biochemical reactions, which is almost identical at every heart beat and involves transient rises and falls of the intracellular calcium (Ca\(^{2+}\)) concentration. However, this regular rhythm can be disturbed to give rise to so-called alternans, which are characterised by beat-to-beat alternations of the whole-cell Ca\(^{2+}\) concentration. Crucially, alternans constitute precursors to severe cardiac arrhythmias, which in turn can lead to sudden cardiac death. A detailed understanding of the origin and formation of cardiac alternans is therefore of vital interest.

In recent years, subcellular Ca\(^{2+}\) alternans have attracted significant attention. Here, different parts of a single cardiac myocyte exhibit different transients of the Ca\(^{2+}\) concentration. While there is clear clinical evidence that these subcellular Ca\(^{2+}\) alternans are pro-arrhythmic, we still lack a complete picture of their genesis and properties. Here we observe that the patterning of subcellular Ca\(^{2+}\) alternans can change abruptly upon parameter variation. Given that these parameters can change dynamically, such sudden transitions have direct implications for the contractile fidelity of cardiac myocytes.

Our findings are based on the insight that cardiac myocytes can be conceptualised as a network of networks. The local network represents the so-called Ca\(^{2+}\) release unit (CRU) and is the fundamental building block of subcellular Ca\(^{2+}\) alternans. The CRUs are coupled to their neighbours via Ca\(^{2+}\) diffusion to form the global cellular network. Traditionally, CRUs are described by coupled nonlinear differential equations, which has led to a large body of computational results. To allow for greater analytical insight, we have combined a piecewise linear (PWL) version [1] of a popular model of a CRU [2] with the master stability function (MSF) approach [3].

Mathematically, the Ca\(^{2+}\) dynamics in a healthy cardiac myocyte correspond to a synchronous period-1 orbit, where all nodes exhibit the same dynamics. The onset of subcellular Ca\(^{2+}\) alternans constitutes a bifurcation where this synchronous state goes unstable. The combination of PWL modelling and MSF allows us to derive explicit equations for the stability of the synchronous state. By using an analytical approach we are able to explore the rich pattern space of subcellular Ca\(^{2+}\) alternans and provide an explanation for the abrupt changes in the Ca\(^{2+}\) patterns that we observe.
2 Results

Figure 1 illustrates the diverse patterns that we observe in a network of 10 × 25 CRUs. Panels A (D) and B (E) show the peak Ca\(^{2+}\) concentration during successive beats. The transition from high to low Ca\(^{2+}\) concentrations, i.e. from red to blue, and vice versa is the hallmark of subcellular Ca\(^{2+}\) alternans.

![Figure 1](https://example.com/figure1)

**Fig. 1.** Peak Ca\(^{2+}\) concentrations, (A, B) and (D,E), respectively, for two sets of coupling strengths (top vs bottom) for successive beats. C and F show the corresponding eigenvectors.

To understand the emergence of these patterns, we used the MSF to calculate the stability of the synchronous network state \(s(t)\), i.e. when all CRUs behave identically, in terms of the eigenvalues of the graph Laplacian of the adjacency matrix \(G\) associated with the network. Linearization and block diagonalization of the system of equations yields a set of decoupled equations for the perturbations \(\xi\) about the synchronous state

\[
\frac{d\xi_i}{dt} = [DF(s) - \sigma \lambda_i DH(s)]\xi_i, \tag{1}
\]

for each eigenvalue \(\lambda_i\) of \(G\). Here \(\sigma\) is the coupling strength, \(DF\) is the Jacobian of the local dynamics, and \(DH\) is the Jacobian of the function describing how a node is coupled in the network. The synchronous state is stable if all the Floquet multipliers of Equation 1 lie in the unit disc.

Since the CRUs are coupled diffusively to their nearest neighbours, we know that the eigenvalues \(\lambda_i\) must be real and negative. This allows us to plot the relationship between the cell’s pacing period \(T_p\) and the parameter \(\beta = \sigma \lambda_i\). Figure 2A shows this dependence for the network used in Fig. 1. For a given coupling strength, alternans occur if at least one \(\beta\) value lies in the region marked “U”. Alternatively, we can determine the critical coupling strength for which the synchronous state loses stability and
Alternans occur as shown in Fig. 2B. For the results in Fig. 1, we fixed $T_p$ and increased $\sigma$ until one eigenvalue crossed from the stable region “S” into the unstable region “U.” In this case, the eigenvector associated with the unstable eigenvalue is predictive of the emergent pattern. This is illustrated in Fig. 1, where panels C and F depict the eigenvector associated with the unstable eigenvalue. As can be seen, the structure of the eigenvector matches very well with the simulation results.

The power of combining PWL modelling with MSF techniques is that all components in Equation (1) are known analytically, and hence bifurcation diagrams as in Fig. 2B can be computed in a semi-analytical fashion. In turn, this allows us to efficiently scan the parameter space and hence predict the emergent patterns of subcellular Ca$^{2+}$ alternans without the need for large-scale numerical simulations.

Acknowledgments

This work was supported by the Engineering and Physical Science Research Council (grant number EP/P007031/1).

References

Threshold activated coupling stabilizes the chaotic states to steady states

Chandrakala Meena\textsuperscript{1}, Pranay Deep Rungta\textsuperscript{2} and Sudeshna Sinha\textsuperscript{2}

\textsuperscript{1} Bar Ilan University, Ramat Gan, 5290002, Israel
meenachandrakala@gmail.com

\textsuperscript{2} Indian Institute of Science Education and Research (IISER), Mohali, 140306, Punjab, India.
pranay.rungta@gmail.com
sudeshna@iisermohali.ac.in

1 Introduction

We review results obtained in [1], on the dynamics of Random Scale-free networks of population patches [2, 3] connected by threshold activated coupling [4, 5]. The broad scenario underlying this coupling is that each population patch has a critical population density it can support, and when the population in the patch, due to its inherent growth dynamics (which may be chaotic) exceeds this threshold, the excess migrates to neighbouring patches. The neighbouring patch on receiving the migrant population may become over-critical too, triggering further migrations. So this form of coupling is pulsatile and inter-patch transport occurs only when there is excessive build-up of population density in a patch, which may initiate a cascade of transport events. In this study we will then aim to obtain broad insights on the dynamics of a complex network under threshold-activated transport, through the specific illustrative example of spatially distributed populations connected by threshold-activated migrations. Our principal question will be the following: what is the effect of threshold-activated dispersal on the dynamical patterns emerging in the network, and in particular, can threshold-activated coupling serve to stabilize the intrinsically chaotic populations in the network to regular behaviour, such as steady states or regular cycles?

2 Model

We consider a Random Scale-free network of $N$ sub-systems and at the local nodes, the dynamics is given by a prototypical map, the Ricker (Exponential) Map. Such a map models population growth of species with non-overlapping generations, and is given by the functional form:

$$x_{n+1}(i) = f(x_n(i)) = x_n(i) \exp(r(1 - x_n(i)))$$

where $r$ is interpreted as an intrinsic growth rate and (dimensionless) $x_n(i)$ is the population scaled by the carrying capacity at generation $n$ at node/site $i$. We consider $r = 4$ in this work, namely, an isolated uncoupled population patch displays chaotic behaviour. The coupling in the system is triggered by a threshold mechanisms. Namely,
the dynamics of node $i$ is such that if $x_{n+1}(i) > x_c$, the variable is adjusted back to critical threshold $x_c$ and the "excess" $x_{n+1} - x_c$ is distributed to the neighbouring patches. The state variable has to exceed from threshold parameter $x_c$ in order to initiate threshold-activated coupling.

2.1 Emergence of steady states and influence of the redistribution time and the number of open nodes on the suppression of chaos

When redistribution time ($T_R$) is large, namely transport processes are fast compared to population dynamics and the critical threshold $x_c$ is small ($0 < x_c < 1$), we have very efficient control of networks of chaotic populations to steady states. This suppression of chaos and quick evolution to a stable steady states occurs irrespective of the fraction of open nodes ($f_{open}$). Open nodes are the nodes in the network, which are open to the environment, and the excess is transported out of the system from such nodes. For threshold values beyond the window of control to fixed states, one obtains cycles of period 2. When $T_R$ is small, namely time-scales of the nodal population dynamics and the inter-patch transport are comparable then fraction of open nodes is crucial for chaos suppression. For small $T_R$, the system does not get enough time to relax to under-critical states and so perfect control to steady states may not be achieved. We demonstrate that when there are enough open nodes, the network relaxes to the steady state even for low $T_R$ (cf. Fig. 1). So more open nodes yields better control of the intrinsic chaos of the nodal population dynamics to fixed populations. When there are very few open nodes, the degree and betweenness centrality of the open node is important, with the region of control being large when the open node has the high degree/betweenness centrality, and vice versa.

![Bifurcation diagrams](image)

**Fig. 1.** Bifurcation diagrams for one representative node in a network of size 100, with respect to $x_c$. Here $T_R = 50$ and the number of open nodes is (a) 1, (b) 10, (c) 30 and (d) 60.
2.2 Quantitative measures of the efficiency of chaos suppression

When, \( f_{\text{open}} > \ln N / T_R \), a global steady state will emerge. This offers an estimate of \( f_{\text{open}}^c \). For instance, for \( N = 100 \) and \( T_R = 50 \), this argument suggests that \( f_{\text{open}} \sim 0.09 \). This also implies that for sufficiently large \( T_R \), or small enough network size, the system can attain steady state even when there is a single open node (i.e. \( f_{\text{open}}^c \to 0 \)). We can also obtain an estimate for the minimum \( T_R \), which we denote as \( T_R^{\text{min}} \), necessary for allowing the network to reach a steady state with just a single open node, namely \( f_{\text{open}} = 1 / N \). When \( T_R > T_R^{\text{min}} \), the range of threshold values that yield steady states (\( \langle R \rangle \)) is always non-zero and when \( T_R < T_R^{\text{min}} \), \( \langle R \rangle \) is zero for \( f_{\text{open}} < f_{\text{c}}^{\text{open}} \) (cf. Fig. 2).

![Graph showing the range of threshold values that yield steady states, \( \langle R \rangle \), as a function of the fraction of open nodes in the network \( f_{\text{open}} \). Here \( \langle R \rangle \) is averaged over different network configurations and initial states and the open nodes are randomly chosen. Results from different redistribution times (\( T_R = 50, 500, 1000, 5000 \)) and system sizes (\( N = 10, 100 \)) are shown.](image)

Summary. We have explored Random Scale-Free networks of populations under threshold-activated transport. Our central result is that threshold-activated transport yields a very potent coupling form in a network of populations, leading to robust suppression of the intrinsic chaos of the nodal populations on to regular steady states or periodic cycles. Further, suppression of chaos is facilitated when the threshold-activated migration is more rapid than the intrinsic population dynamics of a patch. So our study suggests a mechanism by which chaotic populations can be stabilized rapidly through migrations or dispersals triggered by excess population density in a patch.

References

Reconstructing Boolean Functions and Networks from Noisy Observational Data

Jie Sun and Erik Bollt
Clarkson University, Potsdam, NY 13699, USA,
sunj@clarkson.edu, bolltem@clarkson.edu

1 Introduction

We consider the problem of reconstructing the structure and dynamics of a Boolean network from noisy observational data. Boolean networks are common in biological applications [1–6]. To date, several methods have been proposed to infer Boolean networks [7–10], but they suffer from high computational cost and require unrealistically large amount of data and thus are infeasible for large networks. Here we propose an information-theoretic framework and develop algorithms that enable efficient reconstruction of Boolean networks and functions, subject to noise. Instead of relying on a combinatorial search, our method iteratively finds relevant nodes and edges and the best Boolean function that utilizes them. We validate our approach using synthetic data from random Boolean networks as well as in an application of automated disease diagnosis.

2 Results

Problem and solution strategy. In a Boolean network, the state of each node depends on the state of its neighboring nodes. To reconstruct a Boolean network, one needs to reconstruct the Boolean function defined on each node of the network. For a given node, we represent observational data by \( \{(x(t), y(t)) : t = 1, \ldots, T\} \) where \( x(t) \in \mathbb{B}^n = \{0, 1\}^n \) are the state of all nodes in the network, and \( y(t) \in \mathbb{B} \) represents the state of the given node. The goal is to find a “best” Boolean function \( f \) that models the given data. To quantify how good a model is, we consider the following objective function:

\[
\min_{K \subset [n], f : \mathbb{B}^{|K|} \to \mathbb{B}} \sum_{t=1}^{T} |f(x_K(t)) - y(t)| + \lambda \kappa(f),
\]

Here the term \( \sum_{t=1}^{T} |f(x_K(t)) - y(t)| \) counts the number of times \( f(x_K) \) differs from the observed value of \( y \), where \( K \) is a subset of the full index set \( [n] = \{1, \ldots, n\} \) and \( f \) is a Boolean function. The additional term \( \lambda \kappa(f) \) is a product of \( \lambda > 0 \) with \( \kappa(f) = |K(f)| \), the order of the Boolean function. In this framework by a “best model” we mean a Boolean function that minimizes the discrepancy between model and data.

This problem seems intractable at a glance mainly because the underlying model is not linear, and cannot be (linearly) parameterized. Nevertheless, we found that in the
small \( \lambda \) regime the problem can be solved in two steps. The first step is to find the set \( K \), which we show is equivalent as solving the following min-max optimization:

\[
\min_{|K|} \max_{K \subset [n]} I(X_K; Y),
\]

(2)

Here \( I(X_K; Y) \) is the mutual information between \( X_K \) and \( Y \) [11]. In principle one can solve this combinatorial optimization by exhaustive search over all possible subsets of order \( K \). Such strategy clearly does not work for large networks. Instead, we can show that this problem can be solved using optimal causation entropy (oCSE) [12]. The idea is to start with \( K = \emptyset \) and iteratively add components that maximizes the mutual information conditioning on existing components; then, after no more component can be added, each is examined and removed if its removal does not decrease the mutual information. Secondly, once \( K \) is identified, what is left is to fit the best Boolean function of order \( K \). Since \( K \) is typically small when the network is sparse, all is needed is to enumerate all order-\( K \) Boolean functions to minimize the objective function.

**Example of application.** In addition to validating our approach on synthetic data generated by random Boolean networks, we consider a dataset derived from a set of images obtained by cardiac Single Proton Emission Computed Tomography (SPECT) [13]. There is a total of \( T = 267 \) patients, each of whom is classified as either normal \( (y_t = 1) \) or abnormal \( (y_t = 0) \). For each patient’s image set, a total of \( n = 22 \) binary feature patterns were created, defining \( x_i(t) = 1 \) if the \( i \)-th feature is present in the SPECT images of the \( t \)-th patient, and \( x_i(t) = 0 \) otherwise \( (i = 1, \ldots, 22, t = 1, \ldots, 267) \). Finally, this post-processed Boolean dataset is further divided into a training set which contains 87 out of the 267 patients’ features and diagnosis, and a validation set which contains such information for the remaining 180 patients. We apply the proposed approach to learn a Boolean network that represents the decision rules, that is, to use the network and its functioning as a diagnosis tool to determine if a given patient is subject to disease based on a reduced set of Boolean features out of the 22 features. As shown in Fig. 1, our method achieves near 80% of decision accuracy on the validation data across a wide range of parameters. The achieve accuracy, generally using only a subset of the full set of 22 Boolean features, without any fine-tuning of parameters or further optimization, is already comparable to the best known result on such datasets [14].

**Summary.** We proposed a mathematical framework based on information-theoretic measured, to reconstruct a Boolean network from noisy observational data, and developed efficient algorithms to solve such a problem. We validated our method on synthetic data and also in an application problem for automated diagnosis of cardiac diseases.

**References**

Fig. 1. Automated diagnosis of heart disease using 22 Boolean attributes derived from cardiac SPECT. We compute the accuracy of diagnosis as the overall percentage of correct diagnosis in the validation set as a function of $\alpha$ (parameter that controls the significance level when estimating mutual information and conditional mutual information from data). In addition, we also show, for each $\alpha$, false positive ratio (FPR) and false negative ratio (FNR), together with the effective number of Boolean variables inferred by our method (dashed curve).

Gene-gene interaction module identification in single-cell RNA sequencing

Simone Marini¹, Danila Vella², Nelson Nazzicari³, and Riccardo Bellazzi²,⁴

¹ University of Michigan, Ann Arbor, MI, USA, smarini@med.umich.edu,
² Istituti Clinici Scientifici Maugeri, Pavia, Italy
³ CREA - Council for Agricultural Research and Analysis of Agricultural Economics, Lodi, Italy
⁴ University of Pavia, Pavia, Italy

1 Introduction

Single-cell RNA sequencing (scRNAseq) techniques are gaining momentum in the scientific community. Not surprisingly, a plethora of novel scRNAseq computational tools are emerging as well, addressing specific scRNAseq-data challenges. To date, scRNA-tools.org, a catalogue of software packages designed for the analysis of single-cell RNA-sequencing data, reports over 250 single-cell bioinformatic tools. These in silico methods are pipelines for scRNAseq downstream analysis, comprising quality control, imputation, batch effect correction, normalization, scaling, highly variable gene detection, clustering, and cluster characterization.

An unexplored niche in scRNAseq data analysis is gene module identification, i.e., grouping genes representing specific cellular functions. Gene module identification eases the interpretation of gene networks, and it is a key step towards applications such as drug target discovery/repositioning [1]. Traditional bulk tissue RNAseq maps gene networks under the assumption of an average cell type. In contrast, with scRNAseq it’s possible to study a different gene network for each cell type/state composing the tissue. For example, from a single tissue sample it would be possible to isolate gene networks detailing different cell types as fibroblast, immune cells, neural cells, etc., each one with its specific functional modules. This is particularly important if rare cell types are present, since infrequent signals can be diluted and impossible to catch with bulk tissue analysis. Here we present MTGO-SC (MTGO for scRNAseq), an algorithm detecting gene modules in scRNAseq data. Each genetic module is labeled with a Gene Ontology (GO) term help interpreting its function. MTGO-SC is based on MTGO [1], a tool for biological network interpretation recently published by our group, outperforming state-of-the art algorithms on protein networks.

2 Results

MTGO-SC is tailored to work on sparse, extremely large expression matrices typical of scRNAseq data. It is designed to rapidly produce an interpretable gene interaction network for each gene cluster resulting from scRNAseq analysis. In this way, SC-MTGO
facilitates the result interpretation with ad-hoc network visualization for each different cell cluster. The sets of functional modules are selected taking into account the specific topological structure of the network, and each module is associated with a specific biological meaning (such as a protein complex or a biological pathway, indicated by a GO term). The collection of functional modules specific for each cell cluster *de facto* enables network comparison, highlighting the different roles of genes which act in the different cellular contexts.

Fig. 1. Gene network visualization with MTGO-SC. Here the networks are obtained from human dendritic cells and monocytes [2]. For this use-case, we designed a network out of single-cell gene coexpression (i.e., a gene is a node) considering Spearman correlation greater than 0.6 to represent the edges. SC-MTGO not only allows to tune the threshold for a binary edge, but works also with weighted edges, and different similarity measures used to build gene interaction networks [3]. In this dataset, a different network is calculated for each type of monocyte or dendritic cell. The network is composed by the genes distinguishing the cell type cluster from the rest (Wilcoxon test, used in Seurat [4] with default parameters). Panel A shows the global network, where the nodes represent genes, the node colors represent the GO membership and the edges represent the gene expression correlation. Panel B shows a network where the nodes represent the modules, each linked to a specific GO, and the edges represents one or more connections among the genes belonging to the modules. Blue, green, and yellow nodes indicate respectively the membership of the three GO classes, i.e., Cellular Component, Molecular Function, and Biological Processes. Panel C shows a intra-module gene network module, in particular cellular response to interferon-gamma (GO:0034341) module of the network representing the plasmacytoid dendritic cells.
Network visualization. MTGO-SC allows the user to navigate the network on three interlaced levels, i.e., the gene-gene interaction, the inter-module interaction, and the intra-module interaction levels. Each network comes from a single-cell cluster. The gene-gene interaction network (Fig. 1, A) is the traditional network view, where each node is a gene and each edge is a coexpression. In the inter-module visualization (Fig 1, B), MTGO-SC depicts networks where the nodes represent the functional modules named after their characterizing GO-term (and not the genes), and each interaction represents one or more edges between the nodes belonging to the modules. An interactive interface allows the user to highlight the three subnetworks linked to the three GO classes (see Fig. 1, caption). SC-MTGO allows also to visualize in detail the intra-module network (Figure 1, C), i.e., a subnetwork within a single module. The use-case producing Fig. 1 is derived from a human data set of immune system cells [2]. Here, MTGO-SC successfully retrieves several functional modules related to the immune response characterizing these cells, such as peptide antigen binding (GO:0042605), IgG binding (GO:0019864), or adaptive immune response (GO:0002250). More importantly, MTGO-SC retrieves cell-subtype specific modules, such as, for conventional dendritic cells CD141+, nervous system development (GO:0007399) [5]; or, for plasmacytoid dendritic cells, cellular response to interferon-gamma (GO:0034341), defense response to virus (GO:0051607), and toll-like receptor signaling pathway (GO:0002224) [2, 6].

Presented as an R package, MTGO-SC is natively integrated with Seurat[4], a popular tool for scRNAseq downstream analysis which is gaining consensus among the scRNAseq scientific community. All the visualization features are made possible thanks to the use of many R packages specific for network visualization.

Summary. We designed an algorithm to find functional modules gene networks obtained from single-cell data. In the single-cell context, automatic module identification has two main functions: (a) to ease network interpretation (i.e., in the case of single-cell data will help identifying the cell type or cell state composing the cluster); and (b) to help generating or verifying hypotheses regarding the genetic machinery underpinning the the specific cell cluster generating the network.

References

Chromatin Assortativity approaches to study phenotypic variability

Vera Pancaldi
Cancer Research Center of Toulouse, INSERM U1037, Université Toulouse III Paul Sabatier, Toulouse, France vera.pancaldi@inserm.fr

1 Introduction

Heterogeneity pervades biological systems at various levels. At a more global level, the difference between individuals is thwarting our successes in medicine and warrants the search for personalised medicine approaches, especially in cancer research. On the other side, heterogeneity in the different ecological systems within an individual, for example the gut, the bone marrow or the tissue surrounding a tumour, is being increasingly recognized as an important factor in establishing the equilibria that we associate to health and that are often broken in diseases. At an even more focused level, the variability of phenotypes in cell populations with identical genetic profiles, such as cells from the same healthy individual, might also play a role in their responsiveness and adaptation to external changes and stimuli.

Previous works in the literature have highlighted a correspondence between variability of cell phenotypes in time, across individuals, and even across species, through observations of correlations in measures of variability on these three levels. These findings suggest that the regulatory environment of genes, namely their genomic context, can impact gene expression at the three levels of response to changes in environment, inter-individual variability and even propensity for mutation.

We aim to investigate how chromatin properties and its organization inside the nucleus affects variability in these three scenarios. Recently, we have suggested a network formalism to represent and study chromatin 3D contacts. We consider networks where nodes are genomic regions and edges are drawn if two specific regions are found to be in close proximity in the nucleus, as assessed by chromosome conformation capture experiments. Measuring chromatin assortativity allows us to test whether regions of the genome that are in preferential contact with each other also share specific biological properties.

2 Results

Here, we have investigated whether regions containing genes that are highly variable across healthy individuals are found to interact preferentially with each other, suggesting the existence of neighbourhoods in the 3D genome with an accumulation of these variable genes. We also applied the same analysis to single-cell gene expression datasets, aiming to assess whether genes that are more variable across cells of the same
type also accumulate in specific genomic locations that are brought together in 3D. Cells with an identical external environment and genome can still display different gene expression profiles and phenotypes due to differences in the modifications of the DNA (epigenomic marks) that impact gene regulation.

We thus considered two biological systems. Firstly, we studied different types of human blood cells collected from healthy individuals for which gene expression and epigenomic data has been collected within the framework of the BLUEPRINT project (monocytes, neutrophils and T cells)[1]. For these cells chromatin 3D interactions have also been mapped using the Promoter-Capture HiC protocol, a technique that allows to map contacts between promoters and other genomic regions [2]. Projecting the variability of expression data for each cell type on the network of chromatin contacts we were able to assess whether gene expression variability is assortative [3]. We thus defined inter-individual Expression Variability Assortativity (EVA) and found it to be positive and significant compared to randomizations, indicating that chromatin fragments containing genes that exhibit high expression variability across individuals are preferentially connected with each other in 3D (Figure 1). In addition, genes that we found to be hypervariable specifically in neutrophils in a previous study [1] showed clustering in the 3D chromatin contact network.

The second dataset we considered was a collection of single-cell gene expression variability measurements in mouse Embryonic Stem Cells (mESCs), one of the best characterised cell types, for which many epigenomic marks have been mapped. We projected single-cell gene expression noise levels, carefully estimated taking into account the dependence of the variability on mean expression levels [4], onto the corresponding promoter-centred chromatin interaction networks [5]. We thus defined Single Cell EVA as assortativity of variability across single cells. Preliminary results showed that this measure is also be positive and significantly different from random expectation (Figure 1).

Despite the interest of measuring a global genomic property, the recent approaches to measure assortativity at a local level[6] inspired us to investigate whether local assortativity measures could reveal interesting aspects of specific genomic regions. So far, notwithstanding the availability of numerous chromatin structure contact maps generated for different cells and in different conditions, the problem of comparing these structures remains an open challenge, together with the interpretation of the functional impact of any differences found. It could be conceived that changes in assortativity of specific features in specific genomic regions would highlight structures characteristic of one cell state or cell type.

**Summary.** Using the concept of assortativity, we showed that regions of chromatin characterised by higher variability in expression, either across single mouse embryonic stem cells or in blood cell types across a panel of healthy individuals, tend to interact preferentially in 3D. These findings are consistent with the existence of neighbourhoods of the 3D genome in which specific concentration of different regulatory factors could accumulate.
**Fig. 1.** Figure 1. A) PChiC 3D chromatin contact map for Neutrophils. Nodes are gene promoters, colour indicates Expression Variability across 125 individuals. B) Zoomed in region of A. C) Values of EVA for data in A compared to randomizations. D) PChiC 3D chromatin contact map for mouse embryonic stem cells, nodes are gene promoters and colour indicates expression variability (DM values) across single cells. E) Zoomed in region of D. F) Values of SingleCell Expression Variability in D compared to randomizations.

**References**

Identifying Dominant Connection Patterns in Protein Structure Networks

Nora S. Martin$^{1,2}$ and Sebastian E. Ahnert$^{1,2}$

$^1$ Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, UK
$^2$ Sainsbury Laboratory, University of Cambridge, UK
nsm36@cam.ac.uk

1 Introduction

Proteins are biological molecules with a complex three-dimensional structure, which enables them to fulfil a variety of biological functions. Therefore, the description, analysis, comparison and classification of solved protein tertiary structures is a widely studied topic. The representation of protein tertiary structures as protein structure networks has led to promising insights [4] and it has been shown that a fraction of the edges in these networks contains sufficient information for the reconstruction of the original three-dimensional structure [6, 7]. We present a new method of finding a subset of network edges for a minimal structure network. A custom network compression algorithm is used, which encodes protein structure networks more compactly by grouping network edges into bundles. Then a coarse-grained network is defined by selecting one edge per bundle.

2 Results

Fig. 1. A) The network compressibility as a function of distance cutoff shows an optimum, which can be used for the structure network construction. B) Protein structure network: the nodes (black dots along the x-axis) represent amino acid residues. There are edges between pairs of residues, whose $C_\alpha$ atoms are closer than 5.9Å in the native structure. The edges are colour-coded by which bundle they are assigned to after compression. Data from PDB entry 5J4A, chain C [5].
Our analysis starts with protein structure networks of solved protein tertiary structures, obtained from the RCSB Protein Data Bank [1]. Proteins consist of chains of amino acid residues, which fold into a three-dimensional native structure depending on the chemical properties of the residues’ side chains. Protein structure networks represent each amino acid residue by a node. There are edges between all pairs of residues, which are closer than a given distance cutoff in the native structure [4]. As we are interested in the overall shape of the protein chain, edges of short sequence range are discarded according to a distance-dependent condition for our analysis.

These networks contain some redundant information: when two regions of the polypeptide chain are close in the three-dimensional structure, this is represented by a group of edges in the structure network. There is an edge for each individual pair of residues which is close in the folded state. Such groups of edges between regions of the polypeptide chain can occur regardless of the type of secondary structure. For a coarse-grained description of the three-dimensional structure, it is sufficient to specify an edge for each pair of chain segments, which are close in the folded state.

This property can be exploited using network compression. We group the network edges into bundles of edges, which connect the same pair of regions of the chain (Figure 1B). When the edges are sorted into such bundles, the edge list can be encoded more compactly. Therefore, we use a greedy algorithm to minimise the information content of the list of edges by assigning them to bundles. For most proteins in a non-redundant dataset (obtained from the PISCES server [8]), a lossless compression of 15-30% is achieved, indicating that our compression algorithm successfully exploits the underlying structure of protein structure networks. In addition, the network compressibility lets us define the distance cutoff of highest compressibility as the cutoff used for the construction of the structure network (Figure 1A). This cutoff strikes a good balance between low cutoff distances, at which there is little potential for compression due to the small number of edges, and high cutoffs, at which there is little compression due to coincidental edges. The distance cutoff of best compression was found to be at around 6-7 Å for most protein structures, similar to cutoffs used in the literature [2].

The bundles returned by the network compression algorithm are then used to generate a coarse-grained network by representing each bundle of edges by a single edge. Two methods are used to test how descriptive these minimal, coarse-grained networks are of the protein structure. Firstly, the matrix of shortest paths of the minimal network is compared to that of the full network. Secondly, we verify that the similarity between the minimal networks of two proteins reflects their assignment into fold groups by established structural classification databases, such as SCOP [3].

Summary. We present a network compression algorithm for protein structure networks. It groups edges connecting the same regions of the polypeptide chain into bundles. These bundles can be used to coarse-grain the network: a minimal network is obtained by replacing each bundle by a single edge.

References

Highlighting the Complex Network Structure of Epigenetic Regulation using Message-Passing

Abhijeet Sonawane¹, Dawn DeMeo¹, John Quackenbush², and Kimberly Glass¹,³

¹ Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
² Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
³ contact: kimberly.glass@channing.harvard.edu

1 Introduction

The biological processes that drive cellular function can be modeled by a complex network of interactions between regulators (transcription factors) and their targets (genes). One critical influence on these interactions is a cell’s “epigenetic state”, or whether the regulatory region of a gene is physically accessible by a transcription factor in that cell. In our analysis we used epigenetic information to build initial network estimates in several different types of cells and benchmarked the accuracy of these networks using independent validation data. We then applied a message-passing algorithm to these initial networks, re-prioritizing edges based on their contribution to local network structure. We observed a drastic improvement in network accuracy after applying the message-passing approach, including among edges that are uniquely observed in a particular type of cell. We call our approach SPIDER: Seeding PANDA Interactions to Derive Epigenetic Regulation. Further investigation suggests that the epigenetic state of network interactions is exploited by SPIDER to recover missing links (false-negatives) in the networks. SPIDER works by highlighting common structures across multiple networks, thus we believe our findings may point to a previously uncharacterized underlying structure in gene regulatory networks.

2 Background and Methods

Estimating accurate and meaningful networks remains a fundamental challenge in biology. Several years ago we developed PANDA (Passing Attributes between Networks for Data Assimilation) [2], a method that uses a “message passing” approach to integrate multiple types of genomic data and construct directed genome-wide gene regulatory networks. PANDA models gene regulation as communication between “transmitters” (regulating genes, or transcription factors) and “receivers” (regulated genes, or those targeted by transcription factors). The algorithm starts by considering a “prior” network consisting of potential routes for communication between nodes. It then performs a series of network projections to integrate with other sources of data (such as gene co-expression and protein-protein interaction), estimate the information flow along each edge, predict where communication is succeeding or failing, and use this information to deduce the most consistent areas in the networks structure. One distinct advantage...
of the approach is that the structure of model is highly flexible and can be easily modified to use different types of input data. Here we describe SPIDER (Seeding PANDA Interactions to Derive Epigenetic Regulation), an extension of PANDA that includes epigenetic data.

Epigenetic data represents biological modifications that can impact the system without altering DNA sequence. For example, changes in the three dimensional packaging of DNA (chromatin structure) can influence the function of a gene by making it more or less physically accessible within the cell. SPIDER uses epigenetic data to appropriately adjust the strength of the edges in the prior network analyzed by PANDA. In particular, SPIDER starts by identifying all potential locations for transcription factors (TFs) across the genome. It then intersects these locations with open chromatin locations derived from epigenetic data. For each TF location that is within a region of open chromatin, SPIDER creates an edge from that TF to nearby genes (all genes within a pre-specified range). These edges make up an epigenetically-informed prior network. PANDA can then be applied to estimate an epigenetically-informed regulatory network. An overview of the SPIDER algorithm is shown in Figure 1A.

3 Results

We have tested SPIDER using DNase hypersensitivity data (a common source of epigenetic information) collected in six different types of human cells (cell-lines) [1]. To start,
we created a human genome-wide regulatory network by scanning regulatory regions of genes for vertebrate transcription factor sequence-motifs [3]. We then ran PANDA on this “open” prior network. We next used SPIDER to intersect potential transcription factor binding sites with DNase peak regions, resulting in six cell-line-specific network priors. We then used SPIDER to run PANDA’s message passing procedure on each of the cell-line-specific priors. This resulted in fourteen total networks: (1) the “open prior” network, (2) the PANDA-predicted network, (3) six epigenetically-informed prior networks (one for each cell-line), and (4) six epigenetically-informed regulatory networks predicted by SPIDER.

We evaluated the accuracy of each of these networks using cell-line-specific gold standards constructed from transcription factor ChIP-seq peaks. We quantified the accuracy of each of the network models using the area under the receiver-operating characteristic curve (AUC-ROC). We observe that the SPIDER-predicted networks are significantly more accurate than any of the other networks, with an increase in AUC of ≈ 0.2 (Figure 1B). Further evaluation of these results indicates that the relatively poorer performance of both the “open” prior and PANDA-derived network is a result of both false-positives (incorrect edges) and false-negatives (missing edges). Although the epigenetically-informed priors had very few false positives, they also had substantially more false negatives. The improved accuracy of the SPIDER-predicted networks relative to these other networks appears to be a consequence of the algorithm’s ability to remove false-negatives (add back in edges missing from the epigenetically-informed-priors). These results indicate that there is an underlying structure in gene regulatory networks that can be exploited by SPIDER. Not only can this help to predict accurate gene regulatory networks, quantifying these network structures has the potential to lend fundamental insights in the driving principles underlying biological systems.

References

2. Glass, K., Huttenhower, C., Quackenbush, J., Yuan, G.C.: Passing messages between biological networks to refine predicted interactions. PLOS ONE 8(5), e64832+ (May 2013), http://dx.doi.org/10.1371/journal.pone.0064832
Topological variables of habitat networks as predictors of species occurrence

Damian O. Ortiz-Rodríguez 1,2, Maarten J. van Strien2, Antoine Guisan3 and Rolf Holderegger 1

1 Biodiversity and Conservation Biology, WSL Swiss Federal Research Institute, Birmensdorf, Switzerland
2 Planning of Landscape and Urban Systems PLUS, Institute for Spatial and Landscape Development, ETH Zürich, Switzerland
damiano@ethz.ch
3 Spatial Ecology Group, Department of Ecology and Evolution, University of Lausanne, Switzerland

1 Introduction

For biodiversity conservation it is crucial nowadays to have modelling tools that allow prediction of the presence or absence (occurrence) of species in specific areas susceptible to human intervention. Species distribution models [4] successfully predict habitat suitability, but they do not consider the connectedness of habitat patches nor their topological placement within an entire habitat network. Such shortcomings can be addressed by a predictive network modelling approach, but few studies have explored the potential of using habitat networks to predict species occurrence so far. In this study, we develop network-based models using available presence-only records that aim to predict species occurrence in habitat patches following the conceptual function

\[ \psi = f(q, lc, nt) \]  

(1)

where \( \psi \) is the binary occurrence-state of a species in a habitat patch; \( q \) is the quality of the patch, \( lc \) is the ‘local connectivity’, and \( nt \) is the ‘network-scale topology’; \( q \) is defined by the properties of suitability and size of the habitat patch (node); \( lc \) refers to the measures of immediate (up to third order) vicinity between habitat patches; \( nt \) refers to the measures that are taken at the scale of the full network.

2 Methods

We used as focal species the European tree frog (Hyla arborea L.) and as study area the densely populated Swiss Plateau. We followed a multi-step procedure involving first habitat suitability modelling and GIS-processing to delineate patches (nodes) based on geolocalized records of the focal species (from Swiss InfoSpecies: https://www.infospecies.ch). We then defined the edges based on three different cost surfaces, which thus yielded three different networks: 1. Uniform, limited only by the dispersal distance
of the focal species, 2. Traffic, incorporated the road traffic to the cost. 3. HS, included the inverse habitat suitability (HS) index values of the landscape matrix. (Fig. 1).

Fig. 1.

The three habitat networks of *H. arborea* made with different cost surfaces: A) Uniform, B) Traffic, C) Inverse habitat suitability. Insets show in detail the nodes (habitat patches) and edges, the main maps show the networks in full on the Swiss Plateau.

We calculated the following independent variables for our models that predict occurrence: Degree, strength and third-order neighborhood (for lc); betweenness and closeness centrality (with a modification to handle disjoint networks) (for nt); habitat availability (hybrid of lc and q), patch size and habitat suitability index (for q). Our response variable, the patch occurrence-state, was parametrized by a sampling-intensity method based on a threshold of sightings of other species over the total number of visits (based on [1]). The occurrence-state of 73 patches was ‘absence’, while 209 had ‘presence’. We fitted the three network models and a model without topological variables (No-Topo) using Boosted Regression Trees [2], and evaluated them with their cross-validated AUC scores and confusion matrices of prediction accuracy.

3 Results & Discussion

We found that the network based on inverse habitat suitability was extremely fragmented; in contrast, the traffic and uniform networks showed very similar big main components. In all of the networks, the most important variable was the HS Index, and
this was followed by strength, third-order neighborhood, and degree in the best performing model, which according to our evaluation parameters (Cross-validated AUC and residual deviance) was the uniform network model. The model without network measures was clearly outperformed by all the models with topological variables (Table 1).

**Table 1.** Comparison of predictive performance of the four models. Scales are in opposite senses

<table>
<thead>
<tr>
<th></th>
<th>Uniform</th>
<th>Traffic</th>
<th>Habitat suitability</th>
<th>NoTopo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-validated AUC score</td>
<td>0.762</td>
<td>0.734</td>
<td>0.748</td>
<td>0.702</td>
</tr>
<tr>
<td>Mean residual deviance</td>
<td>0.557</td>
<td>0.714</td>
<td>0.762</td>
<td>0.917</td>
</tr>
</tbody>
</table>

The results support the already acknowledged relevance of habitat suitability, while also showing the importance of the topology of the networks, clearly conveyed by the underperforming model without network measures (NoTopo, Table 1). Interestingly, patch size was consistently at the bottom of the variable importance, so our results contradict strongly the habitat amount hypothesis [3], which claims that big enough total area close to a patch is what matters for the occurrence of a species in a habitat patch, irrespective of how the patches are fragmented and connected. In summary, we show that abstract topology measures of habitat networks are relevant for the prediction of species occurrence, which opens the possibility for their use in different contexts around the world that are in dire need of effective and inexpensive conservation management.

**References**

Part II

Community Structure
Evaluating the leximin method for community detection

Arya D. McCarthy¹,² and David W. Matula²

¹ Johns Hopkins University
² Southern Methodist University
arya@jhu.edu matula@smu.edu

We evaluate the leximin method for community detection, a divisive clustering scheme which achieves a higher normalized mutual information score (NMI) than eight popular methods on a common set of benchmark graphs with weak communities. This performance is an artifact of how NMI is computed, rather than an indication that the leximin method better recovers the true communities. We use this to argue that NMI should be replaced with another known measure: adjusted mutual information (AMI). On graphs with strong communities, the leximin method is competitive in its ability to recover ground-truth communities. Finally, it distinguishes itself from popular hierarchical CD methods by early stopping on weak communities, allowing ties, and achieving high performance.

1 Introduction

Community detection (CD) algorithms of recent years are typically stochastic or ad hoc (so that the accident of the labeling affects the result), with some exceptions [7][10]. We formulate community detection by relaxing the sparsest cut problem to a linear program (LP) model. The model routes traffic flow between all pairs of nodes, with an objective that minimizes congestion while fairly satisfying demand [8]. (In applications, the traffic could be goods in a commodity distribution network, gossip in a social network, or signals in a neural network.) When rerouting cannot avoid congestion, the saturated edges define a sparse “bottleneck cut”. Continuing to increase the allocation of flow gives a sequence of bottleneck cuts which dissect the graph [5]. Introduced in the 1980s, this dissection is a divisive clustering method now known as the leximin method [4][6].

The leximin method is the earliest known CD method to use mathematical programming, and it has been repeatedly rediscovered in traffic literature. It is immune to relabeling of nodes, producing a canonical dendrogram for a given graph. Unlike other mathematical programming techniques for CD [1], ours is deterministic, and exact optimization is tractable. At the time of its proposal, it was impractical for even medium-sized networks; advances in LP theory [2] and computing hardware have made it usable for graphs on hundreds of nodes. We evaluate leximin in the context of modern CD algorithms, on benchmarks modeling real-world properties.

³Formally, our LP leverages weak duality between hierarchical sparsest cut and the hierarchical maximum concurrent flow problem [8], which finds a lexicographic maximin (i.e. leximin) allocation of flow between all pairs in a network.

⁴The relaxation will either find the sparsest cut or a multipart cut (a grid) when the entire network (or at least five components) congests at once. This is consistent with LPs being in the complexity class P while sparsest cut is NP-hard.
2 Experimental Methods

We created 25 LFR benchmark graphs [3] for each combination of parameters \( n \in \{80, 100, \ldots, 240\} \) and \( \mu \in \{0.03, 0.09, \ldots, 0.75\} \) (the mixing parameter\(^5\)), then applied the leximin method and the eight CD algorithms in the igraph package. We used NMI and AMI to score each method’s computed communities against the LFR ground truth.

3 Results

Performance for various network sizes in terms of the mixing parameter \( \mu \) is shown in Figure 1.\(^6\)\(^7\) Those graphs which took longer than 3 hours to process are excluded, though all graphs with \( n \geq 180 \) and \( \mu \geq 0.5 \) were processed within this bound.

Under NMI (the de facto standard measure), the leximin method’s performance plummets at \( \mu \approx 0.4 \), then rises again, exceeding all eight algorithms’ scores in the limit as \( \mu \to 1 \). This is misleading: the leximin method’s best computed partition is into

---

\(^5\)The fraction of each node’s edges connected to other communities, \( k_{ov}/k \).

\(^6\)The behavior was sequential, so we only plot \( n \in \{80, 160, 240\} \).

\(^7\)Plots and citations for remaining methods are available in github.com/aryamccarthy/thesis.
single-node communities, missing the mesoscopic structure that other algorithms find. It still receives a higher score. This motivates evaluation by AMI, which discounts NMI by the expected value, accounting for chance clusterings. Both measures have the same maximum (1.0), but AMI gives a constant mean value of zero [9]. When clusters are scored using AMI, most methods’ performance profiles are unchanged, e.g. multilevel, which remains effective on graphs with strong communities. The edge betweenness algorithm’s performance drops faster as $\mu \to 1$, and leximin’s score for high $\mu$ is zero. This is a fairer assessment of the algorithms’ performance. Leximin performs competitively on strong communities. There, it outperforms leading eigenvector [7], the other method capable of producing ties and early stopping.

4 Conclusion

By either measure, the leximin method is competitive for networks with strong communities. While feasible only for small to medium graphs, it is advantaged by discerning ties and producing canonical clusterings. Future work will explain the graph properties of the method’s phase transition for weak communities. With broader implications for evaluating CD, we see that NMI can exaggerate community detection performance. Using AMI mitigates this, which we encourage for the CD community moving forward.

References

Community detection in networks with unobserved edges

Till Hoffmann¹, Leto Peel² Renaud Lambiotte³, and Nick S. Jones¹,⁴

¹ Department of Mathematics, Imperial College, London SW7 2AZ, UK, tah13@imperial.ac.uk
² ICTEAM, Université catholique de Louvain, Louvain-la-Neuve B-1348, Belgium leto.peel@uclouvain.be
³ Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK, renaud.lambiotte@maths.ox.ac.uk
⁴ EPSRC Centre for Mathematics of Precision Healthcare, Imperial College, London SW7 2AZ, UK nick.jones@imperial.ac.uk

Detecting communities in networks provides a means of coarse-graining the complex interactions or relations (represented by network edges) between entities (represented by nodes) and offer a more interpretable summary of a complex system. However, in many complex systems the exact relationship between entities is either unknown or unobserved. Instead, we may observe interdependent signals from the nodes, such as time series, which we may use to infer these relationships. Over the past decade, a multitude of algorithms have been developed to group multivariate time series into communities with applications in finance, neuroscience, and climate research. For example, identifying communities of assets whose prices vary coherently can help investors gain a deeper understanding of the foreign exchange market [2, 3] or manage their market risk by investing in assets belonging to different communities [4]. Global factors affecting our climate are reflected in the community structure derived from sea surface temperatures [5].

Current methods for detecting communities when network edges are unobserved, typically involve a complicated process that is highly sensitive to specific design decisions and parameter choices. In this work, we develop a Bayesian hierarchical model for multivariate time series data that provides an end-to-end community detection algorithm that does not extract information as a sequence of point estimates, but instead propagates uncertainties from the raw data to the community labels.

The variability of high-dimensional time series is often the result of a small number of common, underlying factors [1]. For example, the stock price of oil and gas companies may be positively affected by rising oil prices, whereas the manufacturing industry, which consumes oil and gas, is likely to suffer from rising oil prices. Motivated by this observation, we model the multivariate time series \( y \) using a latent factor model, i.e. the \( n \)-dimensonal observations at each time step \( t \) are generated by a linear transformation \( A \) of a lower-dimensional, latent time series \( x \) and additive observation noise. The entries \( A_{ij} \) of the \( n \times p \) factor loading matrix encode how the observations of time series \( i \) are affected by the latent factor \( q \). Using our earlier example, the entry of \( A \) connecting an oil company with the (unobserved) oil price would be positive, whereas the corresponding entry for an automobile company would be negative.
Fig. 1. Detected communities of stocks are correlated with industry sectors. (a) heat map of the factor loadings inferred from constituents of the S&P100 index. Each row corresponds to a stock and each column corresponds to a factor. (b) a two-dimensional t-SNE embedding of the factor loading matrix with cluster labels including credit card (CC) and fast-moving consumer goods (FMCG) companies.
Our approach naturally supports multiscale community detection as well as the selection of an optimal scale using model comparison. We validate and study the properties of the algorithm using a series of synthetic datasets. We then apply it to daily returns of constituents of the S&P100 index to identify salient communities of similar stocks and to climate data of US cities to identify homogeneous climate zones.

Figure 1 shows the detected communities from the daily returns of constituents of the S&P100 index of the stocks of 100 large companies in the United States. We obtained 252 daily closing prices for all stocks during 2016\(^1\). The community assignments capture salient structure in the data. For example, the three smallest communities each having only two members consist of two credit card companies, two defence companies, and two chemical companies (which have since merged to form the conglomerate DowDuPont). Other specialised communities consist of financial services companies (e.g., Citigroup, Goldman Sachs), as well as manufacturing and shipping (e.g., Boeing, Caterpillar, FedEx, United Parcel Service).

References


\(^1\)https://finance.yahoo.com/
Detectability of Macroscopic Structures in Directed Networks: a Stochastic Block Model Approach

Mateusz Wilinski\textsuperscript{1}, Piero Mazzarisi\textsuperscript{2}, Daniele Tantari\textsuperscript{1}, and Fabrizio Lillo\textsuperscript{2}

\textsuperscript{1} Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
\textsuperscript{2} Dipartimento di Matematica, Porta di Piazza San Donato 5, 56126 Bologna, Italy

1 Introduction

Disentangling network macroscopic structures is one of the founding problems in complexity science [1]. One of the most basic models of communities in networks are the stochastic block models [2]. It was recently shown in [3] that in this case, the detectability of real communities only from the network topology is limited. Even though the results were shown mostly for \textit{planted partition}, where there are only two parameters, the conclusions are universal. The result was later generalised into general stochastic block model with symmetric affinity matrix [4].

Simple graphs are often just an approximation of real-world networks. Neglecting weights, multiple links or self loops may be justified in some cases but it also implies a loss of information. Important generalisation of a simple graph is a directed network. Community detection in such networks is already an active field [5]. In relation to stochastic block models, the main difference is that the affinity matrix does not need to be symmetric anymore. We will show that by introducing an asymmetry of directions, we are able to significantly increase the range of the detectable phase without changing the network assortativity. Importantly, this qualitative change holds for an entire class of hardly detectable models, where both the average in- and out-degree are the same across all groups.

2 Results

In order to analyse the impact of edge direction on the detectability of communities, we examined a more general case of planted partition model. Similarly to the planted partition, it will be a stochastic block model. In the most general case SBM assumes there are $q$ groups. Each vertex of a network has a group label $t_i \in \{1, \ldots, q\}$ and the frequency of these labels is described with a vector of probabilities $n_t$. Edges are generated independently according to a $q \times q$ matrix of probabilities, with $Pr(A_{ij} = 1) = p_{ij}$. In the original planted partition $\forall_{i} n_t = \frac{1}{q}$, $\forall_{i} p_{ii} = p_{in}$ and $\forall_{i \neq j} p_{ij} = p_{out}$. Our generalisation, which we will call the directed planted partition, introduces a third parameter to
the affinity matrix. The resulting change in the affinity matrix is as follows:

\[
\mathbf{p} = \begin{bmatrix}
p_{in}^{(1)} & p_{out}^{(2)} & \cdots & p_{out}^{(2)} \\
p_{out}^{(1)} & p_{in}^{(2)} & \cdots & p_{out}^{(2)} \\
\vdots & \vdots & \ddots & \vdots \\
p_{out}^{(1)} & p_{out}^{(2)} & \cdots & p_{in}^{(1)}
\end{bmatrix}.
\] (1)

Note that this parametrisation is difficult, in the sense that it keeps average in- and out-degree equal throughout the groups, only for odd number of groups. Otherwise the matrix becomes symmetric.

By varying the ratio \( \frac{p_{out}}{p_{in}} \) one can generate, using the planted partition model, networks with different levels of assortativity. Surprisingly, there exists a certain level of the above-mentioned ratio at which the community structure becomes undistinguishable from a random graph. If the ratio is higher than this level, no algorithm is able to infer the real assignments nor the real block structure. We were able to show analytically, using a perturbation approximation, a similar detectability condition for the directed stochastic block model. In the case of directed planted partition it can be rewritten using model parameters:

\[
c \left( \frac{q}{N} \right)^2 < 2p_{in} \left( p_{out}^{(1)} + p_{out}^{(2)} \right)^2 + \left( p_{out}^{(1)} - p_{out}^{(2)} \right)^2 \cot^2 \left( \frac{\pi}{2q} \right),
\] (2)

where \( c \) is the average degree. It is worth mentioning that when \( p_{out}^{(1)} = p_{out}^{(2)} \), the equation simplifies to the one obtained for classical planted partition. We were also able to find, in a similar manner as in [3], a more general result connecting affinity matrix first eigenvalue with the detectability of a given directed stochastic block model.

In the case of directed planted partition there are three parameters, assuming that the number of groups \( q \) is constant. Additionally, if we want to make the networks comparable, we should also fix the average degree \( c \). As a result, one can use two numbers to parametrise the model. We choose a following parametrisation:

\[
\varepsilon = \frac{2p_{in}}{p_{out}^{(1)} + p_{out}^{(2)}}, \quad \gamma = \frac{p_{out}^{(1)}}{p_{out}^{(2)}}.
\] (3)

Numerical results, obtained using belief propagation [6], for particular case of \( q = 3 \) and \( c = 6 \), are shown in FIG 1, together with analytical approximation drawn with a dashed line. There are two distinct phases. An ordered phase with positive overlap and a paramagnetic phase with zero overlap. Interestingly, one can reach the detectable phase by either increasing the assortativity \( \varepsilon \) or by decreasing the symmetry \( \gamma \). In both cases there is a range of values that guarantees detectability, regardless of the other parameter value.

3 Conclusions

We analysed the problem of community detection limits for the directed stochastic block model with asymmetric affinity matrix. We focused on the directed planted parti-
Fig. 1. Two dimensional phase diagram of a directed planted partition with $q = 3$ groups, $N = 90000$ nodes and average degree $c = 6$. The colour scale corresponds to the overlap between the inferred and the real nodes assignment. Dashed black line represents the analytical approximation.

...tion and obtained the detectability threshold as a function of its parameters. Importantly, the results show that detectability can be achieved not only by increasing assortativity but also by increasing asymmetry. Both parameters, above a given threshold, lead to detectability regardless of the value of the other. This was also confirmed by extensive numerical simulations.

References

Comparing Graph Clusterings: 
Set partition measures vs. Graph-aware measures

Valérie Poulin and François Théberge
Tutte Institute for Mathematics and Computing, Ottawa, Canada
vpoulin@gmail.com, theberge@ieee.org

1 Introduction

An impressive number of graph clustering algorithms have been proposed, studied and compared over the past decades. To identify better graph clustering techniques, one needs a way to score the techniques against one another. A typical method is to compare values of some similarity measure between ground truth partitions of given graphs and the partitions produced by the different algorithms on those graphs. However, the choice of the similarity measure used is crucial and has a huge impact on the conclusions made.

In graph clustering comparison studies [7, 11, 10, 4], set partition similarities are used as accuracy measures. Typically, a member of the information-based family [9] such as Adjusted Mutual Information or pair-counting family [6] such as Adjusted Rand Index is used to assess the superiority of a graph clustering algorithm over another. We call these measures graph-agnostic as they ignore the graph structure.

The graph-agnostic measures account for the nodes in a graph but ignore the edges. Should the similarity between partitions A and B, shown in Figure 1, be the same on the left and right graphs? When including the edges, the impact of placing vertex 8 in one part or the other is quite different on both graphs. In this work, we introduce a family of similarity measures with their adjusted forms for graph partitions that take edges into account and prove that both graph-agnostic and graph-aware measures are critical for effectively comparing graph partitions.

Let $G = (V, E)$ be a graph with $V$, the set of vertices and $E$, the set of edges. All graphs considered are undirected. Let $A$ and $B$ denote two connected vertex partitions of $G$ and let $P_A$ denote the pairs of vertices lying in the same part of $A$.

Fig. 1. Partitions A and B are identical as set partitions on both graphs.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
The key value of pair counting similarity measures, $PC(\cdot)$, is $|P_A \cap P_B|$, the number of pairs belonging to the same parts in both partitions. Different normalization functions $f$ (mean–mn, geometric mean, minimum or maximum) are used to ensure values in $[0,1]$ and adjusted versions of those measures, $APC(\cdot)$, must be used in practice [1]:

$$PC_f(A,B) = \frac{|P_A \cap P_B|}{f(|P_A|,|P_B|)}, \quad APC_f(A,B) = \frac{|P_A \cap P_B| - |P_A||P_B|/\binom{\ |P_A| + |P_B| }{2}}{f(|P_A|,|P_B|) - |P_A||P_B|/\binom{\ |P_A| + |P_B| }{2}}.$$ 

2 Graph-aware clustering comparison measures

A connected graph partition $A$ can be described in two ways: a vertex partition or a binary edge classification with endpoints of edges being either in the same part (1) or in different parts (0) of the partition. Hence, a connected partition $A$ induces a binary edge classification $b_A$ in $\{0,1\}^{|E|}$. We use $|b|$ to denote the L1-norm which allows to concisely write $|P_A \cap P_B \cap E|$ as $|b_A \cdot b_B|$. This partition description opens the door to new ways of measuring the similarity between two graph partitions using binary edge classification accuracy measures. Those measures turn out to be equivalent to restricting the pair counting family to an edge counting family. Using a normalization function $f$ as before, we define the graph-aware measures and their adjusted forms as:

$$PC_f(A,B;G) = \frac{|b_A \cdot b_B|}{f(|b_A|,|b_B|)}, \quad APC_f(A,B;G) = \frac{|b_A \cdot b_B| - |b_A||b_B|/|E|}{f(|b_A|,|b_B|) - |b_A||b_B|/|E|}.$$ 

$APC_{mn}$ is effectively the adjusted Rand Index (ARI). We thus refer to $APC_{mn}(\cdot,G)$ as the Adjusted Graph-aware Rand Index (AGRI) and we restrict the rest of our graph-aware/graph-agnostic comparisons to those two measures $PC_{mn}(\cdot)$ and $PC_{mn}(\cdot;G)$.

Different algorithms produce partitions of different sizes and many are known to suffer from the resolution issue [3]. It is therefore interesting to understand how the measures behave on partitions of various resolutions. We define $\mathcal{G}(n,p,q,A)$, a variant of Girvan and Newman model [5] to study a simple family of graphs having community structure. Graphs in $\mathcal{G}(n,p,q,A)$ have $n$ vertices split into a partition $A$: $p$ is the proportion of (randomly selected) pairs of vertices in same parts of $A$ sharing an edge and $q$, the proportion of pairs of vertices in different parts of $A$ sharing an edge.

The following result states that none of the two measures $PC_{mn}(A,B)$, $PC_{mn}(A,B;G)$ directly captures how ‘close’ partition $B$ is to the ground truth partition $A$. Instead, $PC_{mn}(A,B)$ measures how close $B$ is to being a refinement of $A$, whereas $PC_{mn}(A,B;G)$ measures the opposite, how close $B$ is to being a coarsening of $A$. Getting high values with respect to both measures indicates that the partitions are indeed similar. The proof is omitted here.

**Theorem 1** Consider $G_A \sim \mathcal{G}(n,p,q,A)$ with $B_1 > A$ a refinement of $A$ and $B_2 < A$, a refinement of $A$ such that $|P_A|^2 < |P_B_1| \cdot |P_B_2|$. Then

(i) $PC_{mn}(A,B_1) < PC_{mn}(A,B_2)$.

(ii) $E_{G_A}[PC_{mn}(A,B_1;G_A)] > E_{G_A}[PC_{mn}(A,B_2;G_A)]$, if $p > q |P_B_1|/|P_B_2|$. 

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 2. Comparing the similarity curves of a partition and a refinement of it of LFR graphs having varying inter-cluster edge densities. The shaded regions indicate the standard deviations (10 runs).

This shows that the unadjusted versions of the measures penalize refinements and coarsening in opposite ways. In Figure 2, we present empirical evidence that the same is true for the adjusted versions of the measures. We compare the accuracy of two graph partition algorithms: the first level and the last level of a hierarchy of partitions obtained with the Louvain [2] method over a family of LFR [8] graphs that range from clear partition structure (low inter-part edge density) to practically no partition structure (high inter-part edge density). The algorithms compared correspond to two different resolutions, one being a refinement of the other. As one can see, the graph-agnostic and graph-aware measures yield contradicting conclusions. The use of a single measure would yield a false recommendation. However, it is not possible to assess the superiority of any of the two algorithms compared in Figure 2 when using both types of measures: the only valid conclusion is that one produces a refinement and the other a coarsening of the true partition. As a consequence, both measures should be used jointly to assess similarity of graph partitions, using a single measure can lead to wrong conclusions in the study of graph partitioning algorithms.

References

1. A.N. Albatineh, M. Niewiadomska-Bugaj, D. Mihalko, J. of Class. 23
7. A. Lancichinetti, S. Fortunato, Phys. Rev. E 80
8. A. Lancichinetti, S. Fortunato, F. Radicchi, Phys. Rev. E, 78
A Short Introduction to Local Graph Clustering
Methods and Software

Kimon Fountoulakis1, David F. Gleich2, and Michael W. Mahoney3

1 University of Waterloo, Dept. of Computer Science, Waterloo ON N2L 3G1, Canada
2 Purdue University, Dept. of Computer Science, West Lafayette, IN, USA
3 University of California Berkeley, ICSI and Dept. of Statistics, Berkeley, CA, USA

Graph clustering has many important applications in computing, but due to the increasing sizes of graphs, even traditionally fast clustering methods can be computationally expensive for real-world graphs of interest. Scalability problems led to the development of local graph clustering algorithms that come with a variety of theoretical guarantees [1]. Rather than return a global clustering of the entire graph, local clustering algorithms return a single cluster around a given seed node or set of seed nodes. These algorithms improve scalability because they use time and memory resources that depend only on the size of the cluster returned, instead of the size of the input graph. Indeed, for many of them, their running time grows linearly with the size of the output.

In addition to scalability arguments, local graph clustering algorithms have proven to be very useful for identifying and interpreting small-scale and meso-scale structure in large-scale graphs [2, 3]. As opposed to heuristic operational procedures, this class of algorithms comes with strong algorithmic and statistical theory. These include statistical guarantees that prove they have implicit regularization properties [4, 5].

One of the challenges with the existing literature on these approaches is that they are published in a wide variety of areas, including theoretical computer science, statistics, data science, and mathematics. This has made it difficult to relate the various algorithms and ideas together into a cohesive whole. We have recently been working on unifying these diverse perspectives through the lens of optimization [6] as well as providing software to perform these computations in a cohesive fashion [7]. In this note, we provide a brief introduction to local graph clustering, we provide some representative examples of our perspective, and we introduce our software named Local Graph Clustering (LGC).

Local graph clustering

Given a seed node, or a seed set of nodes, the goal of local graph clustering is to compute a cluster “nearby” the seed that is related to the “best” cluster nearby the seed. Here, “best” and “nearby” are intentionally left under-specified, as they can be formalized in one of a few different but related ways. For example, “best” is usually related to a clustering score such as conductance. Formally, local graph clustering can be easily understood as a recovery problem. One assumes that there exists a target cluster in a given graph and the objective is to recover it from one or more example vertices inside the set. We can be more precise for a formulation involving conductance. Assume that there exists a target cluster \( B \) with conductance \( \phi_T \) and we have one seed node in \( B \), our objective is to find a cluster \( A \) that resembles \( B \) with conductance bounded by some function of \( \phi_T \), where resemblance is captured here by normalized precision and recall. Moreover, we want to do this in running time and memory proportional to the size of \( A \).
As a quick example of why local graph analysis is frequently useful in data science applications, we present in Figure 1 the results of finding a good partition of both a random geometric graph and a more typical graph from machine learning and data science. In the random geometric graph, the size of best cluster or community is about half the graph. In this case, an algorithm with runtime that scales with the size of the graph is reasonable. In the graph that is more typical of machine learning and data science applications, the best cluster or community derived from a conductance metric is an exceedingly small fraction of the network. This means that standard graph algorithms whose runtime depends on the size of the graph will do an enormous amount of work to return a tiny portion of the graph. Many other examples of this general phenomenon can be found [2, 3, 9, 10]. Local graph clustering techniques can find this cluster (and other similar clusters that happen not to be globally optimal) while touching not many more edges and nodes than are in the output cluster, greatly reducing the computation time.

Our Software

Local Graph Clustering (LGC) is a Python package that uses C++ routines and brings scalable graph analytics on your laptop. In particular, LGC provides methods that find local clusters, methods that improve a given cluster, tools to compute network community profiles, and multi-class label prediction. The software is on GitHub [7].

Methods in LGC. LGC implements seven local graph clustering methods. Three spectral methods, i.e., approximate PageRank [1], PageRank Nibble [1], $\ell_1$-regularized PageRank [4], and four flow methods, i.e., Max-flow Quotient-cut Improvement [11], FlowImprove [12], SimpleLocal [13] and Capacity Releasing Diffusion [14].

Pipelines. In LGC one can find pipelines which employ the above methods to compute network community profiles (NCPs) [9]. An NCP is a plot that is defined as the the quality of the “best” community as a function of community size in a network. To measure the quality of a community we use conductance. Computing the NCP of a graph is an NP-hard problem, and therefore we compute an approximate version of it using
local graph clustering methods. The same approximation has been suggested also in [2, 9]. LGC also implements multi-class label prediction using local graph clustering as a workhorse [15].

**Scalability.** LGC offers routines to work with graphs that scale to the available memory of your system. We have used these routines to study graphs with billions of nodes on large memory machines, and graphs with 117 million edges on a laptop computer by using about 9.4 GB RAM. Examples can be found in the GitHub repository [7].

**Acknowledgements.** This material is based on research sponsored by DARPA and the Air Force Research Laboratory (AFRL) under agreement number FA8750-17-2-0122, as well as the Army Research Office (ARO). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and the AFRL or the U.S. Government.

**References**

7. Fountoulakis, K., Meng, L., Gleich, D. F., Mahoney, M. W.: Local Graph Clustering (2018). GitHub repository https://github.com/kfoynt/LocalGraphClustering
Community detection by modular decomposition of random walk

Hiroshi Okamoto\textsuperscript{1,2}, Xule Qiu\textsuperscript{3}

\textsuperscript{1}RIKEN Center for Brain Science. Saitama 351-0198, Japan
\textsuperscript{2}Dwango Artificial Intelligence Laboratory. Tokyo 113-0033, Japan
hiroshi.okamoto299792458@gmail.com
\textsuperscript{3}Research & Development Group, Fuji Xerox Co., Ltd. Kanagawa 220-8668, Japan

1 Introduction

Discovering communities is primarily important for understanding functional properties of complex systems described by networks. Therefore, a lot of community detection algorithms have been proposed so far. It is noticeable that many of outperforming ones can be formulated by random walk. The map equation \cite{1, 2} detects communities by searching for the most parsimonious way to describe a random walk on the network. Comparative studies using benchmark networks have demonstrated the best performance of the map equation \cite{3}. The modularity maximization has been most widely used as a standard of community detection. The modularity maximization was first proposed in the framework seemingly different from random walk \cite{4}, but it has later been shown that the algorithm can be more generally formulated by random walk \cite{5-7}. These suggest that using random walk is an effective approach to designing community detection algorithms. Here we put forward an alternative random-walk formulation of community detection. Our formulation can genuinely detect overlapping between communities and their hierarchical organization, which are essential characteristics of community structure of real-world networks.

2 Methods

Let $p(n)$ be the probability that a random walker is at node $n$, which can be easily calculated as the steady state of the Markov chain: $p_n = \sum_{m=1}^{N} T_{mn} p_{m-1}$, where $T_{mn}$ ($m = 1, \cdots, N$) is the transition probability from nodes $m$ to $n$ with $N$ being the total number of nodes. Let $p(n|k)$ be the probability that a random walker is at node $n$ conditioned that he/she is staying in community $k$. The probability distribution $p(n)$ over the entire network can be decomposed as a linear combination of localized probabilities $p(n|k)$ as

$$p(n) = \sum_{k=1}^{K} \pi(k) p(n|k),$$

(1)
where $K$ is the putative number of communities and $\pi(k)$ is the probability of community $k$. Community detection is achieved by solving the decomposition (1) w.r.t. $\pi(k)$ and $p(n|k)$. Following the latent-variable method to solve probabilistic mixture modelling, we can derive the EM algorithm in the following form:

**E-step**

$$r_k = \pi(k) \prod_{n=1}^{N} \left[p_n(n|k)\right]^{\tau_n^{(t)}} \sum_{k=1}^{K} \pi(k) \prod_{n=1}^{N} \left[p_n(n|k)\right]^{\tau_n^{(t)}}, \quad (2)$$

**M-step**

$$p_l(n|k) = \chi \sum_{m=1}^{N} T_{mn} p_n(m|k) + (1-\chi) \sum_{i=1}^{L} p_i r_i \tau_n^{(t)}/2\pi(k) \quad (0 \leq \chi \leq 1), \quad (3)$$

$$\pi(k) = \sum_{i=1}^{L} p_i r_i, \quad (4)$$

where $p_l = T_{mn} p(\mu)$ is the probability of link $l$ with $\mu$ and $v$ being its initial- and terminal-end nodes, respectively; $L$ is the total number of links; $\tau_n^{(t)} = \delta_n + \delta_v$; $\chi$ is the only parameter involved in the algorithm. The $\pi(k)$ and $p(n|k)$ can be solved in the steady state obtained by recursive calculation of the EM step. The probability that node $n$ belongs to community $k$ is given by $p(k|n) = p(n|k) \pi(k)/p(n)$, which defines soft overlapping between communities. The community to which node $n$ mainly belongs is determined by $\arg \max_k p(k|n)$.

Both the map equation and the modularity maximization rely on greedy search (the Louvain method [8]). In contrast, our formulation uses a more theoretically-principled machine-learning approach. Computational cost of community detection by our algorithm (for a fixed value of $\chi$) scales as $\sim O(LK)$, meaning that it belongs to the fastest class of algorithms to discover soft overlapping between communities.

### 3 Results and Discussion

We demonstrate fundamental properties of our algorithm using Zachary’s karate club network. Despite the putative number of communities $K$ chosen initially, the EM step evolves $\pi(k)$ of only two communities to nonzero and decays those of the $K-2$ others to zero (Fig. 1a). Thus, for the current choice of the parameter value (here $\chi = 0.5$), our algorithm automatically determines the number of communities. Soft overlapping between the two surviving communities is indicated by $p(k|n)$ (Fig. 1b). Identifying the main belonging of each node perfectly recovers the factional separation of the karate club (Fig. 1b, inset).

We have confirmed that $\chi$ controls the resolution of decomposition of the network into communities (data not shown). Inspired by these findings, we propose to extract...
hierarchical organization of communities by quasi-statically (namely, very slowly) changing $\chi$ from a very small value (~0.05) to unity while continuing the EM step. Discrete phase transition intermittently occurs as the value of $\chi$ increases (Fig. 1c). At each point of phase transition, the probabilities $\pi(k)$ of some communities sharply increase whereas those of some others drop to zero, indicating that smaller communities merge to form larger ones. Stable $\pi(k)$ in an interval bounded by one and the next phase transitions correspond to a specific layer of hierarchy.

Results of extensive comparative studies using the map equation and the modularity maximization as baselines will be reported in the conference. We will also show that our formulation provides a possible solution to the resolution/field-of-view limit problems, which are associated with many of existing community detection algorithms including the map equation and the modularity maximization.

Fig. 1: (a) The definite number (here, two) of communities survive as the EM step progresses. (b) Soft belonging of each node to the surviving communities. (c) Hierarchical organization of communities emerges thorough a series of discrete phase transitions.

Acknowledgments. This study was partly supported by KAKENHI (00374067).

4 References

Community Structure Based on Circular Flow in a Large-Scale Transaction Network

Yuichi Kichikawa¹, Hiroshi Iyetomi¹, Takashi Iino¹, and Hiroyasu Inoue²

¹ Faculty of Science, Niigata University, Niigata 950-2181, JAPAN
² Graduate School of Simulation Studies, The University of Hyogo, Kobe 650047, JAPAN

1 Introduction

In general, interactions between individual economic agents are considered to play an important role in the economy. For instance, firms are connected to each other directly or indirectly through their business transactions to form a complex directed network.

Recently, firm-level network analyses based on comprehensive database of interfirm transaction relations have begun to appear [1, 2, 4, 7, 9]. Economists as well as physicists have recognized the importance of taking an explicit account of interfirm links when seeking to understand economic issues such as the origin of business cycles and the possibility of a chain-reaction in firm bankruptcies.

Very recently, we have studied [5] topological and flow structures of a nationwide production network in Japan. The objective of this study is to advance the previous empirical analysis on the industrial flow structure embedded in microscopic supplier-buyer relations with a special emphasis on its circularity. Hierarchy of the production network is expected to emerge from self-organization of supply chain in the industrial system. On the other hand, inner loops of production, giving rise to a nonlinear feedback mechanism to complicate dynamics of the industrial system, can be an engine for economic growth.

2 Data and Methods

The present analysis is based on a big data of 4,974,802 transaction relations between 1,066,037 firms in Japan, collected by the Tokyo Shoko Research, Ltd. in 2016.³ This data virtually accommodates whole industrial activities in Japan. We regard firms as nodes and transaction relations between them as directed links spanning from suppliers to customers. Since information on the volume of each transaction is not available, we assume that all the links have the same weight. To focus on the circular flow structure of the network, we deal with only the GSCC (Giant Strongly Connected Component) of the network. Because circular flow is confined almost inside the GSCC in which any pairs of nodes are connected in both ways. The TSR transaction network has 530,174 firms in the GSCC.

³This is the largest connected component in the network obtained from the original data, containing 99.3% of all active firms listed in the data.
First, we decompose the actual production network into hierarchical flow and circular flow components using the Helmholtz-Hodge decomposition [3, 6]. In general, one can write flow $F_{ij}$ running from node $i$ to node $j$ as $F_{ij} = F_{ij}^{(p)} + F_{ij}^{(c)}$. Here $F_{ij}^{(p)}$ denotes the potential flow from node $i$ to node $j$ which is given by $F_{ij}^{(p)} = w_{ij}(\phi_i - \phi_j)$, where $\phi_i$ is the Helmholtz-Hodge potential associated with node $i$ and $w_{ij}$ is a positive weight for linkage between nodes $i$ and $j$. In the potential flow network nodes are perfectly ranked, because the potential flow runs from a node with higher potential to a node with lower potential. On the other hand, $F_{ij}^{(c)}$ denotes the circular flow component in which incoming flow and outgoing flow are exactly balanced at each node, so that there is no hierarchy among nodes in the circular flow network. The Helmholtz-Hodge potential of nodes in a directed network identifies their hierarchical positions in the flow structure. In contrast, the circular flow component illuminates feedback loops built in the system. And then, to identify important loop structure in the TSR transaction network, we apply the map equation method [8] for community detection to the circular flow network which is the GSCC of the network with the circular component alone.

3 Results and Discussion

The GSCC of the TSR transaction network is decomposable into 18,660 communities by the map equation. Fig. 1 is visualization of the top 10 communities using the Helmholtz-Hodge potential. This is a community detection scheme using only the information of the circular flow. However, all of the largest communities are stretched out in the $z$ direction except for the community 4. Such a shape of the communities indicates that they also possess hierarchical nature. Therefore, both hierarchical and circular structures coexist in the major communities.

The communities 1, 2, 4, and 5 are clusters of firms in which manufacture and wholesales dominate, but they are characterized by supply chain relations of their own products. The community 1 includes many manufacture and wholesales firms of textile and apparel. The community 2 includes wholesales and retail trade of seafood, manufacture of food of seafood and fisheries cooperative. The community 4 includes firms of medical and health services together with those of manufacture and wholesales of pharmaceutical products. The community 5 includes many manufacture and wholesales firms of metal products and construction. The remaining communities 3, 6, 7, 8, and 9 are firm clusters in which construction dominates. Industry distribution of the construction communities are similar to each other, but each of the communities is regionally characterized quite well, localized to a specific prefecture. Measuring the magnitude of contribution to the flow structure from individual firms in each community, we find that the dominant industries which characterize the communities are of high hierarchy and low circularity. On the other hand, firms contributing to the reverse flow from downstream to upstream do not belong to the main industries of the supply chain, but to their complementary industries supporting the main business. For example, in the community 1, firms of manufacture and wholesales of textile and apparel are highly hierarchical, and those of road freight transport, equipment installation work and wholesale trade(machinery and equipment) have a high contribution to the feedback loop.
A more detailed analysis is in progress to provide a new insight into the structure of inter-industry relationships by clarifying the circular flow component hidden in the interfirm transaction network.

This research was supported by MEXT as Exploratory Challenges on Post-K computer (Studies of Multi-level Spatiotemporal Simulation of Socioeconomic Phenomena, Macroeconomic Simulations). We used computational resources of the K computer provided by the RIKEN Center for Computational Science through the HPCI System Research project (Project ID:hp170242, hp180177).

Fig. 1. The 10 largest communities in the circular flow network on the GSCC of the transaction network, visualized in three-dimensional space with three different viewpoints. Nodes are aligned in the z direction according to their values of the Helmholtz-Hodge potential; basically, transaction flows from top to bottom. On the other hand, the x and y coordinates of nodes are determined by the energy minimum principle with a spring-electric model.

References
Part III

Diffusion and Epidemics
Network localization is unalterable by infections in bursts

Qiang Liu and Piet Van Mieghem

Delft University of Technology, Delft, the Netherlands,
{Q.L.Liu,P.F.A.VanMieghem}@TuDelft.nl,
WWW home page: https://nas.ewi.tudelft.nl/

1 Introduction

The near-threshold behavior, i.e. the behavior around the threshold where a phase transition occurs, is of great interest in the study of dynamical processes, because many real complex systems may operate near the phase transition point. One of the most extensively studied dynamical processes in network science is the susceptible-infected-susceptible (SIS) spreading process [1]. For some networks, the SIS epidemic remains restricted into a small subnetwork and does not spread over the whole network for infection strength just above the (mean-field) epidemic threshold. This restricted spreading phenomenon is known as the (metastable) localization of the SIS process [2, 3], and has been studied recently.

In this work, we further study the SIS localization phenomenon. In previous studies, localization of epidemic processes means that only a finite number of nodes is infected in the thermodynamic limit, i.e. when the network size \( N \to \infty \). In this work, the definition of epidemic localization is that the average fraction of infected nodes, i.e. the prevalence, tends to zero for an infinitely large network, but the number of infected nodes is not necessarily finite. In the following part, we first clarify some misconceptions about the SIS localization in previous studies and show the availability of mean-field methods. We point out that the order of the near-threshold prevalence as a function of the network size \( N \) is essential for understanding the influence of the network structure on spreading processes. Motivated by the essence of the prevalence order, we confine ourselves to a mean-field approximation and study a bursty spreading effect which maximizes the near-threshold prevalence by a factor equal to the largest eigenvalue \( \lambda_1 \) of the adjacency matrix of the network. Even though the spectral radius \( \lambda_1 \) diverges with network size \( N \), the spreading bursts cannot change a localized spreading to a delocalized one if the principal eigenvector of the adjacency matrix of the network is localized. Thus, we conjecture that the spreading localization is only determined by the network structure but not by the spreading dynamics.

2 Results

The bursty SIS model is still an SIS model and each infected node can still be cured with rate \( \delta \) as a Poisson process, but the infection (infecting all healthy neighbors) only happens at the time points: \( 1/\beta, 2/\beta, \ldots \) with infection rate \( \beta \) and effective infection rate
\[ \tau := \beta / \delta. \] This bursty SIS model is a limit case of a non-Markovian SIS model \cite{4} and the epidemic threshold is \( \tau^{(1)} := 1 / \ln(\lambda_1 + 1). \) The bursty effect leads to a periodical steady-state prevalence with time, and the ratio between the maximum and minimum prevalence equals to the largest eigenvalue \( \lambda_1 \) when \( \tau \downarrow \tau^{(1)}. \) Just above the normalized epidemic threshold \( \tilde{\tau}_c := \tau / \tau^{(1)} - 1 = 0, \) the maximum steady-state prevalence is \( y_{\max}(\tilde{\tau}) = a_{\max} \tilde{\tau} + o(\tilde{\tau}) \) with

\[
a_{\max} = \frac{2}{N} \left( \frac{\lambda_1 + 1}{\lambda_1} \right) \ln(\lambda_1 + 1) \sum_{i=1}^{N} x_i \lambda_1 \sum_{i=1}^{N} x_i^3
\]

where \( x = [x_1, \ldots, x_N]^T \) is the principal eigenvector and the minimum prevalence is \( y_{\min}(\tilde{\tau}) = a_{\min} \tilde{\tau} + o(\tilde{\tau}) \) with \( a_{\min} = a_{\max} / (\lambda_1 + 1). \) Recall that the steady-state prevalence of the traditional SIS process is \( y_{\text{st}}(\tilde{\tau}) = a_{\text{st}} + o(\tilde{\tau}) \) with \( 5 \)

\[
a = \frac{\sum_{i=1}^{N} x_i}{N \sum_{i=1}^{N} x_i^3}
\]

which is fully determined by the eigenvector localization. If the coefficient \( a \) follows an decay as \( O(N^{-\epsilon}) \) for \( \epsilon > 0 \) in a network with localized principal eigenvector \( x, \) then the maximum coefficient \( a_{\max} \) will also converge to zero as \( a_{\max} = O(N^{-\epsilon} \ln N) \) since \( a_{\max} = 2 \ln(\lambda_1) a < 2 \ln(N) a. \) Thus, the bursty effect cannot change a localized epidemic process to a delocalized one, which is fairly counterintuitive in the star graph where a infected hub can infect all the other nodes just after each burst.

The result is evaluated on synthetic and real networks. As shown in Fig. 1(a), the spreading on ER network is delocalized. Figure 1(b) and 1(c) show the coefficient \( a_{\max} \) of networks with power-law and exponential degree distribution decays with the network size \( N \) and the spreading on those networks is localized. For real networks, we generate corresponding synthetic networks (Power-law or exponential) with similar parameters to real networks and then plot \( a_{\max} \) in Fig. 1(d). Since those synthetic networks are localized and \( a_{\max} \) is similar, the real networks are also localized for bursty SIS model.

We show that the periodical burst of the infection cannot change a localized spreading to a delocalized one near the epidemic threshold. The localization near threshold is determined by the network structure but not the specific spreading pattern. The details can be found in \cite{11}.

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. (a): The coefficient $a_{\text{max}}$ and $a_{\text{min}}$ of ER networks; (b): The coefficients $a_{\text{max}}$ of networks with power-law degree distribution converge to zero with network size $N$; (c): The coefficient $a_{\text{max}}$ of networks with exponential degree distributions; (d): The coefficient $a_{\text{max}}$ of some well studied real networks: Email-URV [6], hep-th [7], PGP [8], astro-ph [7], Internet [9], Email-Enron [10], and cond-mat 2005 [7].

The mixed assortativity property of finite Barabasi-Albert networks and its influence on diffusion times

Maria Letizia Bertotti and Giovanni Modanese
Free University of Bolzano, Italy
marialetizia.bertotti@unibz.it
giovanni.modanese@unibz.it

1 Introduction

In spite of the large number of existing studies on Barabasi-Albert (BA) networks, their two-point correlation functions have been completely analysed only recently by Fotouhi and Rabbat [1], who have given the full expressions of the conditional probabilities $P(h|k)$ in the large network limit for any value of the parameter $\beta$ (the number of child nodes in the preferential attachment process).

Concerning the assortativity properties of these networks, in previous work some estimates of the Newman coefficient $r$ were found [2]. According to these estimates, for large $N$ (number of nodes), $r$ vanishes as $-\ln^2 N/N$. It was therefore generally believed that BA networks are almost uncorrelated, and numerical simulations appeared to confirm this. However, more recent asymptotic estimates [3, 4] yield a different result: $r$ vanishes only as $-\ln^2 N/\sqrt{N}$ for large $N$. It should be recalled that for real networks with the same scale-free exponent ($\gamma = 3$), the $r$ coefficient is always small in absolute value, so even this small total disassortativity is significant.

2 Results

But there is more: by computing the function $k_{nn}(k)$ (average nearest neighbor degree) we have shown that it is strongly decreasing for small $k$ and slowly increasing for large $k$. An example can be seen in Fig. 1. This means that the total slight disassortativity measured by the $r$ coefficient is in fact the result of an unexpected mixed assortative/disassortative behavior of these networks.

This peculiar structural property may be related to some features of dynamics on BA networks. In particular, we have shown [4] that for the Bass innovation diffusion model on a network, finite BA networks exhibit the fastest diffusion among scale-free networks with exponent $\gamma = 3$. This result was obtained by solving numerically the coupled nonlinear differential mean-field equations of the model [5, 6] (a system of $n$ equations for a network with maximum degree $n$) and finding the time of the diffusion peak as a function of the network structure.

Comparisons were made between BA networks, uncorrelated networks, disassortative networks built according to a method by Newman [7] and assortative networks built
Fig. 1. Function $k_{nn}$ for a BA network with $\beta = 1$, $n = 100$ (largest degree; the corresponding number of nodes is $N \approx 10^4$). Note the decreasing character for small $k$ (denoting network disassortativity) and the increasing character for large $k$ (assortativity).

Fig. 2. Time of the diffusion peak (in years) for the Bass model on different kinds of scale-free networks with exponent $\gamma = 3$, as a function of the imitation coefficient $q$. All networks have maximum degree $n = 100$ ($N \approx 10^4$). The $q$ coefficient varies in the range $0.3 - 0.48$, corresponding to a typical set of realistic values in innovation diffusion theory. The publicity coefficient is set to the value $p = 0.03$. The lines with $\beta = 1, 2, 3, 4, 5$ correspond to BA networks with those values of $\beta$. Their assortativity coefficients are respectively $r = -0.104, -0.089, -0.078, -0.071, -0.065$. The disassortative network is built with a method by Newman [7, 4] and has $r = -0.084$. The assortative network is built with our recipe [4], with $\alpha = 1/2$, and has $r = 0.863$. 

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
with a recipe we have recently developed [8]. Results are displayed in Fig. 2. The networks employed had typically maximum degree \( n = 10^2 \), which corresponds to \( N \approx 10^4 \) nodes. It turns out that the fastest diffusion process occurs on BA networks with \( \beta = 1 \).

A related result was found earlier, with different methods, by De Agostino et al. [9]. They devised a Monte Carlo rewiring of (initial) BA networks, which maintains the degree distribution and changes the assortativity coefficient \( r \). In this way, they explored a range of \( r \) approximately between -0.2 and 0.45. The diffusion times on these networks were estimated for the SIS epidemic model, which is closely related to the Bass model, looking at the spectral properties of their adjacency matrices. They found that the assortative character of a network lowers the epidemic threshold and the disassortative character tends to make diffusion faster once it has started. This fits well with our findings. BA networks appear to have good diffusion properties because they are both slightly assortative for large \( k \) (i.e. for the hubs), and disassortative for small \( k \).

References

Disease Risk Assessment of the German Cattle Trade Network: A Static and Temporal Network Analysis

Jason Bassett¹, Inia Steinbach¹, Hartmut H.K. Lentz², and Philipp Hövel¹,³,⁴

¹ Technische Universität Berlin, Germany
j.bassett@tu-berlin.de,
² Friedrich-Loeffler-Institute, Greifswald, Germany
³ Bernstein Center for Computational Neuroscience Berlin, Germany
⁴ University College Cork, Ireland

1 Introduction

The German cattle trade network is one of the major ones in the E.U. and plays a centrally strategic role in the potential spread of cattle disease in case of an outbreak. For that reason it is important to understand its dynamics and multiple network behavioural facets to improve the surveillance and existing policies and practices in an emergency.

In this work we present a multitude of complementary network analysis methods applied on an edgelist dataset of cattle movements in Germany with a daily resolution from 01.01.2010 to 31.12.2014, as well as predictions of worst-case disease spreading scenarios of particular dynamics. Lastly, given that different types of farms, premises or nodes might affect the spreading dynamics we introduce an index based on purchases and sales to classify different types of nodes in question for a given node type-agnostic network.

2 Methodology

Following the methodology of [1] we divide the analysis into two parts: the static and the temporal analysis of the German cattle trade network.

For the static analysis we focus on distinguishing the network in its components, namely the giant strongly and weakly connected components (GSCC and GWCC), the giant in and out components and the tendrils and external nodes. Furthermore, we calculate the range distribution (i.e. the maximum possible outbreak size per node) as well as some standard centrality measures (degree and weight distributions). Lastly, we test the stability of the GSCC by means of targeted node removal.

The temporal analysis consists of approaches such as defining the activity patterns of the network on different time-scales, examining the temporal overlap of network node and edge configurations for selected successive periods in time [2] and revealing the available temporal pathways for a spreading process to manifest itself on the network. To that end, we model an epidemic spread process on the given network (SI) as a worst-case demonstration of a potential spread [3].

Regarding the type of farms’ (nodes’) recognition we employ the purchase-to-sale ratio (PSR) to identify the different farm types. The index was introduced in [4] and is defined as
\[ \text{PSR} = \frac{P - S}{P + S}, \]  
\[ (1) \]

with \( P = \sum_{i=1}^{n} p_i \) and \( S = \sum_{j=1}^{m} s_j \) the total purchases (\( n \)) and sales (\( m \)) performed respectively in an aggregated period of time.

3 Results

As a teaser of the results we present in figure 1 the ‘memory’ of the network on a monthly basis. It is a measure (Jaccard index) of overlap on the edge configuration of the network for time aggregations of 28 days.

![Fig. 1. Edge configuration overlaps over 28 day time aggregations.](image)

Moreover, as a temporal spreading index we draw the probability density function of the shortest path durations together with the path density in figure 2. With this methodology the most probable time-frame for a contagion to occur in the network is 84 days.

Last but not least, we have managed to reveal some marginal cases of farm types with the purchase-to-sale ratio. In particular, we distinguished three cases:

1. \( \text{PSR} = 1 \): slaughterhouses as \( P = 1, S = 1 \) and the farm only buys animals.
2. \( \text{PSR} = 0 \): traders, as they buy and sell equally (\( P = S \)).
3. \( \text{PSR} = -1 \): only sellers. Self-sustaining small farms, which virtually never buy animals.
**Summary.** We examined the German cattle trade network over a 5 year period frame with various tools, taking into consideration its static and temporal aspects separately. We further demonstrated a methodology to assess the accessibility of the network and, given a certain source, to evaluate its worst-case epidemics (SI dynamics). Finally, taking into account the effect of the nature of the farm types on the network, we presented a metric that can sort them according to their input and output animal counts.

**References**

Effective epidemic containment using link importance

Joan T. Matamalas¹, Alex Arenas¹, and Sergio Gómez¹

Dept. d’Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili,
43007 Tarragona, Spain

1 Introduction

Epidemic containment is a major concern when confronting large-scale infections in complex networks [1, 2]. Many works have been devoted to analytically understand how to restructure the network to minimize the impact of major outbreaks. In many cases, the strategies consist in the isolation of certain nodes, while less attention has been paid to the intervention on links [3]. In epidemic spreading, links inform about the probability of carrying the contagion of the disease from infected to susceptible individuals. Here, we confront this challenge and propose a set of discrete-time governing equations that can be closed and analyzed, assessing the contribution of links to spreading processes in complex networks [4]. Our approach allows an scheme for the contention of epidemics, based on deactivating the most important links in transmitting the disease.

2 Epidemic link importance

Let us consider a discrete time SIS dynamics that runs on top of a complex network, of \( N \) nodes and \( L \) edges, with adjacency matrix \( A \), and where each node \( i \) can be in one of two different states \( \sigma \), either susceptible (S) or infected (I). We can say that a link \((i, j)\) between nodes \( i \) and \( j \) is in state SI if \( \sigma_i = S \) and \( \sigma_j = I \). The parameters of the SIS dynamics are the infection and recovery probabilities, \( \beta \) and \( \mu \), respectively.

Our objective is to find an effective strategy to contain the SIS epidemic process through bond percolation. To propagate the epidemics, a link must be either in an SI or IS state. Let us suppose we have a link \((i, j)\) in state IS. First, with probability \( \beta \), node \( i \) can infect node \( j \) through this link, changing to state II. Next, infected node \( j \) may transmit the disease to some of its neighbors. Thus, if we had removed link \((i, j)\), we would have cut this path of infections initiated at node \( i \). This means that, the larger the expected number of infected neighbors of node \( j \), the larger the impact of removing link \((i, j)\) for the spreading of the epidemics. The expected number of infected nodes produced in this way can be expressed as

\[
\bar{n}_{ij} = \beta P(\sigma_j = S, \sigma_i = I) \sum_{r=1}^{N} A_{jr} \beta P(\sigma_r = S | \sigma_j = I),
\]

where \( P(\sigma_r = S | \sigma_j = I) \) is the conditional probability that node \( r \) is susceptible when its neighbor \( j \) is infected. Since this measure is asymmetric, even for undirected networks, and removing an edge affects the propagation of the disease in both directions, we define the link epidemic importance of a link, \( I_{ij} \), as

\[
I_{ij} = \bar{n}_{ij} + \bar{n}_{ji},
\]
3 Epidemic link equations

Using an approach similar to the Microscopic Markov-Chain Approach in [5], it is possible to find equations for the probabilities of links being in any of the epidemic states. Let us denote $\Phi_{ij} = P(\sigma_i = S, \sigma_j = I)$, $\Theta_{ij}^S = P(\sigma_i = \sigma_j = S)$ and $\Theta_{ij}^I = P(\sigma_i = \sigma_j = I)$, for all pairs of neighboring nodes. The evolution of the joint probability $\Phi_{ij}$ of one link depends on $\Phi$, $\Theta^I$ and $\Theta^S$ to the rest of the neighboring links, and the infection rules of the SIS dynamics. Thus, we can write the following equation for each link:

$$\Phi_{ij}(t+1) = \Theta_{ij}^S q_{ij}(t) (1-q_{ji}(t)) + \Phi_{ij}(t) (1-\beta)q_{ij}(t) (1-\mu) + \Phi_{ji}(t) (1-\mu) q_{ji}(t) + \Theta_{ij}^I (1-\mu). \tag{3}$$

See [4] for details. The asymmetry of probability $\Phi_{ij}$ means we need two equations per link, one for $\Phi_{ij}(t+1)$ and another for $\Phi_{ji}(t+1)$. Similarly we can obtain an expression for the symmetric probability $\Theta_{ij}^I$:

$$\Theta_{ij}^I(t+1) = \Theta_{ij}^I(t) (1-q_{ij}(t)) (1-q_{ji}(t)) + \Phi_{ij}(t) (1-\beta)q_{ij}(t) (1-\mu) + \Phi_{ji}(t) (1-\mu) (1-\beta)q_{ji}(t) + \Theta_{ij}^I(t) (1-\mu)^2. \tag{4}$$

Finally, $\Theta_{ij}^S = 1 - \Phi_{ij} - \Phi_{ji} - \Theta_{ij}^I$.

The $q_{ij}(t)$ in equations (3) and (4) stands for the probability that a susceptible node $i$ is not infected by any of its neighbors (excluding node $j$):

$$q_{ij}(t) = \prod_{r \neq j}^{N} \left(1 - \beta A_{ir} \frac{\Phi_{ir}}{\Phi_{ir} + \Theta_{ir}^S}\right). \tag{5}$$

We call this system of $3L$ equations and unknowns our Epidemic Link Equations (ELE) model. It can be solved by iteration, starting from any meaningful initial condition, until fixed values are found. Apart from the solution where all nodes are susceptible, $\Theta_{ij}^S = 1$ for all the links, a non-trivial one appears when the system is above the epidemic threshold, which can be shown to be:

$$\beta_c = \frac{\mu}{\Lambda_{\text{max}}(B)}, \tag{6}$$

where matrix $B$ has components $B_{ij} = (1-\gamma)A_{ij} - Y_k \delta_{ij}$, $k_i$ is the degree of node $i$, and $\gamma = \frac{\beta(1-\mu)}{\mu(2-\mu)+\beta(1-\mu)}$.

4 Epidemic containment

Our approach for effective epidemic containment consists in removing the links with largest link epidemic importance. This is possible once we have solved the ELE model, computing the $k_i$ for all the links in the network using equation (2).

For the assessment of the performance of different containment strategies, we show in Fig. 1 their comparison in terms of the required fraction of removed links to attain...
Fig. 1. Fraction of links removed for total epidemic containment on synthetic networks. Comparison between epidemic importance and other four strategies: (A) eigenscore; (B) edge betweenness; (C) node infectivity; (D) random removal. Each point represents a configuration consisting of a network and a set of epidemic parameters.

total containment, $L_{TC}/L_0$, when applied to a large set of synthetic networks (from ER to BA) and epidemic parameters. The results point to a clear advantage of the link importance method over the node infectivity and random approaches, and better or equal results with respect to edge betweenness and eigenscore. In fact, only eigenscore achieves results comparable to link importance, with a slight advantage for our method.

References

Reframing the Gossip Problem as Information Spreading over Complex, Dynamic Networks

Daniel Goldsmith\textsuperscript{1,2}, Yonah Shmalo\textsuperscript{1}, and Yitzchak Shmalo\textsuperscript{1}

\textsuperscript{1} Yeshiva University, Department of Mathematical Sciences, New York, NY 10033
daniel.goldsmith@mail.yu.edu, yshmalol@mail.yu.edu, yitzchak.shmalo@mail.yu.edu,
\textsuperscript{2} Chainalysis Inc., New York, NY 10003

1 Introduction

This research focuses on peer-to-peer information sharing. Specifically, we analyze Gossip-style communication using the perspective and tools developed in the field of complex networks. Consider a set of $N$ nodes labeled \{A\textsubscript{1}, A\textsubscript{2}, ..., A\textsubscript{N}\}, each possessing a unique piece of information $a_i$. If two nodes are connected by an edge at a time $t$ then they will share all of their current information. That is, suppose $A_i$, with information $a_i$, is connected with $A_j$, with information $a_j$, then, as a result of this connection, both $A_i$ and $A_j$ will have information \{a\textsubscript{i}, a\textsubscript{j}\}. Now suppose $A_j$, with information \{a\textsubscript{i}, a\textsubscript{j}\}, is connected with $A_k$ with information $a_k$. While $A_j$ maintains its information \{a\textsubscript{i}, a\textsubscript{j}\}, $A_j$ and $A_k$ will both have information \{a\textsubscript{i}, a\textsubscript{j}, a\textsubscript{k}\}. The question addressed in this paper is, for any given $N$, what is the minimal number of edges required so that all the nodes have all the information \{a\textsubscript{1}, a\textsubscript{2}, ..., a\textsubscript{N}\}. This problem is referred in the literature as the Gossip Problem[1][2], and it has been demonstrated that for a system of $N$ actors $N \geq 4$, the minimum number of connections that leads to total information exchange is $2N - 4$. In considering, for instance, a system of $N = 10$ self driving cars in which each vehicle needs to know the position of all the other 9, this result is significant for instead of each car needing to communicate its location to all the other cars resulting in $N(N - 1)/2 = 45$ exchanges, we can achieve the same result with only $2N - 4 = 16$ communications. Obviously if $N$ is a much larger number, the difference in the number of communications in the two cases needed can be in the orders of magnitude. Computational results for a Random algorithm whose nodes were allowed to fully spread their information are also examined.

2 Proposed Method

2.1 Deterministic Method

To track the spread of information within the network we introduce a matrix at every time step, $I(t)$, whose row $I_i(t)$ is the information attained by node $A_i$ at time $t$. That is, if $I_{i,j}(t) = 1$ then node $A_i$ has node $A_j$’s information after the first time step (first conversation). Thus it is clear that at time $t = 0$, $I$ is nothing but the identity matrix, as each node has its and only its information. To express the transfer of information upon
conversation between nodes $A_{i(t)}$ and $A_{j(t)}$ at time $t$, we need only multiply the matrix $I$ by the matrix
\[
M^{i,j} = \begin{cases} 
M_{i,j}^{i,j} = 1 \\
M_{j,i}^{i,j} = 1 \\
M_{k,k}^{i,j} = 1 & \text{for all } k < N \\
M_{k,l}^{i,j} = 0 & \text{for } k \neq l
\end{cases}
\]

Computationally this process can also be achieved via pairwise row addition, where
\[
I(t) = I((t-1) + I_j(t-1) = I_j(t), \text{ and}
\]
\[
I(t) = M^{i(t),j(t)}I(t-1).
\]

Thus, $I(t) = \prod_{k=1}^{t} M^{i(k),j(k)}I(0)$, which ultimately leads to
\[
I(2n-4) = \prod_{k=1}^{2n-4} M^{i(k),j(k)}I(0).
\]

It is clear that there exists many sequences $\{M^{i(t),j(t)}\}_{i,j=1}^{2n-4}$ resulting in a matrix $I$ which will have no 0 entries, for if at least one exists (which we know to be the case) then, by symmetry arguments, we have that at least $N!$ similar sequences would also work.

2.2 Random Method

Next, we expand this methodology to a truly Random Algorithm [3] that different from the one already described as it does not guarantee a final stable graph, nor is it limited to using only $2N-4$ edges. The idea here is to consider the probability of reaching a network over which most of the information has been transmitted by random attachments between nodes as a function of how many conversations are allowed to take place (note, as this algorithm lacks the wisdom to connect nodes in an efficient way, we expect that the average number of edges needed to achieve a stable graph should be substantially larger than $2N-4$). For a set of $N$ nodes $\{A_N\}$, a parameter $C$ being the number of allowable conversations, and for some time $t < C$ we take two non-equal random integers $a$ and $b$ from $[1,N]$ and establish a connection between $A_a$ and $A_b$ by multiplying the described matrix $M^{i(t),j(t)} = M^{a,b}$ by $I(t-1)$.

3 Results

Figure 1 shows the number of cliques as a function of conversation value and size of clique. During the early stages of conversations we see the formation of many small cliques, this, due to the fact that any early conversation gives rise to a 2-clique. Later we see the death of such small cliques as they get absorbed by larger ones, and, as many cliques overlap, we see a very large number of large cliques. As $t$ approaches $C$ however, we have a formation and domination of one giant clique.
In Figure 2 the blue curve reflects the size of the largest clique (y-axis) as a function of time (x-axis) under the random method with a maximum of 650 Conversations. The purple curve reflects the number of cliques smaller than \( N/3 \) as a function of conversations (the true values for this curve were divided by 10 so that it can be graphed along with the first curve). As expected, these relatively small cliques dominate the graph until a big enough clique arises and begins to absorb the smaller ones. The precise moment at which this happens is a subject of continued study.

4 Conclusion

As shown, the Gossip Problem can be reframed and better analyzed from the perspective of complex, dynamic networks. In considering questions pertaining to final distributions of information, growth of cliques, and attainment of consensus, we have demonstrated that in the early stages of the random algorithm there are many small cliques in the Information Matrix and as time evolves these cliques merge into fewer, larger ones.

References

ViralRank: A new metric for influencers identification

Flavio Iannelli¹, Manuel Sebastian Mariani², and Igor M. Sokolov¹

1 Institute for Physics, Humboldt-University of Berlin, Germany,
iannelli.flavio@gmail.com
2 URPP Social Networks, University of Zurich, Switzerland

1 Introduction

A pivotal idea in network science, marketing research and innovation diffusion theories is that a small group of nodes – called influencers – have the largest impact on social contagion and epidemic processes in networks. Despite the long-standing interest in the influencers identification problem in socio-economic and biological networks, there is not yet agreement on which is the best identification strategy. State-of-the-art strategies are typically based either on heuristic centrality metrics or on analytic arguments that only hold for specific network topologies or peculiar dynamical regimes. Here, we leverage the recently introduced random-walk effective distance – a topological metric that estimates almost perfectly the arrival time of diffusive spreading processes on networks – to introduce a new centrality metric, ViralRank, which quantifies how close a node is to the other nodes. We show that the new centrality metric significantly outperforms state-of-the-art metrics in detecting the influencers for global contagion processes. Our findings reveal the essential role of the network effective distance for the influencers identification and lead us closer to the optimal solution of the problem.

2 Results

Previous works [? ,?] have pointed out that in order to predict the hitting time of a spreading process in geographically-embedded systems, network topology and the corresponding weight flows play a more fundamental role than the geographical distance. The main idea behind ViralRank is to rank the nodes based on the random-walk effective distance $D_{ij}^\text{RW}(\lambda)$ between pairs of nodes which quantifies almost perfectly the hitting time of a reaction-diffusion spreading process on the network [?], for a given set of spreading parameters contained in the constant $\lambda$.

We define the ViralRank score of a node $i$ as the average random-walk effective distance from all sources and to all target nodes in the network,¹

$$v_i = \lim_{\lambda \to 0} \frac{1}{N} \sum_j \left( D_{ij}^\text{RW}(\lambda) + D_{ji}^\text{RW}(\lambda) \right),$$  \hspace{1cm} (1)

where the effective distance [?]

$$D_{ij}^\text{RW}(\lambda) = -\ln Z_{ij}(\lambda),$$  \hspace{1cm} (2)

¹We assume that the network is connected.
can be interpreted as a free energy for the partition function $Z_{ij}(\lambda)$ at temperature $1/\lambda$, defined as

$$Z_{ij}(\lambda) = \sum_{n=1}^{\infty} e^{n H_{ij}(n)} e^{-\lambda n} = \langle e^{-\lambda n_{ij}} \rangle.$$  \hspace{1cm} (3)

In the partition function, $H_{ij}(n)$ is the hitting-time probability of the random walk with transition probability $P_{ij} = A_{ij}/\sum_k A_{ik}$, where $A_{ij}$ is the adjacency matrix, and $n_{ij}$ is the random-walk hitting time playing the role of an internal energy level. By taking a small-$\lambda$ (high-temperature) expansion, node $i$’s ViralRank score reads (up to a normalization constant)

$$v_i \approx \sum_j \left( \langle n_{ij} \rangle + \langle n_{ji} \rangle \right),$$  \hspace{1cm} (4)

where $\langle n_{ij} \rangle$ is the mean-first passage time from $i$ to $j$ defined recursively as $\langle n_{ij} \rangle = 1 + \sum_{k \neq j} P_{ik} \langle n_{kj} \rangle$ if $i \neq j$, zero otherwise. With this choice, a node $i$ is central if a random walk starting at node $i$ is able to quickly reach for the first time the other nodes in the network and, at the same time, it is well reachable from all other nodes.

**Fig. 1.** Illustration of the ViralRank centrality in terms of the random-walk effective distance $D_{ij}^{RW}$ for different seed nodes $i$ (the central red circles in the figure). The clouds of nodes around each seed node $i$ represent the other nodes $\{j\}$ in the network. Their radial distance from the center of the cloud is proportional to their total random-walk effective distance ($D_{ij}^{RW} + D_{ji}^{RW}$) from the source node $i$; their color ranges from dark-blue (low distance) to white (high distance). The average effective distance yields the ViralRank score $v_i$ (horizontal axis). The cases depicted here evolve continuously from a central node (small $v_i$, left side of the panel) which tends to be close to many other nodes, to a peripheral seed node (large $v_i$, right side of the panel) which tends to be far from the other nodes.

To assess the metrics’ performance in the influential spreaders identification using the susceptible-infected-removed (SIR) model, we compare the score they produce with the score of the nodes by their spreading ability $q_i$. The latter is defined as the average number of nodes in the removed state after the infection process has ended, given that the process was initiated by node $i$—i.e., node $i$ was the only infected node at time $t = 0$. At each time step, each infected node can infect each of its neighbors with probability $\beta$, and infected individuals are removed from the dynamics with probability $\mu$.

We find (Fig. ??) that for all the analyzed datasets, there exists a dataset-dependent upper-critical value $\beta_u$ such that ViralRank is the best-performing metric for $\beta > \beta_u$. 

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
among relevant state-of-the-art metrics. The value $\beta_u$ is always larger than the epidemic threshold $\beta_c$, which confirms that ViralRank is the most effective metric for the identification of influential spreaders for spreading processes in the supercritical regime.

Fig. 2. Contact-network spreading model: a comparison between nodes’ centrality and nodes’ spreading ability $q$ in real networks. The metrics considered here for comparison are the following: degree $k$, k-core centrality $k_c$ [?], random-walk accessibility $a$ [?], LocalRank $l$ [?] and the non-backtracking centrality $\eta$ [?]. Pearson’s linear correlation ($r$) between nodes’ centrality and $q$ as a function of the ratio $\beta/\beta_c$ between transmission rate and epidemic threshold for six empirical networks: (a) 9/11 terrorists, (b) emails, (c) jazz collaborations, (d) network scientists co-authorships, (e) protein interactions and (f) Facebook friendships.

Summary. Our results demonstrate that among state-of-the-art centrality measures, there is no universally best-performing metric; the only consistent conclusion is that ViralRank outperforms all the other metrics for processes sufficiently far from criticality, which is actually the relevant region for real biological or computer-virus epidemics. Therefore, the optimal choice of a metric for identifying the influential spreaders critically depends not only on the considered dataset but also on the parameters of the particular spreading process. Remarkably, in most of the analyzed datasets, not only ViralRank outperforms other metrics in the $\beta > \beta_c$ range, but it also approaches the perfect correlation with the spreading ability, $r(-v; q) \simeq 1$, for specific ranges of $\beta$ values within the supercritical region.

References

Modelling epidemic spreading in urban areas with large-scale agent-based transport simulations

Jürgen Hackl¹ and Thibaut Dubernet²

¹ Institute of Construction and Infrastructure Management, ETH Zurich, 8093 Zurich, Switzerland
hackl@ibi.baug.ethz.ch

² Institute for Transport Planning and Systems, ETH Zurich, 8093 Zurich, Switzerland

1 Introduction

Epidemic spreading is strongly correlated with human mobility. Thus, understanding and modelling human mobility patterns is an important component for the development of realistic epidemic spread models. Nowadays, mobility and traffic have reached a complexity and volume of an unprecedented degree. Millions of people travel billions of miles on international flights each week, while hundreds of millions of people commute on a complex web of highways and railroads most of which operate at their maximum capacity. Despite this increasing connectivity and the ability to visit almost every place on this planet in a matter of days, the magnitude and intensity of modern human traffic have made human society more susceptible to threats intimately connected to human travel. Long range human mobility is now responsible for the rapid geographical spread of emergent infectious diseases [1]. In recent years this has been made evident by several epidemic outbreaks, including the severe acute respiratory syndrome (SARS) in 2003, the 2009 H1N1 influenza pandemic, the Ebola virus in 2013 and the Zika virus epidemic in 2016. In all those cases, the disease started locally but spread rapidly around the globe. Consequently, intense research effort has been devoted during the recent decade to the development of quantitative models for epidemic spreading. Hence, numerous epidemic models have been devised in the past with a wide range of complexity levels. One less complex model is the “classic” approach to model the population and disease spread via a dynamical system, where the dynamics of the infected population is described with a system of coupled non-linear differential equations [2, 3]. With the increasing availability of computer resources, agent-based simulation has become a practical method to study epidemics [4]. Agent-based models are stochastic, spatially explicit, discrete-time, simulation models where the agents represent single individuals which interacting in time and space according to prescribed rules [5]. Such an approach enables to model the entire population in a plausible manner, providing a model of community that acts as if it inhabits the city or the whole country [6, 7]. In this work, disease propagation models to simulate epidemic outbreaks are implemented in a large-scale agent-based transport simulation (MATSim [8]). Based on the agents’ movements and activities a spatial-temporal social network is derived. This complex network is used to study seasonal influenza outbreaks in the metropolitan area of Zurich, Switzerland. The results of the model are compared with the “classical” SIR model.
2 Method

Reproducing real-world behaviour of individuals’ daily path in an urban setting is done with the open source agent-based transport simulation called MATSim, which uses an iterative approach for agent-based dynamic traffic assignments. Individuals choose activities in different locations. Sequences of activities are generated and equilibrated in MATSim based on a co-evolutionary algorithm, which alters the agent’s behaviour from iteration to iteration, trying to find optimal routes, modes and departure times and therefore maximize the total utility of their daily activity schedule [9]. Based on this model, individuals can be identified, who share the same location at the same time and thus have a chance to infect each other. This information is extracted to a spatial-temporal social network where the epidemic spread model is applied.

In order to get a valid result, the epidemic spread model is calibrated based on real-world observations of seasonal influenza occurred in Switzerland in the period 2016/2017 [10]. The dataset contains diagnosed influenza cases by practitioners on a weekly basis. The model results are fitted to the data set using least squares estimation. For implementation purposes, a simulation of a seasonal influenza outbreak in a human population located within the metropolitan area of Zurich, Switzerland (see Fig. 1), is used to implement and illustrate the methodological framework.

Simulation of an influenza epidemic spread in a geographic area is computationally intensive and requires the use of georeferenced data sets and a limited number of individuals interacting in the urban space. Due to computational reasons, the entire population of the area cannot be taken into consideration. For illustration purpose only 1% of the population is used, i.e. one agent represents 100 individuals. In total 15'286 agents, are considered to be involved in an influenza epidemic at an urban scale. Since no behavioural changes of the agents are considered, the results of a single day traffic simulation are reused, only updating the stages (susceptible, infected, recovered) of the agents.

![Fig. 1. Temporal-spatial epidemic spread in the metropolitan area of Zurich.](image)
3 Results

Fig. 1 illustrate the temporal-spatial spread of the virus over one day. Initially, only the most influential agent is infected. At 6:30 he arrives at work where the virus spread to another agent at the facility. Around 12:30 some agents leave the workplace for lunch or end their shift and carries the virus to further. Finally, when the agents come home they also infect their families. At the end of the day, one 50 agents are infected. Fig. 1 (bottom) summarize the number of infected agents over the first day.

Fig. 2 shows the agent-based epidemic spread model and the “classical” compartment SIR model fitted to real data of seasonal influenza in the season 2016/2017 observed in the Zurich area. The best fitted agent-based model from the 21’600 simulations has a least square error of $7.96 \cdot 10^{-6}$ while the SIR model has $3.42 \cdot 10^{-6}$. The SIR model approximates the data with a smooth function, caused by the underlying analytical form, while the agent-based model allows reproducing the non-smooth behaviour.

Summary. This study proposes to link epidemiological modelling and transport modelling in a synergistic way to simulate the outbreak of a communicable disease, such as seasonal influenza, in an urban area where different activities take place during a daily citizens’ routine. The results of disease propagation simulation indicate that the model is successfully able to generate various scenarios of an outbreak in complex and realistic urban settings by incorporating movement in the agent entities. The addition of mobility allows realistic emulation of daily behaviours of individuals of a population that interact among themselves and that perform stationary activities in fixed spatially located areas after moving from one place to another.

References

Maxmin-ω: A New Threshold Model on Networks

Ebrahim L. Patel
Mathematical Institute, University of Oxford, Oxford OX2 6GG
ebrahim.patel@maths.ox.ac.uk

1 Introduction

The maxmin-ω system is an intuitive model of asynchronous dynamics on a network. Each node in this system possesses a state, which is updated at the end of a cycle; the cycle comprises the following processes. First, the neighbourhood nodes transmit their state to node \( i \), which takes some non-zero transmission time. Node \( i \) waits for a fraction \( \omega \in (0, 1] \) of the arriving states before generating its new state, which takes processing time \( \xi_i \); this new state is a function of the incoming neighbourhood states, though the state dynamics are not our focus here. Once this is complete, the node updates its state and simultaneously transmits this state to downstream nodes, where the cycles are reiterated. Denote the time of state change of node \( i \) in cycle \( k + 1 \) by \( x_i(k + 1) \). The following recurrence relation gives the \( (k + 1) \)th update time of (the state of) node \( i \).

\[
x_i(k + 1) = x_{(\omega)}(k) + \xi_i
\]

where \( x_{(\omega)}(k) \) represents the \( k \)th time of the (last of the) fraction \( \omega \) of inputs arriving at \( i \). If there are \( n \) nodes in the neighbourhood of \( i \), then \( x_{(\omega)}(k) \) denotes the arrival time of the \( m \)th input where \( m = \lceil \omega n \rceil \). Overall, maxmin-ω is deterministic, making it simpler than traditional asynchronous schemes that often examine such local interactions as being stochastic (e.g., [1, 2]). Therefore, we expect maxmin-ω to be computationally cheaper than such methods. Moreover, these interactions mimic the dynamics of applications such as neuronal networks [3, 4] and virus transmission [5, 6]. Thus, maxmin-ω fits into the class of threshold models whose roots lie in epidemic spreading [7].

2 Outcomes

We apply the maxmin-ω model to three synthetic weighted networks: (i) a ‘regular’ network, whose nodes are arranged in a ring, and the neighbourhood \( \mathcal{N}_i \) of each node \( i \) is symmetric, i.e., \( \mathcal{N}_i = \{ i - r, \ldots, i - 1, i, i + 1, \ldots, i + r \} \), where neighbourhood size \( n = 2r + 1 \); (ii) a network whose nodes take random neighbourhoods with size normally distributed about mean 20; (iii) a scale-free network, where neighbourhood sizes \( n \) are determined by the probability distribution \( P(n) = 0.5n^{-\alpha} \), for various \( \alpha \) values.

Edge weights in these networks represent transmission times. Thus, we consider the monotonic increasing sequence of update times of each node. The differences in these update times form a periodic sequence after a transient time \( K \). That is, for \( k \geq K \),
\(\exists \mu \in \mathbb{R} \text{ and } \rho \in \mathbb{N} \) such that \(x_i(k + \rho) = \mu + x_i(k)\). Figure 1 shows \(K\), period \(\rho\), and cycletime \(\chi = \mu/\rho\), averaged over all nodes, as a function of \(\omega\). Most of these results indicate period and transient time peaking when \(\omega \approx 0.5\), whilst cycletime is almost linearly increasing with \(\omega\) [8]. The one departure from this trend comes from the scale-free case when \(\alpha = 1.5\); here, all three graphs are almost constant, suggesting that maxmin-\(\omega\) dynamics on scale-free networks are less interesting (and more predictable) than networks with a more heterogeneous neighbourhood (in-degree) distribution.

For non scale-free networks, we hypothesise that \(\omega\) is a function of the three variables of transient time, period and cycletime. The resulting maxmin-\(\omega\) system can be shown to model the real world application that yields the aforementioned three quantities.

For illustration, consider the topical application of Twitter. We suppose that people are likely to pass information on (‘retweet’) when a sufficient number (\(\omega\)) of their friends have tweeted them. Thus, if the timings of people’s tweets are periodic, a correlation with the periodic behaviour of the maxmin-\(\omega\) system may potentially be established. Tweet dynamics could therefore be modelled by maxmin-\(\omega\).

Finally, there are intriguing unanswered questions regarding the theory and potential applications, such as neural networks and epidemic spreading. Indeed, the key difference between maxmin-\(\omega\) and traditional threshold models is that of feedback. In threshold models such as [7], once the threshold is achieved, the cells stop processing, whereas once the threshold \(\omega\) is achieved, the maxmin-\(\omega\) system waits for the next cycle before iterating the same process, further iterating until at least some periodic behaviour is reached; this periodicity defines ‘viral’ behaviour in a maxmin-\(\omega\) system. Maxmin-\(\omega\) thus provides a new and compelling view of dynamics on networks of the real world.

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. The maxmin-$\omega$ model applied to three different networks. The horizontal axis represents $\omega$ whilst, from left to right, the vertical axis represents mean transient time, mean period $\rho$, and mean cycletime. (a) Regular network of size 20, with neighbourhood sizes $n$ as indicated. For each $n$, mean results are obtained using 100 different scenarios, each of which takes random transmission times from the positive integers up to 10. (b) ‘Random’ network of size 50, with neighbourhood sizes normally distributed about mean 20 and variable standard deviation $\sigma$. Upon fixing this network for various $\sigma$, mean results are obtained from 100 random transmission times from the positive integers up to 10. (c) Scale-free network of size 50, and fixed transmission times (taken as random positive integers up to 10). For each $\alpha = 0.1, 0.5, 1.5$, we obtain mean results from 100 scale-free networks satisfying $P(n) = 0.5n^{-\alpha}$.
Network Reconstruction from NIMFA Viral State Observations of Multiple Epidemic Outbreaks

Bastian Prasse and Piet Van Mieghem
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands
b.prasse@tudelft.nl
p.f.a.vanmieghem@tudelft.nl

1 Introduction

Modern epidemiology evolved into a research field which encompasses various spreading processes, such as the spread of opinion, trends and innovations. The defining characteristic of epidemic processes is that the spreading may occur from an individual to another individual only if the two individuals have contact, for instance by a friendship (in real life or on online social networks). The network between individuals is thus crucial to understanding the dynamics of the epidemic processes. However, in most real-world scenarios the contact network is unknown, but can be reconstructed from observing the epidemic process over time, which is the focus of this work.

2 System Model

We assume that $K \geq 1$ epidemic outbreaks occur. For a network with a symmetric $N \times N$ adjacency matrix $A$ with the elements $a_{ij} \in \{0,1\}$ and for the $l$-th epidemic outbreak, we denote the viral state of node $i$ at discrete time $k$ by $v_{i}^{(l)}[k] \in [0,1]$. The viral state $v_{i}^{(l)}[k]$ can be interpreted as the probability of the infection of node $i$ or, if node $i$ refers to a group, as a fraction of infected individuals. We model the viral state evolution by the discrete-time version of the NIMFA epidemic model [1]

$$v_{i}^{(l)}[k+1] = (1-\delta)v_{i}^{(l)}[k] + \beta(1-v_{i}^{(l)}[k])\sum_{j=1}^{N} a_{ij}v_{j}^{(l)}[k], \quad i = 1, \ldots, N, \quad (1)$$

where $\delta$ and $\beta$ are the curing and infection rates, respectively. For every epidemic outbreak $l = 1, \ldots, K$ of observation length $n_l$, equal to $n_l$ time slots, we denote the $N \times n_l$ viral state matrix by $V^{(l)} = \left(v^{(l)}[1], \ldots, v^{(l)}[n_l]\right)$. Furthermore, we define the concatenation of the viral state matrices of all outbreaks by the $N \times (n_1 n_2 \cdots n_K)$ matrix $V_{\text{all}} = \left(V^{(1)}, \ldots, V^{(K)}\right)$. Given the concatenated viral state sequence matrix $V_{\text{all}}$, we aim to estimate the adjacency matrix $A$. 
3 Unidentifiability from a Single Epidemic Outbreak

The NIMFA equations (1) are linear in $A$, and we obtain $AV_{all} = M$ for some $N \times (n_1n_2 \cdots n_K)$ matrix $M$, which follows from (1). The adjacency matrix $A$ can be obtained accurately from solving the set of linear equations $AV_{all} = M$ only if the rank of the concatenated viral state matrix $V_{all}$ is sufficiently great. If we observe only one outbreak, i.e. $K = 1$ and $V_{all} = V^{(1)}$, then the rank of the matrix $V_{all}$ is almost always too small for an accurate network reconstruction due to two reasons. First, the network may contain a hierarchical structure that forces the viral state $v^{(1)}[k]$ to remain in a subspace $S$ of $\mathbb{R}^N$, i.e. $v^{(1)}[k] \in S$ for every time $k$. To the best of our knowledge, equitable partitions [2] are the only hierarchical structures that confine the viral state $v^{(1)}[k]$ to stay in a subspace $S$. Second, even if the network does not have an equitable partition, the viral state matrix $V_{all} = V^{(1)}$ has a bad condition number (unless the network size is very small), and thus the problem of solving the set of linear equations $AV_{all} = M$ is numerically ill-conditioned. Only if multiple epidemic outbreaks are observed and used, the viral state matrix $V_{all}$ is sufficiently well-conditioned and, as a result, the adjacency matrix $A$ can be estimated accurately.

4 Network Reconstruction Method

Our network reconstruction method consists of three steps. First, we compute the truncated singular value decomposition (TSVD) [3] of the viral state matrices $V^{(l)}$, which is obtained by setting the smallest singular values to zero. We define $\tilde{V}_{all}$ as the concatenation of the TSVDs of the viral state matrices $V^{(l)}$ of all cascades $l = 1, \ldots, K$. Second, we pose the constrained linear least squares problem

$$\min_A \|A\tilde{V}_{all} - M\|_2^2, \quad \text{s.t.} \quad a_{ij} \in [0, 1],$$

and solve (2) independently for each row of the adjacency matrix $A$, which is beneficial for the computational complexity. In the third and last step, the final estimate for the adjacency matrix $A$ is obtained from the non-binary solution of (2) by rounding the elements to zero or one. Furthermore, we propose an adaptation of the network reconstruction method which estimates both the adjacency matrix $A$ and the spreading parameters $\beta, \delta$. Hence, we generalise the spreading parameter identification method of Paré et al. [4], which assumes the adjacency matrix $A$ to be known.

5 How Many Epidemic Outbreaks are Required for an Accurate Network Reconstruction?

For every epidemic outbreak $l = 1, \ldots, K$, we generate the initial viral state $v_i^{(l)}[0]$ of every node $i$ uniformly in the interval $[0, 0.01]$. Hence, we assume that the initial viral state $v_i^{(l)}[0]$ is close to the healthy state, which is a realistic assumption if the viral state $v_i$ of a node $i$ corresponds to a (geographic) group of individuals. We define the error metric as the scaled one-norm deviation $\varepsilon_A = \frac{1}{2} \|A - \hat{A}\|_1$, where $\hat{A}$ is the estimated adjacency matrix and $L = N(N - 1)/2$ is the maximal number of links for a network of size $N$. Figure 1 depicts the error $\varepsilon_A$ versus the number $K$ of epidemic outbreaks.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. The average fraction $\epsilon_A$ of erroneous links of the estimate of the adjacency matrix $A$ versus the number of epidemic outbreaks $K$ for Erdős-Rényi random graphs with link probability $p = 0.1$ and an effective infection rate $\frac{\beta}{\delta} = 1.1 \epsilon_c^{(1)}$, where $\epsilon_c^{(1)}$ is the epidemic threshold.

6 Conclusions

We introduced a method to reconstruct the network topology from NIMFA viral state observations of multiple epidemic outbreaks. The proposed method is based on singular value decompositions and a constrained linear least squares formulation, which can be solved efficiently. However, since a single epidemic outbreak does not contain sufficient information to reconstruct the network exactly, multiple epidemic outbreaks must be observed for an accurate network reconstruction. Numerical simulations demonstrate that the network can be accurately reconstructed with the proposed method by observing a small number of epidemic outbreaks. The incorporation of a priori knowledge of network properties to improve the performance of the reconstruction method stands on the agenda of future research.

References

Epidemic spreading with awareness and different time scales in multiplex networks

Francisco A. Rodrigues\textsuperscript{1,2,3}, Paulo Cesar Ventura da Silva\textsuperscript{1}, Fátima Velásquez-Rojas\textsuperscript{8}, Colm Connaughton\textsuperscript{2,3}, Federico Vazquez\textsuperscript{7,8}, Yamir Moreno\textsuperscript{4,5,6}

\textsuperscript{1} Institute of Mathematics and Computer Science, Universidade de São Paulo, Brazil.
\textsuperscript{2} Mathematics Institute, University of Warwick, UK.
\textsuperscript{3} Centre for Complexity Science, University of Warwick, UK.
\textsuperscript{4} Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Spain.
\textsuperscript{5} Department of Theoretical Physics, University of Zaragoza, Spain.
\textsuperscript{6} Institute for Scientific Interchange, Turin, Italy.
\textsuperscript{7} Instituto de Física de Líquidos y Sistemas Biológicos, La Plata, Argentina.
\textsuperscript{8} Universidad de Buenos Aires, Buenos Aires, Argentina.

1 Introduction

We propose a mathematical model for the spread of disease with awareness in complex networks. Our model considers the propagation of a disease in a population, simultaneously to the spreading of information about it, by which individuals become aware of the disease and of prevention methods, reducing their contagion probabilities. These two processes run in a double-layer multiplex network: one layer for the disease spreading and another one for the information awareness to hold the disease. As of the definition of a multiplex [3], each layer has the same number of nodes, and there is a one-to-one link between the nodes in different layers. In this sense, we identify each pair of linked nodes from each layer as the same “individual”; the only difference from one layer to the other one lies in the structure of connections inside the layers.

The model for the epidemic spreading adopted here is a reactive SIS (susceptible-infected-susceptible) compartmental model [6]. For the spreading of information awareness to prevent the transmission we use a cyclic Maki-Thomson rumour model in complex networks [4], which we call UARU (unaware-aware-stifler-unaware). In our model, we have six different states, i.e., SU (susceptible and unaware), SA (susceptible and aware), SR (susceptible and stifler), IU (infected and unaware), IA (infected and aware) and IR (infected and stifler). Using these overall states, we define the interaction between the epidemics and awareness by adding two new features. First, a susceptible node that is informed (aware or stifler) will reduce its contagion probability by a factor $\Gamma$ (with $0 \leq \Gamma < 1$) for each contact, meaning that it will get the disease from each of its infected neighbours with probability $\Gamma \beta$ (less than $\beta$). Such feature represents the adoption of prevention methods against the disease. Second, an additional transition called self-awareness is considered: if not informed by a neighbour, an infected-unaware (IU) node can, during the same time step, become aware with probability $\kappa$, by knowing its own condition. This process simulates the case in which an infected subject recognizes the symptoms of the disease and becomes aware of the infection. The
variable $\pi$ controls the velocity of the rumour propagation. With probability $\pi$, only the rumour transitions (awareness, self-awareness, stifling and forgetting) can be promoted during the current time step. With the complementary probability $(1 - \pi)$, the epidemic transitions (infection and recovering) can occur.

## 2 Results

We performed extensive Monte Carlo (MC) simulations of the dynamics described in the last section, where we considered a multiplex network composed by two layers with scale-free organization and $N = 1000$ nodes each. Each layer was generated independently by using the configuration model [1] with power-law exponent $\gamma_{sf} \approx 2.5$ and minimum degree $k_{min} = 4$, with a resulting average degree $\langle k \rangle \approx 7.4$ in each layer. Figure 1 shows these stationary densities for two different values of $\pi$. Each curve is normalized by its value when there is no self-awareness (i.e., $\kappa = 0$). We notice that, in both cases, the self-awareness is beneficial to the disease prevention, as the densities of aware (A) and stifler (R) nodes increase with $\kappa$, thus reducing the density of unaware (U) nodes and the disease prevalence (I). Figure 1 also shows good agreement between Markov chain mathematical formulation (which we do not present here, see [2]) and Monte Carlo simulations.

We can observe an unexpected behavior by making a minor modification to the baseline model. We extend the idea of self-awareness to stifler nodes, considering that a stifler, which is also infected by the disease, is less likely to forget the information. That is, a node who knows about its own infection does not inform other nodes and also impairs the transmission of other nodes, creating additional stiflers around it. This behavior is approximately observed in the case of HIV transmission, in which some infected individuals know about its own infection but do not voluntarily notify their sexual partners [5], acting as infected-stiflers. In figure 2, we can see that for small $\pi$
Fig. 2. Normalized disease prevalence $\rho^*_\kappa(\pi)/\rho^*_\kappa(\kappa = 0)$ vs $\kappa$ for the a) baseline and b) modified models. The values of the timescale parameter $\pi$ increase from the darker to the brighter color, showing how the curves change their behavior with $\kappa$ as $\pi$ increases. Other parameters are set to: $\beta = 1.0$, $\mu = 0.9$, $\gamma = 0.5$, $\alpha = 0.6$, $\Gamma = 0.0$, $\sigma = 0.6$.

(faster epidemics, slower information), the prevalence exhibits its normal decreasing behaviour with $\kappa$ for both baseline and modified models. On the other hand, for larger $\pi$ (slower epidemics, faster information), the curves for the modified model flip their slope for larger $\kappa$ values, whereas they maintain the same behaviour for the baseline model. This means that, when the informational processes are considerably faster than the disease transmission, the self-awareness process can generate too many stiflers and impair the information spreading, increasing the prevalence. Therefore, for both baseline and modified models, the time scale plays an important role on determining the effectiveness of the information on reducing the disease prevalence.

References

Interplay of hubs and cooperation in social contagion

Yongjoo Baek¹, Kihong Chung²‡, Meesoon Ha³, Hawoong Jeong², and Daniel Kim²§

¹ University of Cambridge, Cambridge CB3 0WA, United Kingdom,
² Korea Advanced Institute of Science Technology, Daejeon 34141, Korea,
³ Chosun University, Gwangju 61452, Korea,
msha@chosun.ac.kr (corresponding author)

1 Introduction

As a simple model of social contagion, the generalized epidemic process (GEP), which captures the cooperative effects by changing the strength of infection at the n-th infection attempt on the same node, has been studied on high-dimensional lattices [1] and modular [2] or homogeneous random networks [3, 4] for the special case where the cooperation threshold is as small as possible (n = 2). However, human societies typically feature structurally heterogeneous networks with highly-connected hubs, which can play the role of “superspreaders” [5]. How the presence of such hubs interferes with the cooperative effects still remains poorly understood.

In this work, we address the issue by studying the GEP on random scale-free networks (SFNs) characterized by the degree distribution $p_k \sim k^{-\alpha}$ with $\alpha > 2$. Even after extending the model to general values of n, we show the model to be still exactly solvable in the asymptotic limit. We especially focus on the properties of the GEP near the tricritical points (TCPs), which form the boundary between continuous and discontinuous transition lines, thereby clarifying the conditions for discontinuous transitions and the universal mechanisms of the compact-cluster formation.

2 Model

In Fig. 1(a), we illustrate the GEP for the cooperation threshold n = 3 on a five-node network. Each node is in one of the four states: susceptible (S₁), weakened (S₂), infected (I), and removed (R). In the beginning, all nodes are in the S₁-state, except for a single randomly chosen node (the “seed”) in the I-state, which initiates the contagion. As the dynamics proceed, a randomly chosen I-node attempts to infect all of its S₁- or S₂-neighbors, each of the former (latter) with an independent and identical probability $\lambda$ ($\mu$). Upon success, the target neighbor turns into an I-node. A failed attempt converts an S₁-neighboring to an S₂ if the node has survived $n - 1$ previous infection attempts. After these updates, the I-node immediately deactivates and becomes an R-node, permanently removing itself from the dynamics. The same procedure is repeated until the network runs out of I-nodes, terminating the contagion. The final fraction of R-nodes, which

‡Present Address: Samsung Semiconductor R&D Center, Hwaseong 18448, Korea
§Present Address: Samsung SDS, Seoul 05510, Korea
we denote by \( r \), quantifies the size of an outbreak. In the special case of \( \mu = \lambda \), the process is equivalent to the standard Susceptible-Infected-Removed model, which is a dynamical version of the bond percolation.

3 Results

Phase transitions. — Utilizing the locally tree-like structure of random SFNs with \( \alpha > 2 \), we can exactly calculate the outbreak size \( r \) in the asymptotic limit, see Fig. 1(b) for some examples. According to the theory, an epidemic outbreak occurs at \( \lambda = \lambda_c \equiv \langle k \rangle / \langle k(1-k) \rangle \) as a discontinuous (continuous) transition of \( r \) if \( \mu \) is above (at or below) the TCP, \( \mu = \mu_c \), satisfying

\[
\mu_c = \frac{\min[2, \alpha-2]}{\sum_{m=0}^{n-2} \left( m - \min[3, \alpha-1] \right)} + \left( 1 - \mu_c \right)^{\frac{1}{\lambda}} \left( \frac{1 - \mu_c}{1 - \lambda_c} \right)^{n-1-m}.
\]

The solvability of this equation has the following implications [see Fig. 1(c) for the \( \alpha \)-dependence of \( \lambda_c \) and \( \mu_c \) for different values of \( n \):]

1. If \( \alpha > 4 \), a discontinuous transition at \( \lambda_c \) is possible only for \( n = 2 \). Still, a secondary discontinuous transition may occur at \( \lambda > \lambda_c \), as illustrated in the inset of Fig. 1(b).
2. If \( 3 < \alpha < 4 \), a discontinuous transition is possible for any \( n \geq 2 \).
3. If \( 2 < \alpha < 3 \), \( \lambda_c = 0 \) and \( \mu_c = 0 \).

Table 1. (Tri)critical exponents of the GEP on random SFNs. For \( \alpha > 4 \), \( \beta_t \) and \( \phi \) are defined only for sufficiently large cooperation threshold \( (n \geq 3) \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \beta_t )</th>
<th>( \beta_c )</th>
<th>( \phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha &gt; 5 )</td>
<td>1/2</td>
<td>1</td>
<td>1/2</td>
</tr>
<tr>
<td>( 4 &lt; \alpha &lt; 5 )</td>
<td>1/(( \alpha-3 ))</td>
<td>1</td>
<td>(( \alpha-4 ))/(( \alpha-3 ))</td>
</tr>
<tr>
<td>( 3 &lt; \alpha &lt; 4 )</td>
<td>1</td>
<td>1/(( \alpha-3 ))</td>
<td>4 - ( \alpha )</td>
</tr>
<tr>
<td>( 2 &lt; \alpha &lt; 3 )</td>
<td>(4 - ( \alpha ))/(3 - ( \alpha ))</td>
<td>1/(3 - ( \alpha ))</td>
<td>1/(( \alpha-2 ))</td>
</tr>
</tbody>
</table>
Tricritical properties. — For $\varepsilon_\lambda \equiv (\lambda - \lambda_c)/\lambda_c > 0$ and $\varepsilon_\mu \equiv (\mu - \mu_t)/\mu_t = 0$, the outbreak size $r$ and the outbreak probability $P_\infty$ obey $r \sim \varepsilon_\lambda^{\beta_c}$ and $P_\infty \sim \varepsilon_\lambda^{\beta_c}$. Such tricritical behaviors are dominant in the regime $\varepsilon_\mu \ll \varepsilon_\lambda^{\beta_c}$, beyond which they cross over to critical ($\varepsilon_\mu < 0$) or discontinuous behaviors ($\varepsilon_\mu > 0$). The $\alpha$-dependence of these exponents is shown in Table 1 and Fig. 1(d), whose non-monotonicity reflects the changing role of the hubs as $\alpha$ is varied. In Fig. 2, the crossover is illustrated for the the average number of $R$-nodes, $\langle R \rangle$, obtained using random SFNs of $N$ nodes with $\alpha > 3$. In the limit $N \to \infty$, our theory implies

$$\lim_{N \to \infty} \frac{\langle R \rangle}{N} = rP_\infty = |\varepsilon_\mu|^{\beta_c + \beta_t}/\phi f_\pm \left(\varepsilon_\lambda |\varepsilon_\mu|^{-1/\phi}\right),$$

(2)

where $f_\pm$ (f_-) is the scaling function for $\varepsilon_\mu > 0$ ($\varepsilon_\mu < 0$). The agreement between the theory and the numerics is good for the crossover scaling regime of $\varepsilon_\lambda$ and $\varepsilon_\mu$ as the finite-size effects become dominant in the immediate vicinity of the TCP.

References

Spreading and influence of misinformation and traditional fact-based news in Twitter

Alexandre Bovet\textsuperscript{1,2} and Hernán A. Makse\textsuperscript{3}

\textsuperscript{1} naXys and Department of Mathematics, Université de Namur, Rempart de la Vierge 8, 5000 Namur, Belgium.
\textsuperscript{2} ICTEAM, Université catholique de Louvain, Avenue George Lemaître 4, 1348 Louvain-la-Neuve, Belgium.
alexandre.bovet@uclouvain.be
\textsuperscript{3} Levich Institute and Physics Department, City College of New York, New York, New York 10031, USA

1 Introduction

Recent social and political events, such as the 2016 US presidential election [1], have been marked by a growing number of so-called “fake news”, i.e. fabricated information that disseminate deceptive content, or grossly distort actual news reports, shared on social media platforms. While misinformation and propaganda have existed since ancient times [2], their importance and influence in the age of social media is still not clear.

Here, we characterize and compare the spread of information from websites containing fake news with the spread of information from traditional news websites on the social media platform Twitter using a dataset of more than 170 million tweets covering the five months preceding election day and concerning the two main candidates of the 2016 US presidential election. We compare the networks of users spreading fake and traditional news and find the most influential news spreader of each type of news. To understand the mechanisms of fake and traditional news diffusion and how they influenced the opinion of people during presidential election, we reconstruct the causal network of influence between the time series of activity of top news spreaders and the supporters of each presidential candidate.

2 Results and discussion

We identify 30 million tweets containing a URL directing to a news outlet, sent by 2.2 million users. We classify news outlets among the top 250 shared domain names as spreading fake news and traditional news, from right to left, based on a list of news outlets curated by independent fact-checking and bias ranking organizations. We find that 25% of the tweets linking to news outlets points to websites containing fake or extremely biased news.

Analyzing the information diffusion networks, we find that user diffusing fake news form more connected networks with less heterogeneous degree distributions than users in traditional news diffusion networks.

Influencers of each news category are identified using the collective influence algorithm [3]. While influencers of traditional news outlets are journalists and public figures
with verified Twitter accounts, most influencers of fake news and extremely biased websites are unknown users or users with deleted Twitter accounts.

We use a multivariate causal discovery algorithm [4] to infer the causal effects between the activity of the top news spreaders and the supporters’ activity, identified with a machine learning approach [5]. This causal analysis reveals that influencers of traditional news are driving the activity of the most part of Twitter while fake news top spreaders are, in fact, mostly following the activity of Trump supporters (see Fig. 1).

Our investigation [6] provides new insights into the dynamics of news diffusion in Twitter, namely our results suggests that fake and extremely biased news are governed by a different diffusion mechanism than traditional center and left-leaning news. Center and left leaning news diffusion is driven by a small number of influential users, mainly journalists, and follow a diffusion cascade in a network with heterogeneous degree distribution which is typical of diffusion in social networks [7, 8], while the diffusion of fake and extremely biased news seem to not be controlled by a small set of users but rather to take place in a tightly connected cluster of users that do not influence the rest of Twitter activity. Our results suggest that Twitter activity is mainly driven by traditional center and left leaning news outlets.

![Fig. 1. Graph showing the maximal causal effects (> 5%) between the activity of the top 100 news spreaders of the fake news, center and left leaning media category and the activity of the presidential candidate supporters. Arrows indicate the direction of a the maximal causal effect between two activity time series. The width of each arrow is proportional to the strength of the causation and the size of each node is proportional to the auto-correlation of each time series.](image-url)
References


The effects of trust and influence on the spreading of low and high quality information

Diego F. M. Oliveira¹,², Kevin S. Chan¹

¹U.S Army Research Laboratory, 2800 Powder Mill Rd., Adelphi, MD 20783 USA.
²Network Science and Technology Center, Rensselaer Polytechnic Institute, Troy, NY - USA.

Introduction

The wide number of people adopting online social media as a major communication platform combined with the low cost of information production and broadcasting has changed completely the ways the modern civilization consume and share information. However, such abundance of information, or memes, is exceeding our capacity to consume it, and each piece of information must compete for our scarce attention. As a result, only a tiny fraction of the memes created end up going viral, while the vast majority will simply die out. Such an interesting behaviour caught the attention of the scientific community and the field of information diffusion experienced a significant growth in the past decade. Scientists from a wide range of fields were brought together to better understand the relationship between humans and the new universe of online social media. With more accurate models of the impact of online social media on the user behavior, the designers of such platforms can maximize the performance of their systems. However, if these systems are used without responsibility, they can harm our society spreading panic and helping the spreading of mis-information and fake news. Therefore, understanding how information propagates, and more importantly, how to prevent the spread of low quality information is of vital importance.

Model and Numerical Results

We consider an agent-based model inspired by the long tradition of representing the spread of ideas as an epidemic process where information is passed along the edges of a network. The model consists of a network where each agent is equipped with a memory containing α memes organized in a “first-in-first-out” manner [1]. Additionally, every meme in the system has a numerical value drawn from an uniform distribution representing its fitness or quality. Furthermore, new memes are continuously introduced into the system in an exogenous way and the rate at which this happens ultimately determines the amount of diversity in the system in the sense that the higher information load µ, the higher the diversity and as a consequence, the harsher the competition. We assume that at time \( t = t_0 \) the system is in its state of higher diversity where each node has \( α \) unique memes. At every time step a node \( i \) is selected at random and with probability \( µ \) it introduces a new meme in the system by adding it to its memory and sharing it with all its neighbors. On the other hand, with probability \( 1 - µ \) the selected node chooses a meme from its memory and, than, transmits it to all its neighbors. Once all neighbors receive the meme, it is placed at the top of their memory, and as a consequence, the last meme in each node’s list is removed or forgotten. Additionally, the probability that an agent selects a specific meme \( m \) from its list to transmit is proportional to the meme’s quality \( f(m) \). More explicitly, if node \( i \) has a set of memes \( m(i) \) the probability of meme \( m_k \) being selected is \( P_i(k) = \frac{f(m_k)}{\sum_{j=1}^{f(m)} f(m_j)} \). We let the system run until it reaches steady state.

Once in such a state, we perform measurements to determine the success of a meme recording their quality, popularity, lifetime, cascade size as well as node’s centrality metrics. Here, we define popularity as the number of times a given meme is selected to be transmitted. Lifetime is the time
Figure 1: (a) Average meme’s quality (color scale bar) as a function of the attention $\alpha$ and the information load $\mu$. (b) numbers of memes observed (color scale bar) for each combination of the $\log_{10}(\text{Infection size})$ and the meme’s quality. The white region indicates that no meme was observed. (c) Meme’s average popularity (color scale bar) and (d) the $\log_{10}(\text{Infection size})$ (color scale bar) as a function of the meme’s quality and the node’s centrality. PDF of (e) meme’s popularity and (f) meme’s lifetime for different values of $\delta$. We compared the model predictions $\alpha = 14$ with an empirical distribution obtained by counting the number of occurrences of hashtags and their lifetime from a sample of public tweets. This empirical popularity and lifetime were distributed according to a power-law distribution. Finally, the relationship between meme’s mean popularity and (g) quality and (h) the normalized node’s influence. The control parameters used were $\mu = 0.1$ and $\alpha = 14$.

passed between the meme’s creation and its extinction and finally, the infection size (or cascade size) is defined as the number of unique nodes that were exposed to a given meme. During each simulation, we monitored 4,000,000 memes for each combination of the parameters considered.

We observed that the average quality of the information of the population highly depend on the information load and a high value of $\mu$ in reality hinders the system’s quality (see Figure 1 (a)). The best scenario, of course, is observed when $\mu = 0$ where there is no innovation and the systems converges to a state where only one meme survives, typically, one with very high quality. The quality of the information also plays an important role in determining the number of agents that will be expose to a given meme. While a very large fraction of the them will be consumed by only the immediate neighbors of the source, only the fittest might end up visiting the entire network as shown in Figure 1 (b). Figure 1 (b) shows the behavior of the cascade size as a function of the meme’s quality. The color bar represents the number of memes to reach a given population for different values of meme’s quality. The results indicate that none meme with quality lower than 0.5 visited the entire network, however, as quality increases, the cascade size also increases. Furthermore, to understand the extent to which a piece of is going to be adopted, it is important to understand how influential nodes might affect the dynamics of adoption. Note the presence of highly connected nodes in the network raises the question of whether their content are highly influential and facilitate the spreading mechanism. To do so, we investigated the behaviors of three of the most commons centrality metrics, namely, degree. Figure 1(c-d) shows the behavior of the meme’s quality as a function of the nodes influence. The color in (c) represents the $\log_{10}$ of the average popularity and while the color in (d) represents the $\log_{10}$ of the average infection size (or cascade size) for each combination of the parameters. The results indicate a strong correlation between the success of a meme and the influence of the agent who introduced it. Therefore, in order to attain success, quality by itself does not guarantee that the meme will spread over the entire network, however, a meme coming from a influential node will receive more exposure at the
beginning of its lifetime and as a consequence, it will be shared more often increasing the chances of going viral.

Observe that in the current model, the information transmitted by a given agent will be received by all its immediate neighbors and placed on the top of their attention list. On the other hand, by introducing a parameter \( \delta_{ij} \) representing the trust a user \( i \) has on user \( j \), the neighbors can decide whether or not to accept a meme based on their trust. At every time step, a number \( \psi \in [0, 1] \) is drawn at random and only users with \( \psi \geq \delta \) will pay attention to the memes they were exposed to. Here, we consider two situations, namely (a) \( \delta_{ij} = \delta_{ji} = \delta \) and (b) \( \delta_{ij} \neq \delta_{ji} \). In case (a), we assume that the trust among users is constant while in case (b) the trust is assigned at random. We investigate the behavior of two measures of meme’s success, namely, popularity and lifetime. The distribution of meme’s popularity and lifetime for different values of \( \delta \) are shown in Fig. 1 (e-f). While both distributions have broad power-law tails indicating that only a few memes spread virally through the population. Surprisingly, the introduction of trust does not seem to have a significant effect on the popularity distribution, however, the lifetime exhibits a peak that corresponds to the average time needed for the memes that are not selected to disappear from the network and, as one can see, the lower is the trust between users, the longer it takes. Observe that, our model reproduces perfectly the broad distributions for both popularity and lifetime obtained from empirical data from Twitter. Figure 1 (g) shows the effect of trust on the relationship between quality and meme’s mean popularity. As expected, memes with higher quality have greater chances to spread and as the amount of trust between agents increases, the mean popularity grows more slowly and the curves become more concave, except for the highest values of quality (> 0.8), which indicates a increased selective pressure favoring memes with higher quality. Finally, Fig. 1 (h) shows the relationship between meme popularity and the normalized node’s influence. As one can see, the more connected a node is, the more popular a meme will become. This results reinforce the the likelihood of attain success will be greater for memes with high quality that originated from highly connected or influential agents.

Summary: We considered a very simple agent-based model to study the diffusion of information in an on-line social network. The model allows us to study how the node’s influence as well as the trust between agents affect the dynamics of the system. The results indicate a strong correlation between the success of a meme and the influence of the agent who introduced it showing that quality by itself does not guarantee the success of a meme. When trust was incorporated into the model, we showed that two of the most common metrics of success, namely, popularity and lifetime showed a broad power-law distributions indicating that only a few memes spread virally through the population, while the vast majority will simple die soon after their creation\(^1\) [2].

References


\(^1\)The research was supported by ARL through ARO Grant W911NF-16-1-0524. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.
Inferring properties of the underlying network based on the epidemic prevalence

Long Ma, Qiang Liu, and Piet Van Mieghem

Delft University of Technology, Delft, the Netherlands,
L.ma-2,Q.L.Liu,P.F.A.VanMieghem@TuDelft.nl
WWW home page: https://nas.ewi.tudelft.nl/

1 Introduction

We consider the problem of inferring properties of the underlying network based on the epidemic prevalence $y(t)$, which is the average fraction of the infected nodes. We focus on the susceptible-infected-susceptible (SIS) model, where an infected node can infect each susceptible neighbor with rate $\beta$ and be cured with rate $\delta$. Both the infecting and curing processes are independent Poisson processes. The effective infection rate $\tau$ is defined as $\tau := \frac{\beta}{\delta}$.

We propose a simulated annealing algorithm to generate an estimated underlying network given the effective infection rate $\tau$, the network size $N$, the number of links $L$ and the prevalence sequence $\{y(t_i)\}_{i=1,...,T}$, where $t_i$ is the $i$-th sample point in time and $T$ is the length of prevalence sequence. This estimated network is not the exact reconstruction of the benchmark network, but properties of the estimated network, e.g. the largest eigenvalue $\lambda_1$ and the average clustering coefficient $C_G$, may be close to those of the benchmark network.

2 Results

In the simulated annealing algorithm, the following random rewiring method is adopted: in each step, 5% nodes of the estimated network are randomly selected and their links are removed and then each selected node is reconnected to every other node with probability $p = \frac{2L}{N(N-1)}$ to keep the average degree unchanged.

It is a fact that different types of networks can produce approximately the same prevalence $y(t)$. We generate two networks A and B with network size $N = 1000$ and number of links $L = 5000$. The degree distributions of networks A and B are long-tailed and Gaussian-like, respectively. As shown in Fig. 1, their prevalences are almost the same.

Figure 2 shows the largest eigenvalue $\lambda_1$, the algebraic connectivity $\mu_{N-1}$ and the average clustering coefficient $C_G$ of a benchmark network and 20 estimated networks. The benchmark network in Fig. 2a is an Erdős-Rényi (ER) network with network size $N = 1000$ and number of links $L = 5000$. The result indicates that the largest eigenvalue $\lambda_1$ and the average clustering coefficient $C_G$ are well estimated, but the algebraic connectivity $\mu_{N-1}$ of the estimated network is less accurate. The benchmark network in Fig. 2b is a Barabási-Albert (BA) network with network size $N = 1000$ and number...
of links $L = 5000$. The result shows that the three parameters of the estimated networks are significantly different from the BA benchmark network. The reason for the difference is that the random rewiring method in the algorithm is similar to the generation of ER networks, and thus the estimated network is close to the ER network. As a work in progress, we are also applying other kinds of rewiring methods in the simulated annealing algorithm. For example, the rewiring probability can also be proportional to the degree of each node resembling the construction of scale-free networks. Applying this rich-get-richer rewiring, the estimated network may reflect properties of scale-free networks.

3 Conclusion

Different types of networks, e.g. ER networks and BA networks, may lead to a similar prevalence. If we know the type of the network, then we can choose the corresponding rewiring method in our simulated annealing algorithm to obtain the properties of the network such as the largest eigenvalue $\lambda_1$ and the average clustering coefficient $C_G$ given the epidemic prevalence [1].

Fig. 1. Prevalences $y(t)$ and degree distribution of two different networks with the network size $N = 1000$ and the number of links $L = 5000$.

References

Fig. 2. The algebraic connectivity $\mu_{N-1}$, the average clustering coefficient $C_G$ and the largest eigenvalue $\lambda_1$ of 20 estimated networks and a benchmark network. All the networks are with network size $N = 1000$ and number of links $L = 5000$. a) The benchmark network is an ER network. b) The benchmark network is a BA network.
A three-time-scales random walk on temporal networks

Julien Petit\textsuperscript{1,2}, Martin Gueuning\textsuperscript{2,3}, Timoteo Carletti\textsuperscript{1}, Ben Lauwens\textsuperscript{2}, and Renaud Lambiotte\textsuperscript{4}

\textsuperscript{1} naXys, Namur Institute for Complex Systems, Namur, Belgium, julien.petit@unamur.be,
\textsuperscript{2} Mathematics Department, Royal Military Academy, Brussels, Belgium
\textsuperscript{3} ICTEAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium,
\textsuperscript{4} Mathematical Institute, University of Oxford, Oxford, UK

1 Introduction

With networks increasing their popularity in many fields of science, associated theoretical questions have gained interest. One of these is to obtain the master equation that governs the evolution of the position of a random walker on the nodes of a network. This problem comes with many flavors in the literature, depending on the nature of the support, and on the type of dynamics running on top of it \cite{1}. It has a fair amount of applications, ranging from the theoretical aspects of (anomalous) diffusion to more practical fields such as ecology or neuroscience.

Recently, temporal networks have come more into focus \cite{2}, and random walks and diffusion processes have been studied in such networks where the links activate for an infinitesimal duration \cite{3}. To the best of our knowledge however, the continuous-time random walk on top of a stochastic dynamical graph with finite duration of edges activation is still largely unexplored.

2 Accounting for independent walker and edges time-scales

Based on \cite{4}, we consider a general model where the walker's waiting-time, the up-time and down-time of edges activation are three independent dynamics, as depicted by figure 1. We aim at obtaining a master equation from which we also get the transient dynamics and not only the steady state \cite{5}. To start off, we give a complete analytical and numerical description of this random walk on directed acyclic graphs (DAG’s).

This will in turn serve in the general case, when cycles in the underlying network are allowed. Indeed, we show that the analytical framework for DAG’s can prove applicable even if there are cycles. For this to hold, a proper timescale separation between nodes ($\psi_j$), and edges ($U_{ij}$ and $D_{ij}$) dynamics, or long enough cycles in the network are shown to be required. Indeed, the model on DAG’s is equivalent to the standard active node-centric and passive edge-centric random walks in temporal networks \cite{1}, where edge duration is instantaneous. In the former, the motion is determined by the waiting-time of the walker and, once a jump takes place, all the relevant edges are available. In the latter case, the walker is ready to jump as soon as it arrives on a new node, and it takes the first edge that appears - the walker thus passively follows the appearing edges modeled.
Fig. 1. A directed temporal network. The graph is static between any two random rewiring times. This figure shows the walker observing a random waiting-time on node 1. At some point the walker is ready to jump, but first has to wait until an edge becomes available. Note that the adjacency matrix $A(t)$ is right-continuous and is given by $A(t) = G_i$ for $t_i \leq t < t_{i+1}$. At every random rewiring time $t_i$, almost surely only one edge changes states.

by a renewal process. These two cases correspond to asymptotic regimes described by our model when a timescale dominates over the others. In general, however, the process is determined by the competition of three timescales.

Leaving the applicability domain of the formulas on DAG’s, we discuss the mechanism by which memory arises in the random walk. Even with a naive memoryless jumper choosing its destination node uniformly and with markovian edges (governed by exponential densities), correlations influence the trajectory of the walker, as illustrated with figure 2.

When there are cycles, the state-space becomes the full history of the random walk, which makes it difficult to fully comprehend analytically. We propose a rigorous method that accounts for the presence of short cycles, which can prove to have a dramatic impact on, say, the asymptotic state of the walk. We then confront the analytics to numerical simulations. Eventually, we notice that the proposed model is general enough to tackle a realm of real-world applications, ranging from the theoretical aspects of diffusion to more practical fields such as ecology or neuroscience. It paves the way to generalizing standard algorithms and tools such as PageRank (with a random surfer with an own waiting-time before being ready to click on hyperlinks with fluctuating availabilities) for centrality measures and other methods for community detection in networks. We discuss these applications, including - if time permits - peer-to-peer (P2P) and disruption-tolerant (DTN) networks. A minimal example of these is given by figure 3.

References

The formulas for DAGs are no longer valid if there are cycles, as can be seen from this comparison with Monte-Carlo simulations. On this figure, the stationary state $n_i(t \to \infty)$ in each node for varying values of the rate of the exponential density of the walkers waiting-time $\psi(t)$ is plotted resulting from Monte-Carlo simulation (solid lines with filled markers) and the analytical model on DAGs (dotted lines with empty markers). The width of the shading around the Monte-Carlo curves corresponds to twice the standard deviation of the mean computed on $410^4$ independent trajectories. The up- and down-time durations also follow exponential densities, $U \sim \mathcal{E}(\eta = 1), D \sim \mathcal{E}(\lambda = 1)$, and the initial condition of the walk is $n_1(0) = 1, n_2(0) = n_3(0) = 0$. This example illustrates that when the time-scale of the walker is faster, the memory effect becomes more pronounced and the error with respect to the Monte-Carlo simulations increases. The graph is visible as an inset on the main plot.

A disruption-tolerant network based on mobile wireless sensors. The left schematic represents three independent routes (dotted / blue / dashed), where two any sensors can communicate when located on the same spot $s$. The underlying network of allowed connections is given by the graph on the right-hand side.
Impacts of mutual selection in temporal networks on random walk process

Mei Yang¹ and Bing Wang¹ YueXing Han¹

School of Computer Engineering and Science, Shanghai University, No. 99, Shangda Rd., Baoshan Dis., Shanghai, P.R. China

1 Introduction

With the availability of a larger number of high-resolution data of complex systems, research in the field of complex networks has developed to study the mechanism of network evolution in time-varying networks [1–6]. With the progress in modeling of temporal networks, it allows one to investigate how dynamically evolving networks influence dynamical processes on networks such as random walk process [7, 8], and percolation [11], etc. Among them random walk has always aroused researchers’ interest since it has broadly been applied in various domains, such as searching and routing on networks, calculating vertex’s centrality, detection of network communities [1].

The mechanism of network evolution plays a fundamental role in the dynamical processes. The most simple model of temporal networks is driven by vertices’ activity [11]. Based on this network model, other factors such as individual’s attractiveness [10], memory effect [9] have also been studied. Among them, one typical work is driven by the assumption of individual’s mutual selection [12]. The effect of mutual selection on the network topology has been already well understood on static networks, however, how this mechanism of network evolution takes effect on random walks process has not yet been studied. Here, we investigate the random walk process on temporal networks by considering individuals’ mutual selection mechanism, described by a probability function \( f(x, y) \). We assume that the mutual selection of two individuals follows independent distribution of fitness, i.e., \( f(x, y) = g(x)h(y) \).

Based on the activity driven network model, we incorporate mutual selection mechanism into the model. Then, we analyse the random walk process on this network model. For random walk process, usually we measure the distribution of walkers in the network and the mean first-passage time. We derive analytical solutions for the walkers at the stationary state and the mean first-passage time (MFPT) of the process for each mutual selection function. By theoretical analysis and numerical results, we find that for the independent fitness distribution, the capacity of vertices to gather walkers is not only related to their activity, but also determined by their propensity to receive links. For the case of a threshold fitness function, we find that for vertices with activity larger than a given threshold, the stronger the activity is, the more walkers they will collect in the stationary state. The results contribute to the understanding of the evolutionary mechanism of time-varying networks and give us insight on their effects on random walks.
2 Results

The activity driven network model is described that each vertex is tagged with activity \( a \) following the activity distribution \( F(a) \). Based on the activity driven model of temporal networks [11], we extended the model with mutual selection principle [12]. The destination of a link depends mutually on the fitness of the origin of the link \( x_i \) and that of the destination vertex \( x_j \), expressed in terms of a symmetric linking probability function \( f(x_i,x_j) \). To calculate the distribution of walkers at each class of vertices with activity \( a \), and the mean first-passage distribution, we firstly need to calculate the probability of a random walker at vertex \( i \) at time \( t \), \( P_i(t) \), which satisfies the following condition:

\[
P_i(t + \Delta t) = P_i(t)\left[1 - \sum_{j \neq i} P_{j}^\Delta(t)\right] + \sum_{j \neq i} P_j(t) \prod_{j \rightarrow i}^\Delta,
\]

where \( \prod_{j \rightarrow i}^\Delta \) is the probability that the walker moves from vertex \( i \) to vertex \( j \) in a time interval \( \Delta t \).

The probability that there exists a link between vertex \( i \) and \( j \) is defined as:

\[
\Omega_{i \rightarrow j}^\Delta = \alpha m \Delta t \frac{f(x_i,x_j)}{N \int_0^\infty f(x_i,\xi) \rho(\xi) d\xi},
\]

where \( m \) is the number of links added into the network at each time, \( N \) is the network size. \( \rho(x_i) \) is the fitness distribution of vertex \( i \). Thus, the probability that the walker moves from vertex \( i \) to vertex \( j \) is given by:

\[
\prod_{i \rightarrow j}^\Delta = \Omega_{i \rightarrow j}^\Delta \frac{1}{m + \sum_j \Omega_j^\Delta} + \Omega_{j \rightarrow i}^\Delta \frac{1}{1 + \sum_j \Omega_j^\Delta}.
\]

With the above calculation, we focus on the impacts of mutual selection on the network evolution and investigate their role in the random walk process. We study two kinds of fitness function \( f(x,y) \). We calculate the distribution of walkers at the station state and the mean first passage time for the random walk process for each fitness function.

2.1 Case One: \( f(x,y) = g(x)h(y) \)

We assume that \( f(x,y) = g(x)h(y) \), where the mutual selection of two vertices depends on their own fitness \( g(x) \) and \( h(y) \). By inserting \( f(x,y) = g(x)h(y) \) into Eq.(2), we have

\[
\prod_{i \rightarrow j}^\Delta \approx \frac{\Delta t}{N\langle h \rangle} \left[ a_i h(x_j) + m a_j h(x_i) \right],
\]

The number of walkers at given vertices with activity \( a \) is \( W_a(t) = [N \rho(a)]^{-1} W \sum_{i \in a} P_i(t) \).

By considering the continuous fitness, one obtains the following equation:

\[
\frac{\partial W_a(t)}{\partial t} = -a W_a(t) - \frac{m(a)}{\langle h \rangle} W_a(t) h(a) + \frac{h(a)}{\langle h \rangle} \int d'a' W_{a'} F(a') d'a' + \frac{m a}{\langle h \rangle} \int W_{a'}(t) h(a') F(a') d'a'.
\]

\[ \tag{5} \]
With $\lim_{t \to 0} \dot{W}_a(t) = 0$, one obtains $W_a$ at the stationary state as follows:

$$W_a(t) = \frac{h(a)\phi_1 + ma\phi_2}{a(h) + mb(a)(\langle a \rangle)}$$

(6)

where $\phi_1 = \int aW_a \rho(a) da$ is the average number of walkers leaving from active vertices of class $a$. $\phi_2 = \int W_a(t)h(a)\rho(a) da$ is the average number of walkers who move to vertices of class $a$. $m$ is the number of links added at each time. $\langle a \rangle$ is the average activity.

The total number of walkers in the system is calculated as:

$$W = N \int \frac{h(a)\phi_1 + ma\phi_2}{a(h) + mb(a)(\langle a \rangle)} \rho(a) da$$

(7)

To calculate the mean first-passage time, we first define $p(i,n)$ as the probability that the walker reaches the target vertex $i$ for the first time at time $t = n\delta t$ and $\xi_i$ as the probability that the random walker hops to vertex $i$ in a jump. $p(i,n)$ and $\xi_i$ are given as follows:

$$p(i,n) = \xi_i(1 - \xi_i)^{n-1}.$$  

(8)

At a mean-field level, we can write $\xi_i$ as follows:

$$\xi_i = \sum_j (W_j/W) \prod_{j \rightarrow i} \zeta,$$

(9)

where $\prod_{j \rightarrow i}$ is the propagator. This leads to the mean first-passage time (MFPT) given by:

$$T_i = \frac{\langle h \rangle W}{h(\xi_i)\phi_1 + ma_4\phi_2},$$

(10)

It is interesting to note that the MFPT not only depends on the details of a vertex’s activity but also on its propensity to receive connections. Enhancing the vertex’s activity or its reception capability of connections will attract more walkers to reach it.

### 2.2 Case Two: $f(x,y) = \Theta(x + y - \xi)$

In this case, we assume that links are connected only if their fitness is large than a threshold, that is, $f(x,y) = \Theta(x + y - \xi)$, where $\xi$ is a threshold. With similar calculation as above, we get the number of walkers at vertices with activity $a$ as follows:

$$W_a = \frac{\phi_3 + \frac{ma}{\varepsilon - a}\phi_4}{a + m[\frac{1}{2}(\xi - a) + \frac{1}{4}]e^{2a - \xi}},$$

(11)

where $\phi_3 = \int_{\xi-a}^{\infty} W_a(t)e^{\varepsilon-a}\rho(a) da$, and $\phi_4 = \int_{\xi-a}^{\infty} W_a(t)\rho(a) da$ are constants.

For the vertices with activity higher than the threshold $\xi$, they share the same constant. However, for the vertices with activity less than the threshold $\xi$, the two constants are different for vertices in different activity $a$. Therefore, only if the activity of a vertex
reaches the given threshold, the amount of walkers at the vertex will increase with the enhancement of the activity.

Then, we calculate the MFPT in this case. Again, as calculated in case one, we obtain $T_i$ as follows:

$$T_i = \frac{\Delta t}{\sum_j \left( W_j / W \right) \left( a_j \Delta t \Theta^2 (x_i + x_j - \zeta) + ma_i \Delta t \Theta^2 (x_i + x_j - \zeta) \right) } = \frac{W}{\phi_3 + \frac{ma_i}{\epsilon_i} \phi_4}.$$  \hspace{1cm} (12)

The MFPT for vertices of class $a$ depends on their own activity and the threshold of the system. The larger the threshold $\zeta$ is, the longer the time for a walker to reach the vertex will be. Thus, only if a vertex’s activity is larger than a given threshold, the time a walker takes to reach it will rapidly decrease.

Summary. In this work, we investigate the random walk process on temporal networks with mutual selection. The connectivity probability depends on a symmetric linking probability function $f(x, y)$. We calculate analytical expressions for the number of walkers at a given vertex, $W_a$, and the mean first-message time (MFPT) needed for a walker to arrive at a given vertex. By analyzing two forms of the fitness function $f(x, y)$, we find that both the number of walkers at a given vertex and the mean first-message time depends not only on the vertex’s activity but also its propensity to receive links. Thus, we conclude that mutual agreement mechanism has deep impacts on the random walk process.

References

1 General Introduction

Tackling the burning issue of climate change requires a better understanding of how green technologies spread. Specifically, a "well-below" 2°C warming by the end of the 21st century will not be possible without a large scale deployment of environmental friendly technologies (Mandel, 2018). Although the existing literature on technology diffusion is large, the growing branch of Economic Networks can bring new insights on this issue.

Overall, research on diffusion in networks has focused on a wide range of topics such as diseases (Klovdahl, 1985), rumors (Moreno, 2004) or patenting (Aghion, 2015). Recently, network based approach has been used to evaluate the underlying network’s dynamics of technology diffusion (Mandel 2018; Beaman, 2018). However, questions related to the structure of the network and the associated technology diffusion dynamics are still pending. This paper offers a perspective on this issue.

We investigate how networks’ structure influences the diffusion dynamics of a technology and shed light on relevant policy implications. To this end, we apply the Linear Threshold Model (LTM) introduced by Granovetter (1978) to three broad classes of networks (ie. lattice, small world and random). In addition to the neighborhood threshold, we introduce a second threshold dealing with the technology cost. Regarding the cost function, we assume a learning effect (ie. the cost decreases as the number of adopters increases), matching the case of clean technologies. We then evaluate and compare, as Cowan (2003), the characteristics of the resulting cascades. Finally, we assess the performance of seeding strategies in networks by comparing different level of seed sets (ie. number) and the resulted aggregate diffusion.

Following Teytelboym et al. (2016), irreversibility of our cascade (ie. diffusion) dynamics gives a singular perspective to our research as a considerable part of the literature supposes that agents can switch multiple times (Blume, 1993; Ellison, 1993; Blume, 1995; Young, 2006; Montanari and Saberi, 2010; Adam et al., 2012). Moreover, our paper is the first to consider a second threshold and its associated technology cost function subject to a learning effect.

By assessing and comparing the performance of diffusions (eg. aggregate diffusion, speed of diffusion) in lattice, small-world and random networks, our approach brings new insights for the design of public policies especially regarding the speed of diffusion and the associated aggregate diffusion. Our preliminary simulations answer...
these questions, suggesting that a government intervention should favour a clustered network structure, though the presence of short paths crossing the network is not a barrier to diffusion. Our simulations also suggest that seeding one supplementary agent -compared to a fixed seed set- leads to a more than proportionate increase in technology diffusion (ie. the larger the initial group, the larger the variation in diffusion when seeding one supplementary agent). However, at a certain level (ie. 20), the diffusion increases at a slower rate. This observation illustrates that, for any amount of seeds, it may be relevant for a government to initially target a larger seed set to maximize the diffusion.

2 Methods

To evaluate the diffusion of a technology in networks, we apply the LTM exposed by Granovetter (1978). We call "a switch" an irreversible transition to new state, such as adoption of the technology (Jackson, 2008). All agents in the network are initially switched off. Then, about ten out of one hundred agents are randomly switched, i.e., seeded. Every heterogeneous agent in the network is endowed with two individual thresholds. In the following periods, if the proportion of neighbors that switches exceeds his first threshold and if the cost of the technology falls below his second threshold, the agent also switches (Granovetter, 1978; Schelling, 1978). This process propagates through the network. Once an agent has switched, he remains switched forever. We assume that agents’ thresholds are randomly and independently drawn from a uniform distribution at the start of the cascade (Kempe et al., 2003). We consider this rule to be relevant if the social planner has no reason to believe that some thresholds are more likely than others (Teytelboym, 2016). We test this model on lattice, small-world and random networks following the Watts and Strogatz algorithm (1998).

These networks exhibit singular characteristics (eg. degree of cliquishness, path length) and are well-documented in the literature. We run one thousand simulations with a random seed set of ten agents and their respective random thresholds. We then voluntarily fix the initial seed set to apprehend the impacts that several seeded agents being neighbors have on diffusion. Finally, we evaluate the influence that seeding an additional agent has on the aggregate diffusion level. We follow the same approach as Cowan (2004) for the analysis of our results.

3 Preliminary results and Conclusion

Preliminary simulations suggest that when the probability of rewiring gets closer to the small-world region, the spread of the technology remains nearly the same while moving to random structure decreases the overall diffusion. This result matches expectations that clustering is typically associated with rapid diffusion - as observed in the economic geography literature on innovation diffusion. However, as Acemoglu (2011) pointed, the role of the initial seed set with respect to diffusion is crucial. The existence of clusters in networks makes hard for diffusion to reach a high aggregate level. Therefore, there is a tension between seeding agents in clusters or seeding neighborhood agents. In one hand, seeding an agent in a cluster decreases the clustering barrier and makes it easier for technology diffusion to percolate, but
it does not certify that any neighbors of the seed will adopt. On the other hand, seeding neighborhood agents increases the probability that connected neighbors will adopt, but leaves the cluster issue unconsidered. Interestingly, when the probability of rewiring $p$ gets closer to 1 (ie. random networks), the aggregate diffusion decreases. Moreover, the diffusion gap between lattice/small world and random networks tends to get larger as the number of initial seeds increases. The overall process of diffusion is linear.

With respect to policy implication, the overall diffusion is central. Then, when $p=0$ and $p=0.1$, the network has a highly clustered local structure which maximizes the diffusion. Indeed, aggregate diffusion in random graphs is lower than in clustered graphs because of a very low level of clustering. If small world structures experience the same level of diffusion compare to lattice networks, it is because they still exhibit a high clustering coefficient. Therefore, for the government, maximizing diffusion means targeting a clustered network structure with or without short paths crossing the network (ie. lattice and small world networks). Another policy implication is the necessity to develop regional clustering. For instance, situations such as technology hubs (eg. Silicon Valley) tend to be efficient to deploy the technology. Promoting larger connected hubs of technology would also favour technology diffusion. Finally, our results suggest that at a certain level of seed set (20), the diffusion increases at a slower rate.

4 References

Reachability analysis of discrete state epidemiological models

Gergely Szlobodnyik and Gábor Szederkényi

1 Introduction

We give a computational characterization of the reachability problem of discrete state compartmental epidemiological systems. The problem is reformulated in the form of an integer programming feasibility problem in order to decide the reachability and find feasible state transition sequences. Therefore, by means of the Lenstra algorithm, the decidability can be performed in polynomial time, given the fixed dimension of the problem. The proposed computational approach relies on a discrete state equation for which the existence of a non-negative integer solution is a necessary condition of reachability.

2 Computational model formulation

Since epidemic models can be formally characterized by mass action law chemical reaction networks [3], we formulate the reachability problem in the context of discrete state reaction networks. A discrete state Chemical Reaction Network (d-CRN) is a triple \( \mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R}) \) so that:

\[
\mathcal{S} = \{ s_i | i \in \{1, \ldots, n\} \}
\]

\[
\mathcal{C} = \{ y_j = \sum_{i=1}^{n} \alpha_{ji} s_i | \alpha_{ji} \in \mathbb{Z}_{\geq 0}, j \in \{1, \ldots, m\}, i \in \{1, \ldots, n\} \}
\]

\[
\mathcal{R} = \{ (y_i, y_j) \subset \mathcal{C} \times \mathcal{C} | i \neq j \}
\]

where \( s_i \) is the \( i \)'th species and \( y_j \) is the \( j \)'th complex of the network. A reaction \( y_i \rightarrow y_j \) is represented by an ordered pair \( (y_i, y_j) \). If a reaction \( (y_i, y_j) \in \mathcal{R} \) exists, then \( y_i \) and \( y_j \) are said to be the source and product complexes of the reaction. A complex \( y \in \mathcal{C} \) is said to be charged at state \( X \in \mathbb{Z}_{\geq 0}^n \) if \( X \succeq y \). A reaction \( r \in \mathcal{R} \) is charged if its respective source complex is charged. We say that a state \( X \in \mathbb{Z}_{\geq 0}^n \) reacts to a state \( Y \in \mathbb{Z}_{\geq 0}^n \).
A reaction (vector) sequence \( \sigma_r \) is an ordered set of reaction vectors \( \sigma_r = r_1 \ldots r_l \), where \( r_i \in R \), \( i = 1, \ldots, l \). A state transition sequence \( \sigma_X \) is an ordered set states \( X_0, X_1, \ldots, X_p \) so that \( X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_{p-1} \rightarrow X_p \). A state \( X' \in \mathbb{Z}_{\geq 0}^n \) is reachable from a state \( X \in \mathbb{Z}_{\geq 0}^n \) (denoted by \( X \rightarrow X' \)) if there exists a reaction \( r \in R \) so that \( X \rightarrow X + r = X' \). For each \( (y_j, y'_j) \in \mathcal{R} \), a reaction vector \( r_{ij} \in \mathbb{Z}^n \) can be associated to tracks the net changes of the species upon firing the reaction: \( r_{ij} = y'_j - y_j \). The notation \( r_{ij} \) will be used for denoting both the \( i \)th reaction of the CRN and the associated reaction vector. It is possible to completely describe a d-CRN in an algebraic way using the so-called stoichiometric matrix \( \Gamma = [r_1 \ldots r_m] \) and \( \Gamma^- = [\bar{y}_1^r \ldots \bar{y}_m^r] \), where \( \bar{y}_{rj} \) represents the vector of the source complex associated to reaction \( r_j \).

Using the stoichiometric matrix \( \Gamma \), the necessary condition of reachability is the existence of a \( c \in \mathbb{Z}_{\geq 0}^n \) solution of the state equation \( X_0 + \Gamma c = X' \) where \( X_0 \) and \( X' \) are the initial and target states, respectively. We introduce the following decomposition:

\[
c = \sum_{j=1}^{K} v_j
\]

where \( v_j \in \{0, 1\}^l \), \( \sum_{i=1}^l |v_{ji}| \leq 1 \), \( j = 1, \ldots, K \) and \( K \in \mathbb{Z}_{\geq 0} \) is a prescribed upper bound on the length of reaction sequences. Using Eq. (1), the reachability problem can be expressed in the form of an integer programming feasibility problem [3].

**Problem statement:** consider a subconservative discrete state CRN \( (\mathcal{S}, \mathcal{C}, \mathcal{R}) \) of dimension \( n \geq 3 \). Let us denote the initial state of the system with \( X_0 \in \mathbb{Z}_{\geq 0}^n \). Let \( X' \in \mathbb{Z}_{\geq 0}^n \) be an arbitrary target state. Is it possible to reach \( X' \) from \( X_0 \) along a state transition sequence shorter than or equal to \( K \in \mathbb{Z}_{\geq 0}^n \)?

**Computational solution:** the above problem can be decided in the form of an IP feasibility problems:

\[
\begin{align*}
x_0 + \Gamma c &= X' \\
v_j &\in \{0, 1\}^l & j = 1, \ldots, K \\
\sum_{i=1}^l |v_{ji}| &\leq 1 & j = 1, \ldots, K \\
x_0 + \Gamma \sum_{i=1}^k v_{ji} &\geq \Gamma^- \nu_{k+1} & k = 1, \ldots, K - 1 \\
\sum_{i=1}^l |v_{ji}| &\leq \sum_{i=1}^l |v_{ji+l}^i| & j = 1, \ldots, K - 1
\end{align*}
\]

To decide the feasibility of the reachability problem one needs to find an integer lattice point in the feasibility region characterized by the above set of equalities and inequalities. This can be performed in polynomial time by means of the Lenstra algorithm, given the fixed reachable state space of the problem [4].

We note that the above feasibility problem can be easily equipped with an objective function. E.g. minimizing \( \sum_{i=1}^l [c]_i \) can result in a minimal length state transition sequence. More details on the theoretical correctness of the approach can be found in [3].
3 Example

We consider the SIRS (susceptible-infected-recovered-susceptible) epidemiological model. The state variable $X \in \mathbb{Z}^{3}_{\geq 0}$ tracks the number of individuals in the compartments in a fixed order: susceptible ($S$), infected ($I$) and recovered ($R$).

Fig. 1 Reaction network structure of a SIRS epidemiological model. Note that 0 denotes the zero-complex and the respective transition $I \rightarrow 0$ represents mortality.

Using the above computational approach it can be shown that the discrete state SIRS model can result in an extinction event, i.e. a state where a compartment is empty. We set the initial and target states of the system to be $X_0 = [50 \ 2 \ 0]^\top$ and $X' = [32 \ 0 \ 0]^\top$, respectively. Note that the target state $X'$ represents an extinction because the number of infected individuals is equal to zero. The reachability of $X'$ from $X_0$ is validated. By equipping the feasibility problem with the cost function $\sum_{l=1}^{L} [c]_l$ we determined a minimal-length state transition sequence as depicted in Figure 2.

While the above model is compartmental we note that the IP feasibility based computational approach can be extended to more complex high-dimensional (e.g. individual-level) networks especially to the class of network structures for which the existence of a $c \in \mathbb{Z}^L_{\geq 0}$ solution of the state equation is both necessary and sufficient condition of reachability.

Fig. 2 State transition sequences reaching the same target state $[32 \ 0 \ 0]^\top$ where the disease vanishes. a) the shortest path determined by IP. b) a sample path simulated by the next reaction method.

References

Cost-efficient sentinel surveillance strategies for preventing epidemics on networks

Ewan Colman\(^1\), Petter Holme\(^2\) Hiroki Sayama\(^3\), and Carlos Gershenson\(^1\)

\(^1\) Centro de Ciencias de la Complejidad, Universidad Nacional Autonoma de Mexico, DF, Mexico
ewan.colman@c3.unam.mx
\(^2\) Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan
\(^3\) Center for Collective Dynamics of Complex Systems, Binghamton University, SUNY, USA

1 Introduction

Surveillance can play a crucial role in preventing emerging infectious diseases from becoming epidemic. In circumstances where it is possible to monitor the infection status of certain people, transport hubs, or hospitals, early detection of the disease allows interventions to be implemented before most of the damage can occur, or at least its impact can be mitigated. This presentation addresses the question of which nodes we should select in a network of individuals susceptible to some infectious disease in order to minimize the number of casualties. While some have tackled the problem of optimizing the placement of \(n\) sentinels on a given network [1], here we follow others in considering simple strategies that can be implemented easily and do not require perfect knowledge of the network structure [2–4]. Intuitively, one would imagine that the best strategy would select sentinels that are both well connected and not too close to each other. While most of the previous work has focused on identifying well connected individuals, here we ask whether a strategy that also distributes the sentinels across different regions of the network can be better than one that simply targets the highest degree nodes.

2 Network topology and sentinel performance

The first part of this presentation addresses the question of how network topology affects the performance of different sentinel placement strategies. We generated networks with tunable degree heterogeneity as well as tunable modularity in one case and spatial structure in another. We then simulated the spread of disease on these networks, and measured the performance of three different strategies for placing \(n\) sentinels. To measure performance we counted the number of cases that happened after any one of the sentinels became infected, effectively measuring the number of cases that could potentially be prevented through intervention. We tested the following strategies: \textbf{Modular}: choose the highest degree node in each of \(n\) different modules, \textbf{Spatial}: choose the highest degree node in each of \(n\) different spatial regions, and \textbf{Global}: choose the \(n\) highest degree nodes from the entire network.
The answer to the question what is the best heuristic strategy for placing sentinels on network in order to prevent as many cases as possible? depends on the topology of the network. Figure 1 shows the regions in parameter space for which different strategies perform better; for highly modular or spatially embedded networks it is usually better place the sentinels on nodes distributed across different regions. Whereas, if degree heterogeneity is high, then a strategy that targets network hubs performs better.

![Graph showing performance of strategies](image)

**Fig. 1.** The performance of each strategy is measured by $\Phi$, the number of cases that occurred after detection of the disease by one of the sentinels. Red areas here show the regions in the space of parameter combinations where the modular (or spatial) strategy outperforms the global strategy.

### 3 Sentinel placement with incomplete data

In real-world applications, modular and spatial information might not be available. It is also unlikely that all of the edges in the network would be known. In such cases the network would likely be fragmented into different connected components. We thus introduce the following strategies: **Component:** choose the highest degree node in each of the $n$ largest components of the sample network, and **Proportional:** from each component choose a number of sentinels that is proportional to its size. We apply these strategies to subsamples of empirical networks and present their performance as a function of the sample size. We also explore the hypothesis that a known result from random graph theory can be used as a guide to help decide which strategy to employ.

Our analysis of strategy performance in empirical data shows that as the amount of information we have increases, the performance each strategy improves (see Figure 2). While for the degree strategy this increase is monotonic, the component strategy begins to fail when one component begins to dominate the network. Using a theory proved by Molloy and Reed [5], that states the conditions necessary for the emergence of a giant component in a theoretical network, we can estimate where the component strategy will peak.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
As can be seen in Figure 2, the proportional strategy proves to be at least as good as the component strategy and almost always at least as good as the global strategy. In conclusion, the benefits of sentinel surveillance can be enormous and, done properly, many cases of infectious disease can be prevented. The work we present here utilizes the theory and exploration of complex networks to optimize such decisions in a cost-effective way. We hope that these ideas will eventually be developed and integrated into real-world public health policies.

Fig. 2. Sentinel surveillance strategies applied to subsamples of a high school contact network [6]. The red line shows the value of $\sum_k k(k - 2)p_k$ where $p_k$ is the proportion of nodes that have degree $k$ in the sample network. Based on previous results in random graph theory we hypothesize that the peak sample size for the component strategy corresponds to the sample size for which this expression is equal to zero.

References

An Extension of Autologistic Actor Attribute Models for Multi-Attribute Influence Processes

Brennan Antone\textsuperscript{1}, Alina Lungeanu\textsuperscript{1}, and Noshir Contractor\textsuperscript{1}

Northwestern University, Evanston, IL 60208, USA
BrennanAntone2017@u.northwestern.edu

1 Introduction

Autologistic Actor Attribute Models (ALAAM) are a statistical method for analyzing social influence in cross-sectional network data [3]. We propose an extension to ALAAM to facilitate the analysis of more complex social influence processes. Specifically, we develop Autologistic Actor Multi-Attribute Models (ALAMAM) that estimate simultaneously social influence on multiple node attributes. Allowing multiple binary attributes increases the method’s flexibility, for instance to model the co-diffusion of multiple behaviors or categorical behaviors with more than two values. In addition to allowing for multiple attributes, we extend ALAAMS to include terms for using weighted, directed exogenous networks. This allows for effects such as interaction frequency or tie strength to be incorporated into the model. We also demonstrate how Bayesian estimation can be utilized to quantify uncertainty in the strength of effects estimated by the model.

2 Application

To demonstrate how ALAMAMs cay be used to understand complex social influence processes, we apply the method to data on the use of modern contraception collected from 1281 and 662 individuals in two villages in rural Kenya. Modern contraception use was assessed by asking individuals if they and their partner were currently using modern contraception. Based on previous work applying the Theory of Planned Behavior [1], data were also collected to predict individual’s use of modern contraception based on three individual attributes: their attitudes towards the behavior, perceived norms about the behavior, and perceived control over the behavior. In an extension to the Theory of Planned Behavior, we modeled the impact of an individual’s network on all four node attributes. A network was constructed based upon who a participant reported talked to about modern contraception in the past year, as well as the reported frequency of that contact in the past year. This formed a weighted, directed network.

3 Method

Our model assumes an exogenous network $X$ defined over a set of nodes $N$. It predicts values for multiple attributes, specified by a set $A$, to be exhibited by each node.
The likelihood a node \( n \in N \) exhibits an attribute \( a \in A \) is based on a vector of effect statistics \( Z_{n,a}(y,x) \) and a corresponding vector of weights \( \theta_a \). We define a system state \( Y \) as a unique combination of the attributes \( A \) of all the nodes \( N \). Using an unknown normalizing constant \( \kappa(\theta_a; a \in A) \), we define the probability of a given state \( Y \) as:

\[
P(Y = y | X = x) = \frac{1}{\kappa(\theta_a; a \in A)} \exp\left( \sum_{n \in N} \sum_{a \in A} \theta_a Z_{n,a}(y,x) \right)
\]

(1)

To formulate a specific ALAMAM, we define a set of effect statistics \( Z_{n,a}(y,x) \) for each attribute that will be used to characterize social influence. A \textit{Constant} effect represents a general tendency for all individuals to exhibit that attribute. A \textit{Contagion} effect represents the influence from an alter with an in-coming tie. Additionally, a \textit{Contagion X Frequency} effect considers the influence from an alter taking into account the frequency of the in-coming tie. We estimated a total of twelve effects taking into account the three aforementioned network effects for each of the four nodal attributes: the use of modern contraceptives, and its three predictors based on the Theory of Planned Behavior (attitudes towards the behavior, perceived norms about the behavior, and perceived control over the behavior). Additionally, we estimated the three effects of these predictors on the use of modern contraceptives. This model design makes the simplifying assumption that social influence on attitudes towards the behavior, perceived norms about the behavior, and perceived control over the behavior all occur independently of one another.

To estimate weights \( \theta_a \) for each of the 15 effects, we used Approximate Bayesian Computation [4]. A total of 406,008 samples were generated by using Markov Chain Monte Carlo (MCMC) approaches with a burn-in period of 50,000 steps using the Metropolis-Hastings algorithm [2]. Posterior distributions on the weights were computed assuming a uniform prior spanning from -2 to 2 for each effect, using a subset of 200 samples most accurately replicating the observed data for all four node attributes.

![Fig. 1. Social Influence on Attitude, Perceived Norms, and Perceived Control](attachment:image.png)
4 Results

By utilizing a Bayesian approach, we can distinguish between effects with certain and uncertain direction and magnitude. Figures 1 and 2 depict posterior distributions for the log-odds of effects included in the model. Log-odds that are not uniformly distributed are more certain, and those with values away from zero have higher positive or negative magnitude. Our findings suggest a pattern of anti-contagion effects for attitude, where individual’s attitudes tend to differ from the attitudes of those with whom they discuss modern contraception use. The results indicate that as frequency of contact is higher, there is more likely to be contagion in both perceived norms and perceived control.

5 Discussion

The development of ALAMAM lays groundwork for future models of complex social influence processes. Potential exists for future work to examine how prescriptive analysis, such as stochastic optimization or robust optimization, may use such a model to plan interventions that promote or hinder the diffusion of attributes in a system.

References

Part IV

Dynamics on/of Networks
Synchronization invariance under network structural transformations

Lluís Arola-Fernández¹, Albert Díaz-Guilera²,³, and Alex Arenas¹

¹ Departament d’Enginyeria Informàtica i Matemàtiques
Universitat Rovira i Virgili, 43007 Tarragona, Spain
lluis.arolaf@urv.cat,

² Departament de Física Fonamental, Universitat de Barcelona, 08028 Barcelona, Spain,

³ Universitat de Barcelona Institute for Complex Systems (UBICS), Barcelona, Spain

1 Introduction

The study of dynamical processes running on top of complex networks has become a central issue in many research fields, ranging from the microscopic realm of genes and neurons to the large realm of technological and social systems. However, many times the information we can accede to about the actual topology of interactions is somehow incomplete, because of experimental limitations or because of lags on the details of the system. Moreover, given that the only reflection of the dynamics on networks is usually a certain macroscopic observable, it turns out that many topologies are compatible with the same dynamical output, raising the problem of multi-valuation. Following this perspective, we analyze the relation between function and structure in a novel mapping problem. Given a certain network structure and a dynamical process on top of it, we wonder how to transform the network into a different structural connectivity so that the collective behavior (i.e. the function) remains invariant. Such transformation must adjust the weights of the interactions in the new configuration to achieve an equivalent steady-state functionality to the original structure.

2 Results

Inspired by the derivation of statistical mechanics from information theory as a particular case of statistical inference, see [1], we tackle the functional mapping as an optimization problem for the unknown weights of the transformed network subject to local structural constraints in the system. In particular, we impose an entropy maximization for the weights distribution subject to a detailed balance that constraint, up to a given order, the input strengths of the nodes (the sum of incoming connections).

For the Kuramoto Model of coupled phase oscillators [2], we derive analytical expressions for the resulting weights according to different states of available information, and we show that these transformations are able to preserve the collective behavior of the system (the phase-synchronization between oscillators) even if the mapped networks have very different connectivity patterns and the coupling function between units.
is highly non-linear. Furthermore, we show that the mapping of homogeneous networks into heterogeneous ones is usually less accurate and requires more -costly- microscopic information (higher-order constraints) than the reverse process, unveiling a symmetry-unbalance phenomenon that emerges from the partial impossibility of preserving the main structural properties of the nodes in the transformation.

In Fig.(1), we present the numerical synchronization diagrams (average order parameter \( r^2 \) in terms of the coupling strength between units \( K \)) for homogeneous (Erdős-Rényi) and heterogeneous (power-law with exponent equal to 3) target networks and their respective transformations of weights. All networks are of size \( N = 2000 \), fixed mean-degree \( \langle k \rangle = 10 \) and a uniform distribution of intrinsic frequencies of the nodes.

**Fig. 1.** Synchronization diagrams for target networks \( A_1 \) (Erdős-Rényi) and \( A_0 \) (power-law), which are initially unweighted and symmetric, and its respective transformations \( T(B_1|A_1) \) and \( T(B_0|A_0) \), using a) only local information (local input strengths are preserved) and b) also with first-order information (local and first-neighbours input strengths are preserved).

Summarizing, we present an analytical methodology that successfully produces functional synchronization invariant networks for the Kuramoto Model, by transforming the weights of the interactions, while preserving the underlying topologies, and exploiting only local structural information. The presented formalism can be applied in a wide spectra of existing problems beyond the mapping scenario and provides new analytical insight to tackle real complex scenarios when dealing with uncertainty in the measurements of the underlying connectivity structure. For additional information, check the published work [3].

**References**

Communication and Coordination for DEEP Agents

Carlos Merlos
Air Force Research Laboratory, Rome, New York, USA,
carlos.merlos@us.af.mil

Abstract. We consider a complex hierarchical network made of heterogeneous multi agent system tasked with path planning and task allocation. We begin with partitioning the tasks and agents into sub-cliques based on several criterion. We assume this tasks are in an anti access/area denial environment and thus perfect/constant communication between all agents is not assumed. Our proposed model results in less messaging overhead and is robust against temporary communication loss.

1 Introduction

As we move from the concept of a multi agent system consisting of homogeneous agents with identical characteristics to a more useful but harder to predict Complex Adaptive System with heterogeneous agents possessing varying skills and abilities, we need the methods of comparing costs across agents to be more robust. We take inspiration from nature by considering the various ways that cross species cooperation can lead to mutually beneficial outcomes.

We propose to define a new model to represent a communication paradigm to be used in multi agent simulations. When dealing with a large number of agents, assuming perfect communication creates a network that can grow exponentially. This problem also increases when consensus may be required to achieve an action or decision. Our proposed network paradigm is designed for communication and coordination for Disruptive Environments with Emergent Planning (DEEP) Agents.

The motivation for this new network paradigm stems, in part, from our exploration of including risk-aware stochastic physical search in a heterogeneous multi-agent scenario. Prior work has explored this issue in the single agent case under some specific assumptions, such as rigid, or explicit, travel costs.[1] Agents utilize cooperative game theory for the partitioning of targets having communications modeled using a probability distribution as well.

2 Model

The proposed model begins by partitioning all the agents and tasks in the system into subgroups. See Algorithm 1. There are several task partition criteria that can be applied, e.g. group objects by location. This can be accomplished by creating bounding boxes within a grid world environment or clustering groups with a maximum group size. We can also classify objects by the type of target they are, eg. ISR, cyber, etc. Likewise
we can partition agents by the capabilities they possess. Sub-clique properties can be multidimensional in which they can represent a combination of multiple characteristics. We assume that there is constant communication between all agents of a sub-clique, e.g. they may all be in relatively close proximity to each other or may have a specific communication sensor on board.

**Algorithm 1 Partition**

```plaintext
function Partition(list of objects) ▷ objects can be agents or tasks
    for obj in objects do ▷
        if obj characteristic in is sub-clique property then ▷ length of sub-clique property ≥ 1
            if length of sub-clique < max size then ▷ constant max size or can vary
                sub-clique += obj
            return list of sub-cliques
```

Once the sub-cliques have been formed, agents within each pack divide the tasks they will individually accomplish. This can be implemented by a naïve greedy algorithm or by a more complex approach such as the Consensus Based Auction Algorithm. [2]. One agent is chosen as the pack leader that will disseminate information with other leaders and inform then of the choices made. See Figure 1.

For more complex environments, the hierarchical model shown in Figure 2 can be used where each singular node is representative as a cluster of sub-cliques. Communication is sent via a leader of all pack leaders from a particular cluster.

The information being passed around can be handled differently depending if the simulation environment is permissive or rigid. If the environment is permissive, the information from the leader of leaders can stream upwards to a central node, as seen in Figure 2, for situational awareness. Without a central authority, agents can decide to ex-
ecute tasks automatically, say by following the Asynchronous Consensus-Based Bundle Algorithm paradigm [3]. On the other hand, if the simulation is in a rigid environment where the central node has supreme authority, the consolidated task list must be decided on before action can be taken. The decision will then be disseminated through the same channels and the agents will either carry on with the proposed plan or reorganize. In either permissive or rigid environments, if a pack leader loses communication to a higher level, then either a new leader that can maintain communication is chosen or the sub-clique proceeds with the agreed upon task list until communication is reestablished.

3 Application

This work can be extended to the following scenario: having heterogeneous agents such as UAVs and ground rovers, where the collateral cost may be the loss of an asset that captured a high priority target but did not have enough energy to return to base. For example, a quad-rotor may accept more risk and scout ahead for a trailing rover. A multidimensional cost function that can include energy used by agents, wait times of targets to be captured, or collateral cost of a lost or captured agent can be used. Such collateral cost may vary based on the financial value of the agent or the cost of having an adversary acquire the asset. The acceptable risk taken by a particular agent, based on the inverse of its current standing in the pack or the number of homogeneous agents still alive, will determine if sacrifice for the overall success of the mission is desirable. Agents can use the above mentioned cost function when deciding on a task. Sub-cliques can be established by the type of UAV/rover and their relative position. Running several simulations with varying DEEP agents can yield the minimum number of agents needed to have a probability of success within a certain epsilon.

4 Acknowledgments

We thank the anonymous reviewers for helpful comments and suggestions. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Air Force Research Laboratory, Department of Defense, nor the U.S. Government.

References

Preference in the restructuring mechanisms of time evolving hierarchies

Gergely Palla\textsuperscript{1}, Dániel Zagyva\textsuperscript{2}, Sámuel G. Balogh\textsuperscript{2}, and Péter Pollner\textsuperscript{1}

\textsuperscript{1} MTA-ELTE Statistical and Biological Physics Research Group, Hungarian Academy of Sciences, 1117 Budapest, Pázmány P. stny. 1/A, Hungary
\texttt{pallag@hal.elte.hu, pollner@angel.elte.hu}

\textsuperscript{2} Dept. of Biological Physics, Eötvös University, 1117 Budapest, Pázmány P. stny. 1/A, Hungary
\texttt{zagyva@hal.elte.hu, balogh@hal.elte.hu}

1 Introduction

Hierarchical organisation in complex networks is a wide spread phenomenon, which is supported by numerous studies with subjects ranging from neural networks [1] through flocks of various species [2, 3] to social interactions [4]. Networks representing real systems are subject to constant evolution in most of the cases, and some relevant aspects of the laws forming the shape of networks changing over time have already been uncovered in the scientific literature. Probably most famous is the preferential attachment rule for growing scale-free networks, which is one of the key concepts of the Barabási-Albert model [5], and was detected also by empirical studies of network data [6, 7].

Along the same line, in the present work our aim is to examine the statistical properties of time dependent networks with a hierarchical structure. Our study is based on the data provided by the NCBI about the MeSH terms, which were introduced for helping the search in the PubMed publication database of the NCBI (comprising more than 26 million citations for biomedical literature) at various levels of specificity. Due to the rapidly developing nature of the medical-, biochemical- and biological sciences, the set of available MeSH terms are yearly updated by the administrators of PubMed.

The directed networks we consider are based on the classifications provided by PubMed, specifying at least one parent for any available MeSH term, except for the roots of the hierarchies. There are altogether 16 different roots, and the total number of descendants of the individual roots (the sizes of the hierarchies) varies roughly between a 1,000 and a 10,000 nodes, whereas the time span of our analysis is 14 years. The links in our network representation are pointing from the parents to their children. Since a part of the MesH terms have multiple parents, the studied networks are not strictly tree-like, instead they correspond to a directed acyclic graph. Due to the yearly updates, the structure of these networks is subject to constant evolution: new MesH terms can appear, terms becoming obsolete can be deleted or be merged with other terms, and also already existing parts of the network may be rewired.

Our goal is to examine how do the different topological- and hierarchical properties of the nodes influence the attachment and detachment of links during the restructuring. Understanding the nature of these processes can help the creation of hierarchy evolution,

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
models that can predict which part of the hierarchy is most likely to be rewired in the future, and what is the expected change in the overall features of the hierarchy.

2 Methods

Since we are interested in the possible preference with respect to node properties, we can consider only those changing links where at least one of the endpoints was present already before the given attachment or detachment event, and thus, we have the possibility to compare the properties of the endpoint to the same properties of other nodes in the same time step. The method we use for detecting whether the attachments/detachments are uniform with respect to a given property \( x \), or instead show preference towards high (or low) values of \( x \) is based on comparing the distribution of \( x \) for the chosen nodes during the change event and the distribution of \( x \) amongst the available nodes [7]. Let \( w_{t,t+1}(x) \) denote the number of chosen nodes in the attachments/detachments between \( t \) and \( t+1 \), having a property value at least as large as \( x \). Similarly, let \( P_t(x) \) denote the fraction of nodes in the hierarchy at \( t \) having a property value \( x \) or larger. (Thus, \( P_t(x) \) is basically the complementary cumulative distribution of the examined property at \( t \).)

By aggregating

\[
W(x) = \sum_{t=1}^{t_{\text{max}}-1} \frac{w_{t,t+1}(x)}{P_t(x)},
\]

we obtain a function \( W(x) \) that is constant if the attachment is uniform in \( x \), since in this case \( w_{t,t+1}(x) \) and \( P(x) \) are simply proportional to each other for any \( x \). However, if larger values of \( x \) are preferred, then nodes having large \( x \) value are chosen at a higher frequency, and thus, the shape of \( W(x) \) becomes increasing as a function of \( x \). In the opposite case, when the attachment/detachment prefers lower values of \( x \), the shape of \( W(x) \) becomes decreasing for similar reasons.

3 Results

We examined the attachment and detachment of the links with respect to the following properties: the number of children, the number of parents, the number of descendants and the number of ancestors. In Fig.1a we show the results for the number of children \( n_c \) when considering the attachment of new links to already existing source nodes in hierarchy ’D’ (Chemicals and Drugs). The curve for the growing events, where the target of the new link is a new node (blue circles), and the curve for rewiring events where the target of the new link is an already existing node (green squares) are displayed separately. The region of \( n_c \) values where the number of already existing nodes to chose from is high enough for reliable statistics is indicated by shaded background. The clearly increasing tendency of the curves in this interval indicates preference for larger \( n_c \) values. The results obtained for simulated random choice from the hierarchy are also displayed (orange diamonds), where \( W(n_c) \) is independent of \( n_c \) as expected. In Fig.1b we show the results for the preference analysis with respect to the number of ancestors, \( n_a \). The shape of the \( W(n_a) \) curves indicates an interesting non-monotonous preference.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. a) Preference with respect to the number of children, $n_c$, in the attachment of new links to already existing source nodes in case of hierarchy ‘D’ (Chemicals and Drugs). We show $W(n_c)$ for both growing events (blue circles), where the targets are new nodes, and for rewiring events (green squares), where the targets are already existing nodes. For comparison we also plot the results for simulated random attachment (orange diamonds), where $W(n_c)$ becomes flat. The dark shaded area corresponds to the region where $P_t(n_c) > 5\%$, whereas the lighter shaded area indicates the region where $P_t(n_c) > 1\%$. b) Non-monotonous preference with respect to the number of ancestors, $n_a$, in the attachment of new links to already existing sources in case of hierarchy ‘D’.

Together with similar results for the other hierarchies the above findings show that the growth and rewiring of the examined hierarchies are governed by non-trivial preference in the attachment mechanisms of the links. The fact, that the attachment is non-uniform with respect to multiple different topological and hierarchical node properties, indicates that time evolution of these systems is far more complex compared to simple preferential attachment models, providing very interesting future challenges for modelling and further statistical analysis.

References

Part V

Ecological Networks
High-order nested subsets and the structure of sites-species-species hypernetworks

Albert Solé-Ribalta\textsuperscript{1}, Claudio J. Tessone\textsuperscript{2}, Carlo G. Ferrari\textsuperscript{3}, and Javier Borge-Holthoefer\textsuperscript{1}

\textsuperscript{1} Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Catalonia, Spain,
\textsuperscript{2} URPP Social Networks, Universität Zürich, Switzerland
\textsuperscript{3} Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina

1 Introduction

The characterisation of species co-occurrence patterns is a central question in ecology to understand not only which mechanisms enable the assembly of communities, but also how these will behave under changing environmental conditions. Community ecology largely agrees at large on the fact that such patterns are non-random. The debate remains, however, to sort out what non-random patterns should be observed.

In particular, two opposing architectures of species-species community organisation have been stubbornly observed, namely nestedness and segregated species co-occurrence. Nestedness is a very common property in empirical ecological networks \cite{1}. In systems displaying nestedness, specialists (species with few partners) tend to interact with subsets of the mutualistic partners of generalists (species with many partners). The emergence of this frequent structural arrangement has ignited much research in terms of system stability and biodiversity \cite{2,3}. However, the presence of nestedness implies a high niche overlap, in contradiction with the Volterra-Gause competitive exclusion principle \cite{4}, which states that two species with equivalent niches cannot exist locally and the one with lower fitness is convicted to extinction. Ultimately, strong competition should result in checkerboard-like patterns, produced by pairs of species with mutually exclusive ranges—a sort of anti-nestedness \cite{5}.

Both organisation principles have received an immense amount of empirical and theoretical support. In the middle ground, attempts to reconcile nestedness and negative species co-occurrence (segregation) have remained within the constraints of a bipartite representation. Little has been done to embrace the problem respecting the dimensionality of the system under study, which corresponds to Hutchinson’s inclusive niche concept \cite{6}, formalised as a region in an \(n\)-dimensional hypercube, i.e. a multi-dimensional space where each dimension represents a resource. In particular, we note that, given the limitations to access concurrent geographical, temporal and species interaction data, almost all efforts have been placed on bipartite networks (species-species or species-site) which stand as flat (projected) views of ecological communities, where one or more dimensions have been collapsed.
2 Results

To overcome the aforementioned constraints, recent advances in the field suggest the benefits of considering higher-dimensional connectivity [7], clearing the way for a finer approach to structure-dynamics interplay [8] in ecological networks. Bridging biogeography and community ecology, we show here that high nestedness and low species co-occurrence are compatible within a larger dimensional space (see Figure 1), and their concurrency can be parsimoniously explained when studied under the lens of higher-dimensional networks and a derived new structural pattern: \( n \)-dimensional nestedness. Quantifying such pattern demands a generalisation of the measures that are used to analyse bidimensional ecological systems. In the case of nestedness, we propose here a natural extension of NODF [9]: the measure of connectivity overlap is augmented from rows and columns, to planes (for \( n = 3 \)) and hyperplanes (for \( n > 3 \)). Noticeably, our measure \( \mathcal{N} \) incorporate as well a null model term, along the lines of [10]:

\[
\mathcal{N} = \frac{2}{\sum_{i=1}^{n} |D_i|} \sum_{i=1}^{n} \sum_{\mu, \nu=1}^{|D_i|} \varepsilon_{uv}^{(i)} - \langle \varepsilon_{uv}^{(i)} \rangle \Theta(k_u^{(i)} - k_v^{(i)})
\]

where \( D_i (i \in [1, n]) \) is one of the dimensions of the \( n \)-dimensional system, e.g. \( P \) plants, \( A \) animals and \( L \) locations, for the three-dimensional case; \( \varepsilon_{uv}^{(i)} \) is the connectivity overlap between nodes \( u \) and \( v \) belonging to \( D_i \); \( \langle \varepsilon_{uv}^{(i)} \rangle \) is the expected overlap between those same nodes, if their connections were laid at random; and \( \Theta \) is the Heaviside function that ensures that the measurement is performed only once.

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. Four prototypical examples of 3-dimensional hypercubes that describe tripartite relationships of an ecological system. This figure illustrates four limiting cases: those in which species-site occurrences are highly structured (A and B), against those in which such occurrences are arranged at random (C and D). The first column displays each of the three partial views of the system (plant-animal, $G_{PA}$; plant-location, $G_{PL}$; and animal-location, $G_{AL}$). The second column places in grey the presence of tripartite interactions. The projection of these hyperlinks on the axis planes reduces again to the view in the left column. The scatter plots in the third column inform about the relationship between the newly introduced $n$-dimensional nestedness (with $n = 3$ here), $N_{PAL}$, and the “traditional” bipartite version of it. Among the observed patterns, we highlight that high levels of $N_{PAL}$ may appear despite low values of species-site overlap, i.e. a segregated scenario (panel C). Finally, the behaviour of $N_{PAL}$ against the granularity of the spatial dimension ($L$) is studied in the right-most column. Noteworthy, we observe, for some arrangements, a non-monotonic evolution of $N_{PAL}$—although by definition $N_{PAL} \sim 0$ for very large $L$. 

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Trophic Structure of Directed Networks: Effects on Topology and Dynamics

Samuel Johnson\textsuperscript{1,2}

\textsuperscript{1} University of Birmingham, School of Mathematics, Edgbaston B15 2TT, UK
\textsuperscript{2} Alan Turing Institute, 96 Euston Rd, London NW1 2DB, UK
S.Johnson.4@bham.ac.uk

The three directed networks shown in Fig. 1 – a food web, a network derived from metabolism, and a gene regulatory network – display markedly different ‘trophic structures’\textsuperscript{[1, 2]}. This is evident even to the naked eye because in these plots the height of each node on the vertical axis represents its trophic level (with different scales for each network). Trophic levels have long been used in ecology to characterise species\textsuperscript{[3]}. Thus, in a food web, species with no in-coming edges such as plants and algae will be at level one, species which only consume these will be at level two, and so on. In general, node \(i\)’s trophic level is defined as
\[ s_i = 1 + \sum_j a_{ij} s_j / k_{in}^j, \]
where \((a_{ij})\) is the adjacency matrix and \(k_{in}^j\) is \(i\)’s in-degree, and the level of basal nodes (those with \(k_{in} = 0\)) is \(s_l = 1\) by convention. An interesting quantity is then a network’s mean trophic level, \(<s>\), which in the ecological context is considered as a measure of ecosystem health\textsuperscript{[7]}. In Ref.\textsuperscript{[1]} we defined the ‘trophic coherence’ of a network as a measure of how

![Fig. 1. Examples of three empirical networks plotted so that the height of each node on the vertical axis is proportional to its trophic level (the scale used for each network is different because of the disparity in mean trophic level). Left: Ythan Estuary food web [4], which when compared to its random expectation has a lower mean trophic level (\(<s>/\bar{s} = 0.28\)) and is more trophically coherent (\(q/\bar{q} = 0.15\)). Centre: A network derived from the \textit{C. pneumoniae} metabolism [5], which has a high mean trophic level (\(<s>/\bar{s} = 1.71\)) and significant incoherence (\(q/\bar{q} = 1.62\)). Right: Gene regulatory network of \textit{E. coli} [6], which has mean trophic level close to its random expectation (\(<s>/\bar{s} = 1.01\)) but is slightly more trophically coherent (\(q/\bar{q} = 0.88\)). The food web and the gene regulatory network are in the ‘loopless’ regime (with \(\tau = -1.32\) and \(\tau = -2.54\), respectively) while the metabolic network is in the ‘loopful’ regime (\(\tau = 1.69\)). Reproduced from the SI Appendix of Ref. [2].]
neatly nodes fall into distinct trophic levels. Specifically, we attribute to each edge \((i, j)\) a trophic difference \(x_{ij} = s_i - s_j\), and look at the distribution of differences over all the edges in the network, \(p(x)\). We refer to the standard deviation of this distribution as an ‘incoherence parameter’ \(q\), because \(q = 0\) represents a maximally coherent network (one formed of integer trophic levels), and greater \(q\) implies less coherence.

**Topological effects**

In Ref. [2] we derive expectations for various quantities related to trophic structure in two random graph ensembles: the basal ensemble (which limits in the directed configuration ensemble and can be used as a null-model), and the coherence ensemble, which is the set of all directed networks with given degree sequence and trophic coherence. We show that several quantities related to the numbers of cycles or to the eigenspectrum of the adjacency matrix depend on the ‘loop exponent’:

\[
\tau = \ln \alpha + \frac{1}{2q^2} - \frac{1}{2q^2},
\]

where \(\alpha = \langle k^\text{in} k^\text{out} \rangle / \langle k \rangle\) is the branching factor and \(\bar{q} = \sqrt{L/L_B - 1}\) is the random expectation for \(q\) given the proportion of edges which are connected to basal nodes, \(L_B/L\). Depending on the sign of \(\tau\), a network will either be in a ‘loopless’ regime, with high coherence, few or no cycles and leading eigenvalues close to zero (\(\tau < 0\)), or a ‘loopful’ one, characterised by less coherence, many cycles and large leading eigenvalues. Empirical networks we have studied – including some of neurons, trading nations and word adjacencies, as well as species, genes and metabolites – conform well to these coherence ensemble expectations [2].

We have also found that trophic coherence is closely related both to motif distributions [8], and to intervality, another topological property which can be extended from ecology to generic directed networks [9].

**Dynamical effects**

A fundamental question in ecology is why biodiversity does not lead to instability, and eventually the collapse of large ecosystems [10, 11]. In addressing this question we found that, under certain assumptions regarding dynamics, the stability of a food web depends crucially on its trophic coherence [1]. Furthermore, using the ‘preferential preying model’ to generate synthetic networks, we showed that the relationship between diversity and stability can be either positive or negative, depending on trophic coherence. In other words, if systems are sufficiently coherent, they can become more stable with increasing size and edge density – the opposite of what happens for other network models which do not account for trophic structure correctly.

Other effects on dynamics are also to be expected, given that trophic structure is related as it is to feedback cycles. For instance, we have shown that spreading processes such as models of epidemics or neuronal avalanches depend on trophic coherence [12]. Whether such a process will percolate to the entire system, and whether activity will decay or become endemic, are both, in some settings, determined by \(q\).
Conclusion

Trophic structure is an under-explored property of complex systems. Recent research has shown that trophic coherence, in particular, is related to various topological and dynamical features of directed networks, but there is much more work to be done. The two versions of the preferential preying model of network generation [1, 12], and the coherence ensemble of random graphs [2], provide tools for studying these effects both numerically and theoretically. There remain many open questions: What mechanisms lie behind the trophic structures observed in nature? Is trophic level informative of node function? What is the relationship between trophic structure and community structure? How does trophic coherence determine processes such as synchronization, percolation or computation? Can trophic levels and related concepts be extended to multilayer networks?

Acknowledgements

Many thanks to all my collaborators on this topic: Janis Klaise, Virginia Domínguez García, Miguel Ángel Muñoz, Luca Donetti, and Nick Jones.

References

EXTINCTION CURVES AND ERROR PROPAGATION IN MUTUALISTIC ECOLOGICAL NETWORKS: THE CASE OF FRUGIVORES IN A TROPICAL RAINFOREST

Suresh Babu and Gitanjali Yadav

School of Human Ecology, Ambedkar University of Delhi, Kashmere Gate, Delhi, India 110006.

Department of Plant Sciences, University of Cambridge, Downing Site, Cambridge U.K. CB23EA

Genomics & Systems Biology Laboratory, National Institute of Plant Genome Research (NIPGR), Aruna Asaf Ali Marg, New Delhi, India 110067.

Abstract The study of mutualistic plant animal interactions in their community context is crucial for understanding the functioning of ecological systems and also for effective management of biodiversity worldwide. In this scenario, an objective assessment of how species loss and other damages, such as introduction of alien invasive species, hunting and climate change, can affect the ecosystem is imperative, although limitations of drawing patterns from examination of species-poor datasets and lack of good quality webs has been a constant deterrent in such initiatives. Here we present a comprehensive study of a critical ecosystem service – the dispersal of seeds by birds and mammals in a tropical rainforest, by way of a network analysis and detailed species-loss impact assessment. Seed dispersal networks represent a prominent category of mutualistic interactions in an ecosystem, particularly so in tropical forests, where upto 90% of tree and shrub species produce fleshy fruits that are adapted to animal dispersal, and fruit-eating vertebrates can significantly influence population dynamics of tropical forest communities, including density, spatial patterning and composition of plant-animal communities. We report the structure and topological properties of a new seed-dispersal network from Great Nicobar Island, India, and investigate a variety of cascading response curves describing secondary extinctions, fragmentation, and communication failure between nodes. We explicitly analyse two network properties, namely nestedness and connectance to illustrate the relationship between diversity and stability and to decipher the topological attributes that drive network robustness to attack or random errors. NConnectance was measured as Links per Species squared (L/S²) while Nestedness...
was calculated using the NodF metric, and significance of nestedness values was assessed by comparing the observed NodF with benchmarks provided by three different null models. For each network, a population of $n = 300$ random networks was generated for each null model. As a statistic indicating significance, we estimated the probability, $p$, that a randomization was equally or more nested than the real matrix. Only the significant NodF values were used for this analysis. The nestedness of GNIC was found to be higher than other frugivory networks known to date (NodF value 21.02) and its degree distribution showed best fit to a truncated power law distribution as shown in Figure 1.

Figure 1. The GNIC Frugivory Network (a) layout with distinct geodesics (b) Nested structure, and (c) degree distribution.
The structure and dynamics of non-normal networks

Malbor Asllani∗1,2, Renaud Lambiotte2, and Timoteo Carletti3

1 MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick V94 T9PX, Ireland
2 Mathematical Institute, University of Oxford, Woodstock Rd, OX2 6GG Oxford, UK
3 Department of Mathematics and naXys, Namur Institute for Complex Systems, University of Namur, rempart de la Vierge 8, B 5000 Namur, Belgium
∗malborasllani@gmail.com

1 Introduction

Network science [1–3] has emerged, in the last 20 years, as an essential framework to model and understand complex systems in variety of disciplines, including physics [1], economy [4], biology [5] and sociology [6]. At its core, network science views a system as a set of nodes that may be connected directly by an edge or indirectly by a succession of edges, thereby forming paths of interactions. The bridge between network structure and dynamics is generally unraveled by defining a linear dynamical model on the nodes where the process is determined by a matrix, somehow related to the adjacency matrix of the underlying network. In addition, critical aspects of the system, such as its stability and characteristic time scales, are usually described by the properties of its spectrum [7]. Central network concepts such as the spectral gap, spectral radius, master stability conditions, community detection and network comparison are all build on this interpretation. The characterisation of a linear system by its spectrum is canonical, but it is unreliable in situations when the linear operator is non-normal, namely its eigenvectors do not necessarily form an orthonormal basis and the transformation to eigenvector coordinates may involve a strong distortion of the phase space. Non-normality has a long tradition in linear algebra and dynamical systems, from early studies in hydrodynamics [8] to more recent works on the robustness of non-normal ecosystems [9] and in neuronal dynamics [10]. Yet, these results remain focused on limited areas of science and a systematic study of the prevalence of non-normality in real-world networks, and its potential impact on dynamics, is still lacking. Here we call non-normal network a network whose adjacency matrix \( A \) is non-normal [11]. By definition, \( A \) is non-normal if it verifies \( AA^T \neq A^T A \). It is thus clear that \( A \) needs to be asymmetric to be non-normal, or equivalently the network needs to be directed to be non-normal. Our main result is grounded on the fact that a large collection of empirical networks in a wide spectrum of disciplines are strongly non-normal showing this way that strong non-normality is ubiquitous in network science [12] (see Figure 1b). Furthermore, we analyse their global structure and show asymmetry is not sufficient and certain types of network architectures are necessary to determine a strong non-normality and last we have developed algorithms to describe their growth.
2 Results

Let us illustrate more in detail the influence of non-normality on the prototypical example of a linear (linearised) dynamics on a non-normal network. Without loss of generality, we consider the model \( \dot{x} = Mx \), where \( M \) encodes the linear dynamics on the network and forms a stable matrix. This means that the spectral abscissa \( \alpha(M) \), the maximum real part of the eigenvalues of \( M \) is non positive. In case of normal networks, if \( M \) has a positive numerical abscissa \( \omega(M) \), namely the maximum eigenvalue of the Hermitian part of \( M \), the system can undergo a transient growth before asymptotically converging to zero, as measured by the norm of the state vector \( x \) (see Figure 1 a)). This transient behaviour cannot be explained by the picture provided by the spectrum of the matrix \( M \) [13] and in situations when the dynamics is obtained from the linearisation around a critical point, this initial growth may trigger nonlinear terms and take the system far away from the equilibrium, and thus radically reshape the dynamical behaviour of nonlinear systems [11]. To describe the strong non-normality of real networks we have developed the two following measures: the first is a structural measures which describes the relative degree of asymmetry of the network \( \Delta := |K_\prec - K_\succ| / K \), where \( K_\prec = \sum_{i<j} \tilde{m}_{ij}, K_\succ = \sum_{i>j} \tilde{m}_{ij} \) and \( K = K_\prec + K_\succ \) where \( \tilde{M} \) is the adjacency matrix after being relabelled in such a way to maximise the number of entries in the upper triangular part of it; the other measure is spectral instead and is known as Henrici’s departure from normality \( d_F(M) = \sqrt{||M||_F^2 - \sum_{i=1}^n |\lambda_i|^2} \) where \( ||\cdot||_F \) is the Frobenius norm and \( \lambda_i \) the \( i \)-th eigenvalue [13]. Obviously when both measures reach 1 the network is at maximum of its non-normality.

![Figure 1](image)

**Fig. 1.** a) We present the evolution of the norm of the stable system described in the text where the blue curve corresponds to the normal case and the red to the non-normal one where the transient growth occurs. b) We compare the two measures of the non-normality \( \Delta \) and \( \hat{d}_F \) for a large collection of networks from different domains showing strong non-normality for 95% of them.

**Summary**

Non-normality may strongly affect linear and nonlinear dynamical systems on networks
and, more generally, their behaviour. The contributions of this work are manifold. First, we show that a strong non-normality is widespread in complex networks empirically observed in variety of domains. As a second step, we reveal the organisation behind non-normality and show that non-normality is associated with a combination of absence of cycles, low reciprocity and hierarchical organisation. We also propose a simple model for growing networks based on preferential attachment reproducing our observations. Finally, we consider in detail a Lotka-Volterra model applied to a real-world network and show that the use of network metrics for non-normality helps to understand the dynamics of the system.

References

Inferring social relations from presence data.  
Manta Rays case study

Juan Fernández-Gracia¹, Jorge P. Rodríguez¹, Lauren Peel²,³,⁴, Konstantin Klemm¹, Mark Meekan², and Víctor M. Eguíluz¹

¹ Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), E07122 Palma de Mallorca, Spain
² Australian Institute of Marine Science, Indian Ocean Marine Research Centre (IOMRC), University of Western Australia (M470), 35 Stirling Highway, Crawley, WA 6009, Australia
³ Save Our Seas Foundation DArros Research Centre, Geneva, Switzerland
⁴ The Manta Trust, Dorchester, Dorset, UK

1 Introduction

Social interactions are ubiquitous in groups of animals, although for some species these interactions might be difficult to observe or quantify. These interactions might be of different nature, e.g. competitive, mutualistic, parasitic, commensalistic, etc [1]; and their global structure is naturally studied with the tools of complex network theory[2]. One typical interaction is that of leading/following[3, 4]. Here we propose a method to extract follower-followee networks from presence data at a certain location. (Similar previous works include [5, 6])

2 Results

2.1 Follower-followee interactions from presence data

For this method we have in mind presence data of individuals at a particular location. The data for each individual consists of an ordered set of timestamps at which the individual is located in the vicinity of that particular location. In order to infer if individual \(i\) typically follows \(j\) we hypothesize that the difference between consecutive times of \(i\) being detected and then \(j\) will be longer than the inverse (consecutive times of detecting \(j\) followed by \(i\)). In order to assess that we extract first the waiting times for \(i\) followed by \(j\), \(\{t_{ij}\}\), and \(j\) followed by \(i\), \(\{t_{ji}\}\). We then compute the distance between the distributions of waiting times as the Kolmogorov-Smirnof distance \(D_{KS}\) between them. \(D_{KS}\) is defined as the maximum distance between the cumulative distributions of the two distributions that are being compared. We use a signed version, so that we distinguish which of the two cumulative distributions is above the other. So if \(D_{KS}(\{t_{ij}\},\{t_{ji}\}) > 0\) then individual \(i\) is following \(j\), while \(D_{KS}(\{t_{ij}\},\{t_{ji}\}) < 0\) reflects the opposite \(j\) following \(i\). The strength of the interaction will be given by the absolute value of the distance \(|D_{KS}|\).
2.2 Manta Rays case study

The data Here we make use of an acoustic data set for the presence of manta rays near a receiver station. A total of 24 animals were tagged and their presence in the vicinity of the receiver station was recorded for a period of one year. Globally there were 41607 recorded events.

Follower-followee networks We omitted the pairs of individuals for which less than 100 interactions were found. An interaction here refers to a waiting time between the event of an animal appearing at the station and the other one. We assessed the KS distances between the distribution of waiting times for each pair as explained above. We used a global reshuffling scheme to compute the $p$-values associated to each KS distance and discarded those edges for which $p > 0.002$. By a global reshuffling scheme it is meant a reshuffling where the number of events of each individuals is kept fixed but the times of the events are reshuffled from the pool of all events. This is different to a local reshuffling scheme, where only the two sequences involved in the calculation of the distance are involved in the reshuffling, but both schemes lead to similar results.

Alltogether we obtained a network of 12 individuals, 5 females of which 4 big ones and one small, and 7 males, of which 4 big ones and 3 small ones; and 33 edges forming a single component, as can be seen in Fig. 1 C.

In order to assess the correlation of sex (size) with the topology of the network we compare two basic quantities, namely the number of edges in the network of each type and their average weight, with the results of $10^4$ reshufflings of the sexes (sizes) but keeping the network structure fixed. Regarding the sexes, mixed edges are on average stronger than expected, and there are much less males following females than expected, in contrast to females following males, which are overrepresented. For same sex edges the results are similar, the number of edges is approximately the one expected, while the strength is smaller than expected. As for the sizes, the most salient result is that the average weight of the interactions of males following males is much lower than expected. The other types of edges are only slightly off the values that are expected. So small individuals following big ones follow a bit stronger than expected and there is one link less than expected. For big ones following small ones, the strength is slightly weaker than expected and there is only one extra edge. For big ones following big ones the strength is slightly bigger than expected and there are only a couple more edges than expected.

To further investigate the dynamics we computed the appearance rate of couples of animals that were linked by an edge of the network. The appearance rate is computed for one individual as a function of the time since the last presence event of the other individual. For followers it turns out to be bigger than for followees up to a certain time lag of around 200 minutes. This enhanced appearance rate of the follower individuals a short time after the animal that is leading confirms our measurement of who follows who.

Finally we also define as leadership index of an individual the sum of weights of incoming edges (combined strength of following relations) minus the sum of weights of outgoing edges (combined follower strength). This index shows that females tend to be followers, while males leaders.
Fig. 1. Process of inferring follower-followee networks. A Sample of presence data for 2 individuals. B Signed Kolmogorov-Smirnoff distance between waiting times of the presence data in A. C Complete network of follower-followee interactions among the studied manta rays.

3 Summary

In this work we have proposed a method for detecting follower-followee interactions from presence data. The method captures boosted appearance rates of animals that are following another animal. In the future we would like to extend the methodology for full trajectories, based on the theory of first return times. Another avenue worth exploring are causality measures used for example in the inference of excitatory/inhibitory relations among neurons, as the spiking data is formally equivalent to the presence data here presented.

References

Part VI

Link Analysis and Ranking
A ‘dramatic truth’ in link prediction: SBM inference fails to effectively predict even the structure of synthetic networks generated with the SBM model

Umberto Michieli1,2, Alessandro Muscoloni1, Leonardo Badia2 and Carlo Vittorio Cannistraci1,3

1 Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Dresden, Germany; 2 Department of Information Engineering, University of Padova, Padova, Italy; 3 Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi “Bonino Pulejo”, Messina, Italy

Abstract. Methods for topological link-prediction are generally referred as global or local. The former exploits the entire network topology, the latter adopts only the immediate neighborhood of the link to predict. Stochastic-Block-Model (SBM) is a global method believed as one of the best link-predictors and widely accepted as reference when new methods are proposed. But, our results suggest that SBM, whose computational time is high, cannot in general overcome the Cannistraci-Hebb (CH) network automaton model that is a simple local-learning-rule of topological self-organization proved by multiple sources as the current best local-based and parameter-free deterministic rule for link-prediction. In addition, after extensive tests on many different real complex networks of small size, Structural-Perturbation-Method (SPM) clearly emerges as the new best global method baseline (Fig. 1). However, even SPM overall does not outperform CH and in several evaluation frameworks (in particular on networks of large size) we astonishingly found the opposite (these results are not available here and should be consulted on the original article [1]). At this point of our study, we decided to investigate better the nature of the disappointing SBM’s performance, and the extent to which SBM fails because of inference issues. The Lancichinetti-Fortunato-Radicchi (LFR) model [2] is a special version of the degree-corrected SBM and is typically adopted to generate artificial benchmarks. Interestingly, although LFR networks are generated by a model based on the SBM theory, we show that, in general, a large family of SBM link predictors cannot reach overall performances on LFR networks at the same level as SPM and CH (Fig. 2), revealing clear inference problems of the SBM family on link prediction tasks.

References

Table 1. For each network, 10% of links have been randomly removed (10 iterations for SBM, SBM DC N, SBM DC and SBM N due to the high computational time, 100 iterations for the other methods) and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. In order to evaluate the performance, the links are ranked by likelihood scores and the precision is computed as the percentage of removed links among the top-r in the ranking, where r is the total number of links removed. The table reports for each network the mean precision over the random iterations, the mean precision and mean ranking over the entire dataset. For each network the best method (or methods) is highlighted in bold. The networks are sorted by increasing number of nodes.

<table>
<thead>
<tr>
<th>Network</th>
<th>SPM</th>
<th>CH2 L2</th>
<th>SBM</th>
<th>FBM</th>
<th>RA L3</th>
<th>SBM DC N</th>
<th>SBM DC</th>
<th>SBM N</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse_neural</td>
<td>0.02</td>
<td>0.09</td>
<td>0.10</td>
<td>0.01</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>karate</td>
<td>0.17</td>
<td>0.19</td>
<td>0.28</td>
<td>0.27</td>
<td>0.15</td>
<td>0.13</td>
<td>0.14</td>
<td>0.01</td>
</tr>
<tr>
<td>stmarks_foodweb</td>
<td>0.26</td>
<td>0.13</td>
<td>0.29</td>
<td>0.14</td>
<td>0.25</td>
<td>0.19</td>
<td>0.18</td>
<td>0.07</td>
</tr>
<tr>
<td>dolphins</td>
<td>0.13</td>
<td>0.14</td>
<td>0.16</td>
<td>0.19</td>
<td>0.10</td>
<td>0.08</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>ythan_foodweb</td>
<td>0.27</td>
<td>0.08</td>
<td>0.24</td>
<td>0.09</td>
<td>0.26</td>
<td>0.20</td>
<td>0.19</td>
<td>0.06</td>
</tr>
<tr>
<td>macaque_neural</td>
<td>0.72</td>
<td>0.56</td>
<td>0.68</td>
<td>0.55</td>
<td>0.64</td>
<td>0.38</td>
<td>0.36</td>
<td>0.12</td>
</tr>
<tr>
<td>polbooks</td>
<td>0.17</td>
<td>0.18</td>
<td>0.15</td>
<td>0.18</td>
<td>0.16</td>
<td>0.15</td>
<td>0.14</td>
<td>0.01</td>
</tr>
<tr>
<td>SEA_terrorist</td>
<td>0.45</td>
<td>0.46</td>
<td>0.29</td>
<td>0.33</td>
<td>0.25</td>
<td>0.29</td>
<td>0.25</td>
<td>0.08</td>
</tr>
<tr>
<td>ACM2009_contacts</td>
<td>0.26</td>
<td>0.27</td>
<td>0.25</td>
<td>0.26</td>
<td>0.27</td>
<td>0.19</td>
<td>0.19</td>
<td>0.05</td>
</tr>
<tr>
<td>football</td>
<td>0.31</td>
<td>0.36</td>
<td>0.34</td>
<td>0.25</td>
<td>0.21</td>
<td>0.28</td>
<td>0.26</td>
<td>0.06</td>
</tr>
<tr>
<td>physicians_innovation</td>
<td>0.07</td>
<td>0.08</td>
<td>0.06</td>
<td>0.08</td>
<td>0.05</td>
<td>0.02</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>AQ_terrorist</td>
<td>0.36</td>
<td>0.42</td>
<td>0.22</td>
<td>0.35</td>
<td>0.26</td>
<td>0.16</td>
<td>0.12</td>
<td>0.03</td>
</tr>
<tr>
<td>manufacturing_email</td>
<td>0.51</td>
<td>0.42</td>
<td>0.47</td>
<td>0.39</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.09</td>
</tr>
<tr>
<td>jazz</td>
<td>0.65</td>
<td>0.58</td>
<td>0.47</td>
<td>0.45</td>
<td>0.40</td>
<td>0.37</td>
<td>0.35</td>
<td>0.09</td>
</tr>
<tr>
<td>residence_hall_friends</td>
<td>0.28</td>
<td>0.25</td>
<td>0.18</td>
<td>0.24</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>rhesus_brain</td>
<td>0.31</td>
<td>0.25</td>
<td>0.21</td>
<td>0.24</td>
<td>0.19</td>
<td>0.23</td>
<td>0.22</td>
<td>0.03</td>
</tr>
<tr>
<td>vanderwaals</td>
<td>0.29</td>
<td>0.19</td>
<td>0.08</td>
<td>0.17</td>
<td>0.13</td>
<td>0.09</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>haggle_contacts</td>
<td>0.62</td>
<td>0.57</td>
<td>0.62</td>
<td>0.63</td>
<td>0.45</td>
<td>0.44</td>
<td>0.44</td>
<td>0.10</td>
</tr>
<tr>
<td>worm_nervoussys</td>
<td>0.16</td>
<td>0.12</td>
<td>0.15</td>
<td>0.11</td>
<td>0.12</td>
<td>0.15</td>
<td>0.12</td>
<td>0.03</td>
</tr>
<tr>
<td>USAir</td>
<td>0.46</td>
<td>0.43</td>
<td>0.38</td>
<td>0.38</td>
<td>0.40</td>
<td>0.39</td>
<td>0.37</td>
<td>0.07</td>
</tr>
<tr>
<td>netsci</td>
<td>0.41</td>
<td>0.54</td>
<td>0.13</td>
<td>0.33</td>
<td>0.29</td>
<td>0.25</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>infectious_contacts</td>
<td>0.37</td>
<td>0.35</td>
<td>0.30</td>
<td>0.33</td>
<td>0.26</td>
<td>0.19</td>
<td>0.16</td>
<td>0.07</td>
</tr>
<tr>
<td>flightmap</td>
<td>0.75</td>
<td>0.56</td>
<td>0.64</td>
<td>0.56</td>
<td>0.58</td>
<td>0.55</td>
<td>0.56</td>
<td>0.12</td>
</tr>
<tr>
<td>email</td>
<td>0.16</td>
<td>0.17</td>
<td>0.09</td>
<td>0.16</td>
<td>0.12</td>
<td>0.10</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>blog</td>
<td>0.23</td>
<td>0.17</td>
<td>0.19</td>
<td>0.17</td>
<td>0.18</td>
<td>0.20</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>mean precision</td>
<td>0.34</td>
<td>0.30</td>
<td>0.28</td>
<td>0.27</td>
<td>0.26</td>
<td>0.22</td>
<td>0.21</td>
<td>0.06</td>
</tr>
<tr>
<td>mean ranking</td>
<td>2.1</td>
<td>2.9</td>
<td>3.6</td>
<td>4.1</td>
<td>4.3</td>
<td>5.2</td>
<td>6.1</td>
<td>7.9</td>
</tr>
</tbody>
</table>
Fig. 1. Synthetic networks have been generated using the LFR model with parameters \( N = [100, 1000] \) (networks size), \( k = [20, 24, 28] \) (average degree) and \( \mu = [0.3, 0.5, 0.7] \) (mixing parameter). The minimum and the maximum of the community sizes have been fixed respectively to \( \min c = N/20 \) and \( \max c = 4*\min c \). The maximum degree of a node has been set to \( \max k = 3*k \). Low values of \( \mu \) generate strong clustering and a desired clustering coefficient \( C = [0.7, 0.5, 0.3] \) is attempted to be satisfied respectively to the values of \( \mu \). For each combination of parameters, 100 networks have been generated. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. In order to evaluate the performance, the links are ranked by likelihood scores and the precision is computed as the percentage of removed links among the top-\( r \) in the ranking, where \( r \) is the total number of links removed. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. The plots in (A) compare the methods of the SBM family, whereas the best of them overall, i.e. SBM N, is compared in (B) with the state-of-the-art methods CH-L2 and SPM. Note that in the SBM variations ‘DC’ stands for degree-corrected and ‘N’ for nested.
A parametric approach to information filtering in complex networks: The Pólya filter

Riccardo Marcaccioli and Giacomo Livan

1 Department of Computer Science, University College London, 66-72 Gower Street, London WC1E 6EA, UK
2 Systemic Risk Centre, London School of Economics and Political Sciences, Houghton Street, London WC2A 2AE, UK

1 Introduction

Extracting relevant information on large complex networks has always been an interest of the scientific complex systems community. Such issue has been addressed by a broad class of techniques, which resort to hypothesis testing in order to assess the statistical significance of each link in a network.

The disparity filter [1], which arguably represents one of most widely used filtering techniques, falls under this category, and relies on a null hypothesis of uniform distribution of a node’s strength over its links. Such a method has been adopted as one of the main benchmarks against which the efficiency of novel filtering techniques has been tested. More recently, a procedure based on a null hypothesis of random connectivity (encoded as the urn problem described by the hypergeometric distribution) has been put forward [2]. Since such procedure was the first to systematically make use of the Bonferroni multivariate correction in the context of network filtering, whereas previous works naively tested all links in a network against a univariate threshold, we shall refer to it as the Bonferroni filter.

Both the disparity and the Bonferroni filters provide a top-down approach based on well defined null hypotheses, against which all links in a network are tested individually. While this certainly presents advantages in terms of convenience, at the same time it can lead to a lack of flexibility, as different networks may display different levels of heterogeneity, to which a one-fits-all null hypothesis cannot adapt. Furthermore, being based on null hypotheses of partially random interactions, both filters systematically tend to identify as statistically significant most links associated with large weights. Yet, the presence of heavy links in a heterogeneous networked system is the norm, rather than the exception, as past interactions naturally breed further interaction in a variety of natural and social systems.

Here, we propose a filtering methodology based on a null hypothesis designed to respond to the specific level of heterogeneity in a network. We shall do so through a statistical test based on the Pólya urn, a well known combinatorial problem driven by a self-reinforcement mechanism according to which the observation of a certain event increases the probability of further observing it. Such a mechanism is governed by a single parameter $a$, which allows to tune the null hypothesis tolerance to heterogeneity, and to study a continuous family of network backbones $\mathcal{P}_a$. 

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
2 Results

The statistical methodology we propose, that we call the Pólya filter, has been fully characterised analytically together with the evolution of the resulting set of validated links as the free parameter $a$ changes. All the results are detailed in a recent article [6] and are the following:

1. The disparity filter can be obtained as a zero-order approximation of a particular ($a = 1$) case of the Pólya filter. This can be seen by setting $a = 1$ in the following approximate form of the $p$-value prescribed by the Pólya filter:

$$
\pi_P(w \mid k, s, a) \approx \frac{1}{\Gamma\left[\frac{1}{a}\right]} \left(1 - \frac{w}{s}\right)^{\frac{k}{a}} \left(\frac{w k}{s a}\right)^{\frac{1}{a} - 1}
$$

Where $w$ is the weight of the link we want to test, $k$ is the strength of the node it is attached to and $s$ its strength.

2. The links progressively included in the subsequent backbones $\mathcal{P}_a$ can be approximately identified by means of the ratio $r = \frac{w}{s} k$. The ratio $r$ introduces an interplay between weights and topology that can be used to understand why the Pólya filter (and therefore the disparity filter) is able to retain the multiscale nature of the original network across all reasonable values of $a$.

3. The sensitivity of the Pólya filter to node heterogeneity and to a global rescaling of the weights can be fully characterised.

4. Three possible alternatives to fix the free parameter $a$ can be provided. The first one is based on apriori constrains. The second one is a maximum likelihood estimation that also shows the sensibility of the free parameter to the original network’s own heterogeneity. The third is based on the idea of finding an optimal compromise between the backbone similarity with the unfiltered network and its centrality for diffusion process on the original network; additionally this latter alternative highlights the ability of our filter to progressively select links with higher salience [5] values.

All the results are both analytically justified and empirically verified using two large real world networks: the WIOT network [3] and the US Airports network [4]. Figure 1 shows a summarizing picture representing some of the mechanics underpinning the Pólya filter.

Summary. We propose a novel technique to extract backbones of statistically relevant interactions between pairs of nodes in a network based on the Pólya urn model. In the network context, the parameter $a$ tuning the Pólya models self-reinforcement mechanism effectively becomes a tolerance to a network’s heterogeneity. This, in turn, introduces an element of flexibility, which, to the best of our knowledge, other network filtering techniques do not provide. Indeed, we have shown that the Pólya filter generates a continuous family of network backbones. Depending on the specific application, the null hypothesis underpinning the filter can be chosen so as to have a different tolerance to heterogeneity. The low-tolerance regime ($a < 1$) corresponds to a rather loose filtering, suited to situations where the main goal is to filter out interactions that can be
Fig. 1. Radial tree visualization of the largest connected component of the US Airports Network filtered using two different values of the parameter $a$ ($a = 1$ and $a = 2$). Links included in both backbones are drawn in purple, while links included only in the $a = 1$ backbone are drawn in grey. Link opacities are proportional to the corresponding values of $r$. We can visually confirm the soft dependency of the filter on $r$ (derived in [6]): most of the less opaque links are purple, but some grey links are more opaque than some purple ones.

unquestionably identified as noise. On the other hand, the high-tolerance regime ($a > 1$) corresponds to increasingly restrictive tests, where only links of substantial structural importance survive. The link selection criterion underpinning the Pólya filter is based on the interplay between topology and the local relative importance of a link, quantified by the parameter $r$. This, in turn, guarantees that the filter does not perform a naive link selection merely based on retaining high strength links connecting hubs, but instead ensures a non-trivial scanning of all the relevant scales of a network. This latter result is strengthen by the fact that the proposed methodology is also able to select links that play a pivotal role (quantified in terms of salience) in diffusion process on the original unfiltered network.

References

4. (2017) Us airports network dataset KONECT.
Link Prediction via Indirect Interaction
Duration
Laxmi Amulya Gundala and Francesca Spezzano
Computer Science Department
Boise State University
Boise ID 83725, USA
francescaspezzano@boisestate.edu

1 Introduction

Link prediction is the problem of inferring new relationships among nodes in a network that are likely to occur in the near future. Classical approaches mainly consider neighborhood structure similarity when linking nodes. However, we may also want to take into account if the two nodes we are going to link will benefit from that by having an active interaction over the time. For instance, it is better to link two nodes $u$ and $v$ if we know that these two nodes will interact in the social network even in the future, rather than suggesting $v'$ as new friend, who will never interact with $u$. Thus, the longer is the interaction estimated to last, the higher is the priority for connecting the two nodes.

In this paper, we address a new variant of the link prediction problem: given a pair of indirectly interacting nodes, predict whether or not they will form a link in the future [4]. By “indirect interaction” we mean that there is a particular action involving both the nodes depending on the type of network. For example, in social networks such as, for instance, Facebook, we have users that are not friends but interact through other user’s wall posts. In Wikipedia hyper-link network, it happens when readers navigate from a page $u$ to page $v$ through the search box (on the top right corner of page $u$) and there is no explicit link on page $u$ to $v$. Some of these searches can be casual or because of an ongoing trend in news or social media, while others suggest the demand of a physical link from page $u$ to page $v$ to improve users’ navigation. Examples of such interactions can be found in other types of networks as well. For example, in Amazon co-purchased products network, we can discover future co-purchased products by looking at users’ search logs that may indicate “indirect interaction” among various products with respect to users and help Amazon to better allocate their product warehouses.

We study the cases of pair of nodes $(u, v)$ that are not connected, yet show some indirect interaction as they are involved in multiple interactions during an observational time interval $[t_0, t_1]$. The indirect interaction between nodes $u$ and $v$ during that time interval indicates that $u$ and $v$ may have something in common and maybe it is useful to link them.
2 Framework

We propose a novel approach to solve our link prediction problem in two steps. First, we focus on the problem of predicting how long an indirect interaction from node \( u \) to node \( v \) last. Second, once the duration is estimated, we predict new links according to their descending order.

We performed a two-sample one-tail t-test to test our hypothesis that the longer users \( u \) and \( v \) keep indirectly interacting, the higher is the probability that a link will be added from \( u \) to \( v \). The \( p \)-value for a significant level \( \alpha = 0.01 \) resulted less than 0.001 on the Facebook Network and Wall Interaction dataset considered, suggesting that there is strong statistical evidence to validate our hypothesis. More specifically, users keep indirectly interacting for \( 3.09 \cdot k \) for pairs of users that end up to be linked vs. \( 1.47 \cdot k \) for users that will not be linked (we set \( k \) equal to six months).

We propose two supervised learning approaches for the problem of predicting the duration of indirect interactions. Given a set of features, the basic approach consists of learning a binary classifier to predict whether or not the indirect interaction will last in the future. The second and more fine-grained approach consists of estimating how long the indirect interaction will last by modeling the problem via survival analysis or as a regression task. Survival analysis is a statistical method to estimate the expected duration of time until an event of interest occurs [2]. It is possible that during any study on a group of entities, there might be instances for which the event of interest did not occur within the study duration or may have incomplete or missing or unavailable data about the event occurrence. Such instances are called censored instances and can be effectively approached using survival analysis.

We used the following set of network features in our evaluation: node degree, number of interactions (hits) from node \( u \) to node \( v \), common neighbors, Jaccard similarity, Adamic-Adar similarity, preferential attachment score, local clustering coefficient, Pagerank of both users \( u \) and \( v \), and the Hadamard (or entrywise) product of Node2Vec feature vectors \( F(u) \) and \( F(v) \) [3].

3 Results

We performed experiments on the longitudinal Facebook Network and Wall Interaction dataset provided by [1] to test our approach to the link prediction task. We divided the data in three time periods: observed indirect interactions during the second half of 2006, trained the model to estimate the interaction duration from 2007 to 2008, and tested link prediction during the first half of 2009. We used majority under-sampling to deal with class imbalance.

Our results show that we achieve an AUROC of 0.90 (with random forest) on the problem of predicting whether or not an indirect interaction from node \( u \) to node \( v \) will last (basic approach). In the case of the fine-grained approach, we have that survival analysis is more suitable than regression for predicting interaction duration (Mean Absol. Error of 0.25 vs. 0.38 and Pearson Corr.
Table 1. Comparison of AUROC values for link prediction between our solution and the best baselines. The table reports the values for the best models only.

<table>
<thead>
<tr>
<th>Our Framework</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic Approach</strong> (Linear SVM)</td>
<td>0.800</td>
</tr>
<tr>
<td><strong>Fine-grained Approach</strong></td>
<td></td>
</tr>
<tr>
<td>Survival Analysis (Exponential)</td>
<td>0.854</td>
</tr>
<tr>
<td>Regression (Lasso)</td>
<td>0.790</td>
</tr>
<tr>
<td><strong>Best Baselines</strong></td>
<td></td>
</tr>
<tr>
<td>Jaccard Similarity</td>
<td>0.645</td>
</tr>
<tr>
<td>Classical Approach (Linear SVM)</td>
<td>0.800</td>
</tr>
</tbody>
</table>

This is due to the fact that survival analysis is able to incorporate censored instances in the model and then learn on more information than regression.

Table 1 (top rows) shows how our proposed framework performs on link prediction. In this experiment, we first trained the model (classification, survival analysis, or regression) to predict how long (or “if” in the case of classification) the indirect interaction will last. Then, we computed the Area Under the ROC curve (AUROC) between these predicted times and the class values (1 if the link has been created later on, 0 otherwise). As we can see, the approach based on survival analysis performs the best with an AUROC of 0.85, while the approach based on classification (resp. regression) achieved an AUROC of 0.80 (resp. 0.79).

To compare our results, we considered hits, Jaccard Similarity, Adamic-Adar, preferential attachment, and cosine similarity between Node2Vec feature vectors \( F(u) \) and \( F(v) \), individually as baselines and assumed their value to be an approximation of the duration of the indirect interaction duration from node \( u \) to \( v \). As we can see in Table 1 (bottom rows), our approach outperforms Jaccard similarity, the best performing baseline, that achieves an AUROC of 0.645. Also, our result is better than a classical link prediction approach that considers all the features used in the paper in input to a classifier, do not consider indirect interaction duration, and can predict new links with an AUROC of 0.80.

References


1 For the case of classification we considered the predicted probability of having a link in the future.
Long distance connections for ranking and robustness enhancement in networks

V. Carchiolo, M. Grassia, A. Longheu, M. Malgeri, and G. Mangioni

Dip. Ingegneria Elettrica, Elettronica e Informatica - Università degli Studi di Catania - Italy

1 Introduction

Networks are nowadays extensively exploited to model phenomena that are the result of the interactions between many elements. Due to their nature, they often show emergent effects, some of which, such as the small-world effect [1], are related to the long-distance links, i.e. the connection between two nodes that are far from each other prior to the link creation. With our work, we focus on the two following emergent effects and enhance them through the creation long-distance links.

Ranking Enhancement. Ranking is a widely used technique to classify nodes in networks according to their relevance. Increasing one’s rank is a desirable feature in almost any context, as a higher rank usually means better reputation for that node. We study how a node’s ranking can be effectively improved by adding long-distance in-links and define a new heuristic whose complexity is $N \log N$ for the best attachment problem, which was proven to be NP-hard in [2]. Our heuristic consists in (i) attaching to some random node in the network, (ii) finding the farthest node, (iii) attaching to that node, (iv) repeat from point (ii) until desired rank is reached.

Network Robustness. Fault tolerance (also known as resilience or robustness) is a highly relevant network feature, especially in real networks, where it is essential to know to what extent a network would still work notwithstanding failures in its structure [5]. The common goal of many studies is to increase the network robustness, for example by properly managing nodes and links and/or by adding new ones. We go in this direction and propose a rewiring strategy based on adding long-distance links to the network between the farthest nodes.

2 Results

Ranking Enhancement. For our simulations, we adopt the widely used PageRank (PR) [3] as ranking metric (nodes are sorted in descending PR order) in a scenario with both Erdős-Rényi (ER) and Scale-Free (SF) networks. Also, we average our results over fifty runs (ten runs on five different instances) per network type. First, we compute the rank gain (as a consequence of an increment of the PR value) in function of the distance of the node on both ER and SF networks with 10K nodes. The results, shown in figures 1, show the rank gain of the target node $t$ when we connect a node (at a given...
distance from \( t \) to \( t \). Let us note that the rank value 1 means the top position (i.e., the highest value of PR), so increments of the rank positions are negative values. In the figure, we report the min, average and max values for each point.

**Fig. 1.** Rank gain of the node \( t \) when it is connected by an in-link from a node \( j \) that reaches it in a given number of hops

Secondly, we test the long-distance in-link based heuristic on 100K nodes networks.

**Fig. 2.** Heuristics performance on 100K nodes networks

The results, shown in figures 2, suggest that our strategy is a significative improvement w.r.t. the pure random approach followed in [4] and also highlight the expected dependence of the algorithm’s performance on the network type. With only 1 in-link, there is no rank gain in the \( ER \) networks, while in \( SF \) networks the target node \( t \) can gain more than 28K positions! It means that selecting the right node \( k \) can give a great advantage in a short time to the node \( t \) in \( SF \) networks, while this is not the case in \( ER \) networks. Moreover, with 10 in-links, our algorithm is able to place the node \( t \) in the position 4K in \( SF \) networks, while it ranks 18K in \( ER \) networks. On the other hand, getting to position 1 requires more than 400 in-links in \( SF \) networks, while it requires only 26 links \( ER \) networks, which comes with no surprise since it is known that, due to their structure, obtaining the top rank in \( SF \) networks is very hard.
The experiments confirm our intuition, as the long-distance in-links are a viable solution for the best attachment problem thanks to the high rank gains that can be obtained even with a small number of in-links.

**Network Robustness.** We simulate an attack to SF networks with 1K nodes and then we repeat it after adding $N$ long-distance links between the farthest nodes, which are recomputed after each addition. Our results, averaged over fifty runs (ten runs on five different instances) per network type and reported in figure 3, show that a peak increment of about 17.5% of robustness, measured as the largest component size (LCC) over the one in the original network instance [6], is gained by introducing only 50 of new links (2.5% of the original number). We also compare our strategy with a random one, where new links are added between two non-linked randomly selected nodes, and our strategy outperforms the random one with a peak difference of more than 13%.

![Fig. 3. SF networks robustness gain in the case of long-distance strategy](image-url)

### References

Part VII

Machine Learning and Networks
An Exact No Free Lunch Theorem for Community Detection

Arya D. McCarthy
Johns Hopkins University

1 Introduction

The no free lunch theorem [9] in machine learning is a claim about the universal (in)effectiveness of learning algorithms. Every algorithm performs equally well when averaging over all possible input–output pairs. Formally, the error (or loss) $\ell$ of the method $f$, summed over all possible problems $(g, T)$, is a loss-specific constant $\Lambda(\ell)$.

$$\sum_{(g, T)} \ell(T, f(g(T))) = \Lambda(\ell) \quad \forall f,$$

where $T$ is the true solution and $g$ generates the input (in our case, a graph $G$) from $T$. To reduce loss on some set of problems means sacrificing performance on others—“there is no free lunch”. This is jarring at first; are we not striving to find superior algorithms for our tasks? The remedy: By making specific assumptions about the subset of problems we expect to encounter, we can comfortably tailor our algorithms to those while sacrificing performance on remote cases.

The No Free Lunch theorem argued for community detection [6, Appendix C] is inexact.\(^1\) A No Free Lunch theorem relies on a homogeneous loss function: one which does not assume a priori superiority of some predictions. Due to a mismatch between the loss function and the problem domain, the loss in the proof is only asymptotically homogeneous, as the size of the input grows. By substituting an appropriate loss function, we are able to claim an exact No Free Lunch theorem for community detection. The result generalizes to other set-partitioning tasks when evaluated with this loss function (and others listed later), including clustering, $k$-clustering, and graph partitioning.

2 No Free Lunch theorems and community detection

Ref. [6] used the adjusted mutual information (AMI) out-of-the-box, as proposed by Ref. [8]. This computes an expectation over the wrong distribution of clusterings. This mismatch makes their claim of homogeneity accurate only to within $1/\mathcal{B}_N$ for the trivial partitions into either one community or $N$ communities, where $\mathcal{B}_N$ is the $N^{th}$ Bell number, i.e., the number of partitions on $N$ nodes.\(^2\)

\(^1\)Throughout this work, we assume that we evaluate against a known ground truth, as opposed to some intrinsic measure of cluster properties like modularity.

\(^2\)Correcting this is arguably a pedantic demand, for two reasons: 1. The inhomogeneity converges to 0 superexponentially in the limit as $N$ grows. 2. The deficiency is only present when $T$ is one of the trivial partitions, which suggest a lack of community structure. Otherwise, AMI as used is homogeneous. Nevertheless, Wolpert’s proof doesn’t admit a loss that is mostly homogenous. To fix this, we must select the right random model, a sample space for a distribution.
AMI adjusts NMI by subtracting the expected MI, shown in blue, from both the numerator and the denominator:

\[
AMI(C, T) = \frac{I(C, T) - \mathbb{E}_{C', T'}[I(C', T)']}{\max_{C, T} I(C, T) - \mathbb{E}_{C', T'}[I(C', T)']}
\]  

(2)

where \( I \) is the mutual information. The maximum is 1, and the mean should be 0. Hidden in this equation is the subtle choice of distribution for the expectation. This distribution has been unquestioningly what Ref. [1] call \( \mathbb{M}_{\text{perm}} \): all partitions of the same class (a multiset of cluster sizes) as \( C \) or \( T \). E.g., if \( C \) is clusters of sizes 2, 2, and 3, Ref. [8] compute the expectation over all partitions with cluster sizes of 2, 2, and 3.

For our task, \( \mathbb{M}_{\text{perm}} \) is inappropriate. It assumes that we can only produce outputs within that restricted space, when in actuality \( \Omega \) is the set of all partitions on \( N \) nodes. Also, our ground truth is sacrosanct. A distribution over \( T \)s would blur the relationship to our presumptive generative process of the graph \( G = g(T) \)—we’d examine truths better matching our notions of, say, core–periphery partitioning. Instead, we take a one-sided expectation: one over \( \mathbb{M}_{\text{num}} \). Ref. [1] give closed-form formulas for the computation. Substituting this bespoke distribution allows an exact No Free Lunch theorem.

3 An Exact Proof

Though guided by Ref. [6], our proof relies on a truly homogeneous loss function.

**Lemma 1.** When adjusted mutual information computes its expectation in \( \mathbb{M}_{\text{all}} \), it is a homogeneous loss function over the entire space of partitions of \( N \) objects.

**Proof.** Our goal is to show that \( L(T) = \sum_{C \in \Omega} AMI_{\mathbb{M}_{\text{all}}}(C, T) \) is independent of \( T \), where the universe (event space) \( \Omega \) is all partitions of \( N \) objects.

AMI’s denominator in \( \mathbb{M}_{\text{all}} \) is a constant (the max-term is \( \log N \) [1] and the expectation for some \( T \) is independent of \( C \)); we can restrict our attention to the numerator. Now we will now prove a stronger claim than our goal:

\[
\sum_{C \in \Omega} \left[ I(C, T) - \mathbb{E}_{C, T}[I(C, T)] \right] = 0 \quad \forall T
\]

(3)

In limited space, we give the outline of the proof: \( T \) is fixed, and the expectation is independent of the particular \( C \) because the expectation is uniform over the space \( \Omega \). Our goal is true then, because straightforward algebra gives Equation 3 as

\[
\sum_{C \in \Omega} [I(C, T)] - \sum_{C \in \Omega} [I(C, T)] = 0.
\]

(4)

The proof is valid whenever the AMI expectation is computed over the same distribution as the event space \( \Omega \). Tasks which assume a fixed number of clusterings—like \( k \)-clustering and graph partitioning—would prefer \( \mathbb{M}_{\text{num}} \).

---

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
**Theorem 1 (Wolpert, 1996).** For homogeneous loss $\ell$, the uniform average over all $\gamma$ of $Pr(c \mid \gamma, d)$ equals $\Lambda(c) / |Y|$. (Plainly, “There is no free lunch.”)

**Proof.** See Ref. [9].

**Theorem 2 (No free lunch theorem for set-partitioning tasks).** For a set-partitioning problem with a loss function of AMI using the appropriate random model, the uniform average over all $\gamma$ of $Pr(c \mid \gamma)$ equals $\Lambda(c) / |Y|$. (Plainly, “There is no free lunch.”)

**Proof.** Lemma 1 proves that adjusted mutual information using the appropriate random model is homogeneous. As with Ref. [6], our algorithms do not learn, so applying Theorem 1 with an empty training set $d = \emptyset$ completes the proof.

Six other measures [3][4][7][11][12]—including the well-known adjusted Rand index (ARI) [2]—extended to $M_{\text{ARI}}^k$ are also homogeneous, though we elide the proof for space reasons. Each thus satisfies the No Free Lunch theorem’s precondition.

**Summary.** The No Free Lunch theorem is now complete and exact for community detection and clustering. We affirm the claim: “there is no single community detection algorithm that is best overall” [6]. For fair comparisons, we must document our assumptions about the subset of problems where we hope to improve. A coarse way to check assumptions is estimating network properties for the task at hand [10][5]. Finally, we’ve shown the importance of random models for evaluating community detection.

**References**

Graphlet Count Estimation via Convolutional Neural Networks

Xutong Liu1∗, Yu-Zhen Janice Chen1∗, John C.S. Lui1, and Konstantin Avrachenkov2

1 The Chinese University of Hong Kong 2 Inria Sophia Antipolis, France

1 Introduction. Graphlets are defined as $k$-node connected induced subgraph patterns. For an undirected graph, 3-node graphlets include close triangle ($\triangle$) and open triangle (O.). When $k = 4$, there are six different types of graphlets, e.g., tailed-triangle (\triangle) and clique (\bigtriangleup) are two possible 4-node graphlets. The number of each graphlet, called graphlet count, is a signature which characterizes the local network structure of a given graph. Graphlet count plays a prominent role in network analysis of many fields, most notably bioinformatics [4] and social science [3].

However, enumerating exact graphlet count is inherently difficult and computational expensive because the number of graphlets grows exponentially large as the graph size and/or graphlet size $k$ grow [3]. To deal with this difficulty, many sampling methods were proposed for estimating graphlet count with bounded error [2, 3, 5]. Nevertheless, these methods require large number of samples to be statistically reliable, which is still computationally demanding. Moreover, they have to repeat laborious counting procedure even if a new graph is similar or exactly the same as previous studied graphs.

Intuitively, learning from historic graphs can make estimation more accurate and avoid many repetitive counting to reduce computational cost. Based on this idea, we propose a convolutional neural network (CNN) framework and two preprocessing techniques to estimate graphlet count.1 Extensive experiments on two types of random graphs and real world biochemistry graphs show that our framework can offer substantial speedup on estimating graphlet count of new graphs with high accuracy.

2 Method. Given a set of undirected graphs and a particular type of $k$-node graphlet, our objective is to develop a CNN which will be trained using part of dataset with known graphlet counts. After training, the CNN can quickly and accurately predict graphlet counts of other unseen graph samples in the set. Our framework takes the graph adjacency matrix as input and outputs the graphlet count of the input graph. Let us define some notations for our CNN. Let $O^{(l)} \in \mathbb{R}^{N(l) \times N(l) \times C(l)}$ be the output tensor at layer $l$, where $l = 0, 1, 2, 3, N(l)$ denotes the width (and height) along each channel and $C(l)$ denotes the channel size. Let $O_{i,j}^{(l)}$ be the $(i, j)^{th}$ element along the $r^{th}$ channel. We assign $O^{(0)}$ as the graph adjacency matrix. Mathematically, our CNN structure can be described as follows:

$$O_{i,j}^{(l)} = \text{ReLU}(W_i^{(l)} \cdot O^{(l-1)}[i : i + H^{(l)} - 1, j : j + H^{(l)} - 1, :] + b_i^{(l)}], \quad l = 1, 2 \quad (1)$$

$$O^{(3)} = \text{ReLU}([\text{Flatten}(O^{(2)}) \cdot W^{(3)} + b^{(3)})]$$

Equation (1) corresponds to two convolution layers. Each layer applies $C^{(l+1)}$ filters over the input feature map $O^{(l-1)}$, and the $r^{th}$ filter is parameterized by a trainable 3D

---

* Both authors contributed equally to this work

1 Our code is accessible at https://github.com/jjanicechen/GraphletCountEstimationCNN.git

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
weight tensor $W^{(l)}_t \in \mathbb{R}^{H(l) \times H(l) \times C(l)}$, where $H(l)$ denotes the width (and height) of the filter. $[a : b, c : d, :]$ is a slicing function which extracts subset of elements indexing from $a$ to $b$ in width, $c$ to $d$ in height and all in channel to form a new tensor. $\cdot$ is the sum of element wise product of two tensors. After adding bias term $b^{(l)}_t$, we apply ReLU ($\max(0,x)$) as the activation function to obtain the output feature map $O^{(l)}$. Equation (2) is associated with the fully connected layer. It flattens the output $O^{(2)}$ into a column vector, applies $W^{(3)}$, $b^{(3)}$ and ReLU to obtain the estimated graphlet count. Finally, our CNN is trained with back propagation and mean squared error as the loss function.

The above CNN structure inherits the learning power for local structural information of graphs. However, we still need to address the following challenges: (1) The input adjacency matrix is not consistent because graphs in the training set may have different sizes. (2) In practice, real world network dataset may not contain sufficient amount of graph samples for training, which will cause overfitting problem. To address these challenges, we introduce two preprocessing techniques:

**Adjacency Matrix Zero Padding.** To preserve edge connectivity information of all training graphs, we consider the largest graph in the training set, and use its dimension (say $N$) as the dimension of the input adjacency matrix ($N \times N$). For other graphs in the training set, we take each adjacency matrix and pad it with zero till we have an input matrix of dimension $N \times N$. This solves the varying input size problem.

**Swapping Augmentation.** To acquire sufficient data for training, we take advantage of the graph isomorphism property, where a graph can be expressed by different input adjacency matrices having the same underlying network structure. Our approach is to randomly pick indices $i$ and $j$, then swap the $i^{th}$ row with $j^{th}$ row and $i^{th}$ column with $j^{th}$ column of the adjacency matrix. We can repeat the swapping operation for each graph $m$ times to create $m$ more training data. Analogous to flipping or rotation of images, we improve CNN’s generalization ability and thus improve the accuracy of our model.

### 3 Data and Metric

Here, we introduce our testing datasets, benchmarking works, and evaluation metrics.

**Random Graph.** We synthesize datasets with two random graph models: random geometric graph (RGG) and Erdos-Renyi (ER) graph. A RGG is constructed by placing nodes uniformly at random in a unit cube and connecting two nodes by an edge if and only if their distance is within a given radius $r$. In a ER graph, the edge between every two nodes exists with probability $p$. In each synthetic dataset, we have 3000 training graphs, 300 validation graphs, and 300 testing graphs.

**Empirical Network.** We test on three real world biochemistry datasets: MUTAG [6], NCI1 and NCI109 [7]. MUTAG dataset contains 188 mutagenic compound graphs. NCI1 and NCI109 each has 4110 and 4127 chemical compound graphs tested on lung and ovarian cancer cells respectively. For MUTAG, we use swapping augmentation to increase the number of training samples. We also apply adjacency matrix zero padding to make all graphs in each dataset have the same size.

**Benchmark.** We compare CNN with three existing frameworks: GRAFT [5], CC2 [2], GUISE [1], which are based on edge sampling, color coding, and Markov Chain Monte Carlo method respectively.

**Relative Error.** Let $c_i$ be the ground truth graphlet count of sample graph $i$, $c'_i$ be its estimated count, and there are $S$ samples in the dataset. We compute the mean absolute
error of the estimations, \( \text{mae} = \frac{\sum_{i=1}^{S}|c_i' - c_i|}{S} \), and mean of ground truth counts, \( \mu = \frac{\sum_{i=1}^{S}c_i}{S} \). We take relative error as \( e = \frac{\text{mae}}{\mu} \).

**Speed.** To ensure a fair comparison, we do not choose running time as the performance metric since it highly depends on hardware and implementation (e.g., running on GPU/CPU). Instead, we measure the number of arithmetic operations they use. For CNN model, we compute the number of floating-point operations (FLOPs). For benchmarking works, we calculate the number of comparison operations in the algorithms.

4 Result. We test our framework on random graph datasets. For approximating 4-clique counts, our CNN model achieves less than 8% relative error on 50-node RGGs with radius 0.45 and less than 5% relative error on 50-node ER graphs with edge existing probability 0.5. We also train our CNN models for estimating 4-path (\( \Omega_4 \)), 3-star (\( \Omega_3 \)), 5-path (\( \Omega_5 \)) on the empirical biochemistry datasets. The relative errors on all three datasets are less than 20% of the ground truth counts. For estimating 4-path on MUTAG dataset, our model performs especially well making only 6% relative error.

**Fig. 1.** Comparison of the number of arithmetic operations used for estimating 4-clique counts, tailed-triangle counts on 50-node ER graphs with edge existing probability 0.5. (a, c) The number of operations used by each framework. (b, d) The relative error each framework makes.

To compare the speed of our CNN with existing methods, the number of arithmetic operations used are calculated. For a fair comparison, we tune the number of iterations for all benchmarking sampling methods, so that they obtain as close relative errors to that of CNN as possible. Figure 1 (a, c) shows that the numbers of arithmetic operations used by GRAFT, CC2, or GUISE are significantly more than that used by CNN. This result demonstrates that our CNN based graphlet count estimation approach offers remarkable speedup on predicting graphlet counts while still maintaining high accuracy.

**References**

Towards quantitative methods to assess network generative models

Vahid Mirzaebrahim Mostofi¹ and Sadegh Aliakbary²

¹ v.mirzaebrahim@mail.sbu.ac.ir,
² s.aliakbary@sbu.ac.ir,
Faculty of Computer Science and Engineering, Shahid Beheshti University G.C, Tehran, Iran

1 Introduction

There is a long-standing history with graph generative models [2]. Modeling physical and social interactions, discovering new molecular and chemical structures, and constructing knowledge graphs are some example applications of these models [10]. Traditionally, many human-designed modeling approaches (e.g. Barabási-Albert model [1]) are developed to model a particular kind of graphs (e.g. scale-free networks) [10]. But a recent and modern approach of complex network modelling exploit automatic construction based on deep learning techniques such as Generative Adversarial Networks (GAN) [4] and Variational Autoencoders (VAE) [5]. GraphRNN [10] and NetGAN [2] are some recent efforts in employing deep learning for network generation. Since such generators (e.g., GraphRNN and NetGAN) use different techniques of deep learning and result in different generative characteristics, it is necessary to quantitatively evaluate them in order to effectively compare their results. Generally, it is a challenging task to assess the quality of generative models [7]. While defining a quantitative assessment for images and text generative models has been a hard task [7], it is more challenging for graph generative models. It is common for scholars to evaluate graph generative models using qualitative techniques, but qualitative evaluation can be very difficult in case of graphs, specially large graphs. It is not easy for an individual to judge a graph unless the graphs are fairly simple and planar [6].

In order to design a quantitative approach to evaluate the generated graphs, we studied different approaches to the same problem in other fields such as image and text generation. A common and established approach for evaluating in these fields is based on utilizing classifiers [3]. In this paper, we propose a general quantitative method for assessing graph generative models. The proposed method is used to evaluate and compare modern models which are based on deep learning, but it is not only limited to deep learning techniques.

2 Proposed Method

There are existing efforts in the literature for employing classifiers in order to evaluate a generative model [11]. Generative models are designed to synthesize artificial data which are similar to real data. If the generated samples are realistic (similar to real samples), it would be hard for a classifier to distinguish the generated samples from the real
ones [11]. As a result, if an established and accepted classifier fails to effectively distinguish real and synthesized data, the generative model has performed its job well. In other words, the inaccuracy of an accepted classifier for distinguishing real and artificial samples is a witness of the accuracy of the generation method.

We propose to use graph classifiers in order to quantitatively assess graph generators. In this regard, we employ Deep Graph Kernels (DGK) [9] as an accepted and established classifier in the field of complex networks, but other accepted graph classifiers could replace DGK if necessary. If the generated graphs are similar to real ones, the classifier would fail to distinguish them and consequently, the accuracy of the classifier would tend to 0.5. In other words, the closer the classifiers accuracy is to 0.5 (i.e., 50 percent precision), the better the generative model has done its task. As a result, we consider the distance of the classifier accuracy from the 0.5 value as a quantitative value for scoring the generative model. If the accuracy of the classifier \( c \) is equal to \( acc_c \) for distinguishing graphs generated by the generative model \( m \) from real graphs, we define the error of the generative model equal to \( error_m = |acc_c - 0.5| \).

3 Experiments and Results

As an experiment, we compared NetGAN with different variations of GraphRNN. Both GraphRNN and NetGAN have tried to evaluate their results using traditional graph metrics. Deep Graph Kernels (DGK) [9] has been used as classifier in all experiments. We configured DGK with MLE kernel, graphlets features and left the rest of the configurations as default. GraphRNN can use different generative models including VAE, RNN and MLP, and we used RNN and MLP in this experiment.

We conduct our experiments on two types of graphs: 1- Caveman graphs [8] 2- scale-free graphs generated by Barabási-Albert model [1]. For each graph type, we consider several graph samples as the “real graphs”, and then we synthesize similar graphs (with those real graphs as the target networks) using the generative models, and we label the generated graphs as “fake” samples. Finally, we feed the “fake” and the “real” graphs to the DGK classifier. The classifier goal is to distinguish “real” and “fake” graphs. Therefore, the generative model that results in less real/fake classification accuracy is preferred, because it has been able to fool the classifier, and the classifier has failed to distinguish real and fake graphs.

<table>
<thead>
<tr>
<th>Graph Type</th>
<th>GraphRNN:RNN</th>
<th>GraphRNN:MLP</th>
<th>NetGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barabási-Albert</td>
<td>71.0%</td>
<td>78.4%</td>
<td>57.0%</td>
</tr>
<tr>
<td>Caveman</td>
<td>88.7%</td>
<td>99.6%</td>
<td>74.0%</td>
</tr>
</tbody>
</table>

Table 1. Performance measures of GraphRNN and NetGAN using DGK as classifier, lower is better

Table 1, shows the accuracy of the DGK classifier in different scenarios. As the results show, NetGAN is better than GraphRNN in generating scale-free (Barabási-Albert) and Caveman graphs because the classifier accuracy is more close to 0.5 in both
cases. In other words the classifier has failed to identify “real” graphs (target networks) from the “fake” graphs (synthesized networks) generated by NetGAN. Additionally, it seems that generating Caveman graphs is a harder task than synthesizing scale-free graphs. Since the classifier shows more precision in the case of the Caveman graphs.

4 Conclusion

Generating graphs using deep learning approaches is a new and growing field. Therefore, the need to quantitative evaluation metrics for graph generative models is overwhelming. We proposed to utilize graph classifiers in order to evaluate graph models. Our preliminary experiments show that NetGAN performs better than GraphRNN variations for two considered network types. As the next steps of this research, we will use different generators and classifiers in future. Additionally, real graph datasets will be included in our experiments.

References

Network Embeddings For Graph Classification and Visualization

Leonardo Gutiérrez-Gómez¹ and Jean-Charles Delvenne²

¹ Institute for Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM)
leonardo.gutierrez@uclouvain.be,

² Center for Operations Research and Econometrics (CORE)
jean-charles.delvenne@uclouvain.be,
Université catholique de Louvain, Louvain-la-Neuve, Belgium

1 Summary

A main challenge in mining network based data is finding effective ways to represent or encode graph structures so that it can be efficiently exploited by machine learning approaches. Several methods have focused in network representation at node/edge or substructure level. However, many real life challenges such as time-varying, multilayer, chemical compounds and brain networks involve groups of graphs opening additional challenges in graph comparison and representation. Traditional approaches for learning representations relies on hand-crafting specialized heuristics to extract meaningful information about the graphs, e.g statistical properties, structural features, as well as engineered graph distances to quantify dissimilarity between networks.

Unlike the node embedding approaches, here we provide an unsupervised approach to learn embedding representation for input graphs, e.g embedding the entire graph, so that it can be used in numerous graph mining tasks. By using an unsupervised neural network approach on input graphs, we learn a smaller dimensional representation which is used to compare graphs in a Euclidean space. Our method is tested empirically on synthetic and real life datasets and evaluated on graph visualization and classification tasks. Results reveal that our method outperform well known graph-kernels in classification tasks being highly efficient in runtime.

2 Denosing Autoencoder and Graph Embeddings

Given a set of data examples \( D = \{ x^{(1)}, x^{(2)}, \ldots, x^{(m)} \} \), the purpose of the traditional auto-encoder [3] is to learn a non-linear mapping which encodes an input example \( x \in \mathbb{R}^n \) in a smaller dimensional latent vector \( y \in \mathbb{R}^d \) with \( n \gg d \). The encoding mapping has the form of \( f_\theta(x) = s(Wx + b) = y \), generally through a non-linear function \( s \) such as sigmoid or \( \tanh \). A reverse mapping of \( f \) is used to reconstruct the input from the feature space: \( g_{\theta'}(y) = s(W' y + b') = z \). The parameters \( \theta = \{ W, b \} \) and \( \theta' = \{ W', b' \} \) are optimized by minimizing the average reconstruction error over the training set:

\[
\theta^*, \theta'^* = \arg \min_{\theta, \theta'} \frac{1}{m} \sum_{i=1}^{m} MSE(x^{(i)}, z^{(i)})
\]  (1)
In order to allow a robust extraction of meaningful features, the objective Eq 1 is redefined in such a way that the autoencoder will be able to clean partially corrupted input. This modification leads to the denoising autoencoder (DAE) [3] which is trained to reconstruct a clean or repaired version from a corrupted input. This is done by transforming the original input $x$ in $\tilde{x}$ through a stochastic mapping $\tilde{x} \sim q_D(\tilde{x}|x)$. By doing so the AE is forced to learn better meaningful features, producing hidden representations robust under corruption of the input. See Fig 1.

The corrupted version $\tilde{x}$ is mapped through the autoencoder to a hidden representation $y = f_\theta(\tilde{x})$ from which we reconstruct a clean $z = g_\theta'(y)$. Thus, the learned encoder $f_\theta$ allow us to mapping input graphs from their vectorized adjacency matrices to feature vectors in an smaller dimensional feature space. Optimization of Eq 1 is done with mini-batch gradient descent and back propagation algorithm [1].

3 Applications

We use the DAE model in order to learn graph embeddings for some synthetic and real life group of networks. Our approach is evaluated in two different applications: graph visualization and classification.

**Graph visualization.** Our approach for graph visualization consists in represent groups of graphs as $\mathbb{R}^2$ points keeping certain notion of similarity. Graph embeddings are mapped to 2D points through t-SNE tool.

We generate synthetic networks as follow: we adopt the Barabási-Albert (BA) model to generate sequences of power law networks with 4 different average degree. Similarly, we generate Erdős-Rényi (ER) networks with 4 different probabilities of edge generation, producing random networks with average degrees similar to the BA networks. We also experiment on real life data, the primary school network [2]. This dataset containing temporal face-to-face interactions between children in a primary school in France. We plot network snapshots for one day. Visualization is shown in Fig 2.

**Graph classification.** For this experiment we apply our method on brain network data we built from magnetic resonance imaging (MRI). Structural and diffusion MRI data of 91 healthy men and 113 healthy woman is preprocessed in order to create undirected networks. All graphs have the same 84 nodes representing neural Regions of Interests (ROIs). Weighted edges correspond to the number of neural fibers linking two ROIs. The ROI keeps the same correspondence among graphs. The task is to classify connectomes in either male-female class. We report average classification and standard deviation accuracies after 10 times 10-fold CV. Results are shown in Table 1.
Fig. 2. (Left) Visualization of power-law networks. Each point corresponds to a graph generated from Barabasi-Albert model. (Center) Each point corresponds to a graph generated from ER and BA model, both with similar average degree. (Right) Visualizing primary school networks. Each point indicates a graph snapshot within a time-frame of the day.

Table 1. Mean and standard deviation classification accuracies on brains dataset.

<table>
<thead>
<tr>
<th>Weisfeiler-Lehman</th>
<th>Shortest-paths</th>
<th>DeltaCon</th>
<th>Barnett</th>
<th>Emb</th>
</tr>
</thead>
<tbody>
<tr>
<td>61.20 ± 2.16</td>
<td>65.45 ± 1.78</td>
<td>75.37 ± 5.7</td>
<td>65.95 ± 2.54</td>
<td>81.43 ± 2.31</td>
</tr>
</tbody>
</table>

4 Discussion

Our approach provides an unsupervised manner to learn feature embeddings for groups of networks defined on the same set of nodes. By adding an amount of noise to input graphs, i.e. removing or adding a small fraction of edges, the DAE learns robust representation of the data, extracting meaningful features capturing the underlying distribution of the data. Transforming graphs into 2D points allows a visual inspection of the hidden structure of groups of graphs. Results in networks classification task shown that our approach learns useful representations of graphs outperforming some well known graph-kernels of the literature.

References

Node Embeddings in Dynamic Graphs

Ferenc Béres¹, Róbert Pálovics², Domokos Kelen¹, Dávid Szabó¹, and András Benczúr¹

¹ Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)
² Department of Computer Science, Stanford University

1 Introduction

Based on the success of neural network embeddings for natural language processing introduced by the Word2Vec algorithm in [6], several network embedding methods have been proposed recently [8, 10, 5, 9] that are highly successful in multi-label classification and link prediction in a variety of real-world networks from diverse domains.

We consider edge streams such as Twitter mentions or retweets where edges arrive continuously over time and have no duration [1]. In edge streams with fast dynamics, it is challenging to maintain an embedding for tracking and measuring node properties and similarities as the edges arrive. Our main result is an online updateable node embedding that respects the order of edge creation.

In a supervised experiment, we show that our online updateable embeddings capture node similarities better than static embeddings. We selected a Twitter data set for tennis tournaments where ground truth labels can be defined based on player participation. We are not aware of any other graph data sets with high temporal granularity, dynamically changing ground truth labeling available.

Our algorithm performs online machine learning [2] by continuously updating a model as we read the edge stream. Its key ingredients are online gradient descent optimization [4], fingerprinting for fast access to neighborhood [3], and temporal walks [1].

2 Online Node2Vec

Graph embeddings [8, 10, 5, 9] build node sequences and consider them as sentences for Word2Vec embedding [6]. The simplest variants generate random walks [8]. In [10], node pairs with similar neighborhoods are also considered. Node2Vec [5] considers random walks and neighborhood similarity by second order Markov chains that tend to stay closer to the neighborhoods of the last nodes.

We define online updateable embedding methods first by turning the neural network optimization procedure online. The skip-gram with negative-sampling (SGNS) method of [7] uses gradient descent, which we can turn online updateable as in [4]. For generating random walks online, we use the temporal walk method of [1], and for neighborhood similarity, we use fingerprinting [3]. We describe these algorithms next.

In our Temporal Walk algorithm, we maintain time respecting walks as new edges $uv$ arrive with timestamp $t$. First, we update the weight of all walks ending at $u$ by multiplying with the exponential time decay function $\exp(-c(t - t(u)))$ where $t(u)$ is the last visit time of node $u$. We proceed similarly for $v$; for computational efficiency,
update procedures for neighborhood similarity

Algorithm 1 Update procedures for neighborhood similarity

procedure UPDATE NEIGHBORHOOD(uv)
  Update the fingerprints of v by the new edge uv
  
  for all out_neighbors x ≠ v of u do
    Update the fingerprints of x by the new edge uv
  
  for all i in 1...k with fi(v) = u do
    if fi(x) = u then
      call Word2Vec with v and x
  
we can delete walks whose weight becomes very small. For each walk from s to u, we obtain a new walk from s to v by adding the new edge uv. Hence, we increase the weight of walks from s to v by the weight of walks from s to u. After computing all weights that involve uv, we generate s, v pairs for Word2Vec with probability proportional to the weight of walks from s to v.

In our Temporal Neighborhood algorithm, we generate node pairs whose neighborhood is similar and give as input to Word2Vec. For each node v, we maintain k MinHash fingerprints [3] that we update dynamically as new edges enter the network. Given a new edge uv, in Algorithm 1 we select a node x as input pair for v to Word2Vec if its temporal MinHash fingerprint is equal to u. We can also generate fingerprints for higher order similarity in the sense of SimRank and its variants [3].

We can take computational advantage of the fingerprints by a heuristic procedure. Note that updating all fingerprints takes the same order of running time as computing the overlap of the neighborhood of u, v and all their neighbors x, which is computationally infeasible. Instead of a costly exact update, we update the MinHash fingerprints of the neighbors x of u as follows. First, we maintain t(x), the last update time of node x, and keep each fingerprint independently with probability exp(−c(now − t(x))); otherwise we discard the fingerprint. While heuristic update correctly handles the probability for discarding an old fingerprint, unfortunately the probability of selecting the new vertex as fingerprint is much higher than in the exact naive algorithm. Yet, as we see in our experiments, our heuristic procedure works well to generate input pairs for Word2Vec.

3 Experiments

In our experiments, we use the Roland-Garros 2017 (RG17) and US Open (OU17) Twitter data sets [1]. By representing mentions as directed edges between Twitter accounts, we obtain a dynamic network where the instance of the directed edge uv appears whenever user u mentions another user v.

For evaluation, we define node labels that dynamically change in time based on the tennis championship schedule. On a given day, a Twitter account is relevant if it belongs to a tennis player who participated in a game on that day. For each relevant node, we generate a list of similar nodes by computing the dot product of the 128-dimensional embeddings that we evaluate by NDCG. The higher the NDCG, the closer are players who play the same day in the embedding. In particular, players who are active on the same day are more similar to each other than to general tennis related accounts and other players who do not play on the given day.
Table 1. Average NDCG in time over the RG17 and UO17 mention graphs.

<table>
<thead>
<tr>
<th></th>
<th>RG17</th>
<th>UO17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indegree</td>
<td>0.229</td>
<td>0.196</td>
</tr>
<tr>
<td>Node2Vec</td>
<td>0.191</td>
<td>0.207</td>
</tr>
<tr>
<td>Temporal Walk</td>
<td>0.273</td>
<td>0.215</td>
</tr>
<tr>
<td>Temporal Neighborhood</td>
<td>0.278</td>
<td>0.277</td>
</tr>
</tbody>
</table>

Table 2. Similarity list of Rafael Nadal for both methods based on embeddings generated at 12:00 on May 31 (RG17). Relevant players are highlighted in yellow and media accounts in gray.

<table>
<thead>
<tr>
<th>Node2Vec</th>
<th>Temporal Walk</th>
<th>Temporal Neighborhood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 rolandgarros</td>
<td>rolandgarros</td>
<td>rolandgarros</td>
</tr>
<tr>
<td>2 aquiactualidad</td>
<td>tsonga7</td>
<td>Gael_Monfils</td>
</tr>
<tr>
<td>3 KikiMladenovic</td>
<td>DjokerNole</td>
<td>DjokerNole</td>
</tr>
<tr>
<td>4 tenistaturn</td>
<td>KikiMladenovic</td>
<td>KikiMladenovic</td>
</tr>
<tr>
<td>5 emirates</td>
<td>delpotrojuan</td>
<td>stanwawrinka</td>
</tr>
<tr>
<td>6 tsonga7</td>
<td>andy_murray</td>
<td>tsonga7</td>
</tr>
<tr>
<td>7 Gael_Monfils</td>
<td>Gael_Monfils</td>
<td>FerVerdasco</td>
</tr>
<tr>
<td>8 Maly_Tweet</td>
<td>TennisChannel</td>
<td>Simona_Halep</td>
</tr>
<tr>
<td>9 M1INFOA</td>
<td>stanwawrinka</td>
<td>andy_murray</td>
</tr>
<tr>
<td>10 Simona_Halep</td>
<td>GarbiMuguruza</td>
<td>WTA</td>
</tr>
</tbody>
</table>

In Table 1, we show NDCG for both datasets, averaged for all active players every six hours. As a baseline, we calculated Indegree and Node2Vec [5] on the graph snapshot of the last 12 hours. Both Temporal Walk and Neighborhood outperform the baselines. In Table 2, we show the ten most similar accounts to Rafael Nadal. Our online updateable models perform better at showing daily players rather than popular media accounts.

Our data and codes are available at https://github.com/ferencberes/online-node2vec.

References

Network properties captured by graph embeddings

Rémi Vaudaine¹, Christine Largeron², and Rémy Cazabet³

¹ Université Jean Monnet, Saint-Etienne, France, Laboratoire Hubert Curien, remi.vaudaine@laposte.net
² Université Jean Monnet, Saint-Etienne, France, Laboratoire Hubert Curien, Christine.Largeron@univ-st-etienne.fr
³ Université Lyon Claude Bernard, Lyon, France, LIRIS, remy.cazabet@gmail.com

1 Introduction

Graphs are commonly used to model systems from various topics such as social networks, images, virus propagation or brain neurons. Different tools have been developed to study these graphs. One of them is graph embedding, which consists in embedding nodes of a graph into a low dimensional space, i.e nodes are represented by vectors of the chosen dimension. This topic has attracted a lot of attention in the last 3 years.

Nodes in a graph can have various topological properties which can be similar for certain nodes and different for others. Those properties define a similarity in the graph: if two nodes have the same property then they should have the same vectors in the embedding space. A similarity can also be defined in the low dimensional space by an operation between two vectors. In most cases, the goal of an embedding is to encode the similarity in the graph into the similarity in the embedding space. Thus, computing the later similarity between pair of nodes in the embedding gives information about the graph itself.


In this paper, we evaluate the ability of embedding algorithms to preserve specific properties:

- neighborhood, i.e., first-order proximity of nodes in the graph
- structural equivalence, i.e., the second-order proximity
- community structure

We define these properties in section 2, we present a summary of the experimental results obtained in section 3.

2 Properties

A graph embedding tends to preserve some properties of the initial graph while mapping each vertex to a vector.
Firstly, we present the neighborhood of nodes: given a graph \( G = (V, E) \) where \( V \) is the set of vertices and \( E \) the set of edges, the adjacency matrix is defined as \( A_{ij} = 1 \) if the nodes \( v_i \) and \( v_j \) are neighbors and 0 otherwise. Some embedding method will focus on keeping such nodes close in the embedding space. Neighborhood in the graph is also called first-order proximity.

Secondly, we consider the structural equivalence as another local property \[7\]. The distance (euclidean or cosine) between the rows \( A_i \) and \( A_j \) will measure how structurally equivalent the nodes \( v_i \) and \( v_j \) are.

Finally, we study community structure which consists in the presence of clusters of nodes with high intra-cluster edge density and low extra-cluster edge density.

3 Results

In our experiments, we compare the algorithms mentioned in Table 1 on a set of different datasets. The properties preserved by the algorithms as well as their embedding similarity function are indicated. The results for the different properties are shown in figure 1.

<table>
<thead>
<tr>
<th>Name of the method</th>
<th>Graph Similarity</th>
<th>Embedding similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplacian Eigenmaps [1]</td>
<td>First-order proximity</td>
<td>Euclidean distance</td>
</tr>
<tr>
<td>Locally Linear Embedding [9]</td>
<td>First-order proximity</td>
<td>Euclidean distance</td>
</tr>
<tr>
<td>HOPE [8]</td>
<td>First-order proximity</td>
<td>Euclidean distance</td>
</tr>
<tr>
<td>SVD</td>
<td>Second-order proximity</td>
<td>Dot-product</td>
</tr>
<tr>
<td>Node2vec [5]</td>
<td>co-occurrence probability</td>
<td>Dot-product</td>
</tr>
<tr>
<td>struc2vec [2]</td>
<td>co-occurrence probability</td>
<td>Dot-product</td>
</tr>
<tr>
<td>Verse [10]</td>
<td>Personalized Page-Rank</td>
<td>Dot-product</td>
</tr>
<tr>
<td>Spring layout [3]</td>
<td>First-order proximity</td>
<td>Euclidean distance</td>
</tr>
<tr>
<td>Multi-dimensional scaling</td>
<td>First-order proximity</td>
<td>Euclidean distance</td>
</tr>
</tbody>
</table>

Table 1. Compared methods, the graph similarity they preserve and their embedding similarity

a) The neighbor dissimilarity is computed by counting the number of nodes that are in common between the set of neighbors and the set of the closest vectors in the embedding space, then this value is normalized and finally we take the average over all nodes of the graph. Algorithm LE performs very well on most of the graphs. LLE performs a bit worse. Every other algorithm fail to have relevant results.

b) The structural equivalence is measured using the correlation coefficient between the list of distances between rows of the adjacency matrix and the list of distances between vectors representing nodes in the embedding space. Maximal value is one, corresponding to perfect correlation. On the other hand, a score of zero means that they are completely uncorrelated. Singular value decomposition (SVD) performs the best and with the lowest dimension. HOPE still has very good results.
c) As for community detection, we generate graphs of 100 nodes with communities using Dancer [6]. Then we compare the partitions found using embedding algorithms and the ground truth given by the generator. We observe that n2v, spring, MDS and verse show good performance at this task on this specific graph.

References

Detecting Suspicious Entities in the Panama Papers

Mikel Joaristi, Edoardo Serra, and Francesca Spezzano

Computer Science Department, Boise State University
Boise, ID 83725, USA
mikeljoaristi@u.boisestate.edu
{edoardoserra, francescaspezzano}@boisestate.edu

1 Introduction

The Panama Papers are a set of 11.5M documents leaked from the Panama-based offshore law firm Mossack Fonseca. The papers show how Mossack Fonseca facilitated the creation of numerous shell corporations used for illegal purposes, including money laundering, tax evasion, terrorist financing, and evading international sanctions.

The Panama Papers network (PPN) [1] consists of 3M entities (persons, companies, etc.) and 1.4M labeled edges expressing relationships between entities. Within the set of entities, 708K are persons or companies. The data involves people and companies from more than 200 countries.

In this paper, we address the problem of identifying bad entities (people or companies) by leveraging the information contained in the PPN [2].

As not all the people involved in the PPN have broken the law or acted improperly, we identified a set of bad entities to use as a ground truth. More specifically, we compared the entities present in the PPN with several freely available online blacklists. By using an exact string matching approach, we identified around 576 bad entities and labeled them as bad, while labeling the remaining entities as unknown. All used blacklists were collected by the opensanctions.org website [3].

To identify suspicious entities, we propose a new algorithm, called Suspiciousness Rank Back and Forth (SRBF), that computes a suspiciousness score for each entity appearing in the PPN according to its connections to other known bad entities and by exploiting edge directionality alternation. This rank can be used to prioritize entities for subsequent investigation.

We conduct a comparison of our SRBF with existing techniques for node classification, such as centrality measures and network embedding models [4], [5], on the task of searching the PPN for bad entities. Our experimental results show that SRBF is effective and achieves an Area Under the ROC Curve (AUROC) of 0.85 and an Area Under the Recall Curve of 0.87, and both outperforms and finds bad entities earlier in the rank than existing techniques.

2 Suspiciousness Rank Back and Forth

We propose a new ranking measure, Suspiciousness Rank Back and Forth, that is based on edge directionality alternation to transfer suspiciousness to connected entities in the
SRBF is an iterative formula that is computed for a given number of iterations $K$. The intuition behind this algorithm is to avoid to block the propagation of some knowledge because of the edge directionality. Instead, we propagate the knowledge discovered in the previous iterations back and forth by alternating the edge directions in the computation of the proposed measure.

Given an unlabeled directed graph $G = (V, E)$ and a set $B \subseteq V$ of known bad entities, SRBF is given by the following equations:

$$
SRBF^0(u, B, G) = (1 - \delta)b(u, B) + \delta \sum_{(v, u) \in E} \frac{SRBF^0(v, B, G)}{|\text{out}(v)| \times |\text{in}(u)|}
$$

where $b(w, B)$ = \begin{cases} 
1/(2 \times |B|), & \text{if } w \in B \\
1/(2 \times |V \setminus B|), & \text{otherwise}
\end{cases}

$$
SRBF^t(u, B, G) = (1 - \delta)b(u, B) + \delta(1 - \alpha)SRBF^{t-1}(u, B, G^T) + \\
\delta\alpha \sum_{(v, u) \in E} \frac{SRBF^{t-1}(v, B, G)}{|\text{out}(v)| \times |\text{in}(u)|}
$$

At the first iteration ($t = 0$), the suspiciousness rank of an entity $u \in V$ is computed as the recursive equation $SRBF^0$ inspired by the PageRank where the more a node is connected to suspicious nodes, the higher its initial suspiciousness rank is. $b(w, B)$ is a bias introducing the label of an entity being suspicious or unknown. Since in our setting $|B| << |V \setminus B|$, i.e., the number of suspicious entities is much lower than the one of the unknown entities, the bias results higher for suspicious entities than unknown ones. In contrast with the PageRank, in our formulation, the bias is not the same for each entity.

In subsequent iterations, we reverse the graph $G$ and compute a suspiciousness rank where we introduce two sources of bias instead of just one. The first bias is the same as $SRBF^0$’s bias. The second bias is given by $SRBF^{t-1}(u, B, G^T)$ and it is equal to the $SRBF$ score computed for the node $u$ at the previous iteration, i.e., by considering the reverse graph. In practice, we propagate the information back and forth. In the above formulas, $\delta$ is the damping factor as in classical PageRank and $\alpha$ is a factor deciding the importance of the previous computation. We set $\delta = 0.85$ and $\alpha = 0.8$.

It is possible to show that $SRBF^t$ converges in a finite number of iterations since they can be reduced to irreducible aperiodic Markov chains. Also, there exists a finite number of iterations $K$ such that $SRBF^{K-2} = SRBF^K$.

3 Results

Inspired by our previous work [6, 7], we propose a one-class machine-learning based approach where we do an extensive comparison of all the existing techniques (centrality measures and embeddings) and our proposed ranking measure. We will also explore the importance of the edges’ directionality by computing the measures and the node embeddings in the original graph $G$ and the reversed graph $G^T$. We also tried considering the graph as undirected, but the results were worse than the ones presented here. In addition, we also included the rank computed by the SRBF at the first iteration (suspiciousness rank (SR)), and used it as a baseline to empirically show the benefits of
### Method

<table>
<thead>
<tr>
<th>Graph</th>
<th>TransR</th>
<th>Node2Vec</th>
<th>DC</th>
<th>LCC</th>
<th>EC</th>
<th>PR</th>
<th>ACCM</th>
<th>SR</th>
<th>SRBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original ($G$)</td>
<td>0.798</td>
<td>0.691</td>
<td>0.674</td>
<td>0.512</td>
<td>0.300</td>
<td>0.790</td>
<td>0.682</td>
<td>0.300</td>
<td>0.790</td>
</tr>
<tr>
<td>Reverse ($G^T$)</td>
<td>0.808</td>
<td>0.764</td>
<td>0.674</td>
<td>0.512</td>
<td>0.641</td>
<td>0.821</td>
<td>0.658</td>
<td>0.826</td>
<td>0.851</td>
</tr>
</tbody>
</table>

Table 1: AUROC results. Shaded cells denote the best value for the column. Bolded values are the best in the whole table. DC: Degree Centrality; LCC: Local Clustering Coefficient; EC: Eigenvector Centrality; ACCM: DC+LCC+EC+PR; PR: PageRank; SR: Suspiciousness Rank.

the back and forward process. For each case, we performed a 10-fold cross-validation experiment in which 90% of the blacklist was used as training data and the remaining 10% was used to test. This 10% of users used as test set were marked “unknown” in the training set. As we have one class of ground truth only, i.e., the blacklist of bad entities, we used One-class SVM as our classification model and compared the results according to the AUROC and the Area Under the Recall Curve. The Recall Curve is obtained by computing the recall@k for $k$ varying in $\{1, 2, \ldots, |V|\}$. The idea of behind this curve is to have a budget $k$, corresponding to the number of people that is possible to investigate and show how the recall of the algorithm varies according to this budget.

Table 1 shows a comparison of AUROC values for each method and graph. The best method that maximizes the AUROC is SRBF with a value of 0.85, when initially applied on $G^T$. It is worth noting that the edge directionality is really important in the outcome of the experiments. As it can be observed in Table 1, there exist a substantial performance gap between the results of the original graph $G$ and the reversed graph $G^T$. This difference exists in all tested methods except the ones that do not consider edge direction (e.g., DC and LCC). Thus, the direction of the edges is crucial in order to describe what the particular entity is doing. The experiments for the recall curve show that SRBF is able to detect a greater number of bad entities earlier in the rank than the competitors, achieving an Area Under the Recall Curve of 0.870 compared to the 0.847 of the second best model, i.e. the Pagerank.

Overall, our experimental evaluation shows that SRBF outperforms the well-known PageRank measure (that is the best baseline) on the task of identifying bad entities in the Panama Papers.

### References

Temporal Alignment of Reddit Network Embeddings

Siobhán Grayson and Derek Greene

Insight Centre for Data Analytics, University College Dublin, Ireland.
siobhan.grayson@insight-centre.org
WWW home page: https://graysons.github.io/

1 Introduction

Motivated by the concepts and findings being developed for diachronic word embeddings, in this paper, we explore how the application of the same principles can be leveraged to study structural roles from a temporal perspective. In the same way words with a similar meaning will repetitively appear in the same contexts, structural roles in graphs are also defined by the topological company that they keep. However, structurally equivalent roles may or may not occur in close proximity within a graph. Our goal is to map the participants of the popular social media website Reddit\(^1\), into an embedding space that best represents the similarity of the structural roles that they occupy and to then measure how their roles change over time.

2 Methodology

Our dataset consists of 16 subreddits identified by Hamilton and Zhangs in their work on characterising Reddit communities [2]\(^2\) as exhibiting the most “loyal” user features (teams and sports related subreddits) and 13 subreddits identified as having the highest “vagrant” user patterns (see Table 1). When identifying loyal and vagrant communities, Hamilton et al. considered user commenting behaviour on Reddit over time and defined loyal and vagrant users as follows: Loyal members are users who for two consecutive months have submitted at least 50% of their comments to one Subreddit. Vagrant members on the other hand are defined as users who comment 1 to 3 times within a Subreddit in one month but then do not submit any comments the subsequent month despite still being active on Reddit. For temporal analysis, we partitioned data from late January to October 2014 into three windows each consisting of three months.

<table>
<thead>
<tr>
<th>Class</th>
<th># SR</th>
<th>(# V_{T1})</th>
<th>(# E_{T1})</th>
<th>(# V_{T2})</th>
<th>(# E_{T2})</th>
<th>(# V_{T3})</th>
<th>(# E_{T3})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loyal</td>
<td>13</td>
<td>15,319</td>
<td>89,496</td>
<td>15,193</td>
<td>91,138</td>
<td>14,531</td>
<td>87,149</td>
</tr>
<tr>
<td>Vagrant</td>
<td>16</td>
<td>13,462</td>
<td>22,323</td>
<td>14,030</td>
<td>23,831</td>
<td>13,314</td>
<td>22,247</td>
</tr>
</tbody>
</table>

Table 1: Notation - SR: Subreddits, \(V_{T1}\): Nodes Temporal Window 1, \(E_{T1}\): Edges temporal window 1.

Fig. 1: Cosine Similarity for alignment evaluation.

\(^1\)The url address for this site is: https://www.reddit.com/
\(^2\)Further details can be found on the webpage where the dataset is available to download: http://snap.stanford.edu/data/web-RedditNetworks.html
For the purposes of this study, two users are defined as having corresponding roles if their occurrences within the Reddit networks are structurally equivalent. To assess user role variation over time, we first select the 100 highest frequency participants for each 3 months and then use the overlap of this set that spans all window partitions to extract temporally related networks. Once we have our temporal networks, actors are then described in terms of their roles by applying the directed and weighted version of the graph embedding algorithm, \textit{struc2vec} [4], specifically designed to capture structural equivalence between nodes.

![Embedding Spaces](image.png)

\textbf{Fig. 2:} Loyal subreddit r/ACMillan before and after alignment, and then dimension reduction.

The embedding spaces in this study are then aligned using normalised orthogonal Procrustes, an approach popular for aligning diachronic word embeddings [1, 5], as it derives the optimal rotation of a “source” matrix with respect to a “target” matrix without scaling by minimising the sum of squared distances between elements. Alignments can be evaluated by generating a second embedding matrix for the same time period and comparing the cosine similarity between vectors. Fig.1 displays the average of aggregated cosine similarity results \(1/N \sum_{i=1}^{N} \cos(v_t^i, v_{t+\Delta}^i)\) and the standard deviations computed across all embedding spaces and their duplicates for both before (Baseline) and after alignment. In all cases, rotations reduced the dissimilarity between temporal user embeddings. Fig.2 illustrates the affect of the alignment process by visualising the embeddings for time period 2 (T2) being aligned to time period 1 (T1) using t-SNE [3]. Occasionally, derived anchors were not dispersed throughout the embedding space which resulted in the sign of the eigenvectors being flipped during PCA. To resolve this, further alignment of roles is applied by changing the signs of equivalent principal components to agree if they do not already.

Once embeddings have been aligned, we can compute the cosine distance between an actor's embedding at time \(t\) and \(t + \Delta\): \(1 - \cos(v_t^i, v_{t+\Delta}^i)\) to detect changes in an individual’s role across time. Greater distances indicate a larger deviation in the type of roles a participant occupied during different periods and vice versa. We then aggregate individual results to derive a mean cosine distance score for each subreddit so that comparisons can be made across loyal and vagrant user role fluctuations. In order to observe the variation of community roles over time, we first find the maximum number of clusters present across time periods to be compared by decomposing the 128 dimensional embedding spaces into 2 dimensions using PCA. The Elbow method using Euclidean Kmeans is then applied to determine the number of clusters present. The maximum equal cluster number across two embedding spaces is recorded and 1-Nearest Neighbours is applied to compute the Euclidean distance between the closest aligned centroids. The resulting value provides insight into how much the general roles present within a subreddit community have changed over time. Finally, silhouette scores are also computed for each embedding space to determine whether roles evolve to become more or less acutely defined over time.
3 Results

The results of our analysis are depicted in Fig.3. The first figure, Fig.3(a), illustrates the average cosine distances computed for each subreddit mapped from time period T1 to aligned T2. The majority of user cosine distances continue to remain as dissimilar to each other in the second temporal embedding space, time period T2 aligned with time period T3. Hence, our preliminary findings suggest that although individual users of Reddit may change role frequently, the universal community level roles remain relatively static in comparison. The static nature of community roles in comparison to user roles is further examined by calculating the average Silhouette Score for each subreddit. The average Silhouette scores, Fig.3, indicate that it’s not an isolated scenario.

![Fig. 3: The temporal user and community role dynamics observed via three different metrics for comparing similarity: Cosine distance, Euclidean distance, and Silhouette scores.](image)

4 Conclusion

In this paper, we applied the role embedding algorithm, struc2vec to three consecutive temporal windows of user networks and then aligned the resulting embedding spaces using orthogonal Procrustes. Overall, our findings suggest that while participant roles fluctuate a lot, the ubiquitous community roles present are a lot more static. However, further analysis is required and we hope to extend the current work to explore subreddits such as AskReddits, Debate Redsits, Questions Redsits, where roles are generally quite distinguished to allow for further comparisons to be made.

References

Prediction of the disease controllability in a complex network using machine learning algorithms

Richa Tripathi\(^1\), Amit Reza\(^1\), and Dinesh Garg\(^2\)

\(^1\) Indian Institute of Technology Gandhinagar, Gujarat, 382355, India, richa.tripathi@iitgn.ac.in
\(^2\) India Research Laboratory, Bangalore, India,

1 Introduction
The problem of disease spreading on networks has been studied using various epidemic models such as SI, SIR, SIRS, etc. \cite{2}. On the other hand, there have been several works \cite{5} for determining essential network and disease parameters for a spread of an epidemic on networks. Machine learning algorithms such as regression, deep learning, classification, and clustering have found successful application in networks. For example for optimal graph partitioning into community structure \cite{3}, classification of diseased networks from the control networks\cite{6} using data from brain imaging, classification of networks using deep learning into various model networks such as Small-World, Scale-Free, Random networks, etc. \cite{7}.

The basic reproduction number \(R_0\) of the disease dynamics known as the basic reproduction ratio determines the stage of disease such that if \(R_0 < 1\) disease will die out in the long run and if \(R_0 > 1\) the disease-free stationary state is asymptotically unstable. Hence, to know if the population of individuals is free from infection outbreaks, predicting \(R_0\) correctly for a disease spread is of paramount importance.

In the present work our aim is to train machine learning regression models with features as networks parameters on various model networks and the corresponding \(R_0\) values as output labels. Finally with the trained model, we predict \(R_0\) on test networks. These model networks were of four kinds - Random (\textit{Rand}), Scale Free (\textit{SF}), Small World (\textit{SW}) and Barabasi-Albert (\textit{BA}) networks and were generated in large numbers for appropriate training. We optimize correct prediction of \(R_0\) on test samples irrespective of network model using various regression methods based on metrics such as mean squared error and coefficient of determination\cite{4}. We present the network generation scheme, disease spreading model and dynamics on networks, various regression schemes and corresponding results in the following sections.

1.1 Epidemic model on networks

\[
\begin{align*}
\frac{dS(t)}{dt} &= -aS(t)I(t) + cI(t) + eR(t) \\
\frac{dI(t)}{dt} &= -aS(t)I(t) - (b + c)I(t) \\
\frac{dR(t)}{dt} &= bI(t) - eR(t)
\end{align*}
\]

where \(S\) is the number of susceptible, \(I\) is the number of infected, \(R\) is the number of recovered. \(a\) is the infection rate constant; \(b\) is the recovering rate constant; \(c\) is the death rate constant related to the disease; \(e\) is the death rate constant about natural causes.

The basic assumption of this model is \(S(t) + I(t) + R(t) = N\), i.e. the total number of individuals remains constant.

The basic reproduction number can be defined as \(R_0 \equiv \frac{aN}{b+c}\). From the steady-state analysis of the Eq. 1, it is clear that stability of the disease spreading can be classified based on a different limit of \(R_0\). The stationary state is asymptotically stable if \(R_0 < 1\) and unstable if \(R_0 > 1\); and the endemic stationary state is unstable if \(R_0 < 1\) and asymptotically stable if \(R_0 > 1\).

2 Machine-learning algorithm based learning scheme

2.1 Linear regression
Linear regression (LR) \cite{4} is a statistical method to analyze the linear relation between the observed responses (independent variable) and target value (dependent variables) of a data set. Mathematically, the target value is defined as a linear combination of the observed responses i.e.

\[
\hat{y} = w_0 + w_1x_1 + w_2x_2 + \cdots + w_p x_p
\]
\( w = \{w_1, w_2, \ldots, w_p\} \) represents a coefficient vector which need to be optimized to get a optimal target value. Basically, LR fits a linear model based on optimum weight (coefficient) vector to the residual sum of squares between the observed responses, and the responses predicted in the data set. Mathematically, it is involved in solving an optimization problem of the form:

\[
\min_w ||y_{true} - y_{predicted}||^2
\]  

\[ 2 \]

2.2 Support vector regression

In machine learning, support vector regression (SVR) [6] is a supervised learning model to predict the trend of a data set based on some non-linear transformation (kernel) function. In SVR, the observed responses is first mapped onto another feature space using some nonlinear transformation to construct a linear model: \( f(input, weights) \equiv f(x, w) \) in this feature space. The linear model can be written as the linear combination of kernel (non-linear) function with corresponds weights \( (w_i) \).

\[
f(x, w) = \sum_{i=1}^{n} w_i g_i(x) + b
\]

Where \( g_i(x) \) is the non-linear mapping and \( b \) is the bias term. The accuracy of estimation is measured by a well-defined, distinct loss function \( L(y, f(x, w)) \), called \( \epsilon \)-insensitive loss function. Therefore, SVR performs a linear regression in the transformed feature space using \( \epsilon \)-insensitive loss function and simultaneously tries to reduce the predicted error by optimizing \( w \) vector.

The kernel trick allows the model to fit the maximum-margin hyperplane in the transformed feature space optimally. The transformed space may be high dimensional. Some of the well-defined kernel function is as follows:

1. Linear kernel:

\[
k(x, x') = \langle x, x' \rangle
\]  

\[ 3 \]

2. Polynomial kernel:

\[
k(x, x') = (\gamma \langle x, x' \rangle + r)^d
\]  

\[ 4 \]

3. Radial basis function kernel (rbf):

\[
k(x, x') = \exp(-\gamma ||x - x'||^2)
\]  

\[ 5 \]

3 Data sets

The dataset comprises of a (2000X6) feature matrix and a column of length 2000 of target values \( (R_o) \). The features are the network properties forming the columns of the feature matrix - average degree, shortest path length, clustering coefficient, network diameter, and maximum degree. There were 500 example networks generated for each of the four model networks. The training set consisted of 1500 example networks (permuted over all the models), and the remaining 500 networks were used for testing the regression model. The random networks were generated using NetworkX [1] module by taking 500 values of the fraction of edges in the network, \( p \). Similarly, different networks corresponding to each network model were generated by varying a parameter for network generation such that it takes 500 values in a particular range. The range of network generation parameter values was carefully chosen, such that the network properties fall in more or less the same range for all the models [5]. After these networks were generated, the SIR model was simulated for 100-time steps on each of these, and the quantities \( S, I, R, R_o \) were calculated and saved along with the corresponding network properties. The data matrix was formed with these network properties as variables and \( R_o \) corresponding to disease spread on this network constructed the target variable.

4 Results and Conclusion

The linear regression resulted in a good fit when the networks from the same model were used for training and testing. The mean squared error and variance scores for the Rand, SF, SW and BA networks were (0.01, 0.99), (0.14, 0.87), (0.02, 0.99) and (0.03, 0.99) respectively, indicating a good fit (see Fig 1 (a), (b), (c), (d)). However, the linear regression failed when all the networks from all the models were used for training (see Fig 1 (e), (f)).

SVR with linear, polynomial and rbf kernels show mean squared error and \( R^2 \) as (8.14, 0.46), (19.65, -0.31) and (0.06, 1.00). This shows that rbf is the best kernel for our purpose and predicts \( R_o \) most accurately (see Fig 2). We took six network structural properties: average degree, shortest path length, clustering coefficient, network
Fig. 1. Figures showing results of linear regression fit for $R_o$ for (a) Random Networks (b) Small World Networks (c) Scale Free Networks (d) Barabasi Albert Networks. Figure (e) depicts that that we cannot fit a linear model for all the networks $R_o$, and (f) shows the corresponding true labels and predicted labels $R_o$.

Fig. 2. (a) The figure shows that the linear kernel in SVR does not offer a good fit to the $R_o$, as the data and the model do not fit each other. (b) The figure shows that the polynomial kernel in SVR does not offer a good fit to the $R_o$, as the data and the model do not fit each other. (c) The figure shows that the RBF kernel in SVR does offer a good fit to the $R_o$, as the data and the model fit each other to a good degree.

References


Topological Data Analysis of Attributed Networks using Diffusion Fréchet Functions with Ego-Networks

Warren Keil$^1$ and Mehmet Aktas$^1$

University of Central Oklahoma, Edmond OK 73034, USA,
wkeil@uco.edu, maktas@uco.edu

Abstract. In this paper we study the attributed networks using topological data analysis. We first extract the ego network of each node. We then define the diffusion Fréchet function over ego networks, which takes both network topology and attribute information into consideration, to extract the topological features. Next, we encode this information in persistent diagrams using functional filtrations and finally reach our goal by combining the distances within the persistence diagrams with machine learning algorithms. Our experiment shows that our method can be promising in clustering the attributed networks.

1 Introduction

Topological data analysis (TDA) has been a very active area of research in the past couple of decades. One of the main goals of TDA is to try to extract certain topological features of data, sometimes referred to as the shape of the data [4]. The idea is that these inherent features of the data are embedded in the shape of the data, and so by finding these topological features associated with the data, it may be possible to uncover information contained in the data not accessible by other methods.

A popular and exciting area of TDA has been on network data. Networks are structured data representing relationships between objects, where nodes and edges represent objects and relationships respectively. In this paper, we consider attributed networks, the networks where vertices are affiliated with multidimensional attributes. The main focus of our study is to be able to use the attributes and network topology of an attributed network to find structure and topological feature inside the data. If successful, this can be used to enhance clustering and other machine learning algorithms.

2 Background

While there has been a considerable amount of research performed on network data, we find a few studies very interesting and influential for our own research. In the paper [1], the authors’ research method involved first looking at one vertex in the dataset, and then creating a subnetwork consisting of every vertex connected to the original vertex. The networks used in their study was from Facebook with vertices representing people and edges representing friends and family. This subgraph is referred to as an ego network. They then use a weighting function to assign weights to each edge of the ego network based on the shared attributes of each adjacent vertex. Through this method, they are
able to identify sub-communities within Facebook networks and even predict if a new subgroup would form [1].

As our second main reference, Martinez, et al. [2, 3] study the results of modifying the Fréchet Function by using a heat function as the kernel instead of the Euclidean norm. To illustrate this, the classical Fréchet function is defined as

\[ F_\alpha(x) := \int_{\mathbb{R}^d} \|x - y\|^2 \alpha(dy). \]

The modified function, defined the Diffusion Fréchet Function (DFF) is,

\[ F(x) := \int_{\mathbb{R}^d} d^2(x, y) \alpha(dy), \]

where \( \alpha \) is some probability measure and \( d(x, y) \) is the solution to the heat equation, also known as Green’s function. They are then able to show that the DFF is able to detect features in data where the classical Fréchet function is not able to. Specifically, the DFF is able to detect all the modes multi-modal data, as well as uncover other information contained in data. They are then able to show that the DFF is stable with respect to the \( p \)-Wasserstein distance. The authors also modify their DFF to be able to compute diffusion distances between network data. The DFF on networks as described in [3] is also proven to be stable with respect to the Wasserstein distance.

3 Approach

Our method of analyzing attributed networks is to first define an ego network for each node in the network as described in [1]. We then compute the DFF over each ego network. The DFF assigns values to each node. Next, we perform a functional filtration using the DFF values and the simplical complex structure in the graph. We then compute the persistent homology of each of these filtrations to store this information as diagrams (or barcodes).

The final step before clustering is to store the all of the information in a distance matrix. The distance matrix is constructed by computing the Wasserstein distance between each diagram. We then use established statistical learning techniques to find structure within data.

4 Results

The data we use in our study is the Amazon co-purchasing dataset, which contains detailed product descriptions, ratings, and an addition, directed network, which describes which pairs of items were bought by the same user. We use first 200 items in the dataset in our experiment. The product descriptions, ratings, number of reviews, and subcategories are the attributes of our network.

We modify the weight function in [1] to be defined as the weight between two alter equals the logistic sigmoid function times the number of reviews times the average rating of two items. We also require their group types to be equal and their sales rank.
to be in a similar interval. This function then defines the structure of each ego network and the weights of the edges of each of these networks.

The statistical learning algorithm we find to be most useful is the $k$-means clustering algorithm. We run the $k$-means clustering algorithm on each ego network using the diffusion distances as the metric. We run the algorithm set to detect 2 through 5 means for each ego network. Next, we analyze the features of each ego network based on which cluster the $k$-means algorithm assigned it to.

What we find is that nodes in a cluster have a lot of features in common such as the type of item or movie genre, but some features were not as apparent. Our current and future work on this project is to identify these surprising cluster results and to first see if they are statistically significant. That is, using hypotheses testing, see if the different clusters can be reliably used to predict the purchasing habits of Amazon customers based on their clustering. We then are to go back and verify that the $k$-means algorithm does require our methods of using the ego network weighting algorithm, the DFF, and the homology groups calculation. This will verify with confidence whether our method of utilizing topology data analysis methods is revealing information in attribute network data not currently accessible by statistical or machine learning methods.

References

Part VIII

Modeling Human Behavior
Learning strategic behavior in social and economic networks

Nicolò Pagan and Florian Dörfler

Automatic Control Laboratory, ETH Zürich, CH-8092 Zürich, Switzerland

It is commonly recognized that certain network positions provide benefits to actors. Thus, it is plausible that they strategically choose their relations to optimize their position in an incentive-guided fashion [1]. Strategic network formation models assume actors aim at maximizing a well-defined utility function based on their network position. [2] showed that, under the assumption of homogeneous payoff functions, the Nash equilibrium definition allows to analytically infer the individual preferences from the stability of specific network motifs. For instance, complete networks are associated with high preference for local support and social influence, and with relatively low cost of maintaining relationships. Yet, real-world networks differ from these specific network topologies, and actors have heterogeneous behavior: some might be more willing to create close triads to receive social support, some might prefer to build structural bridges in order to get competitive advantages. Even though the problem of analytical identifying the exact individual preferences becomes then intractable, the Nash stability condition provides a quantitative tool to infer the individual preferences.

1 Model

We consider weighted and directed graphs $\mathcal{G}$, whose nodes are identified with a set $\mathcal{N}$ of agents and whose arcs weights $a_{ij} \in \mathcal{A} \subseteq [0,1]$ denote the strength of the directed relations among agents $i$ and $j$. We assume that each agent $i$ has control on the weights $a_{ij}$ of her outgoing links (followers), while she cannot change her incoming links $a_{ji}$ (followees). Agents are endowed with a payoff function and assumed to be rational, thus they aim at maximizing it. The payoff function $V_i$ of agent $i$ depends on her action, $a_i$, and on the collection of all other agents’ actions, $a_{-i}$, and it is composed by a parametric combination of several individual incentives taken from different strategic network formation model (see [3] for a survey), reading as

$$V_i(a_i, a_{-i}|P_i) = \alpha_i \cdot t_i(a_i, a_{-i}, \delta_i) + \beta_i \cdot u_i(a_i, a_{-i}) - \gamma_i \cdot c_i(a_i),$$

where $t_i(a_i, a_{-i}, \delta_i)$, $u_i(a_i, a_{-i})$ and $c_i(a_i)$ measure respectively the influence, the clustering and the cost to agent $i$. A sketch of the different contributions of the payoff function is shown in Figs. 1a, 1b and 1c. The parameters $\alpha_i, \beta_i, \gamma_i, \delta_i$ are real numbers, with $\alpha_i \geq 0$, $\gamma_i > 0$, and $\delta_i \in [0,1]$, and they model the individual actor’s preferences $P_i = \{ \alpha_i, \beta_i, \gamma_i, \delta_i \}$. For instance, large $\alpha_i$ and $\delta_i = 0$ are symptomatic of a preference for direct social influence while increasing values of $\delta_i$ exhibit more interest towards reciprocity and indirect influence. Yet, low values of $\gamma_i$ represent lower cost of maintaining links and increasing values of $\beta_i$ correspond to higher attention to the clustering coefficient, thus to the social support perceived when a friend of a friend is a friend [4].
Note that $\beta_1$ is allowed to take negative values. In contrast with Coleman’s theory of closed triads [5], this enables us to model a number of contexts in which agents prefer ties with unconnected others, as in Burt’s theory of structural holes [6]. Among the different stability notions available in the game theory literature, we use the Nash equilibrium one, which is consistent with the selfish attitude of agents. Formally, $G^*$ is a Nash equilibrium if for all agents $i$ and for all agents $j$,

$$V_i\left(a_{ij}, a_{-ij}^*, a_{-ij}^*|P_i\right) \leq V_i\left(a_{ij}^*, a_{-ij}^*, a_{-ij}^*|P_i\right), \quad \forall a_{ij} \in \mathcal{A}.$$ 

Under the assumption of homogeneous individual preferences, i.e., $P_i = P$ for all agents $i$, it is possible to analytically derive a correlation between particular network topologies $G^*$ and individual preferences $P^* = \{\alpha^*, \beta^*, \gamma^*, \delta^*\}$, as shown in [2]. Conversely, when dealing with real-world networks of heterogeneous agents, the Nash equilibrium condition enables to estimate the likelihood that a certain actor has a particular individual preference $P_i = \{\alpha_i, \beta_i, \gamma_i, \delta_i\}$. Thus, we construct a likelihood function $L_i$ such that it penalizes user preferences which violate the equilibrium conditions, namely

$$L_i(P_i) = \sum_{j \neq i} \left( \sum_{a_{ij} \in \mathcal{A}} \min \left\{ 0, V_i\left(a_{ij}^*, a_{-ij}^*, a_{-ij}^*|P_i\right) - V_i\left(a_{ij}, a_{-ij}^*, a_{-ij}^*|P_i\right) \right\} \right),$$

and we define the maximum likelihood estimate $\hat{P}_i$ as the centroid of the points which optimize the likelihood function, namely $\hat{P}_i = \arg \max_{P_i} \{L_i(P_i)\}$.

2 Results

We firstly tested our model on a binary dataset describing the trading relationships of manufactured goods among 24 countries in the period 1965-1980 [7]. The model clusters countries with similar trading behavior, as shown in Fig. 2a, indicating that Japan had a similar strategic behavior compared to the USA and Switzerland, namely more inclined towards the social support parameter $\beta$. This evinces its rapid escalation in the GDP per capita ranking during that period [8] compared to other economies, e.g., Spain, which focused more on direct influence $\alpha$. Second, we analyzed the unweighted...
directed network of confiding relations within a large Australian bank [9]. The individual preferences estimation analysis (see Fig. 2b) shows that a more competitive behavior (negative values of $\beta$) is typical of high hierarchical positions. Conversely, low-ranking positions are more inclined towards social support (positive $\beta$). Interestingly, the branch manager exhibits high $\delta$ value, which is symptomatic of high reciprocity.

Summary. We presented a network formation model in which agents strategically control their outgoing ties in order to optimize their own utility function. We used the Nash equilibrium condition to construct a maximum likelihood estimation model capable to learn the individual user preferences. We successfully applied this model to two different datasets, both describing competitive contexts, yet delivering controversial results. Noteworthy, the estimation procedure can be adapted to different descriptions of the payoff function as well as to different definitions of equilibrium.

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Exploring Information Dissemination Strategies to Foster Collective Intelligence

Dirk-Jan van Veen\(^1\) and Hans R. Heinimann\(^1\)

Singapore-ETH Centre, 1 Create Way, CREATE Tower, #06-01, 138602, Singapore,
dvanveen@frs.ethz.ch

1 Introduction

The world is growing into a network of interconnected systems, the behavior of which is getting more complex. This complexity implies that we will experience behavioral patterns in the future that we cannot predict with historical data. As a consequence, decision-making bodies are facing the challenge to look very carefully at what is happening in a specific situation, capturing weak signals, making sense out of it, and framing the problem for decision-making. Herbert Simon emphasized already in the 1980s [1] that this framing process, which precedes the decision-making process, has not been studied and still is not understood well. The purpose of this paper is to develop a proof of concept model to explore how unequal turn-taking and various information sharing strategies affect information dissemination in groups.

2 Model

We developed an Agent-Based Model that comprises a fully connected network of ‘virtual humans’ (agents) that aggregate information in order to solve a decision-making problem. Each agent has a memory that can store \(n\) information units concerning \(m\) decision alternatives, and has the ability to assess its preferred decision alternative. An agent receives information either at the start of the simulation or through interaction with other agents. If she receives new information, she reassesses her preference for a decision alternative. Finally, she casts a vote for her preferred decision alternative. Decisions are made on the basis of majority consensus and agents can try to persuade each other by sharing novel information.

Agents take turns at sharing information. Each turn, one agent is selected stochastically to share a unit of information. Turn-selection is biased, meaning that one ‘dominant’ agent has a consistently higher chance to be selected. The other agents are all equally likely to communicate. This mechanism represents the phenomenon of participation inequality, an unequal distribution of turn-taking, observed in many real-life groups [4]. The selected agent then picks a unit of information from her memory (the information it picks will depend on the information sharing strategy) and shares it with all other agents at once, without distortion.

At the start of the simulation, an equal amount of information is distributed to each agent. One half of an agent’s information is common information - meaning that it is known by all other members of the group. The other half is unique to individual agents.
Common information favors a suboptimal decision alternative, whereas all unique information combined reveals the optimal decision alternative. As a result, the optimal decision alternative is hidden to the group as a whole, but can be discovered if agents share unique information with each other. This paradigm is also known as a hidden profile problem [2].

The simulation stops if a majority of the group members vote for the correct decision alternative-a ground truth- or when no novel information has been shared for 30 consecutive turns.

The information dissemination strategies we impose on the group have two components: on the one hand, we vary the ratio of participation inequality, on the other hand, we impose an information sharing strategy that determines how an agent selects the information it wants to share. The information sharing strategy demonstrated here is random: an agent selects one unit of information at random from the set that was given to her at the start of the simulation. In addition to varying the dissemination strategies, we also vary the amount of total information in the simulation. This acts as a proxy for problem complexity. We performed the simulation experiment with five agents to prevent decision ties and because there are no appreciable decision performance gains for group sizes beyond five members (e.g. [3]).

**Results**

Figure 1 illustrates the performance of the group in terms of speed and accuracy. Accuracy is expressed as the percentage of correct group decisions, speed as the expected number of interactions needed to reach a correct decision.

The contour plot at the top of the graph denotes accuracy. We see that high participation inequality impairs decision accuracy. In fact, high inequality impairs decision accuracy exponentially. For instance, an increase in inequality from 20% to 40% barely affects accuracy, whereas an inequality increase from 70% to 90% sees accuracy slide by almost 20 percentage points. Besides that, we observe that solution accuracy is invariant to the amount of information the group is faced with. This can be explained by the fact that we did not impose a limit on the agent’s cognitive ability to oversee all information and process it without bias. Non-random interaction patterns and cognitively limited agents (not presented here) have demonstrated impaired accuracy with an increasing information load.

The surface plot in gray below depicts the time needed to reach consensus. Here we see hierarchical groups (high participation inequality) make faster decisions than equal-input groups. Furthermore, decision time increases linearly with the number of information units involved in the decision problem. Equal-input groups also face a steeper increase in decision time than hierarchical groups.

**3 Discussion**

The simulation experiment has established that the detrimental effects of participation inequality are non-linear. Granted that this simulation focuses mostly on top-down
Fig. 1. Speed-accuracy trade-off for a fully connected, five-agent network solving a hidden-profile decision problem. Top contour plot depicts what percentage of hidden profile problems have been solved successfully. Bottom surface plot shows the median number of interactions required to solve those hidden profile problems. One data point consists of 20,000 - 150,000 repeat simulations, depending on how quickly the stability criterion was satisfied.

group processes, it supports Woolley’s [4] research on collective intelligence of groups, which has also indicated that participation inequality can lead to a decrease in group performance.

Here we have only have space to present one simulation result. In future work, we will explore various sharing strategies that resemble real-world interaction patterns such as advocating one’s own opinion, groupthink, and devil’s advocacy. Furthermore, we will explore scenarios in which agents have limited cognitive processing capacity.

References

Mobility helps problem-solving systems to avoid Groupthink

Paulo F. Gomes¹,², Sandro M. Reia¹ Francisco A. Rodrigues³,⁴,⁵, and José F. Fontanari¹

¹ Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos, São Paulo, Brazil.
paulofisicajatai@gmail.com
² Instituto de Ciências Exatas e Tecnológicas, Universidade Federal de Goiás, 75801-615 Jataí, Goiás, Brazil.
³ Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Caixa Postal 668, 13560-970 São Carlos, São Paulo, Brazil.
⁴ Mathematics Institute, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK.
⁵ Centre for Complexity Science, University of Warwick, Coventry CV4 7AL, UK.

1 Introduction

Learning through observation and imitation are central to the success of the human species. It has been hinted that agent-based models of imitative learning could reproduce some features of the problem-solving performance of task forces. However, if the agents are too propense to imitate their more successful peers or the group is too large, then the imitative learning search yields a calamitous performance which is reminiscent of the Groupthink phenomenon of social psychology that occurs when everyone in a group starts thinking alike. Here (https://arxiv.org/abs/1808.02931) we examine if this remedy to Groupthink, namely, the mobility of agents, works for the agent-based model of imitative learning too.

2 Model

We study the performance of a group of $M$ agents by calculating the time (number of iterations) needed to solve a problem, which is to find the unique global maximum of a fitness landscape generated using the NK model [1–4]. The agents are randomly placed in a size $L$ square box with periodic boundary conditions. The density of agents $\rho = M/L^2$ is kept constant. Each agent $i$ is the center of its interaction circle with radius $d = \alpha d_0$, where $\alpha > 0$ and $d_0 = 1/\rho$. An agent $j$ will be neighbor (and able to interact with) of the agent $i$ if the distance between them is less than $d$ ($j$ is inside the circle centered on $i$).

Each agent has a string $\phi$ with $N$ bits, a fitness $F(\phi)$ and explores the fitness landscape aiming at finding its global maximum $\Phi$. They can either suffer a mutation with probability $1-p$ or imitate the model string (when it exists) with probability $p$. The model string is the neighbor agent (inside the interaction circle) with highest fitness. The mutation consists of the flip of one random bit from the string. On the other hand,
in the imitation the agent flips one bit which is different from the model string, so that at the end it becomes more similar to the model string.

The algorithm is as follows. One agent \( i \) is randomly selected and then moves a fixed distance \( \delta d_0 \) in a random angle \( \theta \in [0, 2\pi) \). The set of agents inside its interaction circle change. With its new neighbors, the agent \( i \) decides if it will mutate or imitate the model string (if it exists). The new string \( \phi_i \) is generated and the new fitness \( F(\phi_i) \) calculated. The time is incremented by \( \Delta t = 1/M \). If \( F(\phi_i) = \Phi \) (which means agent \( i \) found the global maximum) the search ends and \( t = t_s \) is the required time to solve the problem in this run. The efficiency of the search is measured by the total number of string operations necessary to find that maximum, i.e., \( M_{ts} \), and so the computational cost of the search is defined as \( C = M_{ts}/2^N \). For convenience, we have rescaled \( t_s \) by the size of the state space \( 2^N \).

3 Results and discussion

The parameter \( \alpha \) defines the average connectivity \( \langle k \rangle \), which is solely determined by the set of agents inside the interaction circle and the sistem size \( M \). As expected when \( \alpha < 1 \) the interaction distance is less than the agent size \( d_0 \) and so \( \langle k \rangle < 1 \). Thus the agents explore the landscape practically independent of each other and \( \langle C \rangle \) is in the same as the independent search.

The left panel of the Figure 1 shows the results of \( \langle C \rangle \) vs. \( M \) with no mobility \( (\delta = 0) \), \( K = 4 \) and different values of \( \alpha \). The main detrimental effect of the existence of local maxima (due to \( K > 0 \)) is to uncouple the fitness of an agent from its distance to the global maximum. As a result, agents at local maxima spread unreliable information to their followers that may trap the entire system in a suboptimal solution. The catastrophic performance observed in the case of densely connected networks \( (\alpha \geq 5) \) and large system sizes \( (M \geq 20) \) is akin to the Groupthink phenomenon, when everyone in a group starts thinking alike, which can occur when people put unlimited faith in a leader (the model agent). A way of circumventing Groupthink is to limit or delay the flow of information among the agents and this can be achieved by reducing the average connectivity. There is however a tradeoff between avoiding the local maximum traps and optimizing the search performance. For instance, the choice \( \alpha = 1 \) avoids those traps altogether and always yields a superior performance compared with the independent search \( C < 1.0 \) in the figure), but it misses the optimal performance (minimum at \( C \approx 0.6 \) and \( M \approx 12 \) in the figure) that can be achieved for larger values of \( \alpha \).

The middle panel of the Figure 1 shows the results with \( \alpha = 2 \) and different values of \( \delta \). The mobility is very effective in avoiding the traps of the local maxima \( (M \sim 50) \) without the incurred tradeoff observed in the \( \delta = 0 \) scenario at \( M \approx 12 \). This is so because the mobility does not change the average connectivity of the agents. Measurement of the mean degree of the agents yields the same results for the \( \delta = 0 \) scenario. Hence, the random motion of the agents does not alter the nature of the network. This observation makes the results with \( \delta > 0 \) even more remarkable since it reveals that the change in \( \langle C \rangle \) is a genuine effect of the mobility of the agents and not a consequence of changing the network connectivity.
The effect of mobility is more noticeable in the right panel of Fig. 1 where we fix the system size to \( M = 53 \), which corresponds to the maximum of the computational cost in the middle panel and vary the step size \( \delta \) over several orders of magnitude. Since for this system size the trapping effects of the local maxima are maximized, moving the model agents far away from their clones is an efficient way to mitigate the influence of those maxima, as seen in the case \( \alpha = 2 \). When the influence of the local maxima is already reduced due to the small influence neighborhoods of the agents, as in the case of \( \alpha = 1.5 \), the mobility can actually help their dissemination over the square box, resulting in the increase of the computational cost. In any event, a large step size \( \delta \) guarantees that the imitative search always outperforms the independent search.

The research of JFF was supported in part by Grant No. 2017/23288-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and by Grant No. 305058/2017-7, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). SMR was supported by grant 15/17277-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). FAR acknowledges the Leverhulme Trust, CNPq (Grant No. 305940/2010-4) and FAPESP (Grants No. 2016/25682-5 and grants 2013/07375-0) for the financial support given to this research.

References

Quantifying the Role of Inactive Links in Social Networks

Andres M. Belaza\textsuperscript{1,2}, Jan Ryckebusch\textsuperscript{1} Aaron Bramson\textsuperscript{3,4}, Corneel Casert\textsuperscript{1}, Kevin Hoefman\textsuperscript{2}, Koen Schoors\textsuperscript{2}, Milan van den Heuvel\textsuperscript{1,2}, and Benjamin Vandermarliere\textsuperscript{1,2}

\textsuperscript{1} Department of Physics and Astronomy, Ghent University, Belgium, \textsuperscript{2} Department of General Economics, Ghent University, Belgium \textsuperscript{3} Laboratory for Symbolic Cognitive Development, RIKEN Brain Science Institute \textsuperscript{4} Department of Software and Information Systems, University of North Carolina Charlotte, USA

Contact email: andres.belaza@ugent.be

1 Introduction

Social balance [\textsuperscript{1,2}] is a key concept in the study of political networks where it is used to understand the emergence of intricate patterns like alliances, detentes or polarity. The core principles of social balance can be summarized in popular phrases like “a friend of a friend is also a friend” and “an enemy of my enemy is my friend”. Several approaches exist to model social balance from the network perspective. These models typically build on three-node cycles and on active (i.e. “+” and “−”) relations between political actors. This results in four types of three-node cycles. The standard formulation, however, only covers a small fraction of the information that is embedded in political networks as it overlooks the role of “inactive” (neutral or non-existent) edges. In this work we include inactive edges to study political networks. This adds an additional layer of dynamics as active edges can change to inactive ones and vice versa. After including inactive edges, ten types of three-node cycles can be discerned in the network.

2 Methods

Our proposed model starts from the perspective of statistical physics and assigns specific energies to each of the ten types of three-node cycles in the network. We also develop a generative model for the network. Thereby, the parameter $s_{ij}$ quantifies the relation between agents $i$ and $j$: $s_{ij} = +1$ for friendship, $s_{ij} = −1$ for enmity, and $s_{ij} = 0$ for an inactive relation. We propose a Hamiltonian that associates an energy to the entire system and quantities various aspects of the underlying dynamics

$$
\mathcal{H} (\{s_{ij}\}) = \frac{1}{6} \sum_{i\neq j\neq k=1}^{N} \left[ -\alpha s_{ij}s_{ik}s_{jk} - \gamma (s_{ij}s_{ik} + s_{ij}s_{jk} + s_{jk}s_{ik}) \right] \\
+ \frac{1}{2} \sum_{i\neq j=1}^{N} \left[ +\omega s_{ij} + \mu s_{ij}^2 \right], \tag{1}
$$

The 7\textsuperscript{th} International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
with \( N \) the total number of nodes in the network. The first term corresponds with standard social balance. Among other things, our approach allows to evaluate some important network properties like the expected value of the system’s magnetization \( \langle L \rangle = \langle s_{ij} \rangle \) and its overall activation \( \langle A \rangle = \langle s_{ij}^2 \rangle \). Using a mean-field approximation we derive a non-trivial relation between \( L, A \) and the parameters of the Hamiltonian:

\[
G_{MF}(L, A) = \frac{\mu}{\ln(1 - \frac{L}{A}) - \ln(1 + \frac{L}{A})}{2 \ln(2) + 2 \ln(\frac{1}{A} - 1) - \ln(1 - \frac{L^2}{A^2})} \approx \frac{\omega}{\mu} - \frac{2 \gamma}{3\mu} (N - 2)L.
\]

(2)

3 Results

We put our proposed framework to the test using time-series data for three empirical political networks. The first network covers international relationship during the Cold War (CW) era (1949-1993) and is extracted from the Correlates of War database. The other two networks stem from a virtual world EVE Online and capture the diplomatic relationships between “alliances” of players from March 2015 to April 2016. As they raise taxes and control territory for example, these player-created alliances play a role similar to that of a state. As a robustness check we construct two networks: one consisting of alliances with more than 200 players (EVE(+200)) and one covering sovereign alliances (EVE(SOV)).

With the sketched framework we observe a persistent hierarchy between the ten triadic energy levels across time and networks [2, ?]. The model also has predictive power for the transition probabilities between triadic states [?]. By calculating the magnetization \( L \) for three empirical networks, we learn that \( L \approx 0.01 \). The empirical confirmation of the non-trivial relationship between \( L, A \) and the parameters of the Hamiltonian can be seen in figure ??.

To this end we plot the l.h.s. of equation (??) versus \( L(N - 2) \) for the three empirical political networks as they evolve over time. The values of \( G_{MF}(L, A) \) for the EVE(+200) network tend to cluster along one particular curve. For the EVE(SOV) network two distinct lines can be discerned. The line to the right consists of samples from the first 100 days of the time period, a period of rapid network growth, while the other line covers the remainder of the time period. We see a similar pattern for the Cold War data where the 1950s are markedly different than the following decades. Apparently, our model enables us to identify a change in inter-country relationships induced by e.g. decolonization and the emergence of the Soviet block.

4 Conclusion

Traditional social balance builds on active (i.e. “+” and “−”) relations and three-node cycles in order to capture the major features of the formation and evolution of alliances in political networks. Three-node cycles with active edges typically represent a few percent of all possible triads in the networks. This raises the question whether more information can be included. In this work we have explored and quantified the addition of inactive edges. We use an approach based on a Hamiltonian allowing for additional dynamics as active edges can change to inactive ones and vice versa. The resulting
Fig. 1. The relation between $L$, $A$ and the parameters of the Hamiltonian using a mean-field approximation. Results are shown for samples of the EVE(SOV), EVE(+200) and Cold War networks. The inset shows the relationship between $L$ and $A$. The change in the slope in the Cold War data corresponds to the 1950s.

A mean-field approximation of our model allows us to define and calculate two highly relevant macro properties of the network: namely the system’s magnetization $L$ and average activation $A$. These two global variables allow one to sketch the phase boundaries between ordered and disordered realizations of the political networks. We have put our proposed approach to the test using three empirical networks. The networks are found to be in a partially ordered state and this state evolves over time clearly indicating major political events.

References

Network-Based Mapping of Narrative Structures [1]

Semi Min¹ and Juyong Park¹,²

¹ Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
² Sainsbury Laboratory, University of Cambridge, Cambridge CB4 1YE, United Kingdom
juyongp@kaist.ac.kr

1 Introduction

Human communication is executed in the form of a narrative, an account of connected events composed of characters, actions, and settings. A well-structured narrative, therefore, can be essential for an effective communication; a haphazard telling (or an attempt at) of a story hampering understanding is a common experience. A coherent narrative structure is therefore a requisite for a well-formulated narrative – be it fictional or non-fictional – for informative and effective communication. But the question of what constitutes a well-formulated narrative can open up the possibility of a deeper understanding of a narrative by studying its structural properties. Here we present a network-based framework for modeling and analyzing the structure of a narrative, which is further expanded by incorporating methods from computational linguistics to utilize the narrative text. Modeling a narrative as a dynamically unfolding system, we characterize its progression via the growth patterns of the character network, and use sentiment analysis and topic modeling to represent the actual content of the narrative in the form of interaction maps between characters with associated sentiment values and keywords. This is a network framework advanced beyond the simple occurrence-based one most often used until now, allowing one to utilize the unique characteristics of a given narrative to a high degree. Given the ubiquity and importance of narratives, such advanced network-based representation and analysis framework may lead to a more systematic modeling and understanding of narratives for social interactions, expression of human sentiments, and communication.

2 Methods and Results

We analyze Victor Hugo’s novel Les Misérables using the methods depicted in Fig. 1, sentiment analysis and topic modeling, in conjunction with the co-occurrence social network of characters. Set around the popular uprising in Paris in 1832 CE, Les Misérables is known for its vivid depiction of the conditions of the tumultuous times and intuition into the human psyche via multiple intersecting plots involving richly developed characters [2]. Considered a masterpiece of storytelling and a favorite material for network and textual analysis, its main plot follows fugitive Jean Valjean’s trajectory that shows him transform into a force for good while being constantly haunted by his criminal past. During his journey he interacts with many characters, some mutually helpful and friendly, and others antagonistic and hostile.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
The sentiment analysis can be performed on a given input text by many available packages such as the LIWC to identify the emotional state of the text, and by extension those of the characters appearing in the text. We can also use the measured sentiment polarity (positive or negative) to find the sentimental qualities of the relationships between the characters, shown as colored edges in the character network (Fig. 1).

The topic modeling can be performed by using a matrix decomposition method such as the NNMF (Non-Negative Matrix Factorization) technique. It finds “bags” of words closely related to each other by virtue of appearing in common text units, hence the name topic. A topic could contain (be strongly associated with) a character name, which allows us to interpret the other words in the same topic as descriptors of the characters. For instance, we show in Fig. 2 the two topics (1 and 2) that are respectively the most strongly associated with two lead characters, Valjean and Marius.

Figure 1. (Top) Two textual analysis tools employed in our work. The first, sentiment analysis, computes the positivity and the negativity of an input text. The second, topic modeling, identifies distinct topics in a given text by grouping commonly co-occurring words. (Bottom) The network of characters in Les Misérables. The edges are colored according to the sentiment of the text where the characters appear together. The node groups are found to correspond to subplots associated with the topics identified in Ref. [1].

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
3 Discussion

An important characteristic of well-written drama is that it evokes emotion in the reader, which in the western literary tradition is conventionally represented by the generic division of drama into comedy and tragedy. This had an interesting connection to a modern computational methodology called sentiment analysis. We found that many characters, especially the central ones, showed significant fluctuations of sentiments during the narrative flow, acting as the carriers of mood and emotions of the narrative. This was true of character relationships as well, and we showed how the sentimental fluctuations correlated with the narrative progression that showed detectable patterns of dramatic tension build-up and resolution. Finally, we used topic modeling as a way to define the state of a character via the topics (keywords) with which they are associated at various points in the narrative. This allowed us to trace quantitatively the changes in characters’ states, and quantify and map out the details of an event or an interaction between characters. We also demonstrated that the flow of topics between characters can reflect the actual story in interesting ways, providing us with a way to systematically represent the patterns of character interactions that previously resided in the text of the narrative.

We believe that our paper presents a wide range of ideas for studying narrative structures that merit further exploration using the methods of network science, data analysis, and computational linguistics. Given the ubiquity and importance of narratives, we hope that future developments based on our work will be beneficial for a wide range of fields including literature, communication, and storytelling.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF-20100004910, NRF-2013S1A3A2055285 and NRF-2016S1A2A2911945) and the BK21 Plus Postgraduate Organisation for Content Science.

References

The factors that affect human health occur at several scales, from gene and protein interactions, to the cellular, organism, and social levels. Understanding and controlling human health is especially complex due to the inter-level interactions that cannot be integrated away and thus form true control hierarchies [1]. Yet, the fact that multiple levels participate in the co-regulation of human health, enables us to measure and study it from other levels beyond the molecular. Indeed, Social media, electronic health records (EHR) and mobile application data enable population-level observation tools with the potential to speed translational research. For instance, disease-specific (sub)populations can be identified via machine learning and complex network analysis of EHR, discovering personalized disease trajectories [2]. Social media data can then be used to identify disease phase transitions by complementing clinical information from EHR with a temporally richer characterization of symptoms and behavioral attitudes, providing preliminary evidence for hypothesis-driven clinical and molecular studies. We will present our recent research exemplifying a complex systems approach to the study of multi-level human health. This includes big data studies demonstrating that: 1) social media data (Twitter & Instagram) cohort studies (depression, epilepsy & opioids) enable identification of previously unknown adverse drug reactions (ADR) and drug-drug interactions (DDI) [3, 4]; 2) web searches and lexical sentiment analysis of social media allow us to uncover collective social behavior in phenomena of interest to public health, such as human reproductive behavior at a planetary level [5]; 3) the longitudinal analysis of EHR reveals significant gender and age biases in the prevalence of known---and thus preventable---DDI [6]; 4) a study of Facebook timelines to identify behavioral markers in deceased patients from Sudden Death in Epilepsy (SUDEP). Together, this work shows how complex systems techniques can be used to test novel hypotheses of public-health relevance, and even invalidate long-standing ones---such as our recent results showing that the reigning biological hypothesis for observed human reproduction cycles is incompatible with newly available planetary data of online behavior, and that a cultural explanation of the phenomenon is much more likely [5].
Figure 1: DDI network from city-wide drug administration EHR data in Blumenau (pop. 338,000), southern Brazil [6]. Nodes denote drugs involved in at least one co-administration known to be a drug-drug interaction (DDI). Nodes: Node color show drug class highest hierarchy level of primary action, as retrieved from Drugs.com. Node size represent the probability of a drug of being involved in a DDI given all its co-administrations with other drugs, \( P(i) \). Edges: weights are \( \tau_{ij} \), a normalized measure of the length of co-administration between two drugs. Higher values (thick edges) denote that drug \( j \) is usually co-administered when drug \( i \) is administered, normalized for all patients taking either drug. Edge colors denote the drug pair increased risk of interaction per gender where red (blue) edges denote interactions with increased risk for women (men).

References
Characterization of Ego-network Circles in Mobile Phone Graphs

Christian Quadri, Matteo Zignani, Sabrina Gaito, and Gian Paolo Rossi

Università degli Studi di Milano, Computer Science Dept., Milano, Italy,
firstname.lastname@unimi.it

A well-known result in social science states that, although people have a large number of social relationships, in practice, their interactions are concentrated mostly on a small portion of their neighbors. According to Dunbar’s hypothesis, this phenomenon is due to a limited human cognitive capacity to handle too many social relationships, together with a few time for socializing. These constraints result in an organization of people’s ego-network into four or five groups depending on the strength of the interactions, the so called Dunbar’s circles. The verification of the Dunbar’s hypothesis, the identification of social circles and their characterization are still open questions, although researchers have found evidence of its validity on different social networks, from small offline networks to online social networks, such as Twitter and Facebook. In addition, little is known about the semantic aspects of the circles, i.e. who are the members of each circle. In this paper we cope with the above issues and we identify and characterize the Dunbar’s circles by leveraging a mobile phone dataset. The main findings of the analysis and characterization of the Dunbar’s circles in our mobile phone graph are summarized as follows:

- We validate the Dunbar’s hypothesis on the organization of ego-networks in circles by finding the subdivision of ego-networks into four or five circles, each characterized by a specific tie-strength and a typical frequency of contact.
- By exploiting the multiplexity of the interactions, we show that people maintain the relationships with the members of the closest circles by combining voice calls and texts; whereas, sporadic interactions are more likely to occur through a single channel only.
- By inferring the home and work locations from the mobility traces of each individual, we find that two of the most common social structures, family and workmates, influence the formation of the circles.

Dataset and Methods We leverage a large dataset of Call Detail Records (CDRs) [4] which collects phone activities of about one million subscribers in the metropolitan area of Milano for a time period of 67 days. We model call and text interactions by a directed multigraph (a link can be labeled as “call” or “text”, according to the communication channel adopted to interact). From this multigraph we obtain the ego-networks by extracting, for each subscriber, his/her out-neighborhood. We restrict our analysis to people whose ego-network size is greater than the median value of the degree distribution (43), which is a good trade-off between the generalization of the results - our analysis relies on about 80,000 users - and the expected size of the circles provided by the Dunbar’s hypothesis. To assess the strength of relationships between ego and
alters we adopt an approach based on the link signature [5] which takes into account the intrinsic differences of the two communication channels. We divide the observation period into one-hour bins, and for both calls and texts, we find the bins that contain at least one interaction. The link signature counts the number of the occupied bins, regardless the type of the interaction; thus makes it possible to evaluate the mixed usage of texts and calls. We apply a clustering technique to extract the circles from each ego-network. Specifically, for each ego-network, we use the Mean Shift algorithm [2] to find the optimal number of clusters. To detect the home and the workplace of each individual we employ the methodology proposed in [3], which leverages the regularity of users in visiting a specific location. This methodology is able to capture the most common behavior of workers and students who work during the day, for 5 days a week.

**Results** In order to validate the Dunbar’s hypothesis on a CDR dataset, we apply the clustering methodology described above on the mobile phone ego-networks to detect the circles and we show that their main features, such as the number of circles, their size and the strength of the ties, are definitely in line with Dunbar’s hypothesis. In Fig. 1(a) we report the distributions of the number of circles identified in each ego-network. Remarkably, distributions are centered around 5 circles which is the optimal number of circles according to Dunbar’s hypothesis and in line with previous results on offline and online social networks [1]. As regards the size of the circles, we compute the number of alters that belong to each circle by considering at most 5 circles; if an ego-network has more that 5 circles, we merge all the alters that belong to the outermost circles into the fifth circle. In Fig. 1(b) we report the distribution of the circle size. The innermost circles contain very few alters, whereas most of the alters belong to the outermost circles, in line with the Dunbar’s hypothesis. Moreover, we found that alters in the innermost circles are characterized by the highest values of tie strength, while the ones in the outermost circles are mostly characterized by weak ties.

Each circle has different semantic; for example, in the first circle, namely the support clique, we have family members, whereas the last circle is formed by acquaintances. This semantic distinction is reflected in how regularly the ego contacts each alter across a period of time. To analyze this aspect, we compute the temporal regularity of the communication with each alter as the ratio of the number of days the ego contacts his/her and the number of days the ego was active. Here, we consider a subscriber as active in a day if he/she performs at least one phone activity in that day. In Fig. 1(c) we report the distribution of the temporal regularity. As we can observe, the alters who have the highest value of relevance belong to the first circle, whereas the outermost circles are characterized by the lowest values of regularity.

We verify if tie-strength is reflected by a different use of the multiple communication channels available. To this aim, for each ego-network and for each circle, we compute the percentage of alters with whom the ego uses both channels. In Fig. 1(d) we report the histogram of the mean of the percentage of the multichannel usage grouped by circle. It emerges that the multiplexity is a property of the innermost circles, namely the strongest connected alters.

The last aspect we analyze is the relationship between the circle an alter belongs to and the distance between her home/work location and the ego’s ones. We detect the home/work locations for each user and we compute the GPS distance between
Fig. 1. In (a) the distribution of the number of circles. In (b) the boxplot of the number of alters in each circle. In (c) the boxplot of the relevance of each alter grouped by circle. In (d) the histogram of the mean of the percentage of the multichannel alters grouped by circle. In (e) the box plot of the distribution of the GPS distance between the home/work location of the ego and the home/work location the alters.

home/work location for each pair ego-alter. In Fig. 1(e) we report the boxplot of the GPS distances for both home and work locations. We can observe that the stronger the tie, the closer the home/work locations. This suggests that individuals maintain strong ties with people who live close, e.g. family members, relatives and roommates, neighbors and so on.

References

Fairness in multiplayer ultimatum games through degree-based role assignment

Andreia Sofia Teixeira¹,²,³, Francisco C. Santos¹,², Alexandre P. Francisco¹, and Fernando P. Santos¹,²,⁴

¹ INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Portugal
² ATP-Group, Lisboa, Portugal
³ LASIGE and Faculdade de Ciências, Universidade de Lisboa, Portugal
⁴ Department of Ecology and Evolutionary Biology, Princeton University, Princeton, USA
{sofia.teixeira,fernando.pedro}@tecnico.ulisboa.pt

1 Introduction

Fairness has a profound impact on human decisions and individuals often prefer fair – over payoff maximising – outcomes [3]. This evidence was pointed several times, often resorting to behavioural experiments with the Ultimatum Game (UG): a Proposer decides how to divide a given resource with a Responder and the game only yields payoff to the participants if the Responder accepts the proposal. Strangely, Proposers tend to sacrifice their share by offering high proposals and Responders often prefer to earn nothing rather than accepting unfair divisions [3]. These counter-intuitive results motivated several theoretical models that aimed at justifying, mathematically, the evolution of fair intentions in human behaviour. In most cases, however, the roles in UG are assumed to be symmetric: each node has an even probability of being the Proposer or the Responder. Also often, both roles are played simultaneously. These assumptions are naturally at odds with reality, where being the Proposer or Responder depends on characteristics of individuals. Proposers – such as employers, auction first-movers or investors – are in the privileged position of deciding which divisions to offer. The benefits of Proposers are even increased in multiplayer ultimatum games [4]. In that case, Responders often need to divide the offers, thus increasing the gap between Proposers’ and Responders’ potential earnings. This leads us to two main questions: i) Which criteria should be used to select Proposers within a group? and ii) What is the impact of different criteria on the emerging levels of fairness? Here we analyse Multiplayer Ultimatum Games (MUG) [5] in heterogeneous complex networks, which allows us to test several node properties as base criteria for defining how to select Proposers in a group.

2 Model - Multiplayer Ultimatum Game on complex networks

In the N-player MUG, proposals are made by one Proposer to the remaining N – 1 Responders [5]. Each individual adopts a strategy (p, q). When playing as Proposer, an individual will offer p. When playing as Responder, individuals use their acceptance threshold (q) to accept (pN−1 ≥ q) or reject (pN−1 < q) the offer. Group acceptance will depend on M, the minimum fraction of Responders that must accept the offer. If the
fraction of individual acceptances stands below $M$, the offer will be rejected and no one earns anything. Otherwise, the offer will be accepted: the Proposer will keep $1 - p$ and the group will share the remainder, that is, each Responder gets $p/(N - 1)$ [4]. Degree heterogeneity will imply diversity in the number of collective dilemmas faced and group sizes in which those are played [6]. Such diversity is here introduced by considering $i)$ Barabási-Albert (BA) [1] networks, having a power-law degree distribution combined with low clustering; and $ii)$ Dorogotsev-Mendes-Samukhin (DMS) networks [2], exhibiting the same power-law degree distributions, yet with higher clustering coefficient – a topological feature of relevance in the context of fairness and N-person games [4].

**Network based role selection:** Previous works [9] show that assigning the role of Proposer to the highly connected nodes decreases the levels of fairness in dyadic interactions [7, 8]. Considering MUG opens now space to study the interplay between group characteristics (such as group sizes) and network-based criteria to select Proposers, in unexplored directions. So far we assumed that, in a group with $N$ individuals $(1...i...N)$, where each individual $i$ has degree $k_i$, the probability that $j$ is selected as Proposer is given by $p_j = \frac{\alpha k_i}{\sum_{k} e^{\beta k}}$, where $\alpha$ controls the dependence of degree on role selection.

**Evolutionary Dynamics:** We simulate the evolution of $p$ and $q$ in a population of size $Z$, much larger than the average group size $<N>$. Initially, each individual has values of $p$ and $q$ drawn from $\{0, 0.01...0.99, 1\}$. The fitness $F_i$ of an individual $i$ of degree $k$ is determined by the payoffs resulting from the games occurring in $k + 1$ groups: one centred on her neighbourhood plus $k$ centred on each of her $k$ neighbours. Values of $p$ and $q$ evolve as individuals tend to imitate (i.e., copy $p$ and $q$) those neighbours that obtain higher fitness. Simulations take place for $10^6$ generations, considering that, in each generation, all the individuals have (on average) the opportunity to revise their strategy through imitation. At every (discrete and asynchronous) time step, two individuals $A$ and $B$ (neighbours) are randomly selected and their individual fitness is computed; subsequently, $A$ copies the strategy of $B$ with a probability $\chi$ that increases with the fitness difference $f_B - f_A$, following the pairwise comparison update rule $\chi = \frac{1}{1 + e^{-\beta|f_B - f_A|}}$. The copied $p$ and $q$ values be added a small perturbation drawn from the interval $[-\epsilon, \epsilon]$. For each combination of parameters, the average values of $p$ and $q$ are obtained taken over the last 50% generations (i.e., after a transient) of all the runs.

3 Preliminary Results and Discussion

Fig. 1 shows that preferentially attributing the role of proposer to low-connected nodes $(\alpha = -2)$ significantly increases the fairness levels in the population. We are able to find, however, parameter spaces where the dependence of $<p>$ on $\alpha$ is seemingly affected by $i)$ the average connectivity of the network and thus on the average size of the groups in which MUG is played and $ii)$ the particular values of $M$. We observe that increasing $M$ or the clustering coefficient (from BA to DMS networks) increases the average proposal $<p>$, as found in previous works for different network topologies [5, 4]. Our results suggest that offering the first move to low-degree nodes balances the natural power of highly connected nodes in scale-free networks, leading to a significant increase in the global levels of fairness. Interestingly, we also find that particular voting
rules ($M$) are able to attenuate the negative effect of high $\alpha$ (i.e., privileged high-degree nodes being selected to be Proposers) on fairness. This opens space to test, computationally, ways of engineering new voting mechanisms that efficiently use collective decision making to prevent unfair behaviours from spreading on complex networks.


**References**

DyNAM-i: a statistical model for the analysis of face-to-face interactions

Marion Hoffman¹, Per Block¹ Timon Elmer¹, and Christoph Stadtfeld¹

Chair of Social Networks, ETH Zürich, Weinbergstrasse 109, 8006 Zürich, Switzerland
marion.hoffman@gess.ethz.ch,
WWW home page: www.sn.ethz.ch

1 Introduction

Face-to-face interactions are at the very core of social systems and understanding them is critical in understanding how individuals form social ties, exchange information, transmit diseases or become culturally segregated [1]. However, interaction patterns are often complex as they can be the result of a wide variety of individual, social, and contextual factors. Recent progress in information technologies and human sensors have opened new opportunities to collect those directly [2] and new statistical models allow us to analyze them and understand which factors explain how people choose their interaction partners. Current models for relational event data [3, 4] can be applied to study time-stamped social interaction data but cannot deal with specific features of face-to-face interactions. More specifically, those have a start and end time, and often happen in groups rather than with one partner at a time. An individual’s decision to participate in a group interaction may depend on her attributes, the attributes of members within a group, existing social network ties, and the individuals’ positions in space. Here, we propose a new model that is tailored to face-to-face interaction data and demonstrate its fruitfulness in a straightforward case study.

2 Model

We propose DyNAM-i, an extension of Dynamic Network Actor Models (DyNAM; [4]) for face-to-face interaction data. This model is a generic framework that can be specified flexibly to study various empirical settings (e.g., face-to-face interactions in universities or during conferences). For example, it can be used to investigate whether the presence of different network layers, such as friendship or spatial proximity, affect the formation of group interactions. The model is actor-oriented [5] and thus expresses changes in an interaction network by a) individual tendencies to join or leave group interactions (modeled as a Poisson rate) and b) the decision which group to join (expressed as a multinomial logit model).

An interaction event is represented as a quadruplet \( \omega = (t_\omega, i_\omega, g_\omega, \rho_\omega) \), where \( t_\omega \) is the time of the event, \( i_\omega \) the actor involved in this event (\( A \) being the set of all actors), \( g_\omega \) the group the actor is joining or leaving (\( G(t_\omega) \) being the set of interaction groups at time \( t_\omega \)), and \( \rho_\omega \) a action type indicating whether it is a joining or a leaving event [3].

An interaction network \( x(t) \) at time \( y \) is defined as a bipartite graph where \( x_{ig}(t) = 1 \).
if individual $i$ is part of group interaction $g$ at time $t$. The rates of leaving and joining an interaction as shown in equation 1 depend on network statistics defined for each dyad actor-group at each point in time $(s_{k,i}(x'))_k$ and some associated weight parameters $(\alpha_{k,i})_k$. We associate a superscript $j$ or $l$ whether we consider joining or leaving.

$$
\tau_i^{joining} = \exp(\sum_k \alpha_{k,i}^j s_k(x'))
$$

$$
\tau_i^{leaving} = \exp(\sum_k \alpha_{k,i}^l s_k(x'))
$$

(1)

The probability of an actor becoming active is given by a competing rate ratio as shown in (3):

$$
p(i \text{ gets active}) = \frac{\tau_i}{\sum_{j \in A} \tau_j}
$$

(2)

When an actor who is already interacting in a group becomes active, she leaves this group. When an isolated actor becomes active, her choice to interact with a certain group is represented by (4), with similar notations for statistics and parameters:

$$
p(i \text{ chooses } g_m) = \frac{\exp(\sum_k \alpha_{k,i}^c s_k(x'))}{\sum_{n \in G(t)} \exp(\sum_k \alpha_{k,i}^c s_k(x'))}
$$

(3)

3 Application

We applied this model to interaction data recorded during a social event within a university research team. The dataset includes 11 actors, and 219 events, and includes variables for age, gender, group affiliation and seniority. The study was set up to show the convergence of our model and its validity rather than for testing specific hypotheses. For leaving, we merely tested the influence of an actor’s age on her propensity to leave: the left part of Table 1 shows that older individuals were significantly more likely to leave an interaction. For joining, we also tested the effect of age, along with the absolute value of the difference between the actor’s attributes (i.e. age, gender, and seniority) and the mean value for his/her group. We see here that individuals were less likely to leave groups with a different average age than theirs and an opposite effect for seniority. The group’s size and the presence of a person they knew before had no significant effect. For choosing groups, we took into account similar statistics. The left part of the table shows that the only significant effect we find here is that individuals were more likely to choose a group with a high age difference.

Both rate and choice models converged in less than 10 iterations. Although their specifications were quite simple, both models resulted in a large improvement of the average likelihood of each event compared to a purely random model (for the choice model, we went from a value of 0.1720 to 0.2231, which means a 30% increase, and for the rate model we went from 0.0031 to 0.0113 which represents a 264% increase).
Table 1. Estimates for the parameters in the rate and choice models

<table>
<thead>
<tr>
<th>Effect</th>
<th>Joining rate model</th>
<th>Leaving rate model</th>
<th>Choice model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>standard error</td>
<td>estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>-6.01</td>
<td>0.77 ***</td>
<td>-5.78</td>
</tr>
<tr>
<td>Age covariate</td>
<td>0.08</td>
<td>0.03 **</td>
<td>-0.004</td>
</tr>
<tr>
<td>Difference with group mean age</td>
<td>-</td>
<td>-</td>
<td>-2.18</td>
</tr>
<tr>
<td>Difference with group mean gender</td>
<td>-</td>
<td>-</td>
<td>-0.03</td>
</tr>
<tr>
<td>Presence of an affiliated member</td>
<td>-</td>
<td>-</td>
<td>0.09</td>
</tr>
<tr>
<td>Difference with group mean seniority</td>
<td>-</td>
<td>-</td>
<td>1.98</td>
</tr>
<tr>
<td>Group size</td>
<td>-</td>
<td>-</td>
<td>-0.01</td>
</tr>
<tr>
<td>Presence of a person known before</td>
<td>-</td>
<td>-</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

Rate models: convergence score: 3e-05, AIC: 2565.8, BIC: 2596.4
Choice model: convergence score: 1e-05, AIC: 366.4, BIC: 382.6

4 Discussion and conclusions

The model that we propose can be used to gain deeper insights into the micro mechanisms underlying the complex patterns of face-to-face interactions. The analytical goal of the model is to test specific hypotheses on individuals’ motivations to interact and could be applied in the future to understand the role of individual factors (e.g., personality) or contextual factors (e.g., spatial layouts). The model is straightforward to parametrize and its estimation is very fast and scalable to large datasets. This is because the likelihood terms (that are derived from equations 1–3) are conditionally independent and the implementation can therefore be parallelized. An implementation will be released in the R package goldfish.

Two limitations remain for this model. First, all decisions are made by individual actors, hence forbidding any coordination between them. Second, the choices of leaving and joining interactions are assumed to be independent. Future research can aim at testing the effect of those assumptions and aim at relaxing them. Face-to-face interactions are a meaningful phenomenon underlying the dynamics of various social systems. Recent data collection techniques and models such as DyNAM-i can now be applied to empirical cases to test various micro-level theories from the social sciences.

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Opinion spreading on complex networks with local and global interactions

Arkadiusz Jędrzejewski, Angelika Abramiuk, Bartłomiej Nowak, and Katarzyna Sznajd-Weron
Wrocław University of Science and Technology, Wrocław, Poland
arkadiusz.jedrzejewski@pwr.edu.pl

1 Introduction

The $q$-voter model is one of the flagship agent-based models of sociophysics used to illustrate opinion dynamics. In recent years, it has attracted especially considerable attention while modelling opinion formation on static as well as on evolving in time networks. It was originally proposed by Castellano et al. [1] and later altered by Nyiczka et al. [2]. One of its modified versions may represent opinion dynamics under two types of social response: conformity and anticonformity [3]. Although widely exploited in sociophysics, it also finds application in computational economics as an underlying mechanism of consumers decision making process, sometimes with slight modifications.

In the work, we analyze a class of opinion formation processes with the spreading mechanism based on the $q$-voter model where the social interactions may occur locally or globally. The competition between interactions arising from conformity and nonconformity gives rise to order-disorder phase transitions, which are investigated by means of Monte Carlo simulations and the pair approximation, an enhancement of the standard mean-field approximation, especially useful in problems involving complex networks.

We analyze how the range of interactions impacts the model behavior and its stationary properties like the concentration of agents with a specific opinion. The applied mathematical approach allowed us not only to anticipate the shape of phase diagrams but also to derive formulas for transition points between a phase in which one opinion is dominating, a positive or a negative one, and a phase in which both, positive and negative, opinions are equally popular. The obtained approximation is validated by carrying out Monte Carlo simulations on several structures like random regular, Erdős-Rényi, Watts-Strogatz, and scale-free networks, including the Barabási-Albert model.

2 Model description

We consider an arbitrary network of the size $N$. Each vertex is associated with one autonomous agent characterized by a binary, spin like variable $s_i = \pm 1$. In every elementary time step, we pick at random an agent and a group of influence comprised of $q$ randomly selected agents. The group is called $q$-panel, and depending on the considered range of social interaction its potential members are determined by the local network topology or are selected from the whole system. As a result, we compare four
q-voter models with different combinations of local and global social responses. The social influence occurs only if the group is unanimous, that is to say, all q individuals have the same opinion. Then with probability $p$, the chosen agent acts as an anticonformist and adopts the opposite opinion to the panel. Otherwise, with probability $1 - p$, it behaves like a conformist and embraces the viewpoint of the q-panel. Drawing agents occurs without repetition.

3 Results

The pair approximation predictions exhibit substantial agreement with simulations for some structures. The most interesting is the importance of the interaction range of conformity and anticonformity. It seems that local anticonformity disorders the system less than the global one, and the influence of the network topology on the system is weaker in that case. On the other hand, global conformity more efficiently orders the system and lowers the model sensitivity to the underlying network structure in comparison to its local version. Moreover, the quantitative behavior of the model studied on weakly clustered complex structures depends mainly on the average node degree of an underlying network $\langle k \rangle$. It means that networks, which may have very different arrangements of edges or node degree distributions, lead to the same average opinion level if they have the same value of the average node degree. Moreover, for large average node degrees, we showed that solutions established from the pair approximation converge to those from the mean-field theory.

Acknowledgments

The work was partially supported by funds from the National Science Centre (NCN, Poland) through Grants No. 2016/23/N/ST2/00729 and No. 2016/21/B/HS6/01256.

References

Learning Quadratic Games on Networks

Yan Leng, Xiaowen Dong, and Alex ‘Sandy’ Pentland

1 MIT Media Lab
{leng, pentland}@mit.edu
2 University of Oxford
xdong@robots.ox.ac.uk

1 Introduction

In a network game, the underlying structure of the network carries critical information and dictates the behavior and actions of the players. Typically, graphs are used as mathematical tools to represent the structure of these networks, and the current literature in this area has exclusively focused on studying the characteristics of games on known or predefined graphs [1, 2]. The few works that propose to learn network games have limitations in that they either assume a discrete strategy space [3, 4], or assume a payoff function that favors a specific type of interaction [5].

In this study, we focus on learning quadratic games, i.e., games with linear-quadratic payoffs [1, 2, 6, 7] which allow for continuous actions as well as interactions of both strategic complements and substitutes. Specifically, we propose a learning framework where, given the Nash equilibrium actions of the games, we jointly infer the graph that represents the structure of the network as well as the individual marginal benefits. We further develop the second framework by considering the homophilous effect of individual marginal benefits in the interaction network. We test the performance of the proposed algorithms in inferring graph structures for network games and show that it is superior to that of the classical solutions of graphical Lasso in [8, 9], albeit developed for slightly different learning settings.

2 Methodology

Consider a network of $N$ individuals represented by a graph $G(V, E)$, where $V$ and $E$ denote the vertex and edge sets, respectively. For any pair of individuals $i$ and $j$, $G_{ij} = G_{ji} = 1$ if $(i, j) \in E$ and $G_{ij} = G_{ji} = 0$ otherwise. Following the literature, in a network game of linear-quadratic payoffs, an individual $i$ chooses her action $a_i$ to maximize her utility, $u_i$, which has the following form [1, 2, 6, 7]:

$$u_i = b_i a_i - \frac{1}{2} a_i^2 + \beta a_i \sum_{j \in V} G_{ij} a_j,$$  \hspace{1cm} (1)

where $b_i$ is called the marginal benefit, and $\beta$ captures the nature and the strength of the peer effect: if $\beta > 0$, actions are called strategic complements; and if $\beta < 0$, actions are called strategic substitutes. With first-order condition of Eq. (1), the equilibrium strategy follows: $a = (I - \beta G)^{-1} b$, where $I$ is the identity matrix. We adopt the critical assumption that the spectral radius of the matrix $\beta G$, denoted by $\rho(\beta G)$, is less than 1.
We consider $N$ players, connected by a fixed strategic interaction network $G$, playing $K$ different games in each of which their payoffs depend not only on their own actions but also that of their neighbors. We define the marginal benefits for these $K$ games as $B = [b^{(1)}, b^{(2)}, \ldots, b^{(K)}] \in \mathbb{R}^{N \times K}$, where each column of $B$ is the marginal benefit vector for one game, and the corresponding actions of the players as $A = [a^{(1)}, a^{(2)}, \ldots, a^{(K)}] \in \mathbb{R}^{N \times K}$. We first consider the case of independent marginal benefit where, for each game, the marginal benefits of individual players follow independent Gaussian distributions. Given the observed actions $A$ and the parameter $\beta$, the goal is to infer a graph structure $G$ as well as the marginal benefits $B$:

$$
\begin{align*}
\text{minimize} & \quad f(G, B) = \| (I - \beta G)A - B \|_F^2 + \theta_1 \|G\|_1^2 + \theta_2 \|B\|_F^2, \\
\text{subject to} & \quad G_{ij} = G_{ji}, \ G_{ii} \geq 0, \ G_{ii} = 0 \quad \text{for } \forall i, j \in V \text{ and } \|G\|_1 = N,
\end{align*}
$$

(2)

where $\text{tr}(\cdot)$, $\| \cdot \|_F$, and $\| \cdot \|_1$ denote operators for the trace, Frobenius norm, and entry-wise $L^1$-norm of a matrix, respectively, and $\theta_1$ and $\theta_2$ are two regularization parameters. The problem of Eq. (2) can be solved via convex optimization (Algorithm 1 in Fig. 1).

We further consider a second case with homophilous marginal benefits, where the effect of homophily may be quantified by the so-called Laplacian quadratic form on the graph: $b^T L_b = \frac{1}{2} \sum_{i,j \in E} G_{ij} (b_i - b_j)^2$, where $L = \text{diag}(\sum_{j \in V} G_{ij}) - G$ is the combinatorial graph Laplacian matrix. This leads to the following optimization problem which can be solved via alternating minimization (Algorithm 2 in Fig. 1): 

$$
\begin{align*}
\text{minimize} & \quad h(G, B) = \| (I - \beta G)A - B \|_F^2 + \theta_1 \|G\|_1^2 + \theta_2 \text{tr}(B^T L_b), \\
\text{subject to} & \quad G_{ij} = G_{ji}, \ G_{ii} \geq 0, \ G_{ii} = 0 \quad \text{for } \forall i, j \in V, \\
& \quad \|G\|_1 = N \text{ and } L = \text{diag}(\sum_{j \in V} G_{ij}) - G.
\end{align*}
$$

(3)

3 Experiments

We evaluate the performance of the proposed learning frameworks on synthetic networks that follow three types of random graph models, i.e., the Erdős-Rényi (ER), the Watts-Strogatz (WS), and the Barabási-Albert (BA) models. We apply our algorithm to infer graph structures and compare against the ground truth ones in the scenario of binary classification, i.e., either there exists an edge between $i$ and $j$ (positive case), or not (negative case). Since the ratio of positive cases is small for all the three types of graphs, we use the area under the curve (AUC) for the evaluation of the learning performance. We compare our algorithms with two baseline methods for inferring graph structures given data observations: the sample correlation and the regularized graphical Lasso in [9]. The performance of the three methods in comparison is shown in Fig. 1. It can be seen that for all the three random graph models, our method outperforms the two baselines for different strength of strategic dependencies. More detailed analysis and results on real world data are presented in the longer version of the paper.

4 Discussion

In this paper, we propose two novel learning frameworks for a joint inference of graph structure and individual marginal benefits for network games with linear-quadratic pay-
offs. We show that our algorithms achieve superior performance compared to the baseline techniques. The proposed approach can benefit a wide range of practical scenarios, including community detection using the inferred strategic relationship, targeting central player in marketing scenarios, and designing intervention mechanisms to achieve specific planning objectives.

References

Part IX

Multilayer Networks
Multiplex decomposition of non-Markovian dynamics and the hidden layer reconstruction problem

Lucas Lacasa\textsuperscript{1}, Inés P. Mariño\textsuperscript{2,3}, Joaquin Míguez\textsuperscript{4}, Vincenzo Nicosia\textsuperscript{1}, Édgar Roldán\textsuperscript{5,6}, Ana Lisica\textsuperscript{7}, Stephan W. Grill\textsuperscript{8,9} and Jesús Gómez-Gardeñes\textsuperscript{10,11}

\textsuperscript{1}School of Mathematical Sciences, Queen Mary University of London, Mile End Rd, E14NS London (UK)
\textsuperscript{2}Department of Biology, Geology, Physics & Inorganic Chemistry, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid (Spain)
\textsuperscript{3}Institute for Women’s Health, University College London, London WC1E 6BT, London (UK)
\textsuperscript{4}Department of Signal Theory & Communications, Universidad Carlos III de Madrid, 28911 Leganés, Madrid (Spain)
\textsuperscript{5}The Abdus Salam International Center for Theoretical Physics, Strada Costiera 11, 34151 Trieste (Italy)
\textsuperscript{6}Max Planck Institute for the Physics of Complex Systems, Nöthnitzerstraße 38, 01187 Dresden (Germany)
\textsuperscript{7}London Center for Nanotechnology, University College London, London (UK)
\textsuperscript{8}Biotechnology Center, Technical University Dresden, Tatzberg 47/49, 01309 Dresden (Germany)
\textsuperscript{9}Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauer Str. 108, 01307 Dresden (Germany)
\textsuperscript{10}Department of Condensed Matter Physics, University of Zaragoza, 50009 Zaragoza (Spain)
\textsuperscript{11}Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, E-50018 Zaragoza (Spain)

Elements composing complex systems usually interact in several different ways and as such the interaction architecture is well modelled by a network with multiple layers—a multiplex network—, where the system’s complex dynamics is often the result of several intertwined processes taking place at different levels. However only in a few cases can such multi-layered architecture be empirically observed, as one usually only has experimental access to such structure from an aggregated projection.

A fundamental challenge is thus to determine whether the hidden underlying architecture of complex systems is better modelled as a single interaction layer or results from the aggregation and interplay of multiple layers. Assuming a prior of intralayer Markovian diffusion, here [1] we show that using local information provided by a random walker navigating the aggregated network, it is possible to determine in a robust manner whether these dynamics can be more accurately represented by a single layer or they are better explained by a (hidden) multiplex structure. In the latter case, we also provide Bayesian methods to estimate the most probable number of hidden layers and the model parameters, thereby fully reconstructing its architecture. The whole methodology enables to decipher the underlying multiplex architecture of complex systems by exploiting the non-Markovian signatures on the statistics of a single random walk on the aggregated network.
In fact, the mathematical formalism presented here extends above and beyond detection of physical layers in networked complex systems, as it provides a principled solution for the optimal decomposition and projection of complex, non-Markovian dynamics into a Markov switching combination of diffusive modes. We validate the proposed methodology with numerical simulations of both (i) random walks navigating hidden multiplex networks (thereby reconstructing the true hidden architecture) and (ii) Markovian and non-Markovian continuous stochastic processes (thereby reconstructing an effective multiplex decomposition where each layer accounts for a different diffusive mode). We also state and prove two existence theorems guaranteeing that an exact reconstruction of the dynamics in terms of these hidden jump-Markov models is always possible for arbitrary finite-order Markovian and fully non-Markovian processes. Finally, we showcase the applicability of the method to experimental recordings from (i) the mobility dynamics of human players in an online multiplayer game and (ii) the dynamics of RNA polymerases at the single-molecule level.

email: l.lacasa@qmul.ac.uk

References

Majority-vote dynamics on multiplex networks

Jeehye Choi and Kwang-Il Goh
Department of Physics, Korea University, Seoul 02841, Korea
kgoh@korea.ac.kr

1 Introduction

Majority-vote model is a much-studied model for social opinion dynamics of two competing opinions [1]. With the recent appreciation that our social network comprises a variety of different “layers” forming a multiplex network [2–4], a natural question arises on how such multiplex interactions affect the dynamics of social opinion and consensus formation. In this Extended Abstract, the majority-vote model is studied, as an archetypical model of nonlinear collective social opinion dynamics, on multiplex networks to understand the effect of multiplexity on opinion dynamics.

2 Results

In the majority-vote model, a node choose its state according to the majority-rule by looking up its neighbors, with additional randomness encoded by the noise parameter. We generalize the original majority-vote model into multiplex networks [5]. In doing so we introduce a number of different layer-wise decision rules how to incorporate the neighbors’ states in different layers. We focus on how global consensus is reached by three different types of voters—the normal, AND- and OR-rule voters—and what different microscopic origins are working in these different types. We found that the AND-model reaches the greatest consensus below the critical noise level $Q_c$. However, it needs much longer time to reach consensus than other models and the consensus collapses abruptly in the vicinity of the transition point. The OR-model has smaller level of consensus than the AND-rule but it reaches the consensus more quickly thanks to weak endurance. The OR-model exhibits more active dynamics with more opinion flips as well as more disagreements than the AND-model, which render its consensus transition continuous at the critical point. The numerical simulation results are supported by analytical calculations based on approximate master equation.

References

Fig. 1. (Top panels) Magnetization $M$ vs. the noise parameter $q$ for the AND- (left) and OR multiplex majority-vote model (right) on multiplex ER networks. Lines are numerical solutions from the approximate master equations and symbols are simulation results on ER networks with mean degrees $3 \sim 8$ as indicated. (Bottom panels) Histograms of steady-state magnetization $P(|m_t|)$ of AND-model on ER networks of $\langle k \rangle = 6$ for different $q$ values and a characteristic time course of $m_t$ at $q = q_c$ corresponding to the bimodal histogram at $q_c$. 
Introduction

Emergency logistic system play an important role in the natural disaster rescue and emergency public events, etc. The damage results of earthquakes, floods, landslides, debris flow and other natural disaster reflect the vulnerability and instability characteristics of logistic network. Thus, it is very important and meaningful to design a logistic network with better reliability. In fact, it is very difficult to measure and control the degree of interference and damage that the natural disasters made on logistic system. Many experts and scholars have carried out relevant research on this aspect. However, most scholars’ research on logistic networks is based on a single network for modeling and analysis[1, 2], and does not take the interaction between networks into account[3]. However, in the real world, each functional network is more or less connected with other networks, such as physical dependency, logical dependency, energy or information exchanges, etc. Strictly speaking, entirely isolated network is not exist in reality. Once a network node fails, the impact of the failure will be spread and amplified due to the inter-network connection. Eventually, a small failure may have catastrophic consequences for the entire system. Different from previous studies, this paper proposed the complex multi-layer model for emergency logistics system based on interdependent network theory. During the structural invulnerability evaluation, the inter-relationships between networks has been taken into consideration appropriately, which is more accordance with the actual situation of the system.

2 Multi-layer emergency logistics system model

Emergency logistics network is a complex network system. In the operation process, all levels of members, such as the command agencies, security bases, material supply stations, reserve warehouses, demand units, and various entities, such as airports, stations, ports are interrelated to form an organic whole based on the command relationship, communication lines and transportation lines. According to the command flow, information flow and material flow in the emergency logistics network, the emergency logistics network has been modeled as a multi-layer system composed of command and control network, space communication network and physical transportation network.
3 Numerical experiment

In this paper, a regional emergency logistics system in Sichuan Province, China is selected as the object of numerical experiment. And, the multi-layer emergency logistics system model is established according to the modeling method proposed before. Fig 1 shows the topology of this system, where contain 288 nodes, including 80 command nodes, 96 communication nodes, 112 traffic nodes, and 817 sides, including 102 lines of command relation, 180 connections of communication relation, 235 lines of traffic relation, 300 connecting edges between networks. When the emergency logistic system suffering intentional attack, the attack order may determined according to the node degree \(\{v_1, v_2, ..., v_k\}\). And, the command-control network, the information communication network and the transportation network suffered attacks respectively. In order to obtain more reasonable results, the numerical experiment is repeated 200 times, and the average value of each experimental result is regard as the final result, showing in Fig 2.

In Fig 2, when \(S \neq 0\), facing with same attack strength, it can always find that \(S_r < S_L < S_C < S_E, S'_r < S'_L < S'_C < S'_E\). This conclusion shows that the invulnerability of multi-layer emergency logistics network is superior to that of the single-layer transportation network. The invulnerability of the single-layer transportation is better than that of the single-layer communication network. The single-layer information communication network is superior to the single-layer command control network. The single-layer network model cannot reflect the invulnerability of multi-layer emergency logistics system because of ignoring the interdependence between the networks. \(S_L(i_c = 66) = 0\), indicating that only a selective 66 times attacks can completely make the system paralyzed when the attacker obtain the information about the emergency logistics system accurately. So, it can be concluded that the network is vulnerable to deliberate attacks. Moreover, a comparative analysis of the invulnerability evaluation results based on different evaluation indexes. Moreover, when \(S \neq 0\), it always can be found...
that \( S'_C > S'_T, S'_C > S'_E, S'_C > S'_T, S'_C > S'_E \), which means that the effective connectivity is greater than the maximum connectivity. This is because the effective connectivity rate taking the nodes which in non-largest effective connectivity group into account. In fact, those nodes are still functional in real system operation.

4 Conclusion

This paper proposed the complex multi-layer logistic model consists of command-control network, space communication network and physical transportation network based on interdependent network theory. By analyzing the invulnerability result under attentional attack and random attack, the conclusion can draw as followed. Firstly, the three-layer network model of emergency logistics system based on the interdependent network theory is more coincident with the actual situation of system compared with the network model established by traditional single network model. Moreover, the invulnerability analysis method proposed in this paper is more accurate and detailed than the traditional method. The next step in the study, it is meaningful to extend the single dependency relationship of a node to multiple dependencies, and the invulnerability result can be further optimized.

Reference

Multiplex Complex Contagion Co-Diffusion

Ho-Chun Herbert Chang and Feng Fu
Department of Mathematics, Dartmouth College, Hanover NH 03755, USA,
herbert.hc.chang@gmail.com

1 Introduction

Guilbeault et al. [1] outline three main directions for research in complex contagion diffusion: an ecology of contagions, the mechanisms of diffusion, and population structures. This study addresses the first two, where we examine the synergistic diffusion of two contagions on a multiplex network. Co-infection is a notion in epidemiology that describes multiple contagions interfering with each other, where they can simultaneously infect a host. Co-infection can be extended to model complex contagions as, like infectious diseases, they do not spread in isolation.

We study the interface of three elements of diffusion: the synergy between contagions, the dormancy rate of each individual contagion, and the multiplex network topology. Dormancy is defined as a weaker form of immunity—dormant agents no longer actively participates in diffusion, but are still susceptible to infection. In sum, the proposed model is desirable as it numerically quantifies synergy and shows the intricate inter-dependencies between network layers when percolation occurs [2].

2 Method

Suppose there are two contagions, Contagion A and Contagion B. Each node i can attain four possible states: naïve (0), A, B, and AB. Additionally, they are either active or dormant, represented by a binary variable. Experiments take place on a multiplex network with a lattice layer and a regular-random-graph (RRG) layer. We keep the degree fixed at four, to contrast short-range diffusion via von Neumann neighborhoods (lattice layer) and long-range connections (RRG). Experiments were run on 6400 nodes, with 1 initial seed for both Contagion A and B, with 100 Monte-Carlos iterations. We model inclusive adoption, where nodes can adopt both states (denoted AB).

Each node is assigned a threshold value within (0, 1). The probability of diffusion then follows canonical logistic growth, given by a multivariate Hill Equation (Eq. 1 and Fig. 1). The concavity determines whether its additivity is synergistic or antagonistic. The $K_i$'s denote the attractiveness of each contagion. $[A]$ denotes the density of neighbor nodes with status A. $S_i$'s are indicator functions, which reduce the diffusion probability to uni-variate sub-cases within canonical logistic diffusion.

$$P(i) = \frac{\left(1 - S_A(i)\right)^{\frac{[A]}{K_A}} + \left(1 - S_B(i)\right)^{\frac{[B]}{K_B}}}{1 + \left(1 - S_A(i)\right)^{\frac{[A]}{K_A}} + \left(1 - S_B(i)\right)^{\frac{[B]}{K_B}}}$$ (1)
Dormancy is useful when modeling real world phenomena, as complex contagions rarely diffuse fully. Dormancy is modeled similar to recovery within the epidemiological literature, but does not confer resistance. Dormant notes are still susceptible to infection, but are discounted from the density values $[A]$ and $[B]$. This uni-directional definition of dormancy captures the active role required of agents. Contagions $A$ and $B$ is associated with constant $\tau_A$ and $\tau_B$ respectively, each a probability value that denotes the likelihood that an infected node becoming dormant.

3 Results

Diffusion of Contagion $B$ on the RRG occurs much faster due to the RRG’s long-range connections. If $\tau_B$ is positive, Contagion $B$ induces a branching effect on Contagion $A$, restricting its diffusion in a way reminiscent of ring vaccination. We conclude the uni-directional definition of dormancy is sufficient for ring vaccination as shown in figure 2.

As $\alpha$ increases, synergy decreases, which is associated with a lower rate of diffusion. Thus, dormancy acts over a longer period of time and negatively influences the depth of penetration, which we call the ceiling. In Contagion $B$, a linear combination of $\tau_A$ and $\tau_B$ induces a change in the ceiling mean, as shown in the white line in figure 3.
Along this line also comes the greatest instability (in standard deviation). Contagion A in contrast has the greatest instability when the difference between \( \tau_A \) and \( \tau_B \) is large [4].

![Diffusion Curves](image)

**Fig. 4.** Diffusion Curves while increasing \( \tau_B \). Low levels of \( \tau_B \) induce a dense lower branch in Contagion A (b). As \( \tau_B \) increases, B percolates and the branches of A are distributed more evenly (c, d). In e) and f), Contagion B rarely diffuses and Contagion A retains its upper branch [4].

When \( \tau_B \) is sufficiently high, increasing the dormancy of a Contagion B slows down its diffusion, which decreases the chance that Contagion B diffuses more quickly than A. However, in cases where A diffuses more quickly, A can provide a synergistic boost to B whose diffusion has stagnated.

**Summary.** We study the interface of synergy and dormancy on the diffusion of complex contagions, each restricted to a layer within a multiplex network. Results reveal that the faster RRG contagion induces branching on the lattice contagion. Three outcomes were recorded. If the RRG contagion diffuses fully, the lattice contagion diffuses partially due to ring vaccination. If the RRG contagion does not diffuse because of high levels of dormancy, then the lattice contagion diffuses fully. A third case arises when the lattice contagion acts as critical mass to bolsters the RRG contagion towards full diffusion, which is the only case where both diffuse fully. Greatest instability occurs at a linear combination on the RRG and when the absolute difference is large for the lattice.

**References**

Extracting Multi-Layer Networks from Sentinel-2 Satellite Image Time Series

Roberto Interdonato, Raffaele Gaetano, Danny Lo Seen, and Mathieu Roche
CIRAD - UMR TETIS, Montpellier, France
{roberto.interdonato,raffaele.gaetano,danny.lo.seen,mathieu.roche}@cirad.fr

1 Introduction

Nowadays, modern Earth Observation systems continuously generate huge amounts of data. The Sentinel-2 (S2) Earth Observation mission, developed by the European Space Agency (ESA) as part of the Copernicus Programme, supplies image acquisition at high spatial resolution (up to 10m) with high temporal revisit period (every 5 days). This information can intuitively be organized into time series of high-resolution satellite imagery, used for area monitoring tasks. Moreover, since S2 SITS are publicly available, these type of data is also attractive for research purposes.

The successful application of classification approaches based on deep learning to remote sensing tasks [5] led to significant improvements in the development of land cover maps, i.e., allowing to automatically learn complex feature representations from satellite images. Nevertheless, different representation models for satellite images can be defined, i.e., based on complex network models, which may allow to discover new relations between the spatial objects in an image and to apply existing analysis techniques to a new domain.

The use of network analysis techniques in the remote sensing domain has often been neglected. Our hypothesis is that a network model would lead to a spatial representation of satellite data that goes beyond conventional mapping, taking into account in particular the different types of relationships between objects or segments identified in the images. The aim of this work is to show how Satellite Image Time Series (SITS) can be profitably represented using complex network models, by proposing a method to extract a Multilayer Network from a SITS.

This will enable the application of network analysis techniques to remote sensing data, paving the way for the definition of innovative approaches able to take into account, together with the spectral and textural information, the contextual information deriving from the spatial organization of the entities in a landscape. As next step, we plan to use multilayer community detection approaches to investigate how the spatial organization of the landscape suggested by the community structure evolves over time, i.e., identifying communities which correspond to sets of image segments or specific objects. While global approaches [4] will help studying the organization of the whole area depicted in the SITS, local ones [2] will allow to focus on specific objects or zones of interest.
2 Proposed Model

In this work, we will propose a methodology to build a multilayer network [3] starting from satellite image time series. In our case, we intend to represent in each layer the image coming from a different timestamp in the series, i.e., obtaining an ordered multilayer network modeling the evolution of a spatial area in a certain time range. Given a SITS containing a set \( S = \{ s_0, \ldots, s_n \} \) of \( n \) images representing the same scene at \( n \) different timestamps, the aim is to obtain a multilayer network \( G = (V, E, L) \), where \( V \) is a set of nodes, \( L \) is a set of layers of size \( n \) (one for each image/timestamp), and \( E \subseteq (V \times L) \times (V \times L) \) is a set of edges (i.e., connecting node-layer couples). We will now describe in detail the three main steps of the proposed methodology: Image segmentation, Interlayer edges insertion and Intralayer edges insertion.

**Image segmentation.** At the first step of our procedure, each image \( s_i \in S \) is partitioned into \( k_i \) segments, i.e., coherent sets of pixels. Each segment of image \( s_i \) will then correspond to a node in layer \( L_i \); note that the segmentation of each image is performed independently, i.e., the number of nodes in each layer may be different. In our example, for efficiency and scalability reasons, we used Felsenszwalbs image segmentation [1] in its Scikit-Image python implementation \(^1\), but any other image segmentation algorithm may be used in its place.

**Interlayer edges insertion.** The second step includes the insertion of the edges connecting nodes in different layers. In other cases (e.g., multiplex networks) only coupling interlayer edges may be allowed, i.e., connecting nodes which represent different instances of the same entity. Conversely, in our case it is not possible to have a 1 : 1 correspondence between nodes in different layers, due to the independent segmentation processes. For this reason, we insert an interlayer edge \( (v_i, l_a, v_j, l_b) \) s.t. \( a \neq b \) if the segments corresponding to nodes \( v_i \) and \( v_j \) have at least a common pixel (i.e., there is an overlap between the segments). Due to the temporal nature of the network, interlayer edges are only inserted between a layer and the subsequent one (i.e., \( b = a + 1 \)), following the time series ordering.

**Intralayer edges insertion.** The last step consists in adding the edges between the nodes in each layer. For this step we resort to Region Adjacency Graphs (RAGs)\(^2\), i.e., an edge \( (v_i, l_a, v_j, l_a) \) between two nodes in the same layer exists if the segments corresponding to nodes \( v_i \) and \( v_j \) are adjacent. Edges are weighted based on the difference of average RGB color between the two segments, i.e., lower weights will correspond to most similar segments. In practical cases, simple refinement procedures may be applied, e.g., merging similar nodes (i.e., the ones connected by edges showing lower weights) or disconnecting dissimilar ones (i.e., the ones connected by edges showing higher weights).

3 Results

For this example, we consider a SITS of High Spatial Resolution Sentinel-2 images of 100km \( \times \) 100km, depicting the Koumbia area (Tuy Province in Burkina Faso). The SITS

\(^1\)http://scikit-image.org/docs/dev/api/skimage.segmentation.html
\(^2\)http://scikit-image.org/docs/dev/auto_examples/segmentation/plot_rag_draw.html
consists of 23 images referring to the period 2016/01/16–2016/12/21. Table 1 reports main structural characteristics of the multilayer network obtained with the proposed method, while Figure 1 shows the distribution of nodes, intralayer and interlayer edges over the 23 layers. The number of nodes in each layer is rather stable, except for a lower peak at layer 18, maybe due to an excessive amounts of clouds in the image. Given the adjacency criterion used in the RAGs construction, the number of intralayer edges is clearly proportional to the number of nodes (the variability among the layers here is naturally emphasized). The number of interlayer edges (intended as number of interlayer edges outgoing from each layer) is slightly more unstable. Even if some proportionality with the number of nodes can still be recognized, this characteristic is also biased by the outcome of the segmentation process (which in turn depends from the landscape changes during time). Nevertheless, given the complexity of the process and the nature of the input data, an average interlayer degree of 3.99 is fairly reasonable, considering that it represents the average number of segments at time/layer t + 1 which overlap with a segment taken at time/layer t.

![Fig. 1. Distribution of nodes, intralayer and interlayer edges over the 23 layers.](image)

<table>
<thead>
<tr>
<th>layers</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>tot nodes</td>
<td>3167802</td>
</tr>
<tr>
<td>avg nodes</td>
<td>137730</td>
</tr>
<tr>
<td>max nodes</td>
<td>170398</td>
</tr>
<tr>
<td>tot intra edges</td>
<td>9889167</td>
</tr>
<tr>
<td>avg intra edges</td>
<td>429963</td>
</tr>
<tr>
<td>max intra edges</td>
<td>541649</td>
</tr>
<tr>
<td>tot inter edges</td>
<td>11833726</td>
</tr>
<tr>
<td>avg inter edges</td>
<td>514509</td>
</tr>
<tr>
<td>max inter edges</td>
<td>691431</td>
</tr>
</tbody>
</table>

**Table 1.** Structural characteristics of the multilayer network.

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)

Juan C. S. Herrera¹ and Carolyn Dimitri²

¹ New York University, New York City NY 10003, USA
jsh501@nyu.edu,
home page: https://github.com/juancsherrera
² New York University, New York City NY 10003, USA
carolyn.dimitri@nyu.edu

1 Introduction


2 Dataset

Our study models an organic dairy supply chain and its evolution between 2002 and 2015, which is based on data from two datasets.

– United States Organic Integrity Database [1], an open access database that contains data for every organic production and handling facility recognized by the United States Department of Agriculture, along with contact information and the date the entity was first certified. We used the address and the Google Maps API to obtain GPS coordinates of each entity.

– (2) Whole Foods supermarkets [2]. We scraped the publicly available list of Whole Foods locations in the USA and used the Google Maps API to obtain the GPS coordinates for each one. This database is not dynamic, and we were not able to locate data indicating the opening year of each supermarket. We assumed an existence of a potential critical mass of consumers that purchased organic dairy products in the area around the supermarket.

3 Methods

We used network analysis to study a simulated dynamic supply network constructed from the geographical location and proximity of three types of nodes.

– Producers: This set consists of certified organic dairy producers.
Handlers: This set comprises certified organic dairy handlers, which are processing facilities that manufacture pasteurized milk, cheese, yogurt, and other dairy products.

Whole Foods supermarkets. Whole Foods is a high end supermarket that caters to affluent consumers and focuses on offering a variety of organic products, including dairy.

The procedure for modeling the dynamics of the supply chain follows Antonioni et. al model for networks in geographical space [3]. The model captures two types of connections: 1. Producers to handlers. 2. Handlers to Whole Foods supermarkets.

Producer-Handler Network. The first part of the supply network connects producers to handlers. We used a radius of 50 miles for constructing the connections, so each producer-handler pair is at most 50 miles (Euclidean distance) apart. This consideration is based on the fact that, because dairy is perishable, the product is sold over short distances. As of 2018 federal and state laws restrict the sale of raw milk directly to consumers [4], hence we connect every farmer to a handler facility instead of directly to a supermarket.

Handler - Whole Foods Network. The second part of the Supply Network is made of handlers and consumers. We used Whole Foods to approximate consumer demand. We used a radius of 300 miles to construct the connections. This consideration is based on the fact that processed dairy products are able to maintain quality over longer distances than raw or fluid milk.

4 Results

The organic dairy supply chain, as shown in Figure 1 for 2002 and Figure 2 for 2015 evidences that the transition into organic dairy presents different types of behavior. The Northeast of the country is characterized by supply chain comprised of large numbers of producers and handlers. The South and the Southwest are characterized by a small number of producers and handlers. The findings are highly suggestive of the importance of peer effects in the organic farming sector, as the networks indicate that organic farms are located in proximity to other organic farms. Dairy farming has not emerged in certain regions of the country.

5 Conclusion

The findings provide evidence that the organic dairy supply chain varies in the different regions of the country. In the West, the supply chain is composed of fewer producers and handlers, while the Northeast region is comprised of many producers and handlers. The potential consequences of this differences are of interest for food policy. From a food policy perspective our study suggests incentives encouraging a transition to organic dairy production.
References

Coordination in Open-Source Software Engineering
A Multiplex Network Analysis

Angelika Schmid¹, Thomas Bock², Janet Siegmund¹, and Sven Apel²

¹ PICCARD Junior Research Group, Universität Passau, 94032, Passau, Germany, www.fim.uni-passau.de/en/research/junior-research-group-piccard/
² Chair of Software Engineering I, Universität Passau, 94032, Passau, Germany, www.infosun.fim.uni-passau.de/se/

1 Background

Open-source software development depends on the smooth collaboration of a large number of voluntary software developers. Coordination failure, that is, a lack of congruence of communication and collaboration, can lead to errors and problems. Network analysis can help to understand such problems better [1]. As Papalexakis et al. [4] argue, multiplex networks (that is, networks containing several types of interaction between nodes) help to understand complex group dynamics in collaboration networks. To investigate the quality of coordination among developers, we apply a canonical tensor decomposition to the multiplex network of developers connected by communication and collaboration events. This decomposition allows to visualize and analyze inter-personal relationships and their dynamics while differentiating multiple interaction types.

Fig. 1. Setting

We map coordination and collaboration activities to a dynamic unipartite network in two steps, as illustrated in Figure 1. First, developers are related to project-related e-mail threads (retrieved from the public mailing list archive Gmane, http://gmane.org/) and changes of code artifacts (retrieved from git repositories). In many open-source projects, developers have to send their proposed code changes to the project members via a mailing list, which makes mailing lists a central instrument of coordination. In the second step, the data are aggregated: Developers interact if they contribute to the same e-mail thread, or when they work on the same piece of code, within a discretized time window. This aggregated unipartite graph can be represented as a tensor $Y \in \mathbb{R}^{N \times N \times 2 \times T}$, where $N$ is the number of developers, 2 is the number of interaction modes, and $T$ is the number of time windows.

Coordination among developers is characterized by three aspects: inter-personal, dynamic, and cross-channel dependency. For example, developers may have preferred co-workers, with whom they are likely to work in the future again, and with whom they are (hopefully) likely to discuss via mailing lists. Without further modeling, analyzing the dynamics of all three aspects would require comparing $2 \times T$ adjacency matrices.
of dimension $N \times N$. The canonical tensor decomposition separates the three forms of
dependency and reduces them to a reasonable complexity level. It reduces $Y$ to $R$ rank-
one tensors, where $R$ is the rank of the tensor, and $\otimes$ is the outer vector product:

$$Y = \sum_{r=1}^{R} \lambda_r \cdot (u_r \otimes u_r \otimes c_r \otimes t_r),$$

where $\lambda_r \in \mathbb{R}; u_r \in \mathbb{R}^N; c_r \in \mathbb{R}^2; t_r \in \mathbb{R}^T; r \in \{1, \ldots, R\}.$

As our edges are undirected, the tensor is symmetric in the first and second mode, and
the first and second vectors in Equation (1) are the same. The vectors $u_r$, $c_r$, and
$t_r$ and the weight $\lambda_r$ form the $r$-th rank-one tensor or \textit{latent component} of the
tensor [2]. The single elements of the $r$-th latent component fulfill the following function:
$u_r \otimes u_r \in \mathbb{R}^{N \times N}$ is the strength of the interpersonal relationship of the developers; $c_r$
describes how strong communication and collaboration are affected; $t_r$ tracks the temporary
importance of the component; $\lambda_r$ defines the overall importance of the component.
Multiplying any two elements in latent component $r$ by -1 leaves the results unchanged;
we will therefore not interpret the signs of the vectors. From the maximum number of
latent components $R$, the desired number (usually covering the largest weights $\lambda_r$) is
chosen for visualization and interpretation. Kolda and Bader [2] explain technical de-
tails of the canonical decomposition. We will next apply the model to an open-source
software project for demonstration.

2 Example Application

We analyze the open-source software project \texttt{BusyBox} (www.busybox.net, a UNIX
command-line tool suite). We consider developers in the network who collaborated
but did not communicate with any other developer via the project’s mailing list, but
not vice versa. Our data contains 217 developers and 52 three-month time windows,
which means that $Y$ is of dimension $(217 \times 217 \times 2 \times 52)$. For the decomposition, we
use the \texttt{rTensor} package [3] of the statistics software \texttt{R}. Figure 2 visualizes 5 latent
dimensions. Because of space constraints, we discuss only selected details here.

1. In Figure 2-(a), we see that there are generally more collaboration than communication
   events in \texttt{BusyBox}, but there is joint trending behavior.
2. Figure 2-(b) visualizes the components’ weights. The $\lambda_r$ are 5629, 1836, 26578,
   22206, and 7674. Thus, the third and fourth latent components are the most impor-
tant ones, and the second component is the least influential.
3. Figure 2-(c) visualizes the Spearman rank correlation between the developer effects
   of different latent components, $\text{cor}(u_r, u_r')$. Components 3 and 4 describe the same
developers’ relationships: the correlation of $u_3$ and $u_4$ is very strong (0.93).
4. Figure 2-(d) reveals via $c_{1,3}$ and $c_{1,4}$ that components 3 and 4 affect mainly col-
   laboration interaction. $c_{2,3}$ and $c_{2,4}$ show that collaboration activity in component
   4 is accompanied by more activity on the same mailing threads than in component
   3. $c_{1,2}$ and $c_{2,2}$ reveal that the component 2 describes a group of developers
   with intensive mail exchange.
5. Figure 2-(e) visualizes the dynamic weights of the five latent dimensions, \( t_1, \ldots, t_5 \).

4 out of 5 components are inactive in the beginning, gain in importance from \( t = 20 \) on, and then show a constant activity level afterwards. The dynamic weight of component 2 levels out after a peak of activity in periods 13 and 14.

In summary, the decomposition of mailing and collaboration activity of the open-source project BusyBox indicates that communication and collaboration activities are only weakly simultaneously related. This could either indicate a time lag or a deficit in coordination. Next steps include checking time-lagged dependencies between the interaction modes and comparing the outcomes to different open-source software projects.

### References

Extensivity in infinitely large multiplex networks

Maria Angélica Araujo and Murilo S. Baptista

Institute of Complex Sciences and Mathematical Biology, University of Aberdeen, SUPA
AB24 3UE, Aberdeen
mariaaaraujo13@gmail.com

1 Introduction

The theory of complex networks has been studied across many fields, such as sociology, biology, mathematics, physics and computer science and a wide variety of phenomena are described by complex networks, from a microscopic level as in neural networks to macroscopic systems of social interactions and technological systems, for instance.

An important concept that can play a significant role in the understanding of complex dynamical networks is the extensivity. A quantity is extensive, if it scales directly with the size of the system. The concept of extensive quantities can be extended to the scope of complex systems. Ruelle was the first to show that chaotic systems can be extensive, by studying the curve of Lyapunov Exponents (LEs) $\lambda_i$ [1], when arranged in descending order as functions of the their normalized index $i/N$ and calculated at different system sizes $N$. If these curves collapse onto a single asymptotic curve as the system size grows, then chaos is considered extensive. A direct consequence of this is that the sum of positive LEs is linearly related to the size of the network.

Extensive chaos has been detected in several systems. In globally coupled dynamical systems, although all the elements of the system are subject to the same influences, the nontrivial connection between their components can give rise to collective chaos and nonextensive behaviour [2]. Due to the nontrivial connections between their components, in these type of systems, the extensivity of Lyapunov spectra has been questioned [3]. In contrast to this, extensivity has been commonly observed in generic models of spatiotemporal chaos in one dimensional space and in locally coupled systems [4]. It was also detected in different classes of random networks with sparse connectivity [5].

In this paper, we study analytically how extensivity of entropy in multiplex networks can be maintained by smartly changing the configuration of the network as it grows in size. A main point of interest in this work is to understand the role of the intra and inter-couplings in the extensive behaviour of large multiplex complex networks. In general, calculating the entropy of a chaotic network of arbitrary size is an unfeasible task, and therefore, we use the sum of the positive Lyapunov exponents, $H_{KS}$, instead, since this sum is an upper bound for the Kolmogorov-Sinai entropy. At variance with Baptista et al. [6], which have explored this relationship in multiplex networks formed only by two layers of nodes, we analyze in this work infinitely large multiplex networks. We start with an initial network formed by two layers, as described in [6], and then, in each step, we duplicate the network that we had before, and connect the two subnetworks, maintaining the same intra and inter structure. We show that extensivity only depends on the coupling strengths, and that other quantities which proved to be important to maintain
extensivity in [6], such as the sum of the intra and inter-degrees of the nodes of the layers, do not interfere in this behaviour. Moreover, we also show that, given an initial dynamical network with an arbitrary intra-connectivity topology, it is always possible to construct an infinitely large extensive network with an infinite number of these initial networks connected with rescaled inter-link strengths. Since all our calculations are exact, we provide a reliable and practical method to achieve extensivity.

2 Methods

Consider, initially, a network $\Omega^{(0)}$ with $N_0$ nodes where the dynamics of the nodes given by the shift map $F (x^{(i)}_n) = 2x^{(i)}_n \pmod{1}$, i.e.,

$$x^{(i)}_{n+1} = 2x^{(i)}_n - \varepsilon \sum_{j=1}^{N_0} A_{ij} x^{(j)}_n \pmod{1},$$

where $\varepsilon$ represents the coupling strength of the intra-connections and $A = (A_{ij})$ denotes the Laplacian matrix. Now, let $\Omega^{(1)}$ be the network constructed by coupling two equal subnetworks $\Omega^{(0)}$, then its number of nodes is $N_1 = 2N_0$ and it can be represented by

$$x^{(i)}_{n+1} = 2x^{(i)}_n - \varepsilon \sum_{j=1}^{N_1} G_{ij} x^{(j)}_n - \gamma \alpha \sum_{j=1}^{N_1} L_{ij} x^{(j)}_n \pmod{1},$$

where $\gamma$ is the coupling strength of the inter-connections and $\alpha = \frac{l_{12}}{N_0}$ is the ratio between the number of inter-connections, $l_{12}$, and $N_0$. The Laplacian matrices of the intra and inter-connections are represented by $G = (G_{ij})$ and $L = (L_{ij})$, respectively.

The network $\Omega^{(k)}$ with a number of nodes given by $N_k = 2^k N_0$ is constructed by coupling two equal subnetworks $\Omega^{(k-1)}$ with $N_{k-1}$ nodes. The intra and inter-connections are undirected, and we consider a diagonal interlinking configuration, which means that each node in a subnetwork is only connected to the corresponding node in the other equal subnetwork, consequently $l_{12} = N_0$ and $\alpha = 1$. Extensivity implies that the sum of the positive LEs $\sum_{\lambda_{i>0}} \lambda_i$ is a linear function of the size of the system, $N_k$.

3 Results

The LEs of $\Omega^{(k)}$ can be divided into two sets, $A_\varepsilon$ that represents the LEs that only depend on the intra-coupling strength, and $A_{\gamma \varepsilon}$, formed by the ones that depend on both, the intra and inter-coupling strengths, [7]. We have explored analytically the conditions for the network $\Omega^{(k)}$ to have the sum of positive LEs extensive, in three different situations. In the first case, we consider networks having only positive LEs, i.e., all elements of $A_\varepsilon$ and $A_{\gamma \varepsilon}$ are positive. Then, in order to maintain extensivity as the network grows, the intra and inter-coupling strengths $\varepsilon$ and $\gamma$, respectively, must satisfy

$$\varepsilon < \frac{1}{N_0} \quad \text{and} \quad \gamma < \frac{1 - \varepsilon N_0}{2k}.$$
In the second case, we assume that $\Omega^{(k)}$ has positive as well as negative LEs and that the inter-coupling strength $\gamma$ is responsible for the change in the sign of the LEs, i.e., all elements of $\Lambda_{\varepsilon}$ are positive and we do not have any restrictions for the ones that come from $\Lambda_{\gamma,\varepsilon}$. Then, we conclude that in order to grow the network extensively, the coupling strengths must satisfy
\[
\varepsilon < \frac{1}{N_0} \quad \text{and} \quad \gamma < \frac{2 - \varepsilon N_0}{2k}.
\]

In the last case, we suppose that all LEs that come from $\Lambda_{\gamma,\varepsilon}$ are positive and we do not impose any conditions for the LEs in $\Lambda_{\varepsilon}$. Then, the coupling strengths must satisfy
\[
\varepsilon < \frac{1}{N_0} \quad \text{and} \quad \gamma < \frac{1 - \varepsilon N_0}{2k},
\]
which are the same conditions found in the first case studied. It shows that if all the LEs from $\Lambda_{\gamma,\varepsilon}$ are positive, then all the LEs of $\Lambda_{\varepsilon}$ will also be.

4 Discussion

Multiplex networks are fundamentally different when compared to single layer networks and their sophisticated nature can provide a more realistic picture to model large real-world systems, since most of real life phenomena are described by multiple types of connections. This paper explores the role of the intra and inter-connections to maintain extensive behaviour in infinitely large multiplex networks. We do not only demonstrate which are the relevant parameters leading to extensivity, but also provide an exact formula into how they are related. We find that it is always possible to construct infinitely large extensive networks by only rescaling the inter-coupling strengths, without needing to change the intra-coupling strengths of the fundamental moduli composing the network. So, extensivity can be achieved in large multiplex networks regardless of the initial network considered. Thus, for single layer networks, the broadly used rescale $\varepsilon = \frac{1}{N}$ (where $N$ stands for the system size) to study collective behavior of networks of different sizes would not produce an extensive network. Our findings also reveal that networks that are not “super” entropic (i.e., have at least one negative LE) are easier to be made extensive. With regards to the generality of our results, a task left for future work would be to explore networks with different values of the parameter $\alpha$.

References

Part X

Network Analysis
Multiscale core-periphery structure in a global liner shipping network

Sadamori Kojaku¹, Mengqiao Xu², Haoxiang Xia², Naoki Masuda¹,²,†

¹ Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom,
² Dalian University of Technology, Dalian, China

1 Summary

The global liner shipping network (GLSN) is a maritime transportation network dedicating to transporting container cargoes worldwide primarily by high-capacity container ships that sail regular routes on fixed schedules. Nowadays, the trade by container ships accounts for 70% of world trade by value each year [1], making the GLSN indispensable to the world economy. In the present work, we represent ports as nodes and inter-port shipping connections as edges. Many transportation networks such as airport networks, railway networks and road networks have been shown to have core-periphery (CP) structure [2]. CP structure is composed of cores and peripheries, where a core is a group of densely interconnected nodes, whereas a periphery is a group of sparsely interconnected nodes (Fig. 1). The core can be densely interconnected with the peripheral nodes or not. A network with CP structure is suggested to be robust against random attacks on nodes [3] and also to be economically efficient [4]. Here we study CP structure in the GLSN. We extend our previous algorithm for finding CP structure in networks [5, 6]. We identify two dominant CP pairs in the GLSN (i.e., a CP pair consists of a group of core ports and a group of peripheral ports), which contain ports in many geographical regions but few ports Oceania and South America, suggesting that the latter regions may be segregated from world’s central markets for international trade.

2 Methods

Our dataset contains information on the calling ports and the maximum possible volume of containers (i.e., capacity) on 1,631 shipping routes in 2015. A route consists of calling ports. Containers can be transported between any pair of ports in the same route. Therefore, we construct a network of \( N = 977 \) ports by connecting every pair of calling ports of each route by a weighted undirected edge.

We extend the previous algorithm [5, 6] in the following three ways. First, we adopt a multiresolution method for community detection [7]. We represent CP pairs in the GLSN by \( c_i \) and \( x_i (1 \leq i \leq N) \), where \( c_i \) denotes the index of the CP pair to which node \( i \) belongs, \( x_i = 1 \) or \( x_i = 0 \) indicates that node \( i \) is either a core node or a peripheral node,

¹These authors equally contributed to this work
†naoki.masuda@bristol.ac.uk

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
respectively. We seek multiple CP pairs (i.e., \((c_i, x_i)\)) by maximising quality function

\[
Q^\text{CP}_\gamma = \frac{1}{2\Omega} \sum_{i=1}^{N} \sum_{j=1}^{N} (W_{ij} - \gamma E[\hat{W}_{ij}]) (x_i + x_j - x_i x_j) \delta(c_i, c_j),
\]

where \(W_{ij}\) and \(E[\hat{W}_{ij}]\) denote the weight of the edge between ports \(i\) and \(j\) in the GLSN and the corresponding expected value for a null model, respectively, \(\Omega = \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij}/2\), and \(\delta(\cdot, \cdot)\) is Kronecker delta. Note that \(Q^\text{CP}_\gamma\) is a fraction of the weight of edges within core nodes and those between the core and peripheral nodes in the GLSN minus the corresponding expected value for a null model. A larger \(\gamma\) value (\(\gamma \geq 0\)) (so-called resolution parameter) tends to yield CP pairs composed of smaller number of nodes.

Second, the previous algorithm often provides different CP structures in different runs even if the initial condition is the same. We mitigate fluctuation in the results using consensus clustering. Specifically, we run the algorithm 100 times for a given value of \(\gamma\). Then, we test the significance of each CP pair [6]. We construct an undirected and unweighted network of ports, where two ports \(i\) and \(j\) are adjacent if the probability that the two ports belong to the same significant CP pair is greater than or equal to 0.9. We regard each connected component of the new network as a consensus CP pair (we call it CP pair for short in the following).

Third, to compute \(E[\hat{W}]\) in Eq. (1), we use a variant of the configuration model [8] that respects the fact that we obtain the GLSN using a one-mode projection of the network composed of ports and routes.

3 Results

At \(\gamma = 0.1\), the present algorithm identifies a dominant CP pair composed of the ports across different geographical regions (Fig. 2(a)). At \(\gamma = 2.1\), the present algorithm detects two CP pairs (Fig. 2(b)). CP pair 1 contains ports in various geographical regions but few ports in some regions such as Oceania, Japan, South and Central America, and West and East Africa, suggesting the segregation of these regions from world’s major trading regions. At \(\gamma = 4\), the present algorithm identifies four CP pairs, two of which, CP pairs 3 and 4, are based mainly in the European Mediterranean and Northwest Europe, respectively. For all resolution values, a majority of the ports are classified as core ports, indicating that most ports are densely interconnected within each CP pair, i.e., community structure. Note that we did not impose the algorithm to find communities, suggesting that the GLSN has community structure rather than CP structure.

References

Fig. 1. Schematic illustration of two CP pairs and its adjacency matrix. The colour indicates the index of the CP pair. Core and peripheral nodes are indicated by filled and open circles, respectively, in panel (a) and filled and open bars, respectively, in panel (b).

(a) $\gamma = 0.1$

(b) $\gamma = 2.1$

(c) $\gamma = 4.0$

Fig. 2. CP pairs in the GLSN. The filled circles represent the ports that are classified as core ports with probability larger than 0.5. The open circles represent the ports that are classified as core ports with probability less than or equal to 0.5. The open squares indicate the ports that do not belong to any significant consensus CP pairs.
Characterising vorticity interactions during the intermittency route to thermoacoustic instability

Abin Krishnan and R. I. Sujith
Indian Institute of Technology Madras, Chennai, India, abin.roja@gmail.com

1 Introduction

Thermoacoustic instability refers to the phenomenon of self-excitation of large amplitude acoustic pressure oscillations in a confinement due to the positive coupling between the unsteady heat release rate oscillations and the acoustic field of the confinement [1]. Such instabilities have been encountered in rocket motors and gas turbine engines used in aerospace and power generation industries. The large amplitude oscillations during thermoacoustic instability induce severe mechanical and thermal stresses, which may lead to the sudden failure of system components. Thus, even after decades of research, understanding the origin, control and prediction of thermoacoustic instability still continues to attract researchers from all over the world.

In a turbulent combustor, the transition from stable combustion (also known as combustion noise) to thermoacoustic instability occurs via a dynamical state known as intermittency [2]. In the present study, we characterise the vorticity interactions during combustion noise, intermittency and thermoacoustic instability. To that end, we construct weighted time-varying turbulent networks using the vorticity field obtained from the analysis of particle image velocimetry (PIV), a laser based optical diagnostic technique used to obtain the 2D velocity field of the turbulent reactive flow inside the combustor.

2 Methodology of network construction

We divide the turbulent reactive flow field into Cartesian cells. We obtain the vorticity at each of the cells from the analysis of PIV. We use Biot-Savart law to characterise the vorticity interactions occurring in the turbulent combustor[3]. We construct vorticity networks at each instant of time. The fluid elements in the Cartesian cells are the nodes of the network, while the links of the network represent the vortical interactions among the fluid elements. Here, we consider the mean of the induced velocities between two fluid elements as the weight of the link between them.

3 Results

Combustion noise is characterised by low amplitude aperiodic oscillations and small discrete vortices in the flow field. Figure 1 shows (b) the vorticity field and the velocity vectors obtained from the analysis of PIV, (c) the spatial distribution of the node
strength of the vorticity network with the velocity vectors superimposed and (d) the node strength distribution in log-log scale at some arbitrary points (A-E) marked on (a) the time series of the acoustic pressure fluctuations during combustion noise. Regions with high values of node strength ($s$) coincides with the regions of high vorticity ($\omega$), which, in turn, corresponds to the small vortices in the flow field. The power law exponent ($\gamma$) of the node strength distribution suggests a scale-free structure for the vorticity network. This implies that there is a large number of vortices with weak interaction strength and a few vortices with very high interaction strength in the flow field.

![Diagram](image_url)

**Fig. 1.** The spatial distribution of (c) the node strength of the vorticity network at the corresponding time instants (A-E) marked on the (a) time series of fluctuating pressure during combustion noise. The velocity vectors and the vorticity field is shown in (b).

Thermoacoustic instability is characterised by high amplitude periodic oscillations and periodic formation of coherent structures in the flow field. The flow field (b), the spatial distribution of the node strength (c) and the node strength distribution with the power law exponents (d) at five different time instants of an acoustic cycle during thermoacoustic instability (a) are shown in figure 2. During the formation of the coherent structure, the vorticity network displays small world property ($\gamma > 3$,[4]) while, during other time instants, the vorticity network displays scale-free nature ($2 < \gamma < 3$). Thus, during one acoustic cycle of thermoacoustic oscillations, the vorticity interaction network oscillates between small-world and scale free networks periodically.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Summary. We construct weighted time-varying spatial turbulent networks based on Biot-Savart law to characterise the vorticity interaction during the different regimes of combustor operation. We observe that the vorticity interaction during combustion noise is characterised by scale-free networks, whereas during thermoacoustic instability, the network topology oscillates between small world and scale-free behaviour in one acoustic cycle. The future work will try to relate the flame dynamics during the different dynamical states to the corresponding changes in the network topology.

References

Unpacking maritime trafficking from the complex network approach

Yi-Chun Chang¹, Wei Ueng² Yen-sheng Chiang³, Seng-Cho T. Chou¹, and Chia-te Chiang³

¹ Department of Information Management, National Taiwan University, Taiwan
d03725001@ntu.edu.tw
² Investigation bureau, Coast Guard Agent, Taiwan
victor@cga.gov.tw
³ Institute of Sociology, Academia Sinica, Taiwan
chiangys@gate.sinica.edu.tw

1 Introduction

Illegal trafficking imposes a great amount of social cost to a country. Understanding how trafficking activities are organized and how they can be disrupted is an important issue not only to researchers, but also to policing officials. Mining a large database administrated by the coast guard agency, we tracked the records of more than 1,500,000 vessels in and out the harbors in Taiwan to pin down the collaboration networks among traffickers. We found that the trafficking collaboration network fits a power-law distribution and the key players of the crimes take more central position in the network than other periphery players. We develop a risk assessment model to identify suspicious smugglers for future surveillance. The complex network approach allows us not only to identify who are the key players in the crime, but also guides us with respect to how to disrupt the network to mitigate its development.

2 The Maritime trafficking network

For different types of crime, the law enforcement should realize the crime network structure and make an appropriate strategy to disrupt the crime group. Duijin et al found that the trafficking network is a scale-free network with power law degree distribution. Therefore, the removal of higher degree nodes works to collapse the trafficking network [2]. Besides, trafficking network has high clustering coefficient and short average path [3]. In this study, we examined the network topology of the maritime trafficking network in Taiwan. We found the network fits the power law distribution. The results show in Fig. 1. The average path length is 2.549, and the clustering coefficient is 0.809. Moreover, we calculated the transitivity that proposed by Krackhardt to observe the efficiency of node removal in the network [4]. In Fig. 1(b), the transitivity suddenly drops after the removal of the top 30 percent high degree nodes.
3 Dataset

We mine the datasets of two information systems, the CGIS (Coast Guard Information System) and the CGVRS (Coast Guard Violation Record System) in Taiwan. The CGVRS records the violations of all the regulations of the ocean, and the CGIS keep track of the personnel records of vessels. By the integration of these two datasets, we create a maritime trafficking networks. At first, we selected 247 smugglers to construct the smugglers network. The smuggler is who committed five type of trafficking cases: 1 guns and weapons; 2 drugs; 3 agricultural, forestry, fishery and livestock products and other goods; 4 illegal entry or exit; 5 human trafficking. We then expanded the network by connecting to their cohorts who committed trafficking cases or violated minor illegal fishing regulations along with the smugglers. Finally, we combined the sailing records and cooffending records and extract the abnormal vessel record related to these trafficking groups. With this approach, we found 26 potential cohorts who have direct or indirect connection with the smugglers (Fig. 3). Further examination shows that 11 of these potential cohort commit-ted trafficking before, and 15 of them violated other regulations like illegal fishing.
4 Results and Conclusion

We rank the nodes of the network by their degree centrality, betweenness centrality, eigenvector centrality, and the measure of constrains by Burt (Fig. 3). It shows that constrains is the best measure to identify smugglers; in contrast, high degree is not associated with the importance of the role in the crime. There findings help investigators consider whom to pay attention to for future crimes. For further work, we plan to collect more personal information, such as prior crime records, the post on the boat and restriction of exit, to have deeper understanding of how these variables are related to the trafficking network. These variables would help us have a broader evaluation of the risk of each individual in the crime.

References

Time series analysis in earthquake complex networks

Denisse Pastén\textsuperscript{1,2}, Zbigniew Czechowski\textsuperscript{3}, and Benjamín Toledo\textsuperscript{1}

\begin{itemize}
\item \textsuperscript{1} Universidad de Chile, Santiago, Chile, denisse.pasten.g@gmail.com,
\item \textsuperscript{2} Advanced Mining Technology Center,
\item \textsuperscript{3} Institute of Geophysics, Polish Academy of Sciences, Warsaw, Poland
\end{itemize}

1 Introduction

The spatial distribution of the hypocenters has been studied using complex networks, converting the spatio-temporal distribution of seismic events into a complex network [1–3]. It has been shown that these type of earthquake complex networks have features of scale-free and small world networks [4–8]. On the other hand, in recent years, some procedures of constructing time series from complex networks have been proposed. One of them generates the time series of node connectivity following a finite-memory random walk on nodes of an undirected complex network [9]. In this work, following the method of Weng et al. [9], we generate a time series from the earthquake hypocenter complex network, but instead of using the random walk we follow a real time path according to the occurrence of seismic events. The nodes of the earthquake hypocenter complex network [1, 3] are formed by cubic cells that contain hypocenters. Nodes in the network are connected according to the time sequence of seismic events in a time period under consideration. In this case, we consider the undirected complex network only. Nodes are characterized by the value of connectivity. The connectivity $k_n$ is the number of links or connections for the $n$th node. After this procedure, following the time sequence of the seismic events in these complex network, we construct the connectivity time series and characterize its multifractal behavior. This approach enables the investigation of the seismic process in space and time simultaneously. We applied the procedure to four data sets in Chile with different seismic activity (different number of seismic events in a chosen time period).

2 Data

We have studied four seismic data sets measured in Chile, in three different zones of Chile. The first set contains data from the central zone of Chile, considering 33160 seismic events. The second set contains data from the far north of Chile near Iquique, and includes the $M_w = 8.2$ earthquake of April 1, 2014. The third data set was measured in the near north of Chile and contains 3752 seismic events. The fourth data set was measured in the central zone of Chile. This data set does not contain events greater than $M_w = 6.0$, it has 14 events greater than $M_w = 5.0$. We called this data set the “calm period” (without large earthquakes). The total number of events in this zone is 5568. These four seismic data sets were measured by the National Seismological Center (Centro Sismológico Nacional) of Chile [10] using local measurements.
3 Results: Singularity spectra and wavelets analysis

In order to construct the complex network for each of the four seismic data sets described in Section 2 we divide the three dimensional volume, in which all events of the zone reside, into cubic cells with side $\Delta$. Each cubic cell can form a node if one or more hypocenters of a seismic event are inside the cell. Nodes in the network are connected according to the time sequence of the seismic events, i.e., if the first event of the data set occurs in node, say, 10 and the second event occurs in node 450, a connection from node 10 to node 450 is generated, and both nodes are thus connected. When the last event is considered then the undirected network is completed (self-connections and multiple connections between nodes are not taken into account). Each node $n$ can be characterized by its final connectivity $k_n$, i.e., the number of links with other nodes. Now, it is natural to construct the time series $k_i = k_{0i}$ of connectivity by walking on the network according to the real time order $t_i$ of seismic events.

In this work we will analyze multifractal characteristics of four connectivity time series constructed from the four earthquake complex networks. Two methods will be applied: the Multifractal Detrended Fluctuation Analysis (MFDFA) and the Wavelets Analysis.

**Multifractal Detrended Fluctuation Analysis**

The Multifractal Detrended Fluctuation Analysis (MFDFA) proposed by Kantelhardt et al. [11, 12] is the widely applied procedure of calculating the multifractal spectra. Therefore, here we present only a short view on the algorithm. For the given time series $x_i$ with average value $\langle x \rangle$ the profile...
is divided into $M_s$ disjoint segments $\nu$ of lengths $s$. For each segment, the polynomial $P^{(m)}_\nu$ of order $m$ is fitted to estimate the trend. Hence, the variance of detrended data in each box $\nu$ has the form

$$F^2(\nu, s) = \frac{1}{s} \sum_{k=1}^{s} [Y((\nu-1)s+k) - P^{(m)}_\nu]^2.$$  \hspace{1cm} (2)

If the $q$th-order fluctuation function $F_q$: 

$$F_q(s) = \left[ \frac{1}{M_s} \sum_{\nu=1}^{M_s} [F^2(\nu, s)]^{q/2} \right]^{1/q}$$  \hspace{1cm} (3)

scales within some range of $s$ according to a power law

$$F_q(s) \propto s^{H(q)},$$  \hspace{1cm} (4)

then $H(q)$ denotes the generalized Hurst exponent.

In multifractal analysis, apart from differences of generalized Hurst exponents $dh$, the singularity spectra are usually estimated [11]. Singularity spectra (Fig. 2) calculated from MF DFA give the following widths of spectra: 0.45, 0.73 and 1.04, for Illapel, Central Zone and Central Zone “calm period” respectively. For Iquique data set we obtained a small value for the width, 0.20, that suggests monofractal features of these data set.

![Fig. 2. In left panel, singularity spectra $f(\alpha)$ calculated from MF DFA for the four data sets: Iquique (red), Illapel (green), Central Zone (black) and Central Zone “calm period” (blue). In right panel, multifractal spectrum dimension, $D(h)$ versus $h$ for wavelets analysis: Illapel (green), central zone of Chile between 2000 and 2015 (black) and central zone of Chile between 2004 and 2009 (blue).](image)

The wavelets analysis leads to plots for multifractal spectrum dimension, $D(h)$ versus $h$, for Illapel, Central Zone and Central Zone “calm period” (see Fig. 2). The plot for Iquique is not presented because the wavelet method is not proper for monofractal data -it leads to overestimation of the spectrum width [13].
Summary. The main objective of our study is to present a novel method of spatiotemporal description of seismic data sets. It is based on relations between complex networks and time series. In this work we were interested in nonlinear correlations hidden in the time series and manifesting by the level of multifractality. In order to estimate the level we applied two methods: MFDFA and the wavelets analysis. The results show the presence of a certain regularity-multifractality decreases with increasing seismic activity (or decreases when seismic activity increases), in this sense we have shown that there is a relation between strong earthquakes and the level of multifractality in the time series of the spatial seismic complex network. Considering these results, we could suggest that the occurrence of great earthquakes destroys nonlinear correlations, established earlier in a premonitory period of the arrangement of seismic system, and turn the system into a simpler monofractal state. Therefore, our approach reflects spatio-temporal reorganization of fault systems in fractal and multifractal description of connectivity time series. To the best of our knowledge, this is the first study that presents a method of characterizing spatio-temporal behavior of seismic complex systems using the procedure of transformation from complex networks into time series. The new method of analyzing seismic catalogues need more testing by using an earthquake model.

References

Assortative mixing in weighted directed networks

Uta Pigorsch and Marc Sabek
University of Wuppertal, Wuppertal NRW 42119, Germany,
pigorsch@uni-wuppertal.de, sabek@uni-wuppertal.de
https://www.statistik.uni-wuppertal.de

1 Introduction

Assortativity is the tendency of a vertex to bond with another based on their similarity, with similarity being usually measured via vertex degree. The most popular assortativity measure is the assortativity coefficient, which is defined as the Pearson correlation coefficient between the excess degrees of both ends of an edge. This very popular assortativity measure has been originally proposed by [1] for unweighted and undirected networks and extended to directed networks by [2] by computing the correlation between the excess out- and in-degree of both ends of an edge. [3], instead, suggest to compute the correlation between the excess out-degrees or between the excess in-degrees providing separate measures for out-assortativity and in-assortativity, and, as such, additional information with respect to the topology of the analysed network.

So far, as already pointed out by [4], assortativity in weighted networks has been insufficiently studied, which is surprising as many real-world networks exhibit weighted edges. [5] consider weighted, but undirected networks. In such networks, the assortativity coefficient of [1] can be computed, but it neglects important information about the intensity of the interaction between two vertices. [5] therefore propose the weighted assortativity coefficient which takes edge weights into account. However, as we will detail below, this weighted assortativity coefficient falls short in the sense that it does not consider the strengths of vertices, but focuses solely on their degrees.

In this paper we propose a more general coefficient of assortativity that nests the aforementioned assortativity measures as special cases, and that can be applied to unweighted/weighted, undirected/directed networks. Moreover, we show that the use of our assortativity coefficient together with its expected statistical error, which we obtain by using the jackknife method [2, 6], enables us to determine the underlying assortative structure of the network more precisely. Furthermore, a comparison of the weighted and unweighted correlation coefficients allows to obtain an accurate characterization of the assortative structure of a network.

2 Method

According to [5] the assortativity coefficient for undirected weighted networks is

\[ r_{LC} = \frac{H^{-1} \sum \omega_i (j_i) - \left[ H^{-1} \sum \frac{1}{2} \omega_i (j_i + k_i) \right]^2}{H^{-1} \sum \frac{1}{2} \omega_i (j_i^2 + k_i^2) - \left[ H^{-1} \sum \frac{1}{2} \omega_i (j_i + k_i) \right]^2}, \] (1)
where $j_i$ and $k_i$ are the excess degrees of the ends $j$ and $k$ of edge $i$, the weight of the $i$-th edge is denoted by $\omega_i$ and $H = \sum \omega_i$ is the sum of edge weights where the sum is over the edges. Obviously, if all edge weights equal one, i.e. the network is unweighted, the coefficient in (1) reduces to the original assortativity coefficient $r^N$ proposed by [1].

The underlying mechanism of this assortativity coefficient, can easily be illustrated. For the ease of exposition suppose integer-valued weights. Then incorporating edge weights is equivalent to replacing each $\omega$-weighted edge by $\omega$ edges with weight one. Thus, high-weighted edges amplify the impact of their connections and therefore contribute more to the overall assortativity. This is reasonable, however, a more obvious generalisation of assortativity to weighted networks would have been to compute the correlation between the excess strengths of both ends of an edge, since this is what degrees generalise to in weighted networks, see [8], who defines the strength of a vertex to be the total weight of its connections, i.e. $s_u = \sum_{v \in V} w_{uv}$ where $V$ is the vertex set and $w_{uv}$ is the weight of the edge between $u$ and $v$.

In fact, note that the emergence of assortativity in a weighted network consists of two mechanisms. The first one is the just mentioned amplification effect. The second one is the connection effect, which is either assortative or disassortative. More precisely, consider two arbitrary adjacent vertices and suppose they have the same degrees but one is the connection effect, which is either assortative or disassortative. More precisely, consider two arbitrary adjacent vertices and suppose they have the same degrees but different strengths. The connection between them is assortative if degrees are used as vertex values, but is disassortative if strengths are used. The second mechanism is not considered in (1).

In the following, we propose a generalised weighted assortativity coefficient that takes both of these effects into account. To this end, let $s'_u = \sum_{v \in V} \omega_i^\alpha w_{uv}^\alpha$, $\alpha \in \{0, 1\}$, be a modified version of vertex strength. Clearly, if $\alpha = 1$ then $s'_u = s_u$, whereas for $\alpha = 0$ it reduces to ordinary vertex degree. Our generalised weighted assortativity coefficient is then defined as

$$r^\alpha_{(\alpha, \beta)} = \frac{\sum_i \omega_i^{\beta} l_i m_i - \Omega^{-1} (\sum_i \omega_i^{\beta} l_i) (\sum_i \omega_i^{\beta} m_i)}{\sqrt{\left[\sum_i (\omega_i^{\beta} l_i^2) - \Omega^{-1} (\sum_i \omega_i^{\beta} l_i)^2\right] \left[\sum_i (\omega_i^{\beta} m_i^2) - \Omega^{-1} (\sum_i \omega_i^{\beta} m_i)^2\right]}}$$

where $l_i$ and $m_i$ are the excess (in- or out-) strengths of the ends $l$ and $m$ of edge $i$. For example, $l_i = s'_l - \omega_i^\alpha$ is the excess strength of end $l$ of edge $i$. Furthermore, $\Omega = \sum_i \omega_i^{\beta}$ with $\beta \in \{0, 1\}$. Obviously, if $\beta = 1$ then $\Omega = H$, whereas for $\beta = 0$ it reduces to the total number of edges in the network. The generalisation is achieved by introducing $\alpha$ and $\beta$, which account for the two different mechanisms, the connection effect and the edge amplification effect, respectively. As such the previous definitions of assortativity are nested as special cases, in particular $r^0_{(0,1)} = r^L$ and $r^0_{(0,0)} = r^N$. In contrast to (1), which is based on the formulation of [1], our coefficient in (2) is based on the formulation by [2, 7], and thus it is capable of handling directed networks as well as undirected networks.

3 Results

We analyse the assortativity of two networks. The first is a toy network introduced by [5], whereas the other is the often studied C. elegans neural network, see [9]. Tab. 1...
The coefficient that best describes the assortative structure of the C. elegans network is $r^\beta_{(1,0)}$, whose actual value depends on the interesting mode of assortativity. The network is out-assortative but in-disassortative and again disassortative if the correlation between out- and in-strengths is considered. However, the fact that $\beta = 0$ indicates that the amplification effect does not become important here. As before, a comparison of weighted and unweighted coefficients will help to identify the connection effect.

Future work will provide a more thorough discussion of our newly proposed coefficient for different networks and further applications. Moreover, we aim to introduce a weighted link rewiring algorithm in order to obtain bounds for the maximum assortativity as well as for the random case to put the observed assortativity into perspective.

### References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Network property and learning model of weblike equation system in a scientific category

Tetsuya Kojima¹, Jun Ishida¹, Shigeyuki Miyagi¹, and Osamu Sakai¹

¹ The University of Shiga Prefecture, 2500 Hassaka-cho, Hikone, Shiga 522-8533, Japan
oh23tkojima@ec.usp.ac.jp

1 Introduction

One category of physics or science in general is composed of huge number of elements, forming a complex weblike system, and an introductory-course student who learns it feels difficulties when one manages to obtain its concepts and meanings. Recently various research activities have been conducted about suitable educational methods, environments and system organization of knowledge. For example, several reports pointed out importance of interactions among students like active learning to arrange educational environments for understanding physics [1].

Another approach is supply of views on conceptual understanding of knowledge, where elements of physical contents are described using network structure, which enables us to obtain a macroscopic view of knowledge [2]. Following such previous work, here we focus on systematic understanding of equations in one scientific category, and consider a equation system in the textbook of electromagnetism [3].

Using derivation relations among equations, we proposed a graphical display method of an equation network with nodes of equations and edges of derivation relations in the textbook [4]. Similarly in studies on various complex networks for objects like document citations [5], we investigated networks of complex systems of chemical reactions [6, 7], and, in this study, the reaction equations are replaced by mathematical ones; in a typical textbook of physics, hundreds of equations form one system of knowledge.

2 Methods of network formation and result of statistical property

In a textbook, starting from very fundamental equations, other equations follow them, and we can set a directed edge from X to Y if equation Y is derived from X. For example, if Faraday’s law \((\nabla \times E = -\partial B / \partial t)\) is under the electrostatic condition, we can derive \(\nabla \times E = 0\), which is clearly described in the given sentences [3]. Almost all derivation processes are easy to clarify in the context [4]. After careful reading of the textbook of electromagnetism [3] from chapter 4 to the end, we obtain a network as shown in Fig. 1 (a), where the total number of nodes or equations \(N\) is 538 with the total edges as 530.

Figure 1 (b) shows degree distribution of the network in Fig. 1 (a). The average degree is 1.97, and the data plot is between the power-law distribution (degree exponent: 2.38) and the Poisson’s one. This result indicates that, although most equations are just connected one or two other equations, some of them play important roles for bundling others by many relations and stabilizing the total system.
3 Models of equation learning and numerical results

To simulate learning processes of equations by students, we perform numerical simulation in which probability processes proceed along the branches of the edges in Fig. 1 (a). We set initial 21 equations that are hypothetically understood by students in advance, where the number of equations that are classified as important equations in this textbook is also 21 [3]. Along directed edges, the working points move over nodes until this hypothetical student does not understand an equation.

We compare results of methods based on two algorithms shown in the following. In case A, the 21 equations are the same as indicated in the textbook as important ones, and the trace of one algorithm is (1)(a), (2), (3) and (4). In case B, the 21 equations are selected automatically by the network topology, and the trace is (1)(b), (2), (3) and (4). (1) Choose preferentially 21 starting equations (a): which are selected as important ones in the textbook, or (b): which are selected from network topology by the sub-algorithm from (i) to (iv), and start a simulation run. (2) When one equation \( Y \) is directed by one edge from that previously understood \( X \) \((X \rightarrow Y)\), one can understand an equation with probability \( p \). (3) If one has already understood equations \( X \) and \( Y \), one can also understand equation \( Z \). (4) The run performed as in (2) or (3) continues until the process stops when one cannot understand an equation with the probability \( 1 - p \). (5) (1) to (4) are carried out for 10000 runs from the probability \( p = 0 \) to 1 at every increment of 0.01, and we store records of the understood equations.

The sub-algorithm for case B is in the following; here, to optimize learning efficiency, we select the 21 initial equations by considering distances among them.

(i) We calculate values of reverse simple pagerank index \( c_{SPR} \) [7] for all nodes in the network. (ii) We set the threshold value as the minimum distance \( L \) \((0 \leq L \leq 5)\) among the initial equations. (iii) We select one equation whose distance to the already selected nodes is larger than \( L \), in descending order of \( c_{SPR} \), and continues until 21 equations are fixed. (iv) The procedure from (i) to (iii) is performed when we change \( L \).

Figure 2 shows numerical results of the learning simulations. The derivation-degree \( D_p(= N_p/N \times 100) \) is the rate how many equations were derived with a certain probability \( p \), where \( N_p \) is the average number of nodes at \( p \) over 100000 runs. The inset figure in Fig. 2 is an example of one simulation run. These results indicate that, the selection of equations that professors put emphasis on are effective for fractions of equations un-
derstood by students, and suitable distances around 2–3 exist which make the number of understood equations maximum.

![Graph showing the averaged derivation-degree $D_{p,\text{avg}}$ as a function of probability $p$ in numerical simulation in cases A and B (left figure), and one shot of derivation procedure of equations in network when $p = 0.2$ and $L = 2$ (right figure). Green, yellow and blue nodes indicate equations initially understood, derived ones in processes, and non-understood ones, respectively.]

4 Conclusion

We created network and investigated its topology by use of derivation relations for the equations of one textbook on electromagnetism. Based on its configuration, we performed numerical runs to simulate student’s learning. The optimization of the starting equations by centrality index (reverse simple pagerank values) and by distance among them can enhance understanding of equations, in comparison with the case of the starting equations preferentially selected by the authors of the textbook. This method contributes to analysis on system of knowledge as well as efficient education in classrooms.

References

Part XI

Network Models
Random Network Models Based on Density

Glencora Borradaile\textsuperscript{1}, Theresa Migler\textsuperscript{2}, and Gordon Wilfong\textsuperscript{3}

\textsuperscript{1} School of EECS, Oregon State University  
\textsuperscript{2} California Polytechnic State University, San Luis Obispo  
\textsuperscript{3} Bell Labs

1 Introduction

We present random graph models based on the density distribution of a graph. The density decomposition of a graph is a hierarchical partition of the vertices into regions of uniform density \cite{1}. The density decomposition is unique in the sense that a given network has exactly one density decomposition. The number of vertices in each region defines a distribution of the vertices according to the density of the region to which they belong, that is, a density distribution. Although density is closely related to degree, the density distribution of a particular network is not necessarily similar to the degree distribution of that network. For example, in many synthetic networks, such as those generated by popular network models (e.g. preferential attachment and small worlds), the density distribution is very different from the degree distribution. On the other hand, in real networks, the density and degree distributions are measurably similar \cite{1}.

Based on this observation, that real networks have similar density and degree decompositions and that many synthetic networks have dissimilar density and degree decompositions, we develop an abstract model, that, given a particular density distribution, produces a network having that density distribution. Applied naively, given a density distribution of a real network, this model generates networks with realistic average path lengths (average number of hops between pairs of vertices) and degree distributions; that is similar to the given real network. In addition to having short average path lengths, large-scale, real networks also tend to have high clustering coefficients \cite{3}. The clustering coefficient of a vertex \(v\) is the ratio of the number of pairs of neighbors of \(v\) that are connected to the number of pairs of neighbors of \(v\); the clustering coefficient of a network is the average clustering coefficient of its vertices. Our model, naively applied, unfortunately, but not surprisingly, results in networks with very low clustering coefficients. However, we show that applying the abstract model in a more sophisticated manner, using ideas from the small world model of Watts and Strogatz \cite{5}, results in much higher clustering coefficients suggesting that real networks may indeed be hierarchies of small worlds. Our hierarchies of small worlds specification is just one way to tune our abstract model; our model is quite flexible allowing for the easy incorporation of other network generation techniques.

A key observation that distinguishes our model from other network models is our qualitatively different treatment of vertices. That is, our model begins by assigning vertices to levels of the density decomposition. This sets vertices qualitatively apart from each other; for example, a vertex assigned to a dense level of the decomposition is treated very differently from a vertex assigned to a sparse level of the decomposition.
2 Results

The density decomposition partitions a graph into rings, $R_0, R_1, \ldots, R_k$. These rings divide the graph into regions of uniform density in the following sense: The graph obtained by identifying the vertices in $\bigcup_{j \geq i} R_j$ and deleting the vertices in $\bigcup_{j < i} R_j$ will always have density between $i - 1$ and $k$ [1]. The density decomposition is summarized by the density distribution $\rho = (|R_0|, |R_1|, \ldots, |R_k|)$, the number of vertices in each ring.

Given a density distribution $\rho$, we can generate a network with $n$ vertices having this density distribution using the following abstract model:

**Input:** density distribution $\rho$ and target size $n$

**Output:** an network $G$ with $n$ vertices and density distribution $\rho$

1. Initialize $G$ to be a network with empty vertex set $V$
2. for $i = |\rho|, \ldots, 0$ do
3. $R_i \leftarrow$ set of $\rho_i$ vertices
4. add $R_i$ to $V$
5. for each vertex $v \in R_i$ do
6. connect $i$ vertices of $V$ to $v$

Using this generic model, we propose two specific models, the random density distribution model (RDD) and the hierarchical small worlds model (HSW), by specifying how the neighbors are selected in Step 6 of the abstract model.

For the RDD model, we choose $i$ vertices from $V$ uniformly at random in Step 6. We use this to model four varied networks: AS, DBLP, EMAIL, and TRUST (Figure 1). For each given network, we generate another random network having the given network’s number of vertices and density distribution. Remarkably, although we are only specifying the distribution of the vertices over a density decomposition, the resulting degree distributions of the RDD networks are very similar to the original networks they are modeling. Further, the average path lengths of the RDD networks are realistic, within 2 of the average path lengths of the original networks However, the clustering coefficients of the RDD networks are unrealistically low (Figure 1).

We provide a more sophisticated model which addresses the unrealistically low clustering coefficients of the RDD model by generating a small world (SW) [5] network among the vertices of each ring of the density decomposition. In the hierarchical small worlds (HSW) model, for vertices in $R_i$, we create a small world network on $|R_i|$ vertices: Order $R_i$ cyclically; for each $v \in R_i$, with probability $p$, connect each of the $i$ vertices before $v$ in this order to $v$; if $c \leq i$ neighbors for $v$ are selected in this way, select $i - c$ vertices uniformly at random from $\bigcup_{j > i} R_j$ (or $R_i$ if this is the densest ring) and connect these to $v$. Clearly, this is a specification of neighbor selection for Step 6 of the abstract model. As with the SW model, the HSW model provides a similar trade-off between clustering coefficient and average path length, although the relationship is less strong. In addition, we observe a similar trade-off between $p$ and degree distribution: as $p$ increases, the degree distribution approaches that of the original network. This is in sharp contrast to the SW model which have degree distributions far from the original (normal vs. close to power law).
Fig. 1. We model four varied networks: In the AS network vertices represent autonomous systems and two autonomous systems are connected if there is a routing agreement between them [7]. In the DBLP network, vertices represent computer scientists and two computer scientists are connected if they have at least one co-authored paper [6]. In the EMAIL network vertices represent Enron email addresses and two addresses are connected if there has been at least one email exchanged between them [2]. In the TRUST network vertices represent epinions.com members and two members are connected if one trusts the other [4]. We give the clustering coefficient versus average path length for RDD, HSW, SW (Small Worlds) and DS (Degree Sequence) models. For HSW we let \( p = 0.1, 0.2, \ldots, 0.9 \). Colors indicate the network being modelled. Squares denote the data for the original networks.

References

A Bayesian approach to reference Exponential Random Graph Models (ERGMs) for comparing networks of different size and composition: Models and empirical tests

Alberto Caimo¹, Alessandro Lomi², and Francesca Pallotti³

¹ Dublin Institute of Technology, Ireland, alberto.caimo@dit.ie
² University of Lugano, Switzerland, alessandro.lomi@usi.ch
³ University of Greenwich, United Kingdom, f.pallotti@greenwich.ac.uk

Abstract

Exponential-family Random Graph models (ERGMs) are generally considered as the most promising statistical framework for the analysis of networks observed at a single time point [13]. During the last decade, the increased generality and flexibility of model specifications for ERGMs have made this class of statistical models relevant to an increasingly wide variety of empirical problems and theoretical issues [1, 9, 12]. Considerable progress has been made in extending ERGMs to larger and more complex samples of network data [2, 5]. However, the implications of differences in network size and composition have received comparatively little attention. In consequence, comparing models estimated on networks potentially coming from different populations remains an open challenge [10].

Uncertainty about how ERGM parameters estimated on networks of different size, with different node sets, and with different structural features may be compared and assessed is making the accumulation of empirical results based on these models problematic, and is limiting the diffusion of ERGMS in empirical research. The objective of this paper is to provide a modeling framework that may help to address this problem. In essence, our proposal involves the development of a Bayesian approach to the definition of a reference model for ERGMs.

In the standard ERGM setting, parameters for terms added to a uniform reference Bernoulli graph model can be interpreted as effects biasing the probability distribution of network graphs relative to a uniform distribution where edges between any dyad arise with constant probability equal to 0.5. However, in most social networks, changes in population size have a limited effect on mean degree, but a considerable effect on density. The recently proposed Krivitsky reference ERGM approach is based on the assumption that, for a set of otherwise comparable networks, the mean degree will be roughly constant in the size of the node set [10, 4]. Krivitsky reference ERGMs are homogeneous Bernoulli graph models where edges between nodes arise independently with a fixed probability
that is not necessarily equal to 0.5. This reference modelling framework is obtained from standard ERGM specification by adding a single edge term, whose associated parameter is the logit of the expected density.

We develop a Bayesian approach [6] for the Krivitsky reference ERGMs in order to compare the posterior parameter estimates of the same ERGM effects on networks of different sizes. The growing interest in Bayesian techniques in network analysis may be attributed to the development of efficient computational tools and the availability of fast user-friendly software [7]. Bayesian analysis is a promising approach to statistical network analysis because it yields a rich picture of the uncertain quantities, which is essential when dealing with complex and heterogeneous relational data. Adopting a Bayesian approach allows direct inclusion of prior information about the network effects into the modelling framework, and provides fully probabilistic quantification and interpretation of uncertainty through the evaluation of the posterior distribution of the parameters associated with the network effects, given the observed data.

We illustrate the empirical value of our methodological proposal in a comparative empirical analysis of data that we collected on two communities of health care organizations (with \( N = 110 \) and \( N = 35 \), respectively) connected by collaborative patient referral relations [8]. We discuss how the Bayesian approach we propose to the empirical specification and estimation of ERGM parameters facilitates the comparison between samples of network data of different size and with different structural and individual characteristics. We frame the implications of our illustrative case study in the context of classic problems in the analysis of social networks [3, 11] involving the discovery of structural elements that are common across different networks, and hence across different empirical settings.

References

A geometrically equivalent transformation between time series and complex networks

Xiaoyi Peng and Yi Zhao

Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China
zhao.yi@hit.edu.cn

1 Introduction

Transformations between time series and complex networks have received great attention in recent years. The majority of the existing methods to convert time series to complex networks can be classified into three categories: proximity networks, visibility graphs and transition networks [1]. Proximity networks are constructed by mapping a time series into a set of nodes and assigning links between nodes according to some measure of similarity. Visibility graphs are established by mapping each point in the time series into a node and linking the nodes by their convexity of observations. Transition networks are created by mapping each subspace that corresponds to a partition of the time series into a node and connecting the nodes by the temporal succession. The methods to transform complex networks to time series are relatively rare, typically, random walk [2] and spectral factorization [3]. But there are very few attempts to build the equivalence of the transformations between the complex network domain and the time series realm. We consider the amplitude difference method which could be viewed as one type of proximity methods to transform time series into complex networks. The results show that the transformation is quasi-isometric, which is consistent with the mathematical evidence that we provided.

2 Results

The transformation method from time series to complex network is very simple. For a given time series of length $L$ denoted as $\{x_t, t = 1, 2, \cdots, L\}$, the network’s adjacency matrix $A$ is defined by the property that the elements of $A$ satisfy $a_{ij} = 1$ if $|x_i - x_j| < \epsilon$, and $a_{ij} = 0$, otherwise, where each data point is regarded as a network node and $\epsilon$ is the transformation threshold. According to this principle, we can transform a time series into a complex networks. To show that the amplitude difference method is quasi-isometric when the transformed complex network is connected, we apply the classical multidimensional scaling (CMDS) [4] to the shortest path matrix which can be obtained from the adjacency matrix $A$. The first eigenvector of the CMDS method, i.e. the eigenvector corresponds to the biggest eigenvalue, can represent the original time series from which the network adjacency matrix is obtained. The original time series and the one reproduced from the CMDS method are quite similar.

The mathematical definition of quasi-isometric mapping is as follows. Let $M$ and $N$ be two closed subspaces of $\mathbb{R}^n$ for some positive integer $n < \infty$. Fix positive real
values \(\varepsilon_1\) and \(\varepsilon_2\). Let \(\phi\) be a surjection from the subspace \(M\) to the subspace \(N\) with the property that two points are \(\varepsilon_1\)-close in \(M\), i.e., \(d_M(x, y) < \varepsilon_1\), if and only if their images are \(\varepsilon_2\)-close in \(N\), i.e., \(d_N(\phi(x), \phi(y)) < \varepsilon_2\). The property also can be rewritten as \(\varepsilon_2 d_M(x, y) - \varepsilon_2 < d_N(\phi(x), \phi(y)) < \varepsilon_2 d_M(x, y) + \varepsilon_2\). Here \(d_M\) and \(d_N\) are some metric on space \(M\) and \(N\), respectively [5].

We assume that the subspace of the univariate time series \(M\) satisfies a loose separation condition, which is related to \(\varepsilon_1\). For each pair of elements \(x\) and \(y\) in \(M\), if \(d_M(x, y) \geq \varepsilon_1\), there exists an element, between \(x\) and \(y\), \(z \in M\) such that \(d_M(x, z) > \varepsilon_1\) and \(d_M(y, z) < \varepsilon_1\). According to the separation condition, for any pair of elements with distance bigger than \(\varepsilon_1\), there is a sequence of elements, in which the distance between any two successive elements is less than \(\varepsilon_1\). By the amplitude difference method, the time series is converted into a complex network whose number of nodes is equal to the length of the time series, which ensures the transformation is a one to one mapping, i.e., a injective and surjective mapping. We define the mapping as \(\phi\), the metric in the time series space \(d_M\) as the Euclidean metric, the metric in the complex network space \(d_N\) as the shortest path length.

Now we prove that \(\phi\) is a quasi-isometry. For two arbitrary points \(x\) and \(y\) in the space \(M\), there is an integer \(k\) such that

\[
ke_1 < d_M(x, y) < (k + 1)e_1. \tag{1}
\]

Because of the time series satisfies the loose separation condition that ensures the complex network connected, we can find a path \(P\) with length \(k + 1\) in \(N\) connecting the nodes \(\phi(x)\) and \(\phi(y)\) which corresponds to \(x\) and \(y\), respectively, i.e. \(d_N(\phi(x), \phi(y)) = k + 1\). According to equation 1, \(ke_1 < \frac{e_1}{2}d_M(x, y) < (k + 1)e_1\), choosing \(e_2\) in \((1, 1 + \frac{1}{k + 1})\), we have \(\frac{e_2}{e_1}d_M(x, y) - e_2 < \frac{e_1}{2}d_M(x, y) < \frac{e_1}{2}d_M(x, y) + e_2 > (k + 1)e_2 > k + 1 = d_N(\phi(x), \phi(y))\). Hence the transformation \(\phi\) is a quasi-isometric mapping, i.e. the time series and the complex network are quasi-isometric.

The transformation from complex networks to time series is a little more complex than its contrary process. The idea of the CMDS method is spectral factorization. Denote the shortest path matrix as \(D = \{d_{ij}\}_{j = 1}^{N}\), then the square-distance matrix \(S = \{d_{ij}^2\}_{j = 1}^{N}\) can be transformed to a centralizing gram matrix [4] \(G^* = -\frac{1}{2}HSH\), where \(H = I - \frac{1}{N}E\), \(I\) is the \(N \times N\) identity matrix and \(E\) is the \(N \times N\) matrix of ones. We have the eigenvector that corresponds to the biggest eigenvalue by applying spectral decomposition to \(G^*\). The reproduced time series, i.e. the first eigenvector, and the original time series are almost the same after normalization, which means the two time series are quasi-isometric.

To show that the time series and the complex network are quasi-isometric, the whole transformation process is shown in figure 1 [5], in which the original time series and the reproduced time series by the CMDS method are almost the same. Based on the quasi-isometric transformation, the underlying geometric features are preserved. As a result, the measurements related to geometric properties, such as correlation sum and correlation dimension, are conserved (see details in ref [5]).

There are two different angles to apply the amplitude difference method, which corresponds to given threshold and given time series, respectively. For a given threshold, any time series that satisfies the loose separation condition can be analyzed by the
The $x$-component data of a Rössler system (a) in chaotic regime ($\dot{x} = (y + z), \dot{y} = x + 0.2y, \dot{z} = 0.2 + z(x^{5.7})$) with the iterative step size 0.2, is transformed into a network (b) provided that $\epsilon$ is small enough to ensure that the constructed network is a connected graph without isolated nodes or sub-graphs. The time series (c) is then reproduced from the previous network.

amplitude difference method. In the other case, we can always choose a proper threshold to analyze, which ensures the connectivity of the transformed network. The CMDS method is based on similarity matrices. Here we choose the unweighted shortest path length to describe the similarity between each pair of nodes. The weighted shortest path also works, or even other measurements which represent the similarities properly also could be used to reproduce time series. Based on the quasi-isometric transformation between time series and complex networks, all the existing analysis methods in both realms can be applied to analyze different properties.

Summary. The amplitude difference method is a quasi-isometric transformation from time series to complex networks. The method to transform the complex network to the original time series is the classical multidimensional scaling. This method ensures the geometrical equivalence between the time series and the complex network, which also provides fundamental theory for the transformation between the two domains.

References

A Monte Carlo method for Large Deviations applied to Erdős-Rényi random graphs

Cristian Giardinà1, Claudio Giberti2, and Elena Magnanini3

1 University of Modena and Reggio Emilia, via G. Campi 213/b, 41125 Modena, Italy. cristian.giardina@unimore.it
2 University of Modena and Reggio Emilia, Via G. Amendola 2, 42122 Reggio Emilia, Italy. claudio.giberti@unimore.it
3 University of Modena and Reggio Emilia, via G. Campi 213/b, 41125 Modena, Italy. mgnlnel@unife.it

1 Introduction

We are interested in applying a population dynamics algorithm introduced in [1] and [2] for investigating large deviations, to the dense, Erdős-Rényi model. The method, which is called Cloning algorithm, is a Monte Carlo scheme which, in its original setup, aims at approximating the scaled cumulant generating function for an additive observable \( F_T := \sum_{t=1}^{T-1} f(X_t, X_{t+1}) \) where \( \{X_t\}_{t \geq 1} \) is a discrete Markov chain on a finite space \( S \) and \( f \) is a real valued function. The method yields \( \mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \) with \( \alpha \in \mathbb{R} \).

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]

When it is possible to apply the Gärtner-Ellis theorem ([3], chapter 5), one can recover the large deviations rate function of \( F \), which is a real valued function. The method yields

\[
\mu(\alpha) = \lim_{T \to +\infty} \frac{1}{T} \log \mathbb{E}[e^\alpha F_T] \quad \text{with} \quad \alpha \in \mathbb{R}.
\]
where $\mathbb{E}_p^{ER}(X^{(n)})$ denotes the expectation with respect to the Erdös-Rényi measure. It has been proved (see [5] and [4] for a further analysis on negative $\alpha$) that limit (1) has the following expression: $\mu^{Tr}(\alpha) = \frac{2}{3} + I_p(u^*)$, where $I_p(x) := p \log \frac{x}{p} + (1 - x) \log \frac{1 - x}{1 - p}$ and $u^*(\alpha)$ solves $\frac{e^{u^2 - h_p}}{e^{u^2 - h_p + 1}} = u$, with $h_p := \log \frac{p}{1 - p}$.

2 Results

The mathematical formula (1) is implemented by the Cloning algorithm via a population dynamics approach. We start with $M_0$ adjacency matrices which represent a family of initial systems (called clones) which evolves according to a tilted Markovian process followed by a branching process. To avoid the explosion or extinction of the family, a resampling step is also added. All the matrices are symmetric and have no self loops: at the starting iteration they have size two ($X^{(2)}$) and the two vertices are connected with probability $p$. At each instant of time, from the algorithmic point of view, the method follows three steps: the evolution, the cloning, and the uniform resampling:

1. each individual evolves according to the transition probability $U'X^{(\xi)}, X^{(\xi+1)}$;
2. each individual reproduces itself according to an average rate $K(X^{(\xi)}$) (if the rate is null, the clone is killed, otherwise, it leaves offspring). Each member reproduces in an independent way with respect to the others;
3. a number of individuals equal to the starting one, $M_0$, is uniformly chosen from those survived after the reproduction phase, then the cycle restarts.

At the end of the cycle (so at size $n$), $\mu(\alpha)$ is approximated by relation $\mu_n(\alpha) = \frac{1}{(\frac{1}{2})} \log \frac{M_n}{M_0}$, where $M_n$ is the size of the population at iteration $n$. Figure 1 shows that

![Fig. 1. Simulations for triangles observable: graphic of the scaled cumulant generating function for different values of $p$. The star line represents the output of the method for different $n$, up to size 15, whereas the continuous line represents the analytic result. As the size grows, the star line approaches the limit curve. For $n = 3$, an exact computation of the scaled cumulant generating function is possible and we use it as a comparison term.](image)
the method converges to the expected curve despite the small size of graphs (note however the scale $n^2$ in (1)).

\[ \mu_{Edg}(\alpha) = \log(p e^\alpha + 1 - p) \quad \forall \alpha \in \mathbb{R}. \]

The algorithm can be also applied using as observable the number of edges, $E(X^n)$. In this case a direct computation of the scaled cumulant generating function is possible, since $\{X_{ij}^{(n)}\}_{ij}$ are i.i.d Bernoulli random variables and yields $\mu_{Edg}(\alpha) = \log(p e^\alpha + 1 - p) \quad \forall \alpha \in \mathbb{R}$. The algorithm perfectly matches the expected curve, as shown in Figure 2.

**Summary.** We used the framework of random networks as a setting to develop an extension of a known Monte Carlo scheme for reproducing the scaled cumulant generating function of two observables, edges and triangles. The algorithm has been tested giving an evidence of the convergence to the expected curves. Future work will investigate the scaled cumulant generating function in the *Replica Breaking phase*: in this region, characterized by $\alpha < 0$, the analytical expression of the function is not known and the graph stops behaving like an Erdős-Rényi exhibiting particular structures which the algorithm can potentially highlight.

**References**

Network construction: A learning framework through localizing principal eigenvector

Priodyuti Pradhan$^1$ and Sarika Jalan$^{1,2}$

1 Complex Systems Lab, Discipline of Physics, Indian Institute of Technology Indore, Khandwa Road, Simrol, Madhya Pradesh 453552, India, priodyutipradhan@gmail.com, sarikajalan9@gmail.com
WWW home page: http://iiti.ac.in/people/~sarika/

2 Discipline of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Khandwa Road, Simrol, Madhya Pradesh 453552, India,

1 Introduction

Networks furnish a mathematical framework to model and decipher the collective behavior of the complex real-world systems [1]. Scrutiny of principal eigenvector (PEV) and the corresponding eigenvalue of the networks are known to provide an understanding of various structures as well as the dynamical properties of the networks [2]. Recently, scrutiny of eigenvector localization has received tremendous attention in network science due to its versatile applicability in many different areas which includes analyzing eigenvector centrality, spectral partitioning, development of approximation algorithms, epidemic spreading phenomenon and in machine learning. One key factor of our interest is to understand the properties of networks which may help in spreading or restricting perturbation in networks captured by the PEV localization. For a network, PEV, $x_1 = ((x_1)_1, (x_1)_2, \ldots, (x_1)_n)^T$, is said to be localized when most of its components are near to zero, with few entries taking very high values [3]. We measure the eigenvector localization using the inverse participation ratio and denoted as

$$Y_{x_1} = \sum_{i=1}^{n} (x_1)_i^4$$

where $(x_1)_i$ is the $i^{th}$ component of $x_1$ in the Euclidean norm. Network properties such as the presence of hubs, the existence of dense subgraph, and a power-law degree distribution are few factors known to make a PEV more localized than the corresponding random network [3]. However, we show that the presence of these features in a network does not guarantee a highly localized PEV of the corresponding adjacency matrix [4]. In other words, if we construct a network which possesses one or more of these structural features, the PEV of the underlying matrix may be more localized than the corresponding random network, but may not be the most localized one. Here, we ask a reverse question: for a given network size and some connections, if we construct a network which has the most localized PEV, what particular structural and spectral properties the network will possess? Instead of analyzing the properties of a network, we construct a network structure by optimizing some specific behavior of the PEV. We perform network evolution based on optimized edge rewiring to obtain the network structure having highly localized PEV [4].
2 Results

We reveal that PEV localization is not a consequence of a single network property and preferably requires a collective impact of several structural features [4]. The final optimized network possesses a special structure independent of the initial network. It contains two graph components of different sizes which are connected via a single node (in Fig. 1). Moreover, the optimized structure contains a special set of edges, rewiring any one of them leads to a complete delocalization of the PEV from a highly localized state (Fig. 2(a)). This sensitivity of the PEV at the most localized state turns out to be related to the behavior of the largest ($\lambda_1$) and the second largest ($\lambda_2$) eigenvalue of the network (Fig. 2(b)). Precisely when the network becomes most localized, the second largest eigenvalue of the adjacency matrix become very close to the largest eigenvalue (Fig. 2(b) and (c)). Further, to illustrate dynamics on the artificially constructed network structures having highly localized PEV, we use the standard susceptible-infected-susceptible (SIS) disease spreading model. In the SIS model, each susceptible vertex becomes infected with the infection rate of $\gamma$, and infected vertices become susceptible to the unit rate. With probability $\rho_i$, a vertex $i$ infected by its neighbours, and the prevalence is given by $\rho = \sum_{i=1}^{N} \rho_i / N$. We know that when the infection rate $\gamma$ cross the epidemic threshold, i.e., $\gamma > \gamma_c$ (where $\gamma_c = \frac{1}{\lambda_1}$), the disease will spread over the networks. However, if the PEV of the adjacency matrix is localized, in the vicinity of the epidemic threshold $\gamma_c + \epsilon$, $\epsilon > 0$ the disease infects a small number of vertices and spreading process becomes slow. As a result, it requires a larger value of $\gamma$ to spread the disease over the network. Fig. 3 manifests that for the initial random network, disease

![Fig. 1. Portrait the schematic diagram of the initial and the most optimized networks.](image)

![Fig. 2. (a) IPR as a function of edge-rewiring ($\tau_{\text{evolution}}$). The networks with large IPR value in the saturation region consists of few edge-rewiring, which leads to a sudden drop in the IPR value. (b) Behavior of $\lambda_1$ (•) and $\lambda_2$ (■) and (c) the gape between $\lambda_1$ and $\lambda_2$ during the network evolution.](image)
infects a large number of vertices for a given $\gamma$ which is slightly larger than $\gamma_c$. Whereas, for the networks corresponding to the intermediate and the optimized states, there exist very few vertices which get infected.

Further, we explore the impact of the optimized edge rewiring for the PEV localization in multilayer networks (MNs). We show that by rewiring a single-layer, one can regulate the contribution of the node weights of the other layer to the PEV of the entire MNs [5]. We analyze several structural and spectral properties during the network evolution process for the single-layer as well as both-layer rewiring protocols of the MNs. For both the protocols, we find that there is an emergence of various structural features as PEV gets localized. Moreover, for both the protocols, there is a noticeable difference present in the spectral properties in the saturation region. For both-layers rewiring protocol, in the saturation region, PEV is sensitive to a single edge rewiring as also observed for the optimized evolution of the monolayer networks. However, interestingly, we get rid of the sensitivity in the PEV in the saturation region by implementing a single-layer rewiring of the MN [5].

Acknowledgement. SJ acknowledges DST, Govt of India grant (EMR/2016/001921) and DAE, Govt of India grant (37(3)/14/11/2018-BRNS/37131) for financial support. PP thanks to members of CSL at IIT Indore for several discussions and fruitful suggestions.

References

Local-community network automata modelling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more

Alessandro Muscoloni¹, Ilyes Abdelhamid¹,² and Carlo Vittorio Cannistraci¹,³

¹ Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany; ² Université Paris VII, 5 rue Thomas Mann, 75013, Paris, France; ³ Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi “Bonino Pulejo”, Messina, Italy

Abstract. From nests to nets intricate wiring diagrams surround the birth and the death of life. Here we show that the same rule of complex network self-organization is valid across different physical scales and allows to predict protein interactions, food web trophic relations and world trade network transitions. This rule, which we named CH2-L3, is a network automaton that is based on paths of length-three and that maximizes internal links in local communities and minimizes external ones, according to a mechanistic model essentially driven by topological neighborhood information. More in details, in this study we provide the theoretical formulation for generalizing two mechanistic models for link prediction, Resource Allocation (RA) and Cannistraci-Hebb network automata (CH) (Fig. 1), by considering paths of arbitrary length. Extensive tests adopting paths of length two (L2) and three (L3) show that L3 methods seem to provide a significant improvement with respect to L2 methods for prediction of protein interactions, however different evaluation frameworks are in disagreement, therefore further investigations are needed. On the contrary, on food webs it seems clear that L3 is a key rule of self-organization to reproduce the network topology. The same can be confirmed for world trade networks, although the improvement is of smaller entity than on food webs, because the networks are very dense and therefore all the link predictors offer by default a good performance. Furthermore, the most important finding is that the local community paradigm theory and the derived CH network automata provide a significant modelling and performance improvement on RA network automata both in L2 and L3, when applied to diverse types of complex networks (Fig. 2).

References

Fig. 1. The figure shows an explanatory example for the topological link prediction performed using the L2 or L3 Cannistraci-Hebb epitopological rationale. The two black nodes represent the seed nodes whose non-observed interaction should be scored with a likelihood. The white nodes are the L2 or L3 common-neighbors (CNs) of the seed nodes, further neighbors are not shown for simplicity. The cohort of common-neighbours and the iLCL form the local community. The different types of links are reported with different colors: non-LCL (green), external-LCL (red), internal-LCL (white). The set of L2 and L3 paths related to the given examples of local communities are shown. At the bottom, the mathematical description of the L2 and L3 methods considered in this study are reported. Notation: $u, v$ are the seed nodes; $z$ is the intermediate node in the L2 paths; $s, t$ are the intermediate nodes in the L3 paths; $d_z$ is the degree of $z$; $d_{iz}$ is the internal degree (number of iLCL) of $z$; $d_{ez}$ is the external degree (number of eLCL) of $z$. For further details please see the Reference.

<table>
<thead>
<tr>
<th>L2</th>
<th>L3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$RA_{L2}(u, v) = \sum_{z \in L2} \frac{1}{d_z}$</td>
<td>$RA_{L3}(u, v) = \sum_{s, t \in L3} \frac{1}{\sqrt{d_s \cdot d_t}}$</td>
</tr>
<tr>
<td>$CH1_{L2}(u, v) = \sum_{z \in L2} \frac{dl_z}{dz}$</td>
<td>$CH1_{L3}(u, v) = \sum_{s, t \in L3} \sqrt{dl_z \cdot dl_t}$</td>
</tr>
<tr>
<td>$CH2_{L2}(u, v) = \sum_{z \in L2} \left( \frac{dl_z}{de_z} \cdot \frac{1 + dl_z}{1 + de_z} \right)$</td>
<td>$CH2_{L3}(u, v) = \sum_{s, t \in L3} \sqrt{dl_z \cdot dl_t} \cdot \sqrt{de_z \cdot de_t}$</td>
</tr>
</tbody>
</table>
Fig. 2. (A) Synthetic networks have been generated using the nonuniform PSO model with parameters $N = 1000$, $m = \{6, 10, 14\}$, $T = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\gamma = 3$ and $C = 5$. For each combination of parameters, 100 networks have been generated. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. In order to evaluate the performance, the links are ranked by likelihood scores and the area under precision curve (AUP) is computed for the top-$r$ removed links in the ranking, where $r$ is the total number of links removed. The plots report for each parameter combination the mean AUP and standard error over the random repetitions. (B) The barplots report the mean AUP and standard error computed over all the networks and repetitions for each dataset of real networks: protein-protein interactions (PPI), food webs, world trade networks and other mixed real networks. Note that for the PPI networks the AUP is computed for the top-100 removed links, whereas for the other datasets for the top-$r$ removed links in the ranking, where $r$ is the total number of links removed. The percentage of improvement of each CH2-method with respect to the corresponding RA-method is reported. For further details please see the Reference.
Minimum curvilinear automata with similarity attachment for network embedding and link prediction in the hyperbolic space

Alessandro Muscoloni\(^1\) and Carlo Vittorio Cannistraci\(^{1,2}\)

\(^1\) Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany; \(^2\) Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi “Bonino Pulejo”, Messina, Italy

Abstract. The idea of minimum curvilinearity (MC) is that the hidden geometry of complex networks, in particular when they are sufficiently sparse, clustered, small-world and heterogeneous, can be efficiently navigated using the minimum spanning tree (MST), which is a greedy navigator. The local topological information drives the global geometrical navigation and the MST can be interpreted as a growing path that greedily maximizes local similarity between the nodes attached at each step by globally minimizing their overall distances in the network. This is also valid in absence of the network structure and in presence of only the nodes geometrically located over the network generative manifold in a high-dimensional space. We know that random geometric graphs in the hyperbolic space are an adequate model for realistic complex networks: the explanation of this connection is that complex networks exhibit hierarchical, tree-like organization, and in turn the hyperbolic geometry is the geometry of trees. Here we show that, according to a mechanism that we define similarity attachment, the visited node sequence of a network automaton - which navigates the network with a growing MST in compliance with the minimum curvilinearity strategy - can efficiently approximate the nodes’ angular coordinates in the hyperbolic disk, that actually represent an ordering of their similarities. This is a consequence of the fact that the MST, during its greedy growing process, at each step sequentially attaches the node most similar (less distant) to its own cohort. Minimum curvilinear automata (MCA, Fig. 1) displays embedding accuracy which seems superior to HyperMap-CN and inferior to coalescent embedding (Fig. 2), however its link prediction performance on real networks is without precedent for methods based on the hyperbolic space. Finally, depending on the data structure used to build the MST, the MCA’s time complexity can also approach a linear dependence from the number of edges.

References

Fig. 1. (A) Visual representation of the mechanism that we define similarity attachment: the visited node sequence of the growing MST using the Prim’s algorithm on a pre-weighted network represents an ordering of their similarities. (B) The algorithmic steps and the intermediate input/output of the MCA embedding algorithm are illustrated, reporting the possible variants. The example network has been generated by the Popularity-Similarity-Optimization (PSO) model ($N = 50$, $m = 2$, $T = 0$, $\gamma = 2.5$) and the nodes are colored according to their angular coordinates in the original network. Description of the variables: $x_{ij}$ value of $(i,j)$ link in adjacency matrix $x$; $\epsilon_i$ external degree of node $i$ (links neither to $CN_{ij}$ nor to $j$); $CN_{ij}$ common neighbours of nodes $i$ and $j$; $N$ number of nodes; $\zeta = \sqrt{-K}$, we set $\zeta = 1$; $K$ curvature of the hyperbolic space; $\beta = \frac{1}{\gamma-1}$ popularity fading parameter; $\gamma$ exponent of power-law degree distribution. For further details please see the Reference.
Fig. 2. (A) The three subpanels show the original network generated with the nonuniform Popularity-Similarity-Optimization (nPSO) model ($N = 100$, $m = 2$, $T = 0$, $\gamma = 2.5$, $C = 8$), the similarities of the nodes inferred using the RA1-MCA1-RAA algorithm and the corresponding embedding of the network in the hyperbolic disk. In all the panels the nodes are colored according to their angular coordinates in the original network. (B) Synthetic networks have been generated using the nPSO model ($N = [100-500]$, $m = 4$, $T = [0, 0.3, 0.6, 0.9]$, $\gamma = 2.5$, $C = 8$, realizations = 100). For each network the hyperbolic embedding methods have been executed and the performance has been evaluated using the HD-correlation (Pearson correlation between all the pairwise hyperbolic distances of the nodes in the original and inferred network) and GR-score (greedy routing score). The plots report for each parameter combination the mean and standard error over the 100 iterations. The methods compared are one MCA variant (RA2-MCA1-RAA), two main coalescent embedding techniques (RA2-ncISO-EA and RA2-ncMCE-EA) and HyperMap-CN. The GR-score evaluated on the original coordinates of the network is also reported. For further details please see the Reference.
Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction

Alessandro Muscoloni¹ and Carlo Vittorio Cannistraci¹,²

¹ Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany; ² Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi “Bonino Pulejo”, Messina, Italy

Abstract. Advances in network geometry pointed out that structural properties observed in networks derived from real complex systems can emerge in the hyperbolic space (HS). The nonuniform PSO (nPSO) is a generative model recently introduced in order to grow random geometric graphs in the HS, reproducing networks that have realistic features such as high clustering, small-worldness, scale-freeness and rich-clubness, with the additional possibility to control the community organization. Generative models allowing to tune the structural properties of ‘realistic’ synthetic networks are fundamental, because they offer a ground-truth to investigate how predictive algorithms react to controlled topological variations. Here, we discuss how to leverage the nPSO model as a synthetic benchmark to compare the performance of methods for community detection (Fig. 1) and link prediction (Fig. 2); and we prove that the nPSO offers a reliable and realistic testing framework which can complement other existing benchmarks not based on latent geometry. Furthermore, we confirm that network embedding information can improve community detection, whereas boosting link prediction in HS still needs further investigations. Indeed, we find that the presence of communities in nPSO significantly modifies the performance of link predictors (compare Fig. 2B and 2C respectively) and is fundamental for the reproducibility of results observed on real networks (Fig. 2A). The nPSO can trigger valuable insights to understand the intrinsic rules of link-growth and self-organization that connect topology to geometry and that are encoded in link-prediction algorithms differentiating their performance.

References

Synthetic networks have been generated using the nPSO model with parameters $N = [100, 500, 1000]$, $m = [10, 12, 14]$, $T = [0.1, 0.3, 0.5]$, $\gamma = 3$ and $C = 4$. For each combination of parameters, 10 networks have been generated. For each network, the community detection methods Louvain, Infomap, Walktrap, and Label propagation have been executed and the communities detected have been compared to the annotated ones computing the Normalized Mutual Information (NMI). The plots report for each parameter combination the mean NMI and standard error over the random repetitions.

The results indicate that overall Louvain appears as the strongest approach, with an almost perfect detection over different values of network size, average node degree, and temperature. Infomap highlights problems in correctly detecting the communities when there are too many inter-community links, as can be seen for $N = 100$ and increasing temperature. The higher temperature in fact leads to a higher number of links between nodes that are geometrically far in the disk, which increases the mixing between the communities. The performance is more stable for bigger networks, although in general slightly worse than Louvain. Walktrap results as robust as Louvain to the increase of network temperature, but the NMI is slightly lower for $N = 100$ and $N = 1000$. As last, Label propagation, which is the fastest approach, but the one with lowest accuracy, performs worse than the other methods and presents the same problem as Infomap for $N = 100$. The results are in agreement with previous studies, hence the nPSO model here proposed seems to provide a good benchmark to test community detection algorithms on networks generated using a latent geometry model which is based on the hyperbolic space. For further details please see the Reference.
A set of 12 large-size real networks has been collected. Synthetic networks have been generated using the PSO and nPSO models with parameters $N = 1000$, $m = 12$, $T = [0.1, 0.3, 0.5]$, $\gamma = 3$ and, for the nPSO, $C = 8$. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks (100 repetitions for all the methods, only 10 repetitions for SBM due to the high computational time). In order to evaluate the performance, the links are ranked by likelihood scores and the precision is computed as the percentage of removed links among the top-$r$ in the ranking, where $r$ is the total number of links removed. The table report the mean precision for each network, as well as the mean precision and mean ranking over the whole dataset. The plots report for each parameter combination the mean precision and standard error over the random repetitions. We compared the performance of state-of-the-art approaches (CRA, RA, SPM, SBM). The results obtained on nPSO networks (panel C) better resemble – in respect to PSO (panel B) - the ones obtained on real networks (panel A) due to the introduction of the communities. For further details please see the Reference.

![Table](image1)

![Graphs](image2)

Fig. 2. A set of 12 large-size real networks has been collected. Synthetic networks have been generated using the PSO and nPSO models with parameters $N = 1000$, $m = 12$, $T = [0.1, 0.3, 0.5]$, $\gamma = 3$ and, for the nPSO, $C = 8$. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks (100 repetitions for all the methods, only 10 repetitions for SBM due to the high computational time). In order to evaluate the performance, the links are ranked by likelihood scores and the precision is computed as the percentage of removed links among the top-$r$ in the ranking, where $r$ is the total number of links removed. The table report the mean precision for each network, as well as the mean precision and mean ranking over the whole dataset. The plots report for each parameter combination the mean precision and standard error over the random repetitions. We compared the performance of state-of-the-art approaches (CRA, RA, SPM, SBM). The results obtained on nPSO networks (panel C) better resemble – in respect to PSO (panel B) - the ones obtained on real networks (panel A) due to the introduction of the communities. For further details please see the Reference.
The microstructure of the giant component in configuration model networks

Ofer Biham\(^1\), Eytan Katzav\(^1\), and Reimer Kühn\(^2\)

\(^1\) Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel,
\(^2\) Mathematics Department, King’s College London, Strand, London WC2R 2LS, UK

biham@phys.huji.ac.il

There is a broad range of phenomena in the natural sciences and engineering as well as in the economic and social sciences, which can be usefully described in terms of network models. One of the central lines of inquiry in the field has been percolation theory, where a major question is the existence, under suitable conditions, of a giant component, which occupies a finite, non-zero fraction of the graph in the thermodynamic limit of infinite system size. Critical parameters for the emergence of a giant component were identified for Erdős-Rényi (ER) networks and more generally for configuration model networks, i.e. networks that are maximally random subject to a given degree sequence, by Molloy and Reed [1]. These authors also established a duality relation between the degree-distribution, restricted to nodes which reside on finite components of a configuration model network, to the degree distribution of the whole network [1], thus generalizing a previous result for ER networks [2]. However, there are no results concerning degree distributions when restricted to the giant component of a random graph ensemble, which is the main aim of the current work [3]. In particular, we show that while configuration model networks are uncorrelated, their giant component exhibits a degree distribution \(P(k|1)\) which is different from the overall degree distribution \(P(k)\) of the network. Actually, it is disassortative and exhibits degree-degree correlations of all orders. We present exact analytical results for the degree distributions as well as higher-order degree-degree correlations on the giant components of configuration model networks. We argue that the degree-degree correlations are essential for the integrity of the giant component, in the sense that the degree distribution alone cannot guarantee that it will consist of a single connected component. In Fig. 1 we show \(P(k|1)\) and \(P(k)\) for ER network with \(c = 2\) and for a scale-free configuration model characterized by \(P(k) \sim k^{-\gamma}\) with \(\gamma = 3\). As can be seen, the analytical results agree with the simulations. We prove that near the percolation transition the assortativity coefficient, \(r\), of the giant cluster is negative, and is typically around \(r \approx -0.2\) there.

Clearly, the knowledge of degree distributions and degree-degree correlations restricted to the giant component of a network can be very useful when investigating dynamical processes on complex networks. It would help to obtain results pertaining only to the giant component of such systems, without the contributions from finite components which often amount to trivial contaminations or (unwanted) distortions of results. We analyze a few examples, such as the localization phenomena in sparse matrix spectra [4] (where finite components of a random graph support eigenvectors that are trivially localized), or the spread of diseases or cascading failures (where an initial failure or initial infection occurring on a finite component will never lead to a global system failure or the outbreak of an epidemic). Finally, the distribution of shortest path
lengths (DSPL) between pairs of nodes in a network [5–7] will contain contributions from pairs of nodes on different components whose distance is, by convention, infinite. To eliminate such unwanted contributions, numerical studies using message passing algorithms or straightforward simulation methods, are often performed directly on the largest component of a given random network. The results may be difficult to compare with theoretical results if the latter do not eliminate finite component contributions (or suppress them by taking the density of links to be sufficiently large to make them effectively negligible). In Fig. 2 we present application of our results to the important SIR processes of epidemic spreading. In particular, we study the probability, $\sigma$, of a random node to remain on the giant component that corresponds to the fraction of the individuals which have become infected. In Fig. 3 we show that the random shells approach (RSA) [6] designed to calculate the DSPL of ER networks is greatly improved by using $P(k|1) - a$ method which we now denote as gRSA.

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 2. Approximate evaluations of the heterogeneous infection probabilities $\pi_4(\sigma)$ (green dashed histograms) for an ER network with $c = 2$ (left) and a configuration model network with a scale-free degree distribution with $\gamma = 3$ (right). The approximation scheme takes the degree-degree correlations correctly up to the 4th shell. Also shown in each figure is the full distribution (black solid line) obtained using a population dynamics approach to solving the self-consistency equations arising in this problem. The agreement is remarkable given the complexity of these functions.

Fig. 3. Illustration of the shell structure around a reference node in a random network (left), as well as analytical and simulation results for the tail distribution of the shortest path lengths, $P(L > \ell)$, of an ER network of $N = 1000$ nodes with $c = 2$ (right). The RSA results (dotted line), are accurate for small distances but greatly under-estimate the tail distribution for large distances. The gRSA results (solid line), which take into account the degree distribution on the giant component, $P(k|1)$, are found to be in very good agreement with the simulation results (circles).
Reconstructing networks with unknown and heterogeneous errors [1]

Tiago P. Peixoto1,2

1 Department of Mathematical Sciences and Centre for Networks and Collective Behaviour, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom
2 ISI Foundation, Via Chisola 5, 10126 Torino, Italy

1 Summary

The vast majority of network datasets contains errors and omissions, although this is rarely incorporated in traditional network analysis. Recently, an increasing effort has been made to fill this methodological gap by developing network reconstruction approaches based on Bayesian inference. These approaches, however, rely on assumptions of uniform error rates and on direct estimations of the existence of each edge via repeated measurements, something that is currently unavailable for the majority of network data. Here we develop a Bayesian reconstruction approach that lifts these limitations by not only allowing for heterogeneous errors, but also for single edge measurements without direct error estimates. Our approach works by coupling the inference approach with structured generative network models, which enable the correlations between edges to be used as reliable error estimates. Although our approach is general, we focus on the stochastic block model as the basic generative process, from which efficient nonparametric inference can be performed, and yields a principled method to infer hierarchical community structure from noisy data. We demonstrate the efficacy of our approach with a variety of empirical and artificial networks.

2 Bayesian network reconstruction

The contemporary study of network systems is motivated in large part by the surging availability of network data during the past couple of decades, which describe the detailed interactions among constituents of large-scale complex systems, such as transportation networks, cell metabolism, social contacts, the internet, and various others. Despite the widespread growth of this field, its relative infancy is still noticeable in some aspects. In particular, even though sophisticated and successful models of network structure and function have been proposed, as well as powerful data analysis methods, most studies of empirical data are performed without taking into account measurement error. Most typically, real networks are represented as adjacency matrices, sometimes enriched with additional information such as edge weights and types, as well as various kinds of node properties, the validity of which is simply taken for granted. But as is true for any empirical scenario, network data is subject to observational errors: parts of the network might not have been recorded, and the parts that have might be wrong. Although this problem has been recognized in the past in several studies [2,3,4,5], the
practice of ignoring measurement error is still mainstream, and robust methods to take it into account are underdeveloped. This is in no small part due to the fact that most available network data contain no quantitative error assessment information of any kind, thus preventing primary experimental uncertainties to be propagated up the chain of analysis.

In this work we formulate a principled method to reconstruct networks that have been imperfectly measured. We do so by simultaneously formulating generative models of network structure — that incorporate degree heterogeneity, modules and hierarchies — as well as models of the noisy measurement process. By performing Bayesian statistical inference of this joint model, we are able to reconstruct the underlying network given an imperfect measurement affected by observational noise. Importantly, our method works also when a single measurement of the underlying network has been made, and the noise magnitudes are unknown. This means it can be directly applied to the majority of network data without available error estimates. In addition to this, our method is capable of extracting hierarchical modular structure from such noisy networks, thus generalizing the task of community detection to this uncertain setting.

Our method is equally applicable when information on measurement error is available, either as repeated measurements or as estimated edge probabilities. For this class of data, we construct a general model that allows for heterogeneous errors, that vary in
different parts of the network. We show strong empirical evidence for the existence of this kind of heterogeneity, and demonstrate the efficacy of our method to include it in the reconstruction.

In Fig. 1 we show the results of our reconstruction method for a network of association among 62 terrorists responsible for the 9/11 attacks [6], for which the method correctly predicts a missing association that was confirmed in further investigations.

Fig. 2 shows the result of the reconstruction method applied to the neural network of the *Caenorhabditis elegans* worm, measured via the manual and error-prone analysis of around 8000 cross-sectional images [7]. Our method predicts 428 missing edges with a high posterior probability. In particular, we note that our reconstruction connects the isolated nodes in the data to the main hub in the network, which is an important neuron situated in the head of the worm.

### References

Modified box dimension of trees and hierarchical scale-free graphs

Júlia Komjáthy1, Roland Molontay2, and Károly Simon2

1 Dept. of Mathematics and Comp. Sci., Eindhoven University of Technology, Netherlands
2 Dept. of Stochastics, Budapest University of Technology and Economics, Hungary

j.komjathy@tue.nl, molontay@math.bme.hu, simonk@math.bme.hu

1 Introduction

The study of complex networks has received immense attention recently, mainly because networks are used in several disciplines of science, such as in Information Technology (World Wide Web, Internet), Sociology (social relations), Biology (cellular networks) etc. Understanding the structure of such networks has become essential since the structure affects their performance, for example the topology of social networks influences the spread of information and disease. Many networks were claimed to show self-similarity and fractal behaviour, for a review see [2]. Heuristically, fractality and self-similarity of a network means that the network looks similar to itself on different scales: if one zooms in on a sub-network, one is expected to see the same qualitative behaviour as in the whole network.

Identifying fractality and the presence of self-similarity of complex networks beyond the heuristics was initiated by Song, Havlin and Makse [10]. They proposed that the procedure for networks must be similar to that of regular fractal objects: using the box-covering method. In this respect, fractality stands for a polynomial relation between the number of boxes needed to cover the network and the size of the covering boxes, defining a finite box dimension. The fractality of many real-life examples was verified in e.g. [11]. A number of other notions of fractal dimension have been investigated throughout the years, for a recent survey we refer to [7].

Sometimes the polynomial relation between the number of covering boxes and box size is too restrictive: many networks show exponential growth of local neighborhoods, e.g. the Internet at router level and most (online) social networks [2], and for these networks the notion of fractality above yields trivial (infinite) dimension. To compensate this effect, in this work we introduce a new concept: the modified box dimension of graph sequences. Our definition of modified fractal dimension takes into account the exponential growth, and with the modified definition, the fractal dimension becomes a proper parameter for networks with exponentially growing neighborhood sizes.

We test our definition on network models that contain hierarchical structure and thus intuitively they should be called ‘fractal networks’. Indeed, based on rigorous proofs, we show that our definition of modified box dimension yields nontrivial values for the hierarchical graph sequence model introduced by Komjáthy and Simon [4], and the Song-Havlin-Makse model [11]. The model in [4] is a generalisation of the model introduced by Barabási, Ravasz and Vicsek [1],[6], can be constructed by a method...
which is common in generating fractals, and can be derived from an actual (graph-directed) self-similar fractal [4].

In case the network is a tree, we find that the modified box dimension is related to the growth rate [5]. We investigate rigorously the modified (Cesaro) box-dimension of spherically symmetric trees, and supercritical Galton-Watson trees and relate it to the growth rate.

2 Notions of fractal scaling

To detect the presence of fractality in complex networks the following method was introduced by Song et al. [10]: For a given network \( G \) with \( N \) vertices, partition the vertices into subgraphs (boxes) with diameter at most \( \ell - 1 \), called \( \ell \)-boxes. The minimum number of \( \ell \)-boxes needed to cover the entire network \( G \) is denoted by \( N_B(\ell) \). Determining \( N_B(\ell) \) for any given \( \ell \geq 2 \) is an NP-hard problem but various algorithms have been proposed to obtain an approximate solution [9]. In accordance with regular fractals, Song et al. proposed to define the box dimension \( d_B \) of \( G \) by the approximate relationship: \( N_B(\ell)/N \approx \ell^{-d_B} \), i.e., the required number of boxes scales as a power of the box size, and the dimension is the absolute value of the exponent.

According to this method, the power form (with a finite \( d_B \)) can be verified in a number of real-world networks such as WWW, actor collaboration network and protein interaction networks [11]. A large class of complex networks (called non-fractal networks) is characterized by a sharp decay of \( N \) with increasing \( \ell \), i.e., has infinite fractal dimension, for example, the Internet at router level or most of the social networks [2] falls into this category.

Here we give a more rigorous mathematical definition of box dimension of graph sequences. Several real-world networks grow in size as time proceeds, therefore it is reasonable to consider sequences, denoted by \((G_n)_{n \in \mathbb{N}}\). Let \( N_B^n(\ell) \) denote the optimal number of \( \ell \)-boxes needed to cover \( G_n \) and \( |G_n| \) its number of vertices.

**Definition 1.** The box dimension \( d_B \) of a graph sequence \((G_n)_{n \in \mathbb{N}}\) is defined by (if the limit exists)

\[
d_B \left( (G_n)_{n \in \mathbb{N}} \right) := \lim_{\ell \to \infty, n \to \infty} \frac{\log (N_B^n(\ell)/|G_n|)}{-\log \ell}. \tag{1}
\]

It is not hard to see that \( N_B^n(\ell) \) does not scale as a power of \( \ell \) when the network has exponential growth rate of neighborhoods, and then Definition 1 yields infinite dimension. On the other hand, the box-covering method gives rise to another interesting parameter if we modify the required functional relationship between the minimal number of boxes and the box size (similarly to the transfinite fractal dimension by Rozenfeled et al. [8]). Namely, we might consider finding \( \tau \) that satisfies \( N_B(\ell)/N \approx e^{-\tau \ell} \). Hence, we define the modified box dimension of graph sequences similarly:

**Definition 2.** The modified box dimension \( \tau \) of a graph sequence \((G_n)_{n \in \mathbb{N}}\) is defined by (if the limit exists)

\[
\tau \left( (G_n)_{n \in \mathbb{N}} \right) := \lim_{\ell \to \infty, n \to \infty} \frac{\log (N_B^n(\ell)/|G_n|)}{-\ell}. \tag{2}
\]

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. The third iteration of an instance of the hierarchical graph sequence model, called "cherry" model: HM$_3$. The boxing of the graph is also highlighted: the green boxes illustrate an optimal 3-boxing and the dashed boxes show an optimal 7-boxing of the graph, i.e. $N_B(3) = 9$ and $N_B(7) = 3$. The modified box dimension of the model is $\tau(\{HM_n\}_{n \in \mathbb{N}}) = (\log K)/2$, here the base graph is on $K = 3$ vertices.

We call $\tau$ the modified box dimension or ‘growth-constant’ since it captures how spread-out neighborhoods of vertices are on an exponential scale. For some models with exponentially growing neighborhood sizes (e.g. for spherically symmetric trees) the limit in (2) does not exist but the limit of the Cesaro means does:

\begin{equation}
\tau^*((G_n)_{n \in \mathbb{N}}) := \lim_{\ell \to \infty} \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{N_B^{i+\ell}(\ell)}{|G_i+\ell|} \right) - \ell.
\end{equation}

### 3 Results

We determine the optimal box covering and modified box dimension of various models: the hierarchical graph sequence model $(HM_n)_{n \in \mathbb{N}}$, [4], Song-Havlin-Makse model $(SHM^p_n)_{n \in \mathbb{N}}$ [11], spherically symmetric trees, and supercritical Galton-Watson trees.

**Theorem 1.** [3] The fractal dimension (Def. 1) of the hierarchical graph sequence model $(HM_n)_{n \in \mathbb{N}}$ and the Song-Havlin-Makse network generation model with "assortativity parameter" $p = 1$ is infinite. The modified box dimension (Def. 2) exists for both models. With $K$ denoting the number of vertices in the base graph $G$ of $(HM_n)_{n \in \mathbb{N}}$, $\tau(\{HM_n\}_{n \in \mathbb{N}}) = (\log K)/2$, and with $m$ being a parameter that drives the number of the newly added nodes in each iteration, $\tau((SHM^p_n)_{n \in \mathbb{N}}) = \log(2m+1)$.

Let $T_f^\infty$ be a spherically symmetric tree: an infinite rooted tree where for each $h$, each vertex at distance $h$ from the root has $f(h)$ many children, $f(h) \in \mathbb{N}^+$. Let $L_n$ denotes
the number of vertices at distance $n$ from the root, and $|x|^+ := \max\{x, 0\}$. Under a mild condition on the growth of the degree sequence $f(h)$, we show the following:

**Theorem 2.** [3] Let $T_f^\infty$ be a spherically symmetric tree with degrees $f = (f(h))_{h \geq 1}$ that satisfies some regularity conditions, then $\tau^*(T_f^\infty) = \log(\text{gr}(T_f^\infty))/2$, where the growth rate of a tree [5] is defined as $\text{gr}(T^\infty) := \lim_{n \to \infty} L_1/n$, whenever this limit exists.

**Theorem 3.** [3] Let $GW_\infty$ be a supercritical Galton-Watson tree with offspring distribution $X$ with mass function $q = (q_0, q_1, q_2, \ldots)$. such that $q_0 = 0$, $q_1 \neq 1$ and $\mathbb{E}[X|\log(X)]^+ < \infty$, and $\mathbb{E}[X] = \mu > 1$. Then, $\tau(GW_\infty) \rightarrow \log(\mu)/2$ almost surely.

**References**

A generalized class of thresholding functions for the hidden variable model of scale-free networks

Sámuel G. Balogh1, Péter Pollner2, and Gergely Palla2

1 Dept. of Biological Physics, Eötvös University, H-1117 Budapest, Hungary
balogh@hal.elte.hu

2 MTA-ELTE Statistical and Biological Physics Research Group, Hungarian Academy of Sciences, H-1117 Budapest, Hungary

1 Introduction

There has been a quickly growing interest in the theory of networks as they can capture the most essential properties of complex systems possibly composed of interacting sub-units, ranging from the microscopic level to the level of society. In the past decade as more real networks’ degree distribution were found to follow the $P(k) \sim k^{-\eta}$ form of decay (scale-free networks) a lot of different growing mechanisms and models were proposed to generate networks displaying the scale-free property [1,2]. Later on, however, it turned out that not all networks emerge directly from a growing mechanism [3]. The concept of a static network with intrinsic node weights (fitnesses, hidden variables) was first investigated by Caldarelli et al. [3] in order to give explanation for the emergence of non-growing scale-free networks whose wiring mechanism is related to the intrinsic weights of the vertices. Subsequently, Boguña et al. extended the formalism and proposed a rigorous analytical framework for the classes of random networks with hidden variables [4].

2 The non-geographical threshold model with intrinsic vertex weights

The concept of hidden variables (intrinsic parameter of the nodes) requires us to assign a fitness parameter $x_i$ to each element of a given set of nodes according to an arbitrary $\rho(x_i)$ probability density function. After distributing these intrinsic parameters, we also have to define a linking function $0 \leq f(x_i, x_j) \leq 1$ giving us the probability of finding a connection between a pair of nodes with fitness values $x_i$ and $x_j$. Note, that the properties of the obtained networks will be completely encoded in the specific forms of $\rho$ and $f$, thus with the help of this framework we can reproduce wide range of degree distributions with the appropriate choices of $\rho$ and $f [5,6,7,8]$.

A striking result was investigated in [3] where specifically an exponential distribution of fitness ($\rho(x) \sim e^{-x}$) was chosen together with a $f(x,y) = \Theta(x + y - \Lambda)$ threshold linking form, where $\Theta(x)$ refers to the Heaviside step function. In this special case, sometimes referred to as non-geographical threshold model, the obtained networks were found to display the $P(k) \sim k^{-2}$ decay of degree distribution and by thus having shown the first evidence that even peaked distributions can generate scale-free networks.
3 Results

By following the continuous approximation presented in [5,6] our work focuses on an analytic extension of the non-geographical threshold model to classes of \{\{\rho, f\}\} constructions which result in generating networks with \sim k^{-2} decay of degree distribution.

We have considered the more general cases where the fitness distribution is \rho_1(x) = H'(x) \exp[-H(x)] or \rho_2(x) = H'(x) H^{-\alpha}(x) with \ f_1[H(x) + H(y)], \ f_2[H(x) H(y)] having lower cut-offs (\Delta_1, \Delta_2). All pair of \{\{\rho, f\}\} functions covered by previous forms can be related to the non-geographical case, and be interpreted as an indirect manifestation of a mathematical symmetry under an arbitrary monotonous transformation of the hidden variables and the probability density distribution defined over them. This result actually generalizes the threshold model of Caldarelli et al. [3], now being extended to a broader set of \{\{\rho, f\}\} functions where the inverse square decay of the degree distribution is satisfied.

By introducing an external, tuneable \beta parameter of the linking function (which can be associated with the modulation of the threshold function) \ f(\ldots) = f_\beta(\ldots) similarly to [9], the scaling exponent of the degree distribution can be modified \eta = \eta(\beta) for each and every particular pair of the generalized \{\{\rho, f\}\} set. (See Fig. 1 for illustration).

Fig. 1: The degree CCDF of four different networks each of whom was generated by distributing the fitness according to \rho(x) = 3x^2 \exp(-x^3) and defining the linking probability in the form of \ f_\beta[H(x) + H(y)] with the same \ H(x) = x^3 but with different \beta values. The corresponding scaling forms \bar{F}(k) \sim k^{-\gamma} (\gamma = \eta - 1) are plotted with red linear curves and the particular values of their exponents are also given. As seen above if we decrease the temperature the range on which the scale-free behavior is present grows narrow while the scaling exponent will have larger value.
It might also be remarkable that according to these results a general mapping can be established between \( \rho \) and \( f \), meaning that for any fitness distribution \( \rho^* \) (assuming additionally that the probability can be normalized and \( H \) being defined above is an increasing and once differentiable function with respect to the fitness parameter) there always exists a linking function \( f^* \) with a lower cutoff which generates scale-free networks with \( \eta = 2 \) scaling exponent and vice versa.

**Acknowledgements**

The research was partially supported by the European Union through projects 'RED-Alert' (grant no.: 740688-RED-Alert-H2020- SEC-2016-2017/H2020- SEC-2016-2017-1) and by the Hungarian National Science Fund, grant no. OTKA K128780.

**References**

1 Introduction

Empirical data collection is often an imperfect process affected by some degree of uncertainty. Uncertainty can come from different sources, for example because of noisy measurements, e.g., in biological experiments, or because of missing information and indirect measurements, as in the case when we infer social ties or influence relationships between individuals based on their interactions. To model uncertain information in networks, probabilistic models in which each edge is associated with an independent probability are the typical choice in the network science literature.

With deterministic networks, a wide range of measures can be used to understand structural and functional properties of the system modeled by the network. With probabilistic networks, we can measure two different aspects. As in the case of deterministic networks, we can use measures to understand the modeled system. At the same time we can use other measures to study the statistical properties of the probabilistic network, that is, the model itself, an aspect that is rarely studied to the best of our knowledge. For example, we can study a probabilistic network to estimate the centrality of some nodes, and at the same time we can compute how much uncertainty is present in the model, e.g., how much confidence we have in the estimated centrality value.

While several measures for probabilistic networks have been proposed, existing works have focused on the characterization of the measures, using one (or rarely a few) probabilistic network datasets to test one specific measure at a time. In this work we follow an opposite approach, and start focusing on the data; in particular, we present a first systematic study of empirical probabilistic graphs, comparing them using a combination of existing measures and also estimating the amount of uncertainty in the models. This gives us the opportunity to identify some missing measures, that we introduce and also use for our study of empirical datasets. More specifically, we provide the following four contributions.

Sources of probability Uncertainty can be generated because of many different reasons. Our hypothesis is that different types of sources of uncertainty may correspond to specific types of probabilistic network models. Therefore, we start our contribution by identifying the different types of sources of uncertainty present in the literature. For example, in some works uncertainty occurs due to the noisy experimental environment and imprecision in the measurement [7]. Other sources of probability include prediction of edges based on past interactions [2], in combination with temporal patterns of
interaction [5] or conditional on the type of the interaction [3]. We also see, of course, edge probabilities reflecting the functional nature of the underlying network.

**Network Measures** To gain insights into the structural and functional properties of probabilistic networks, we computed several existing measures for probabilistic networks on a collection of real datasets. Applying network measures from deterministic networks on probabilistic networks often involves averaging over the possible network topologies arising from the edge probabilities, or require some adaptation in their computation, which we will discuss further in the full paper. An excerpt from the measurements including expected degree, $\eta$-degree, probabilistic clustering coefficient and entropy is shown in Table 1.

**New Measures for Probabilistic Networks** Our compilation of existing measures highlights some need for more work in this area. First, averaging over instances of probabilistic networks is not only computationally expensive in large networks and for some selected measures, but it also loses information. One example of a measure that is not always well described by averaging is degree distribution, where for two nodes with the same expected degree one of the two can have a higher probability of having a higher degree than the other. Then, for some very popular deterministic network measures there appear not to be any counterpart for probabilistic networks. One such case is degree assortativity, that we introduce in this work and also use to characterize real networks. More details are presented in Section 2.

**Impact of Probabilistic Information** In many real studies, probabilistic networks are not currently used. Instead, despite the uncertainty in the sources of the data, a deterministic network is constructed, based on some assumptions about what edges to keep. But what is the result of performing the analysis neglecting the presence of uncertainty? In the last part of our contribution we study in which topological models of the underlying networks the probabilistic measures and the deterministic measures are correlated and in which ones they are not. With topological models of the underlying network we refer to the probabilistic network after the removal of the probabilities. The provided results of this contribution aid us to know which topological types of probabilistic networks can be studied as deterministic networks without significantly affecting the results, and in which topological types of probabilistic networks applying the probabilistic measures is necessary not to obtain different results, e.g., different centrality rankings.

Table 1: A selection of network measures and probabilistic networks. $\mathbb{E}x\bar{D}$: average expected degree, $\overline{\eta-{degree}}$: average $\eta$-degree (for three values of $\eta$), $\mathbb{E}xCC$: average expected clustering coefficient.

| Dataset  | $|V|$ | $|E|$ | $\mathbb{E}x\bar{D}$ | $\overline{\eta-{degree}}$ | $\mathbb{E}xCC$ | Entropy | Ref |
|-----------|------|------|-------------------|-----------------|----------------|---------|-----|
| Fruit Fly | 3751 | 3692 | 0.29              | 1.22            | 0.13 0.04 0.0061 | 607     | [4] |
| Enron     | 78105| 269002| 0.39             | 0.78            | 0.33 0.25 0.0235 | 14535   | [5] |
| Facebook  | 19023| 72041| 0.79             | 1.62            | 0.72 0.47 0.0318 | 5406    | [5] |
| DBLP      | 684911| 2284991| 0.96            | 2.26            | 0.78 0.47 0.1313 | 3737    | [4] |


2 Assortativity

Degree assortativity in deterministic networks represents the tendency of nodes to be adjacent to nodes with similar degrees. Several methods have been proposed to describe degree assortativity in deterministic networks. The simplest way is forming a degree correlation matrix $C$ in which $C_{ij}$ is the number of edges connecting nodes with degrees $i$ and $j$. Similarly, in probabilistic networks we may want to know to what extent nodes with high (resp., low) expected degrees tend to be associated with nodes with high (resp., low) expected degree.

However, in probabilistic networks nodes are connected by probabilistic edges; therefore, each pair of nodes has a specific probability of being adjacent, that can be very high or very low, and should thus offer a different contribution to the network assortativity. To do so, we introduce the average expected degree of neighbors which is the generalization of the average nearest-neighbors degree [6] and weighted average nearest-neighbor degree [1]:

$$k_{fn,i} = \frac{1}{E_i} \sum_j p_{ij} E_j$$  \hspace{1cm} (1)

where $k_{fn,i}$ is the average of the expected degree of the first-hop neighbors of node $i$, $E_i$ is the expected degree of node $i$ and $p_{ij}$ is the probability of the edge incident to nodes $i$ and $j$. Therefore, the average expected degree of the neighbors of all nodes with expected degree $e$, i.e. $k_{fn}(e)$, is the mean of $k_{fn,i}$ restricted to the class of all nodes with $E = e$. If $k_{fn}(e)$ is an increasing function of $e$, then nodes with similar expected degrees tend to have more probable edges.

In the full paper we present a first systematic study of empirical probabilistic networks using a combination of existing measures and identify missing measures, that we introduce and apply to a collection of data sets.

References

Part XII

Network Neuroscience
Internal connectivity and topological roles of nodes in functional brain networks

Onerva Korhonen¹, Elisa Ryyppö², and Jari Saramäki²

¹ Université de Lille, CNRS, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, Lille, France,
onerva.korhonen@gmail.com,
² Aalto University, Department of Computer Science, Espoo, Finland

In the human brain, information processing is distributed among groups of specialized neurons. Interaction and information transfer between these neuronal assemblies are essential for the normal brain function [7]. Therefore, a complex network is a natural model for the brain. Nodes of brain networks are brain areas and links represent structural or functional connections between these areas. The structure of brain networks is characterized by internally tightly connected modules that are spatially separated and connect to each other by weaker long-range links [9, 8]. Brain areas play different topological roles in the network depending on their function. For example, hubs of the network can be divided to local and connector ones: local hubs orchestrate the activity of their own module, inside which they are strongly connected, whereas connector hubs mediate connections between different modules [1].

Structure of functional brain networks strongly depends on how the network nodes are defined. In functional magnetic resonance imaging (fMRI) studies, it is common to use Regions of Interest (ROIs) as network nodes. ROIs are collections of fMRI measurement voxels that cover the entire brain. They are typically defined a priori by e.g. anatomical landmarks or activation observed in earlier studies. The fMRI BOLD time series of each voxel reflects its activation, and the average of the voxel signals of a ROI is commonly used as a ROI time series.

The ROI approach assumes functional homogeneity: all voxels of a ROI are supposed to have similar dynamics and can therefore be averaged without significant information losses. However, we have shown that this assumption does not hold for a majority of ROIs in parcellations based on both anatomy and function [4]. In particular, ROIs show low static spatial consistency, defined as the average Pearson correlation coefficient between voxel time series of ROIs. Besides, rich voxel-level connectivity structure is visible inside ROIs [5]. Furthermore, both spatial consistency and internal connectivity structure of ROIs fluctuate in time [5].

Here, we investigate how the internal connectivity structure and time-dependent functional homogeneity of ROIs relate to network topology. To this end, we quantified the internal connectivity of ROIs by several measures, including static spatial consistency [4] and spatiotemporal consistency [5] that measures the temporal stability of functional homogeneity. We used these measures and ROI size to predict the topological roles of ROIs in three different classification tasks: hub versus non-hub and local versus connector hub classification with a logistic ridge regression model, and a seven-class classification introduced in [3] (Fig. 1) with a linear discriminant analysis model.
The definition of link weights significantly affects the structure of functional brain networks [6]. Here, we adopted the most common pipeline for constructing a functional brain network and used Pearson correlation coefficient to define link weights.

The internal-connectivity-based prediction models outperformed random classification in all three classification tasks (see Table 1). Taking into account the important predictive role of ROI size, we repeated the classification using only predictors independent from the ROI size and still observed a classification accuracy significantly higher than that of a random classificator. High static spatial consistency predicted hubness in general, whereas among hubs low static spatial consistency was a predictor of a connector role.

The low and fluctuating functional homogeneity is often interpreted as a technical flaw that should be eliminated by introducing a more sophisticated approach for defining the ROIs. However, our results show a clear connection between ROIs’
Table 1. Performance of the predictive models. The internal-connectivity-based models outperformed a random classifier (Random) in both the training set (80% of the data) and the test set (20% of the data). Results remained similar when ROI-size dependent predictors were removed from the model (without ROI size).

<table>
<thead>
<tr>
<th>Task</th>
<th>Training %</th>
<th>Test %</th>
<th>Random %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hub vs non-hub</td>
<td>74.43</td>
<td>74.07</td>
<td>55.01</td>
</tr>
<tr>
<td>Hub vs non-hub (without ROI size)</td>
<td>64.22</td>
<td>62.31</td>
<td>55.01</td>
</tr>
<tr>
<td>Local vs connector hub</td>
<td>67.47</td>
<td>66.54</td>
<td>50.20</td>
</tr>
<tr>
<td>Local vs connector hub (without ROI size)</td>
<td>61.26</td>
<td>60.85</td>
<td>50.20</td>
</tr>
<tr>
<td>Multi-class</td>
<td>36.10</td>
<td>38.12</td>
<td>23.45</td>
</tr>
<tr>
<td>Multi-class (without ROI size)</td>
<td>32.02</td>
<td>32.89</td>
<td>23.45</td>
</tr>
</tbody>
</table>

functional homogeneity and their roles in the network topology. Although the exact mechanisms connecting the internal structure of ROIs to their functional roles remain unknown, our results suggest that the variation of functional homogeneity is not a flaw but a genuine feature required for normal brain function. Excluding internal connectivity of ROIs from brain network analysis may mean omitting meaningful information. Therefore, network neuroscience would greatly benefit from new methods that are able to include in the analysis the connectivity both inside and between ROIs.

References

Can MRI brain networks reproducibly distinguish psychosis patients from control subjects?

Sarah E Morgan*1, Jonathan Young*2, Ameera X Patel3, Kirstie J Whitaker1,4, Cristina Scarpazza2, Therese van Amelsvoort5, Machteld Marcelis5, Jim van Os5, Gary Donohoe6, David O Mothersill6, Aiden Corvin7, Martijn van den Heuvel8, Edward T Bullmore1, Michael Brammer2

1 Brain Mapping Unit, Department of Psychiatry, Cambridge University, Cambridge, UK
2 Institute of Psychiatry, Psychology and Neuroscience, King’s College, London, UK
3 Microsoft Research, Cambridge, UK
4 Alan Turing Institute, London, UK
5 Maastricht University, Maastricht, Netherlands
6 NUI Galway, Ireland
7 Department of Psychiatry, Trinity College Dublin, Ireland
8 VU Amsterdam, The Netherlands

*Authors contributed equally.

Algorithms to classify individual psychosis patients from control subjects using MRI scans have proved elusive, in large part because it is often difficult to compare results between studies which use different datasets, machine learning methods and pre-processing techniques [1]. For example, a recent review found that reported classification accuracies from 64 MRI-based schizophrenia studies ranged from 56-100% accuracy [1]. Here we tackle these challenges by using two independent datasets and three imaging modalities to assess which MRI features show reproducible control/patient differences and how accurately those differences can be used to distinguish the groups. Given the importance of understanding the biological basis for diagnoses, we also examine which features drive the classification accuracies and how reproducible their selection is between datasets.

The two datasets we use are from Dublin (N=108, collected as part of a Science Foundation Ireland-funded neuroimaging genetics study-'A structural and functional MRI investigation of genetics, cognition and emotion in schizophrenia') and the Maastricht GROUP consortium [2] (N=124). These datasets were chosen because they had T1w MPRAGE, resting state (rs) fMRI, DWI and DTI images from both psychosis patients and healthy control subjects. Most of the patients had a diagnosis of schizophrenia (diagnosed using DSM-IV criteria), although some had other forms of psychosis, namely psychotic disorder, brief psychotic disorder, schizophreniform disorder or schizoaffective disorder. In both datasets, subjects with a diagnosis of major depressive disorder or bipolar disorder were excluded. The data was pre-processed to obtain regional sMRI and DWI measures [4] (cortical thickness, grey matter volume, surface area, mean diffusivity and fractional anisotropy), DTI networks [5] and rs-fMRI networks [3]. We used each of these measures in turn as inputs to machine learning to assess how well they could classify control subjects and patients. To perform the machine learning, we first applied Gaussian process regression to remove the effects of age and gender (which
varied between groups in some cases). We then performed a 200 fold randomised cross-validation where 10% of subjects were allocated to testing and 90% to training, using a linear SVM kernel classifier.

We find that rs-fMRI connectivity matrices give high case/control classification accuracies (83% in the Dublin dataset and 72% in the Maastricht GROUP dataset), see Table 1. These accuracies are based on a remarkably reproducible underlying regional pattern of control/patient connectivity differences, as shown in Figure 1. Many of the regions where we observe the largest, most reproducible control/patient differences are frontal and temporal regions which are known from the literature to be implicated in schizophrenia [6, 7]. Cortical thickness, mean diffusivity and DTI measures are also predictive, although importantly the underlying pattern of differences is less reproducible between datasets. Interestingly, the absolute classification accuracies vary between datasets, perhaps due to differing symptom levels or data quality, although which modalities and types of data give higher accuracies is broadly consistent between the datasets.

<table>
<thead>
<tr>
<th>MRI measure</th>
<th>ML</th>
<th>Maastricht</th>
<th>Dublin</th>
</tr>
</thead>
<tbody>
<tr>
<td>fMRI connectivity matrix</td>
<td>AUC</td>
<td>0.77</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>LOO bal acc</td>
<td>0.72</td>
<td>0.83</td>
</tr>
<tr>
<td>fMRI connectivity matrix- degree centrality</td>
<td>AUC</td>
<td>0.75</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>LOO bal acc</td>
<td>0.70</td>
<td>0.76</td>
</tr>
<tr>
<td>Regional cortical thickness</td>
<td>AUC</td>
<td>0.67</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>LOO bal acc</td>
<td>0.63</td>
<td>0.76</td>
</tr>
<tr>
<td>Regional mean diffusivity</td>
<td>AUC</td>
<td>0.48</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>LOO bal acc</td>
<td>0.53</td>
<td>0.82</td>
</tr>
<tr>
<td>Regional DTI weighted by no. of streamlines</td>
<td>AUC</td>
<td>0.59</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>LOO bal acc</td>
<td>0.57</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 1. Summary of machine learning results. We report the area under the curve (AUC) and leave-one-out balanced accuracy (LOO bal acc) for machine learning classifiers using different MRI inputs.

Ultimately, the rs-fMRI connectivity differences presented here are an important candidate biomarker for psychosis. This work also provides clearer insight into which MRI data types are best able to discriminate psychosis patients from control subjects and how reproducible those results are between two independent datasets. Future work will build on the understanding developed here to determine whether similar approaches can be used to predict outcome for first episode psychosis patients.
Fig. 1. Regional machine learning classifier weights for the rs-fMRI data from a) the Maastricht GROUP dataset and b) the Dublin dataset. These two independent results correlate significantly with each other ($r=0.31, p<0.001$), suggesting that the machine learning classifier finds similar predictive features in both datasets. They also correlate with a regional t-statistic for control-patient differences in the rs-fMRI data.

References

Effects of spatial smoothing on group-level differences in functional brain networks

Ana Triana\textsuperscript{1}, Enrico Glerean\textsuperscript{1,3,4} Jari Saramäki\textsuperscript{1}, and Onerva Korhonen\textsuperscript{2}

\textsuperscript{1} Aalto University, Department of Computer Science, Espoo, Finland, ana.trianahoyos@aalto.fi
\textsuperscript{2} Université de Lille, CNRS, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, Lille, France,
\textsuperscript{3} Aalto University, Department of Neuroscience and Biomedical Engineering, Espoo, Finland,
\textsuperscript{4} Helsinki Institute for Information Technology, Helsinki, Finland

Normal brain function is based on interactions between spatially separated groups of specialized neurons that produce complex behaviors. Therefore, during the last decades, it has become increasingly popular to model the brain as a complex network, leveraging on multiple neuroimaging techniques [4]. Out of those, brain connectivity using functional Magnetic Resonance Imaging (fMRI) is a popular approach to investigate differences between healthy and clinical groups. Before creating a functional network, fMRI time-series are subjected to several preprocessing steps to control for noise sources, such as physiological artifacts and head motion. However, these steps may have undesired side effects. For example, it has been shown that spatial smoothing induces changes in the functional network structure [1]. Nonetheless, its effects on group-level differences are still unknown.

Here, we investigate the effects of spatial smoothing on the difference between Autism Spectrum Disorder (ASD) patients and healthy controls (TC) with resting-state fMRI data from the ABIDE initiative (N=33 male, matched subjects pairs) [2]. We consider spatially un-smoothed fMRI data as well as filtered data with Gaussian kernels with full width half maximum (FWHM) spanning from 4mm to 18mm. Functional link weights are computed as pairwise Pearson’s correlation between time series from 264 regions of interest (Brainnetome atlas [3]). We identify groups differences using Network Based Statistics (NBS) [6].

We find that network differences between groups are affected by smoothing. In particular, the number of significant functional links increases with increasing smoothing filter width (F-statistic $>16$). However, different brain regions are involved at different levels of smoothing. Few links are prevalent at all smoothing levels; these links are found between sensorimotor and subcortical areas such as thalamus and striatum. This effect is seen independently of the ROI size or physical link length.

Given that we cannot be completely certain about the differences between ASD and TC, it is difficult to say whether spatial smoothing is emphasizing existing differences in the networks by improving the signal-to-noise ratio, or if its effects are highlighting non-existing differences as significant findings. However, there is a clear increasing trend between the spatial smoothing kernel size and the number of nodes and links found in the networks.

In general, the effects of spatial smoothing are diverse, non-trivial and hard to predict. This has important consequences for group comparisons of functional networks:

The 7\textsuperscript{th} International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
spatial smoothing might distort network differences between groups. Hence, it is important to consider these effects when preprocessing resting-state fMRI for studies of group-level differences analysis using functional networks.

References

Fig. 1. Smoothing affects network comparisons between two groups for unthresholded, full weight matrices. Colors indicate functional areas of the brain [5] and bins show the Brainnetome ROIs (the names are omitted due to space constraints). The image is symmetrical around the vertical middle line to show the right and left hemispheres. The level of smoothing affects comparisons by increasing the number of nodes, links, and functional areas detected as different between the two groups.
Identifying the Coupling Structure in Complex Systems through the Optimal Causation Entropy Principle, Information Flow and Information Fragility

Erik Bollt, Jie Sun

Inferring the coupling structure of complex systems from time series data in general by means of statistical and information-theoretic techniques is a challenging problem in applied science. The reliability of statistical inferences requires the construction of suitable information-theoretic measures that take into account both direct and indirect influences, manifest in the form of information flows, between the components within the system. In this work, we present an application of the optimal causation entropy (oCSE) principle to identify the coupling structure and jointly apply the aggregative discovery and progressive removal algorithms based on the oCSE principle to infer the coupling structure of the system from the measured data. We will include discussion of examples such as the functional brain network as inferred by fMRI functional magnetic imaging.

Identifying connections in a complex process manifest as causal direct information flow suggests a new way of detecting and understanding fundamental changes in the dynamical process of a complex system. The question of fragility and robustness concerns how the macroscopic behavior of a system will change in response to local perturbations. We interpret the phrases robust and fragile as a global descriptor of the system, in terms of the change of the information carrying capacity of paths between states of a complex system, due to the loss of a state, or connection, with a corresponding descriptor in terms of information betweenness. Stated more broadly about the interdependencies of complex systems, consider a large-scale process in which minor changes frequently occur, and the question is, can we define and, hence detect, those changes which would render the system effectively different and likewise significantly alter the system performance, before the system might fail. Thus, here we suggest a fragility-robustness duality to detect a tipping point whereby even a minimal detail change can cause a catastrophic systemic outcome.

Department of Mathematics, Department of Electrical and Computer Engineering, Department of Physics, Clarkson Center for Complex Systems Sciences, C3S2, Clarkson University, Potsdam, NY 13699-5815, e-mail: bolltem@clarkson.edu

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Robust Navigable Cores in the Human Brain Networks

Zalan Heszberger\textsuperscript{1,2}, Andras Majdan\textsuperscript{1}, Andras Gulyas\textsuperscript{1,2}, Tibor Gyimothy\textsuperscript{3}, Vilmos Bilicki\textsuperscript{3}, and Jozsef Biro\textsuperscript{1}

\textsuperscript{1} Budapest University of Technology and Economics, Budapest, Hungary \{heszi, majdan, gulyas, biro\}@tmit.bme.hu
\textsuperscript{2} MTA-BME Information Systems Research Group, Budapest, Hungary
\textsuperscript{3} University of Szeged, Szeged, Hungary \{gyimi,bilicki\}@inf.u-szeged.hu

1 Introduction

Given a set of network nodes with coordinates in an arbitrary metric space, the greedy navigational core (GNC) can be identified and calculated, as the minimum set of links between these nodes which provides 100\% greedy navigability [2]. The level of presence of the Greedy Navigational Core is high in various real networks, and can also be disclosed in structural networks of the Human Brain. In [2] we have shown that the GNC precision (the ratio of the number of the GNC links included in the real network and the total number of GNC links, sometimes referred to as true positive rate) is 89\% in a five-subject based averaged structural brain network.

2 Results

Here we perform results on a follow-up investigation of structural greedy navigability in 200 structural brain networks from 40 individual subjects at 5 different scales (these scales correspond to resolutions of 83, 129, 233, 463, 1015 nodes in the brain structural networks) [3]. For the GNC network generations only the physical (3D Euclidean) coordinates of the brain parcels were used, no other anatomical data or considerations were utilized. We found that the level of GNC precision (the number of GNC links in the brain network) is quite consistent among the 40 brain networks within all scale, in spite of the fact that the 40 brain networks significantly differ from each other. The mean (and the standard deviation) of the GNC precisions in different scales (with increasing resolutions) are 0.85 (0.025), 0.88 (0.019), 0.81 (0.017), 0.70 (0.017), 0.51 (0.025), respectively. GNC precisions turned out to be also robust against weak (possibly spurious) connections. For instance, when 50\% of the links are removed from the brain networks the precisions remain close to the original values: 0.72 (0.030), 0.75 (0.019), 0.68 (0.020), 0.57 (0.019) 0.42 (0.020). One can also observe that larger part of the navigational core is missing from higher resolution brain networks, however, the precisions are still consistent.

In scale 5 we have constructed a sequence of average networks, based on averaging networks over link weights (inferred from measured anatomical strengths of fiber paths), and cutting out links with small average weights. We have also generated an
average GNC network based on averaging the centroids of the brain parcels. The size of the average GNC network (the number of links) is 2652. Regarding the average network sequence, the following table shows the GNC precision in the function of the size of the average network:

<table>
<thead>
<tr>
<th>Threshold</th>
<th>(10^{-8})</th>
<th>(10^{-7})</th>
<th>(10^{-6})</th>
<th>(10^{-5})</th>
<th>(10^{-4})</th>
<th>(10^{-3})</th>
<th>(10^{-2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaged Brain network size</td>
<td>101468</td>
<td>101450</td>
<td>95929</td>
<td>67553</td>
<td>39406</td>
<td>15996</td>
<td>2381</td>
</tr>
<tr>
<td>GNC Precision</td>
<td>0.925</td>
<td>0.925</td>
<td>0.924</td>
<td>0.907</td>
<td>0.874</td>
<td>0.769</td>
<td>0.263</td>
</tr>
</tbody>
</table>

Note, that if none of the links have been cut out, the average brain network contains 101468 links (full mesh would have 514 605 links) and in this case 0.9253 fraction of the 2652 links of the average GNC are in this network. If the number of links in the average network is quite comparable to the sizes of individual networks (Threshold=0.001, size of average network=15996), the GNC precision is still as high as 76.8%. When the size of the average network is comparable to the average GNC network, the GNC precision is still amazingly 26.28%.

Besides anatomical weighting strategies, the link prevalence score (the number of networks containing the link) can also be used to identify possibly existent (high prevalence score) and non-existent (low prevalence score) links in the inferred brain structural networks, and based on this one can compromise the false positives and false negatives in pruning the networks [1]. Here we also present results on GNC precisions in individual networks (the above mentioned 40-subject-5-scale networks with the number of nodes 83, 129, 234, 463, 1015 in scale 1, 2, 3, 4, 5, respectively) thresholded by the link prevalence scores. Thresholding means in this case that in an individual network only that links are kept, which are present at least \(T-1\) other networks too, where \(T\) is the threshold. Within a resolution, every network is thresholded by all possible values of LPS’s (1,...,40), then the GNC precisions are measured in all resulted networks. The GNC precisions are then averaged over the subjects for every LPS threshold. The important observation is that GNC precisions are consistent (low variations across subjects) and robust against LPS thresholding in all scale (see Figure 1). For example in scale 1 for LPS=1 (no link is removed) GNC precision is 0.85 and for LPS=30 (about 30% of links are removed from every network) GNC precision is still as high as 0.80.

In all the 5 scales it can be observed that for lower values of LPS thresholds the GNC
precision remains almost intact while for higher values its decrease is fastening. The fastening decrease measurably coincides with the right hand side (consisting of possibly existing links) of the link prevalence distribution. This means that most of the true positive links in GNC networks are also possibly existent. The consistency and robustness of the greedy navigational core network precisions are remarkable in the light that these networks are inferred only from the physical coordinates of the brain parcels by straightforward geometric computations. Based on these we think navigational cores and their precisions are possible candidates for auxiliary testing and qualifying sets of structural networks of the Human Brain.

![Fig. 2. Greedy Navigational Core precision and Link Prevalence Score distribution in the function of Link Prevalence Score](image)

**Acknowledgment** Project no. 123957, 129589 and 124171 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FK_17, KH_18 and K_17 funding schemes respectively. Z. Heszberger and A. Gulyas have been supported by the Janos Bolyai Fellowship of the Hungarian Academy of Sciences and by the UNKP-18-4 New National Excellence Program of the Ministry of Human Capacities.

**References**

Communicability indicates structural and functional connectivity changes in children with autism spectrum disorder

Roma Siugzdaite\textsuperscript{1}, Hannelore Aerts\textsuperscript{2}, and Daniele Marinazzo

\textsuperscript{1} Experimental Psychology Department, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium
roma.siugzdaite@ugent.be,

\textsuperscript{2} Data Analysis Department, Ghent University, Henri Dunantlaan 1, 9000 Ghent, Belgium

1 Introduction

Autism spectrum disorder (ASD) is a complex developmental brain disorder, characterized by a spectrum of impairments in social functioning and communication, often accompanied by stereotypic behavior. It hinders a person’s ability to communicate and develop social relationships, and it creates repetitive behavior and sensitivity to changes in the environment.

A recent review of fMRI studies on children with ASD proposes that discrepancies between findings of autism regarding hypo- and hyper-connectivity might be reconciled by taking into account developmental changes \cite{3}. From birth until about 11 – 12 years old, the brain makes a lot of new connections. Meanwhile, a process called synaptic pruning begins. Neuron connections that are damaged or not used just stop functioning and disappear. It makes synaptic connections more clear and efficient. Children and adolescents with autism have a surplus of synapses in the brain, and this excess is due to a slowdown in a normal brain ‘pruning’ process during development \cite{6}. In particular, resting-state functional connectivity MRI studies provide evidence for widespread hyper-connectivity in children with ASD, whereas hypo-connectivity is observed in adults with ASD. However, what happens during puberty is not well known and there are two potential scenarios \cite{3}: 1) the ASD group shows a less steep developmental increase in functional connectivity over the age span compared with the control group; 2) the ASD group displays an anomalous pattern of connectivity across the pubertal period. In this study, we investigated a group of children aged 9 to 12, looking at functional connectivity networks (functional connectome) and morphological differences (structural connectome) between children with ASD and typically developing (TD) children. Applying the graph theoretical perspective as well as two novel methods related to information flow throughout the network, we attempt to identify the method best suited to distinguish between ASD and TD children in early puberty. Given that current research on ASD adolescents is scarce, and taking into account the developmental model of functional brain connectivity in ASD with hyper-connectivity in early childhood and hypo-connectivity in adulthood, we hypothesize that functional connectivity is not different during our investigated age if the changes are linear, but communicability alters in ASD during age from 9 to 12.
2 Results

Demographic analysis showed significant group differences in Full IQ (CON: 112.0 ± 14.16; ASD: 106.1 ± 17.18; \( p = 0.013 \)) and VIQ (CON: 111.9 ± 14.31; ASD: 104.1 ± 16.26; \( p = 0.001 \)) that indicates language use and communication impairment. Examining the connectome topology of ASD and TD children using standard graph measures (clustering coefficient, characteristic path length, smallworldness) on both raw and MST optimized connectivity matrices (where all the vertices connected together, without any cycles and with the minimum possible total edge weight using minimum spanning tree algorithm), we did not find any significant differences between both groups in structure nor function (all \( p > 0.05 \)). Similarly, semi-metrics percentage that looks into alternative paths having greater strength than the direct edge, yielded no significant group differences (ASD: 0.8436 ± 0.0199; CON: 0.846 ± 0.0201; \( p = 0.4675 \)). However, communicability a measure of the ease with which information can travel across a network [2] revealed significant differences between ASD and CON children’s structural and functional connectomes. Structural connectome showed decreased communicability in ASD group though functional connectome demonstrated increased communicability in ASD group.

![Communicability in SC of ASD](image1)

![Communicability in SC of CON](image2)

![Communicability in FC of ASD](image3)

![Communicability in FC of CON](image4)

**Fig. 1.** Communicability measure in structural and functional connectome

(1)
Summary. The preliminary literature in autism spectrum disorder (ASD) shows an emerging picture of atypical function and structure in brain networks responsible for the social communication, and language [1]. Taking advantage of a relatively large open-access dataset of children with ASD, we explored the structural and functional organization of the brain connectome in autistic children and typically developing control group. In this study we have used anatomical images and resting-state functional magnetic resonance imaging data from 69 children with ASD and 79 age-matched controls (9-11 years old) from the ABIDE open-access database. From structural and functional imaging data we constructed connectomes and first of all we applied standard graph measures (clustering coefficient, characteristic path length and smallworldness). Additionally, semi-metrics [4] and communicability measures [2] were computed to investigate bottlenecks of the information flow throughout the network. Results showed no significant group differences in standard graph measures nor semi-metrics computed on structural and functional connectivity matrices. But communicability measures, in contrast, were found to differ significantly between both groups. We conclude that the structural and functional topology of the ASD connectome does not differ significantly from that of TD children. However, using more sophisticated measure that measures information flow throughout the ASD connectome at a global level appears to be altered, showing more pruned (reduced in strength) structural connections and stronger connected functional connectome in ASD compared to TD children.

References

Part XIII

Networks in Finance and Economics
Transition from plasticity to instability in the structure of economic networks

Celian Colon\textsuperscript{1} and Jean-Philippe Bouchaud\textsuperscript{2}

\textsuperscript{1} Environmental Research and Teaching Institute, Ecole Normale Supérieure, F-75230 Paris Cedex 05, France, celian.colon@polytechnique.edu
\textsuperscript{2} Capital Fund Management, 23 rue de l'Université, 75007 Paris, France

Abstract. This paper examines the spontaneous emergence of structurally stable economic networks. Market-balanced production networks are simulated, in which firms can replace suppliers to increase profits. These systems exhibit very large set of stable configurations that are far apart in configuration space, as found in many Constrained Satisfaction Problems. Rewiring cascades allow the system to move from one configuration to another in response to a shock. This plasticity is however reduced when the productivity of firms and inputs become more heterogeneous. A transition occurs where networks fail to stabilize, and permanently wander through unstable configurations.

1 Introduction

The use of networks in economics has the potential to transform our understanding of the link between microeconomic behaviors and macroeconomic observables. The ongoing trend of increased globalization and interconnection between firms and financial institutions makes this approach even more relevant. Here, we propose a theoretical model to assess the macroeconomic stability of microeconomic rational decisions.

In a market-balanced supplier–buyer network, firms looks for higher profits by changing their suppliers, thereby gradually transforming the network topology. This rewiring dynamics drive the system to stable configurations, in which it is in the interest of no firm to change supplier. This topological equilibria are path dependent, and the system move from one equilibrium to another in response to node removal. When the network becomes more heterogeneous—i.e., more distinct firms attributes, more differences between input's productivity—, it generally fails to stabilize and keeps on changing configurations forever.

2 Model: an input–output network with a rewiring dynamics

Our model is based on Acemoglu et al.[1]. We simulate a production network, which consists of \( n \) firms meeting the demand of households. Households have a fixed budget to buy goods, which are delivered by the firms. Each firm produce a distinct good using labor, supplied by households, and intermediary inputs, supplied by the other firms. The production process is modeled by Cobb–Douglas functions.
Each firm has suppliers and clients among the other firms. These input–output relationships form a production network. We choose as starting topologies the class of scale-free networks proposed by Colon and Ghil [2], which exhibits some empirical features of real production networks, studied by Fujiwara and Aoyama [3] on a dataset of over 1 million Japanese firms.

Production levels and prices are set so that (i) firms maximize their profit, (ii) households maximizes their utility, represented by a log-utility function, and (iii) supply meets demand in all markets. This market equilibrium exists whatever the network topology and is unique.

The originality of our model is that the topology evolves over time. Firms aim to increase profits by replacing existing suppliers by new ones. We model this behavior by preselecting, for each firm, a fixed set of possible suppliers among the other firms. This set corresponds to the feasible technologies the firm can utilize. At each time step, a firm assesses whether replacing a currently-used supplier by a currently-unused one can increase its profit. If it does, the firm perform the most profitable switch, and all prices and production levels are adjusted accordingly. Note that firms have full information over the network and know how all quantities would adjust after replacing a supplier.

This discrete-time dynamics reach a fixed point when no firm has any interest in changing suppliers. To ensure that each firm gets its turn to rewire, we organize rounds. In a round, each firm—one by one and in a random order—tries to rewire, in a random order. If a firm rewire, the topology is immediately updated. If no firm rewire during a whole round, then the topology has reached a stable configuration, which corresponds to a Nash equilibrium, in the game theoretical sense, a typical problem in a large class of so-called “Constrained Satisfaction Problems”; see e.g., [4]. We characterize the rewiring dynamics and examine the path of the system in the network space. We assess the influence of network size $n$, and of the distributions of productivity factors. Finally, we test network resilience to removal of firms.

3 Results

*The rewiring process redistribute profits* When all firms and all inputs have the same productivity, the network topology reaches a fixed point. Although all firms are similar when taken in isolation, their embeddedness into a web of relationships generates differentiated behavior. Some firms rewire more than others, some succeed in increasing profits, others incur losses. In the absence of technological innovation, the total value added of this economy remain constant. Profits is simply redistributed among firms.

*Stabilization gets increasingly slow in large networks* The average number of rounds needed to reach an equilibrium grows with $n$. A 10-firm network stabilizes after less than 4 rounds, a 100-firm one needs 10 rounds. In addition, for $n \geq 100$, we increasingly encounter extreme cases, in which the number of rounds required for stabilization is several times higher than the median. With growing interdependencies, the profit of a firm is more likely to get impacted by rewiring decisions made by others. Stabilizing the system takes longer. This results is linked to the small-world feature of the class of networks used in our model, which is empirically observed.
Topological equilibria are path dependent For the same initial network, the rewiring dynamics lead to a wide range of stable configurations. For instance, 1,000 simulations of a 50-firm network almost always lead to 1,000 distinct topologies. The distance between these topologies, defined by the proportion of different edges, is about 100 in average. Only very small systems falls back to a reduced number of configurations: when \( n = 10 \), 1,000 simulations lead to no more than 500 equilibria. The order according to which firms rewire at each round is the source of path dependency.

Shutting down a single firm triggers waves of reconfiguration If we force a firm to disappear from a system standing at a topological equilibrium, we observe a cascade of rewirings. This cascade drives the network to a new stable configuration. For a 200-firm network, this cascade is of about 50 rewirings. The space of topological equilibria is rather patchy. A localized perturbation sparks a wave of rearrangements.

Topological equilibria vanishes in heterogeneous networks These smooth topological readjustments may however become more turbulent. As the distribution of firm or input productivity become heterogeneous, we observe a transition from fixed-point equilibria to an unstable dynamics; see Fig. 3. First, limit cycles emerge, in the system dynamics sense: the network keeps on switching between two topologies forever. Next, strange attractors appear, in which the system alternate between a limited set of already visited configurations in a seemingly erratic manner. Last, trajectories diverge: the network wander in an ever increasing set of topologies. The heterogeneity of firm or input productivity generates contradicting opportunities among firms that they are unable to resolve.

Fig. 1. Proportion of runs reaching one of the three following dynamical regimes—fixed point (green), 2-round-period limit-cycle (orange), other trajectories, including larger period limit-cycle, strange attractors and diverging trajectories. The bifurcation parameter on the x-axis is the standard deviation of the lognormal distribution used to generate the input–output coefficients. These coefficients determines the nominal productivity of intermediary inputs in the production functions of the firms. Results obtained for 50-firm networks with an average connectivity of 4 and 4 extra potential suppliers per firm.
4 Discussion

The globalization and interconnection of supply chains and financial flows are ongoing economical trends. Our findings question the existence of network configurations that satisfies all agents. Even if such configurations would exist, the magnitude and complexity of these interactions may prevent economic network to actually reach them. The entry and exit of new firms cause the system to constantly evolve.

This dynamical feature of economic network could be accentuated by information limitation. In our model, firms perfectly know the network structure, and are able to forecast the impact of their rewiring on all prices and production levels, under the hypothesis of balanced markets. We expect that partial knowledge of the network structure would impede the stabilization process. On the other hand, perfectly forward-looking agents, who fully anticipate all future decisions of the other firms triggered by its own decisions, may lead to more stable networks, but this is probably difficult to realize in a complex world.

References

Optimising Systemic Risk in Interbank Networks

Christian Diem¹, Anton Pichler² and Stefan Thurner³

1 Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria, c.diem@wu.ac.at
2 Institute for New Economic Thinking, University of Oxford, Oxford OX2 6ED, UK
3 Complexity Science Hub Vienna, Josefstädter Straße 39, A-1080, Austria

1 Introduction

Systemic risk in financial markets is a network phenomenon, since the default of a single bank can adversely impact other institutions through financial interlinkages. Recently, the topic of reducing systemic risk by reorganising networks appeared in the context of overlapping portfolios [1]. We demonstrate a more general method for minimising systemic risk, which is applicable for any type of exposure network as long as systemic risk is quantified with DebtRank [2]. The suggested framework reorganises the network topology, but keeps important economic features of the optimised result comparable to the original network. We apply this approach for the first time to an interbank market. Although in reality interbank networks are not formed by a deterministic optimisation procedure, this approach provides an educated guess of the prevailing systemic risk reduction potential. To test our method, we apply the optimisation to twelve real world interbank networks from Austria and find that the systemic risk can be reduced on average by about 60%.

2 Method

We quantify the systemic risk of a given interbank network as the sum of all single bank DebtRanks (DR) in the network [2]. Since DR is not representable in closed form, we approximate it by the sum of all direct impacts only, i.e. when a bank defaults, only its impact on direct neighbours is considered. The objective function (DirectImpact) in the optimisation problem (1) is based on this first-order approximation,

$$\min_{L \in \{M \in \mathbb{R}^{n \times n} | M_{ii}=0\}} \sum_{i=1}^{n} \sum_{j=1}^{n} \min \left( \frac{L_{ij}}{e_j}, 1 \right) \frac{a_j}{L},$$

(1)

where $L_{ij}$ indicates the liability from $i$ to $j$, $e_j$ is the equity and $a_j$ the interbank assets of $j$ and $L$ is the total size of the market. The constraints (2) keep the total interbank assets $a_i$ and liabilities $l_i$ of each bank constant and ensure that the probability-of-default ($\kappa$) weighted exposure ($R_i$) of each interbank loan portfolio remains as in the original network. This generally reduces the flexibility for reorganising the network, but still allows for sizeable reductions of systemic risk.

$$l_i = \sum_{j=1}^{n} L_{ij}, \quad a_i = \sum_{j=1}^{n} L_{ji}, \quad R_i = \sum_{k=1}^{n} L_{kj} \kappa_k \quad \forall i.$$  

(2)
We can reformulate this optimisation problem as a Mixed Integer Linear Problem (MILP) by using SOS2 type constraints and solve it with CPLEX’s R interface and the ROI package [3]. Solving the MILP yields a network with minimal direct impacts, but not necessarily a financial network with minimised systemic risk in terms of total DR. However, our computations demonstrate the effectiveness of this heuristic method in reducing overall systemic risk. We apply our optimisation to a quarterly time series of real world interbank networks from 2006 to 2008. The networks contain the 50 largest Austrian banks with respect to total assets in the corresponding quarter. Since $\kappa$ is not available directly and sufficient data for its estimation is neither we use the banks’ leverage ratios as a proxy for their credit riskiness.

The reduction in systemic risk due to our optimisation procedure is summarised in Figure 1. From Figure 1a) we see that the DRs of the optimised network lie substantially below the empirical ones during the whole sample period. The average reduction of the network DR amounts to 60% from 9.4 to 3.6. Figure 1b) indicates two interesting findings. The first observation is that also in the optimised network at least one bank always stays highly systemic exhibiting a DR of above 0.5, apart from the first two quarters of 2007. In quarter two and three of 2008 two banks have a DR greater than 0.5. The second observation is that in all quarters the reduction of medium sized DRs in the range of 0.1 to 0.3 seems to work very well. Thus, it seems that for the largest players in the interbank market the DR reduction potential is limited. However, there appears to be a considerable reduction potential for DRs of medium sized banks. In Figure 2a) we observe that the total DR and direct impacts of the empirical networks...
Fig. 2. The left figure compares the DRs and DirectImpacts of the twelve empirical and optimised networks on the network level. The right figure shows the same comparison on the single bank level. The $R^2_{emp} = 0.75$ and the $R^2_{emp} = 0.73$ on the network level and $R^2_{emp} = 0.86$ and the $R^2_{emp} = 0.79$ on the single bank level.

exhibit a fairly linear relationship. This indicates that optimising the direct impacts in a network should be a reasonable proxy for reducing the DebtRank of these networks. The optimised networks still exhibit this linear relationship, although with a larger slope parameter. This indicates that our optimisation achieves a stronger reduction in direct impacts than in DebtRank, which was to be expected. When considering Figure 2b) we observe a similar pattern on the single bank level.

Summary. We find strong evidence that the introduced optimisation approach achieves a sizeable reduction of DebtRank in real world interbank networks, while keeping economic conditions for banks similar. The optimised financial network could be a useful tool for regulators and serve as a best-case benchmark for systemic risk in a given empirical financial network. It can also give an educated guess of the prevailing systemic risk reduction potential. Furthermore, we find evidence that really large players in the interbank market also remain systemic after the optimisation, whereas the DebtRanks of medium sized banks can be reduced substantially. Thus, medium sized banks could be an important leverage point for reducing overall systemic risk within an interbank network.

References

Stability in Core-Periphery Production Networks

Manish Sarkhel¹ and Nagarajan Krishnamurthy¹

Indian Institute of Management, Indore, MP, 453331, India
f14manishs@iimidr.ac.in, nagarajan@iimidr.ac.in

1 Introduction

The core-periphery structure is a prominent feature of industrial clusters [1], [2], [3]. In manufacturing and production clusters, the periphery is often comprised of small firms who supply to larger firms in the core [4]. For instance consider the bicycle cluster in Ludhiana, India where large firms who make finished goods (bicycles) source parts such as tyres, bells, etc. from smaller ancilliary firms and assemble them [6]. In such industrial clusters, network stability is vital to the sustainability of the network [5]. In this paper, we model a supply chain network with a core-periphery structure where firms (players) in the core source inputs from firms (players) in the periphery, add value to the products and resell them. Additionally, core firms network among themselves to reduce their costs. Our analysis brings forth several interesting insights such as a denser core leads to increased profits for suppliers in the periphery. Furthermore, as the number of firms increases, the profits of the suppliers and the manufacturers come down.

2 Model

We consider an oligopolistic market with core-periphery structure selling a single type of homogeneous good. There is a set of \( m \) identical producers in the periphery who incur a fixed marginal cost of production \( c_p \). These peripheral players do not have direct access to the market. They depend on a set of \( l \) identical sellers in the core for the same. Each peripheral player \( i \) sells quantity \( q_i \) to a player \( j \) in the core who has market access. For maintaining links, a marginal cost \( c_t \) is encountered from the core to the periphery. This cost is shared by the players in the core and periphery. Players in the core bear \( \theta \) fraction of the cost while peripheral players bear \( (1 - \theta) \). Within the core, the cost per link is \( c_c \). The players in the core on receiving the goods from the periphery add value to these products by finishing, advertising, packaging, etc. They incur a fixed marginal cost \( C_j \) for doing the same, where \( j \) is the index of the core player. The marginal cost of the \( j^{th} \) player in the core decreases when she collaborates with other players in the core to reduce the cost of value addition. After adding value, the players in the core sell their goods to the customers who approach them with the objective of maximising their utility in a Cournot competition setting. The objective of the peripheral players similarly is to maximise their utility by selling to the players in the core. We assume that all parameters are of common knowledge to the agents who are self-interested and care only about their utilities.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
With respect to the network, let $M = \{1, 2, 3, \ldots, i, \ldots, m\}$ denote set of peripheral players who supply to a set of $L = \{1, 2, 3, \ldots, j, \ldots, l\}$ players in the core. Let these set of players be connected by a graph $G$. Let $g^{L+M}$ denote the complete network. Then, $G = \{g : g \subseteq g^{L+M}\}$ is the set of networks containing $L + M$ nodes. Let $g \in G$ and $(ij) \in G$ denote a link. The neighbourhood size of the $j^{th}$ core player with respect to the peripheral players is denoted by $\gamma_j(g)$. Similarly, the neighbourhood size of the $i^{th}$ peripheral player with respect to the core players is denoted by $\eta_i(g)$. There is also a network between the core players. The neighbourhood size of this network for the $j^{th}$ player in the core is denoted by $\delta_j(g)$.

The $j^{th}$ core player’s utility which is to be maximised may be written as:

$$\pi_j(g) = (\beta - b_c q_j - b_p \sum_{k \neq j}^l q_k - p - C_j(g))q_j - (c_i \gamma_j(g) \theta + c_i \delta_j(g))$$

Where, $C_j(g) = c_0 - c \delta_j(g)$, and $c_0$ denotes the per unit cost of value addition without any links among the core players while $c$ is the marginal reduction in cost of value addition per link in the core. Thus, a complete graph among the players in the core would mean $\sum_{k=1}^l \delta_k(g) = l(l - 1)$. Similarly, a star network where the $j^{th}$ player is the centre of the star would have $\delta_j(g) = l - 1$.

The $i^{th}$ peripheral player’s utility which is to be maximised can be written as:

$$\pi_i(g) = (\alpha - b_p q_i - b_p \sum_{n \neq i}^m q_n - c_p q_i - c_i \eta_i(g)(1 - \theta))$$

### 3 Results

**Theorem 1.** The unique optimal profit of the $j^{th}$ core player and $i^{th}$ peripheral player is:

$$\pi_j^*(g) = \left(\frac{\left(\frac{c \gamma_j(g)}{\langle j \rangle} \sum_{j=1}^l \delta_j(g) + \beta - c_0 - c_p \right)}{\left(m + 1\right)}\right)^2 \frac{l}{b_c(l + 1)} - c_i \eta_i(g)(1 - \theta)$$

$$\pi_i^*(g) = \frac{1}{l^2 b_c} \left( c \left(\delta_j(g)(l - 1) - \sum_{k \neq j}^l \delta_k(g)\right) + \frac{mc \left(\sum_{k=1}^l \delta_k(g)\right) + (\beta - c_0 - c_p)l}{(m + 1)(l + 1)}\right)^2 - (c_i \gamma_j(g) \theta + c_i \delta_j(g))$$

Theorem 1 describes the equilibrium solution for both the type of agents. We observe that the degree of connections $\delta_j(g)$ of the core players among themselves figures in the utility functions of both the core and peripheral players. Thus, as the number of links in core goes on increasing, the profits of the peripheral players also increases. This indicates a denser core is better for the peripheral players as it positively affects their profits. Further, the total number of players in the network affect both the core player’s and peripheral player’s utility.
Theorem 2. The neighbourhood size of the core players resulting in minimum profits is:
\[ \delta^*_j(g) = \frac{b_c c_c}{2(mc)^2} - \frac{\beta - c_0 - c_p}{mc(l+1)(m+1)} \]

The optimal profits of the players within the core decreases in \( \delta_j(g) \) up to a unique point. After which it again starts to increase. This is because \( \pi^*_j(g) \) is convex in \( \delta_j(g) \).

Fig. 1. \( \pi^*_j(g) \) and no. of links \( \delta_j(g) \).
(Parameters values:
\( c = 3; l = 11; \beta = 10; c_0 = 30; c_p = 0.2, m = 15; c_t = 0.2; \theta = 0.3, b_c = 0.2, g_j = 10; c_c = 0.4, \sum_{k \neq j} \delta_k(g) = 60) \)

4 Conclusion

Our contribution to the literature relates to the formation of core-periphery networks. We find that in core-periphery manufacturing networks, a densely connected core is preferred by the periphery. Further, we find that optimal profits of players in the core is convex and decreases with the number of links up to a unique point. From a broader perspective, we try to understand the impact of network structure on supply chains which has been mostly missing from the literature.

References

Hierarchical Identification of Key Firms of International Tax Avoidance in Global Ownership Network

Tembo Nakamoto¹, Abhijit Chakraborty², and Yuichi Ikeda³

¹ Graduate School of Advanced Integrated Studies in Human Survivability, Kyoto University, 1, Yoshida-Nakaadachi-cho, Sakyo-ku, Kyoto, Japan, nakamoto.tembo.75w@st.kyoto-u.ac.jp, ikeda.yuichi.2w@kyoto-u.ac.jp,
² Graduate School of Simulation Studies, University of Hyogo, 7-1-28, Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo, Japan, abhiphyiitg@gmail.com

1 Introduction

We examine whether firms with a high risk of international tax avoidance can be detected from a global ownership network. Corporate ownership is studied using the methodology of network analysis [1]. In addition, research on international tax avoidance from the viewpoint of jurisdictions has been conducted using this concept [2], but it is indispensable from our research interest to analyze international tax avoidance from the viewpoint of an individual multinational group for clearing the relationship between ownership structure and tax. In our study, we propose a model to analyze international tax avoidance from the viewpoint of individual multinational groups and apply our model to representative multinational groups in order to verify the effectiveness of our model. After the effectiveness is verified, the analysis will be performed for all global ownership data consisting of 30 million firms for clarification characteristics of multinationals holding identified firms and relationship between the global ownership network and the identified firms.

2 Hierarchical Base and Conduit Firms Identification Model

Multinationals are often said to utilize intermediary firms for tax purpose. The intermediary firms are key firms to international tax avoidance and classified as base or conduit firms. The base firms gather and accumulate multinationals’ profits around the world through the conduit firms, which can be said that the base and the conduit firms have hierarchical relationships. Moreover, base firms using differences in criteria for tax residence in each jurisdiction have drawn attention these days. A basic idea of our research is to try to identify such intermediary firms by calculating base or conduit centralities hierarchically, taking each jurisdiction’s different criterion into consideration.

We define the base and the conduit centralities, focusing on the outline of the network and putting a bold premise that multinationals’ profits (especially dividends) flow along ownership links because our dataset records only less than half of the financial information as for the multinational groups targeted. Since base firms accumulate profits from many affiliates, we thought that base firms should have many ownership links
through which profits flow. Based on this idea, the base centrality is defined as $B_a = (k^{in}_a - k^{out}_a)/(k^{in}_a + k^{out}_a)$ and the conduit centrality is defined as $C_b = k^{in}_b/(k^{in}_b + k^{out}_b)$. Here, $a$ denotes each firm, $b$ denotes a firm of which $b$ holds shares. $k^{in}_a$ represents the number of firms with which $a$ holds shares, $k^{out}_a$ represents the number of firms holding shares of $a$, and $k^{in}_b$ represents the number of firms with which $b$ holds shares. The total of $k^{in}_b$ and $k^{out}_b$ reflects a scale of $a$ in its multinational group.

The algorithms to calculate centralities hierarchically as follows. At first, we calculate the base centrality for each firm $a$, and if $B_a$ is larger than 0, then we calculate $C_b$ and identify the layer. These calculation are conducted hierarchically. When $a$ has shares of $b$ whose $C_b$ are larger than 0, we regard $a$ as base-firm and $b$ whose conduit centrality is larger than 0 as conduit-firm. When a firm is identified as both of base-firm and conduit-firm, we count the firm as base-firm.

3 Datasets and Analysis

The data used in this study is “Orbis 2015 database” by Bureau Van Dijk. This dataset comprises financial information and ownership structures about 30 million firms across more than 20 jurisdictions, and the newest recorded data is about 2014.

We analyzed six multinational groups, which are Amazon, Apple, Google, Microsoft, Starbucks, and Uber, in order to verify our model’s effectiveness, because the names of key firms which is important for international tax avoidance are reported in the articles [3]. In our analysis, it is considered as an affiliate that has a holding relationship of more than 10% where most of the tax treaties start to lighten their withholding tax.

![Network relating to Google](image)

(a) Google group  (b) Layer 1  (c) Layer 2  (d) Layer 3

**Fig. 1.** Network relating to Google. (a) includes all Google’s affiliates and (b) - (d) include base-firms colored by red and conduit-firms colored by yellow. Rounded rectangles colored by green, orange, and blue show hierarchical relationships among (a), (b), (c) and (d).
Table 1. Number of base and conduit firm identified

<table>
<thead>
<tr>
<th>Multinationals</th>
<th>Total</th>
<th>Base</th>
<th>Conduit</th>
<th>Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>312</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Apple</td>
<td>152</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Google</td>
<td>190</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Microsoft</td>
<td>266</td>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Starbucks</td>
<td>274</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Uber</td>
<td>18</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

4 Results

Figure 1 shows the ownership network of Google and the network of their base-firm and conduit-firm, as an example. We identified 3 layers about Google and it shows importance of a hierarchical analysis. Table 1 shows the number of firms identified as base-firm and conduit-firm by our model (at the column of Base and Conduit) and the number of firms known as base or conduit firms within our identifications (at the column of Reported). Total means the number of affiliates in each multinationals group. Taking Google as an example, 3 firms known as base or conduit firms are included within 9 firms we identified as base-firm and conduit-firm. Our model detects 2 or 3 known base or conduit firms. These results show that our model is capable of identifying the key firms of international tax avoidance.

5 Conclusion

The effectiveness of our model is confirmed because key firms identified through our model. Currently, the analysis for the whole of our dataset recording global ownership is in progress. We will present the result in COMPLEX NETWORKS 2018 conference.

Acknowledgements

The present study was supported by the Ministry of Education, Science, Sports, and Culture, Grants-in-Aid for Scientific Research (B), Grant No. 17KT0034 (2017-2019) and by MEXT as Exploratory Challenge s on Post-K computer.

References

Hierarchical structure of the cryptocurrency market

Uta Pigorsch and Marc Sabek

University of Wuppertal, Wuppertal NRW 42119, Germany,
pigorsch@uni-wuppertal.de, sabek@uni-wuppertal.de
https://www.statistik.uni-wuppertal.de

1 Introduction

Since the Bitcoin price reached its all-time high of $19,665.39 in December 2017 cryptocurrencies have attracted the attention of mainstream media and their popularity has increased. As reviewed by [1] it is likely that Bitcoin and alternative blockchain-based tokens emerge as their own asset class that may become an interesting investment alternative and diversification instrument. There exists a plethora of cryptocurrencies and the question arises whether and how they are related to each other. In this paper we therefore aim at analysing the network topology of cryptocurrencies in order to derive a meaningful taxonomy of economic or common factors affecting specific groups of cryptocurrencies.

For financial data a series of papers studies the network topology by means of the minimal spanning tree (MST), c.f. [2–6], where a network is created by linking two companies based on their stock return correlation. [2, 5] show that the resulting network structure exhibits groups corresponding to industry sectors. Moreover, using the MST methodology [4] reveals that popular models of portfolio dynamics fail to capture topological properties of real financial markets. Furthermore, the MST of such a stock network is scale-free [5, 6], i.e. insights gained from studying scale-free networks carry over. We adopt these methods to analyse the cryptocurrency network.

2 Data and Method

Our dataset ranges from 01/01/2014–20/08/2018 ($T = 1693$ trading days) and consists of daily closing prices of $n = 100$ cryptocurrencies, $P_i(t)$, $i = 1\ldots n$ and $t = 1\ldots T$, exhibiting the highest market capitalization at the beginning of August 2018 [7]. In our analysis we consider log returns of the $i$-th currency at time $t$, i.e. $r_i(t) = \log(P_i(t)/P_i(t-1))$. In order to assess the degree of similarity of the return series of a pair of currencies the Pearson correlation coefficient $\rho_{ij}$ is computed [9]

$$\rho_{ij} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{(\langle r_i^2 \rangle - \langle r_i \rangle^2)(\langle r_j^2 \rangle - \langle r_j \rangle^2)}}$$

(1)

where $\langle \cdot \rangle$ indicates an average over time. An appropriate metric is formed by using (1) and computing $d_{ij} = \sqrt{2[1 - \rho_{ij}]}$ [2]. Subsuming the distances $d_{ij}$ into a $n \times n$ matrix $D$ yields an adjacency matrix of the cryptocurrency network, that is used to construct the
MST via Kruskal’s algorithm [10]. We compute the MST for each year of our sample based on the corresponding cumulative moving correlation. The resulting MSTs are depicted in Fig. 1.

It is reasoned by [2, 5] that a suitable space for linking stocks in a financial network is an ultrametric space, in particular the space endowed with the subdominant ultrametric distance \( d_{ij}^* \), which is the maximum value of a single step occurrence if moving through the shortest path from \( i \) to \( j \). This metric is opted for due to its simplicity and its preferable properties, i.e. it is directly obtainable from the MST and a hierarchical tree (HT) can unambiguously be built without further assumptions [11]. Thus, we also use it for linking cryptocurrencies in an agglomerative clustering procedure with single linkage [12] to produce the HT in Fig. 2.

3 Results and Discussion

Our analysis shows, that the MST in 2018, presented in Fig. 1, exhibits a similar tiering as the stock networks studied in [2, 4], which implies that standard models of portfolio dynamics will also fail to reproduce the topological properties of the cryptocurrency network, such that more sophisticated models are needed.

Moreover, from Figs. 1 and 2 the hierarchical structure of the cryptocurrency market becomes evident. The HT reveals the existence of four groups of cryptocurrencies namely (i) VET, VERI (ii) ONT, CTXC (iii) KMD, ELF and (iv) ETH, ICX and WAN.

Furthermore, a power law has been fitted to the degree distributions of the MSTs resulting in exponents \( \alpha = \{2.4, 3.1, 4.8, 2.3, 2.6\} \) for the years 2014–2018. This implies that the cryptocurrency networks’ MSTs are also scale-free.
Hierarchical tree (HT) of the ultrametric space associated with the minimal spanning tree from 2018 of Fig. 1. The leaves of the tree are marked by the respective cryptocurrency’s tick symbol. The ultrametric distances at which currencies are clustered are given by the y-axis.

In summary, we thus find that there exist topological similarities between stock returns and cryptocurrency exchange rate networks. This result may not have been expected a priori since stocks are naturally related to each other via common market and industry specific factors, similar firm characteristics etc., while for cryptocurrencies such a relationship is not that obvious. We will therefore further investigate the meaningfulness of the emerging taxonomy from an economic point of view. Similarities in technology, attractiveness or other macrofinancial factors may be important determinants.

References

7. Data provided by Coinranking and CryptoCompare
Point-process network of firms bankruptcies

Yoshi Fujiwara^1 and Shinsuke Koyama^2

^1 Graduate School of Simulation Studies, University of Hyogo, Kobe, Japan, yoshi.fujiwara@gmail.com,
^2 Department of Statistical Modeling, Institute of Statistical Mathematics, Tokyo, Japan, skoyama@ism.ac.jp

1 Introduction

Point-process network is a statistical estimation about hidden relationships among elements from observation of events that took place in space and time (see Fig. 1). Examples include earthquake and its aftershocks in seismology [1–3], neural spiking in neuroscience [4], and human interactions in social activities [5, 6]. Such a phenomena can be observed also in economic systems [9]. For example, one can observe chained failures. As a typical example, when one or more of a firm’s customers go into bankruptcy, the firm suffers sudden drop of sales possibly resulting in chained bankruptcy. See [10, 11] for empirical studies.

We utilize the widely applied methodology of multivariate self-exciting point-process [7, 8] in order to find background rate of failures, due to some entire economic shock, and also to uncover hidden network structure among industrial sectors, namely, suppliers and customers relationship. See [12] for the model of event cascades using networks of additive count sequences.

![Fig. 1. Schematic illustration of multivariate self-exciting process and an estimated network](image)

2 Data and Method

We use an exhaustive list of bankruptcies in Japan during the years from 2013 to 2016. The number of bankruptcies is more than 28,000 observations. One can estimate the
probability of bankruptcy as 1% per year; i.e. $10^3$ firms goes into bankruptcy among $10^6$
of all active firms in Japan. The database includes information about date of bankruptcy,
its size (debt when bankrupted), and the failed firm’s industrial sector.

Let $n_i(t)$ be the event number of failures at time $t$ and space (sector) $i$. See Fig. 2
for the space-time diagram of failure events with their sizes. We assume that it follows
a Poisson or other types of counting process with instantaneous rate for the occurrence
of event, $\lambda_i(t)$. $\lambda_i(t)$ depends on past history of events and sizes of the failures; self and
mutually exciting process.

$$\lambda_i(t) = \mu_i(t) + \sum_j \sum_{s \geq 0} K_{ij}(s) z_i(t-s) \quad (1)$$

where $\mu_i(t)$ is a background rate. $K_{ij}(s)$ represents an influence from sector $i$ to $j$ with
time delay $s$. $z_i(t)$ is the total sum of failure sizes in sector $i$ at time $t$. Size of event
follows a Pareto distribution, as is well-known. We assume that the kernel $K_{ij}(s)$ can be
expressed by

$$K_{ij}(s) = a_{ij} \phi_{ij}(s) \quad (2)$$

where $a_{ij}$ is an adjacency matrix of the hidden relationship, and $\phi_{ij}(s)$ is a decaying
function of time delay $s$. One can calculate the probability for the observable of events
and their sizes, and the likelihood function to estimate the parameters in the model.
Moreover, one can perform statistical estimation to do “de-clustering”, i.e.

$$n_i(t) = y_i(t) + \sum_j \sum_{\ell} y_{ij\ell}(t) \quad (3)$$

where $y_i(t)$ represents spontaneous events due to the background rate of failures, and
$y_{ij\ell}(t)$ corresponds to triggers from $j$ to $i$ with time delay $\ell$. In summary, from the
observables $n_i(t)$ and $z_i(t)$, one can estimate $K_{ij}$ under a parametric assumption for the
function $\phi_{ij}(t)$, and also estimate $y_i(t)$ and $y_{ij\ell}(t)$ as a result. See [12] for complete
discussion.

3 Results

The contribution of this paper is:

- We show that cascades in bankruptcies can be modeled with a multivariate self-
  exciting point-process and its network by employing large empirical data of chained
  bankruptcies (Fig. 2). The estimated network $a_{ij}$ implies that there is an influence from
downstream to upstream of production among firms. This holds especially for manu-
facturing sector. Construction has relatively weak ties, while retail and wholesale has
strong links, implying vulnerable paths of chained failures.

- We establish a method to estimate probabilities for background level of occurrence
  and lead/lag relationship among events for each occurrence in time and space, and also
to estimate the temporal change of critically of the process so that one can monitor the
stability of economic systems.
Fig. 2. Space (industrial sector) and time (daily) diagram of bankruptcies which took place in Japan during fiscal year 2014. Size of each circle is logarithm of bankruptcy size (debt when bankrupted).

References

Anomaly detection in networks using spectral methods and network comparison approaches

Andrew Elliott¹, Paul Reidy³, Milton Martinez Luaces³, Mihai Cucuringu¹,², and Gesine Reinert¹,²

¹ Alan Turing Institute
² Department of Statistics, University of Oxford
³ Accenture Plc.

1 Introduction

Financial fraud is a global challenge. According to [3], financial fraud losses in the UK payment industry across payment cards, remote banking and cheques amounted to £768.8 million in 2016. Some of this fraudulent behaviour, such as money laundering, may manifest itself through unusual patterns in financial transaction networks. In such networks, customers are denoted by vertices, and two vertices \( u \) and \( v \) are linked by a directed edge if there is a money transfer from \( u \) to \( v \); the edge is annotated with the transferred amount. Unusual patterns could include long heavy paths and large cliques, with these patterns changing and many patterns likely to remain undetected. Not only is detecting such anomalies in networks of financial transactions a pressing problem, but there are also applications in other domains such as cyber security. Reviews on anomaly detection are, for example, Refs [2, 4].

Here we propose a novel method which is based on a combination of two approaches, namely the NetEMD network comparison method from Ref. [6], and spectral localisation statistics. Our method computes 140 features for each network and uses these features to classify nodes as anomalous or non-anomalous. These features are processed in two different ways. Our first method, which we call the Feature Sum method, aggregates the features as an unweighted sum. Our second method, which we call the Random Forest method, is a classifier constructed using a synthetic model which is inspired by anti-money laundering, and described in the next section (see Fig. 1A for the anomalies). We measure performance by the average precision score, for example see [5]. We also graphically consider Precision and Recall independently against the ranking in our statistic (results not shown).

2 Data and methods

We use two synthetic data sets. The first data set is a collection of weighted ER graphs. This data set is split up into a training set for the Random Forest method, and a test set for the Random Forest method. The second data set is used to evaluate the performance of the methods.

Data Set 1: weighted ER graphs We construct anomalous networks which are based on a directed ER graph model with 10,000 nodes using \( p = \frac{1}{1000}, \frac{2}{1000}, \ldots, \frac{5}{1000} \). In each network, anomalies as detailed in Fig. 1(A) are planted at random, so that between 25 and 400 nodes are anomalous. Each edge in the anomalous subgraph is given
a weight which is uniform \((w, 1)\) distributed, with \(w = \frac{990}{1000}, \frac{991}{1000}, \ldots, 1\), whereas each other edge is given a weight which is uniform \((0,1)\) distributed. We only consider parameter combinations \((p, w)\) for which the expectation of the planted anomalies in the ER graph are less than 1. For each such combination of parameters we construct 100 networks. The first 70% of these networks form the training set and the remaining 30% the test set.

**Data Set 2: Accenture Data** The second data set comes from a generating model developed by Accenture, to which the Turing team had limited access during the main development of the method. The model generated 100 networks each with 55,000 nodes, of which 49 are anomalous.

**Methods.** In total, 140 node-based features are evaluated on the networks; these include basic statistics such as degree and strength, but also features which aggregate triad counts, and the top 20 eigenvectors of a set of graph-based matrix operators which are associated with the network, such as the popular graph Laplacians. Many of these features take into account what would be expected at random.

For computational considerations, we split the network into smaller communities. This split has as drawback that long paths may be cut, and thus no longer detectable. Therefore, we enrich our methods by features which are specifically designed to find paths in networks (not described here).

The *Feature Sum* method then uses the sums all of these features to rank the nodes, and the higher its rank, the more anomalous the node is assumed to be. Our second method trains a random forest on the training data using the standard parameters in the RandomForestRegressor from Sklearn. We fit a separate random forest to each parameter regime and construct a consensus ranking, resulting in 44 features. Among these 44 features, basic statistics and measures based on localisation measures appear prominently, with statistics based on strength and short path lengths being well represented amongst the most important statistics. This is not surprising, as it is likely that the basic statistics would do well on the easier structures.

### 3 Results

**Performance on the ER Test Set:** For Data Set 1 we report the performance of our methods and compare them to *Oddball*, a popular alternative anomaly detection method from Ref. [1].

The results for Data Set 1 are displayed in Fig. 1(B). The median average precision score of both *Feature Sum* and the *Random Forest* outperforms random allocation and compares favourably to Oddball, with the fitted model achieving near perfect scores in some parameter regions. All three methods considerably outperform random classification. Even the *Feature Sum* method, which does not use any training networks, performs favourably to Oddball, indicating that the underlying features have sufficient information to capture to anomalies.

**Performance on the Accenture Model:** With respect to the distributions of the average precision score in this model (Fig. 1C), both the *Random Forest* and the *Feature Sum* methods outperform random allocation. Moreover, the *Feature Sum* method outperforms the *Random Forest* method on this data set, indicating that the random forest may have been over-fitted to the task of anomaly detection in general networks.
Fig. 1. A The types of anomalies which are embedded in Data Set 1. B Data Set 1 results of the Random Forest, the Feature Sum methods, and Oddball. C Data Set 2 results of the Random forest and the Feature Sum method.

Finally, with respect to the precision (figure not shown), the Feature Sum method performs better than Random Forest on the nodes with top Feature Sum ranks, with precision close to 1. Within the first 1024 nodes (1.8% of nodes), both methods have an average recall of ≈ 0.93, indicating good performance of both models.

4 Discussion and conclusion

Our work provides a general anomaly detection framework which can outperform the popular Oddball method. The method can be specialised to achieve higher performance or faster running time via a reduced feature set.

References

Triadic and quadratic closure as mechanisms for inter-firm transactional network dynamics

Tomomi Kito¹ and Petter Holme²

¹ Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 165-8555, Japan, kito@waseda.jp
² Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8503 Japan, holme@cns.pi.titech.ac.jp

1 Introduction

In general, complex networks are interconnected systems whose interconnections are neither random, nor completely regular. The deviation from randomness—the network structure—can, in most cases, be attributed to a contingent growth process. When the positions of nodes in the network affect the performance of agents and the agents have an opportunity to influence the attachment and deletion of links, the network structure will be strongly influenced by the agents trying to optimize their positions. One celebrated example is preferential attachment mechanism, where a propensity for new nodes to attach to high-degree nodes induces a fat-tailed degree distribution. A yet more classic mechanism for the emergence of network structure is triadic closure, where the addition of a link completes a triangle [1, 2]. There can be many reasons completing a triangle—or, more generally, forming a short cycle—can optimize the network position. In this work, we first show that triadic and quadratic closure exists in the evolution of a large, inter-firm transactional network. Then we proceed to investigate the microscopic causes of these closure mechanisms.

2 Data

In this study, we use comprehensive annual data on inter-firm transactional relationships (i.e. which firm sells product(s) to which) in Japan, from 2000 to 2015. The data was provided by Teikoku Databank, Ltd., which is a leading credit agency in Japan, gathering information on most of the Japanese firms. The data is gathered through investigation of financial statements and corporate documents, as well as by interview-based surveys. For every year, the number of firms is about 1 million, and the number of transactional relationships is 3–4 million.

3 Methods and Results

First, we investigate the graph distance between firm i and j in year t, d_t(i, j) (i.e. the number of links in the shortest path between i and j), when we observe a new trade
relationship between $i$ and $j$ in year $t + 1$. This means that firm $i$ sells products to firm $j$ at year $t + 1$, but did not sell anything to $j$ at year $t$.

Fig. 1 shows the average value of $d$ for nodes that will be connected in the next year, relative to the corresponding value for random pairs of nodes. Note that in this calculation, the network is treated as undirected. We see that there is a significant tendency for triadic and quadratic closure (new trade relationships happening between nodes separated by two or three links) [3]. Triadic closures are more than 100 time more common than expected by chance, while the corresponding number for node-pairs three steps away is around 20. When nodes are separated by more than three links, this effect is gone. To get a more detailed picture, we further investigate the cases of $d = 2$ and $d = 3$, with the directionality of trade relationships taken into account. Given a new trade relationship, $\gamma = (i, j, t)$, there are four possible paths of length two, and eight possible paths of length three (see Fig. 2). We can code the paths of length two as OI, OO, II and IO—OI meaning that the path from $i$ to $j$ goes via an intermediate $k$, such that there is a trade relationship from $i$ to $k$ and from $j$ to $k$.

![Fig. 1](image.png)

**Fig. 1.** (A) Graph distance between a given pair of nodes. (B) Relative frequency of graph distances $d$, when the network is treated as an undirected graph. Note the logarithmic y-axis.

Fig. 2 shows the values of $f_{OO}$, $f_{OI}$, $f_{IO}$ and $f_{II}$ for new trade relationships, relative to random node pairs. We find the OI and OO paths overrepresented. OI corresponds to a firm starting to sell to a provider of a previous customer; OO corresponds to a firm starting to sell to the customer of a customer. In the same manner, we also investigated paths of length three. We see a consistent overrepresentation of OIO cycles—completing a square where the origin $i$ of $\gamma$ is in the same positions as the node two steps away. Somewhat intriguingly, there is also an overrepresentation of III cycles, contradicting the idea of a strictly hierarchical organization.
Fig. 2. Path statistics for directed networks. Panel (A) shows statistics for the four possible types of directed triadic closure; Panel (b) shows the corresponding plots for paths of length four.

4 Conclusion

We have studied a large data set of the trade patterns of firms. This data set is time-resolved enabling us to study the formation of new links. We found that indirect network effects like triadic and quadratic closure are important. For directed networks there is an overrepresentation of situations where the common suppliers to one firm starts trading with each other. We are currently examining the triadic and quadratic closure mechanisms closer. As mentioned, since the firm networks are formed by careful decision processes in the firms, we can expect their structure to be a result of an optimization process. While these decisions have been studied for individual actors in the management literature (e.g. [4]), their relation to the triadic and quadratic closure mechanisms are unknown. To understand these mechanisms, we need to understand how the node(s) previously connecting the nodes of the new link \((i, j)\), are related to \(i\) and \(j\). If they are, for example, far from \(i\) and \(j\), \(i\) or \(j\) might have acquired the abilities of the intermediate node(s) to cut transportation costs. We plan to perform similar analyzes to study the profit, industrial categorization, and firm size of the intermediate firms.

References

Effects of industrial and geographical proximities of firms on the Japanese inter-firm network

Akihiro Hamamoto¹ and Tomomi Kito²

¹ Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tenodai, Tsukuba, Ibaraki, Japan,
P s2017515@s.tsukuba.ac.jp,
² Department of Industrial and Management Systems Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo, Japan,
kito@waseda.jp

1 Introduction

In economics, a growing strand of the literature has highlighted the role of an economy’s input-output (I/O) network – a network constructed based on publicly available data on inter-industrial exchanges of products – to obtain insight into shock propagations/amplifications [1][2]. However, in an I/O network, firms belonging to a given industry are all agglomerated as one node, regardless of other attributes (such as geographical location, size, connectivity to others, …). On the other hand, it is well known that firms often form regional clusters, but most existing studies have focused on specific regions or industries (e.g. [3][4]).

How many firms’ industrial and geographical proximities impact the composition and structure of firm clusters? In this study, we pursue this question by applying network analysis methods to large-scale empirical data on the inter-firm transactional network in Japan.

2 Data

In this study, we used empirical data on inter-firm transactional relationships in Japan, in 2015. The dataset was provided by Teikoku Databank, Ltd. (TDB), one of the leading private credit research company. Their uniquely-collected rich data has been to date used for various research. The number of transactional relationships the dataset covers is 4,153,912, and firms involved in these relationships are in total 1,136,203.

3 Methods, Results, and Future Work

We first applied to the data the community detection algorithm proposed by [5], which categorized more than 90% of the firms into 7 communities. (Note that we also conducted another community detection method [6], which resulted in similar communities with no significant difference.) The number of firms and the number of within-community edges in each community are summarized in Table 1-(a).
Table 1. (a) Size of each community, (b) Distribution of their industries of firms in each community.

(a) Size of each community:

<table>
<thead>
<tr>
<th>Community</th>
<th>Firms</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>464333</td>
<td>2942690</td>
</tr>
<tr>
<td>2</td>
<td>199430</td>
<td>1546627</td>
</tr>
<tr>
<td>3</td>
<td>144393</td>
<td>886827</td>
</tr>
<tr>
<td>4</td>
<td>93069</td>
<td>497605</td>
</tr>
<tr>
<td>5</td>
<td>67928</td>
<td>468406</td>
</tr>
<tr>
<td>6</td>
<td>43304</td>
<td>840582</td>
</tr>
<tr>
<td>7</td>
<td>47231</td>
<td>371346</td>
</tr>
</tbody>
</table>

(b) Distribution of industries of firms in each community:

<table>
<thead>
<tr>
<th>Industry</th>
<th>Community</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>0.39</td>
<td>0.41</td>
<td>0.20</td>
<td>0.07</td>
<td>0.11</td>
<td>3.25</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Forestry</td>
<td>0.95</td>
<td>1.13</td>
<td>0.98</td>
<td>0.96</td>
<td>0.16</td>
<td>0.90</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>Fishing</td>
<td>0.76</td>
<td>0.04</td>
<td>5.11</td>
<td>0.09</td>
<td>0.27</td>
<td>3.41</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>Mining</td>
<td>1.28</td>
<td>1.13</td>
<td>0.16</td>
<td>0.09</td>
<td>0.89</td>
<td>2.61</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>Construction</td>
<td>1.13</td>
<td>1.73</td>
<td>0.14</td>
<td>0.23</td>
<td>0.10</td>
<td>1.14</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>Manufacture</td>
<td>1.34</td>
<td>0.41</td>
<td>1.04</td>
<td>0.86</td>
<td>1.25</td>
<td>0.61</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Wholesale</td>
<td>0.84</td>
<td>0.53</td>
<td>1.85</td>
<td>0.83</td>
<td>2.94</td>
<td>0.50</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>Finance</td>
<td>1.49</td>
<td>0.41</td>
<td>0.44</td>
<td>0.72</td>
<td>0.38</td>
<td>0.40</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Real estate</td>
<td>0.69</td>
<td>0.68</td>
<td>0.92</td>
<td>0.88</td>
<td>0.60</td>
<td>0.88</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Transportation</td>
<td>1.09</td>
<td>0.63</td>
<td>1.50</td>
<td>0.84</td>
<td>0.33</td>
<td>1.67</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>Public Service</td>
<td>1.16</td>
<td>1.15</td>
<td>0.25</td>
<td>0.36</td>
<td>0.23</td>
<td>1.18</td>
<td>1.33</td>
<td></td>
</tr>
<tr>
<td>Service</td>
<td>0.82</td>
<td>0.84</td>
<td>0.91</td>
<td>1.78</td>
<td>0.56</td>
<td>1.01</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Government</td>
<td>1.20</td>
<td>1.37</td>
<td>0.08</td>
<td>0.09</td>
<td>0.24</td>
<td>1.44</td>
<td>1.04</td>
<td></td>
</tr>
</tbody>
</table>

Table 1-(b) shows the distribution of industries to which firms in each community belong to. In this table, each value indicates the ratio of the number of firms in the community belonging to the industry, compared to when firms are randomly distributed (red cells show high values, whereas blue show low values). That is, in Community 3, for example, there are firms that belong to agriculture 4.28 times more than expected. As this table shows, some communities seem to be comprised of firms in particular industries, while others are not.

In the same manner, we investigated how firms in each community are geographically distributed, and found that some communities are comprised of regionally-clustered firms, while others show no significant regionality (see Fig.1-(a) in Communities 2 and 3, as examples).

In order to further investigate the composition and structure of the communities, we conducted Functional Cartography [7], which identifies nodes' universal roles in the communities. Moreover, we conducted the multiple core-periphery (CP) pair detection, proposed by [8]. Some part of the results are shown in Fig.1. The method [8] identified three CP pairs in Communities 2 and 3, respectively. The geographical distributions of firms that were identified as 'core' and 'periphery' are visualized in figures (b, c, d). Our findings through this analysis can be summarized as follows:

1. Firms' industrial and geographical proximities indeed impact on the formation of clusters, but in complex ways, resulting in heterogeneous compositions and structures of communities.
2. Hub firms (identified by [7]) are not necessarily large-scale firms, and not those with high 'coreness' (identified by [8]).
3. Core firms (by [7]) are geographically spread across the country, whereas firms in peripheral areas of multiple CP pairs are somehow locally clustered together (see Fig.1 (b, c, d), for both communities).

Some of the above results (and more results we obtained that are not explained here)
require a deeper investigation. We are currently working on it, with the hope of elucidating the complex mechanism of inter-firm business clusters.

Fig 2. Examples of the results obtained via community detection [6] and multiple core-periphery pair detection [8]. (a) Geographical distribution of firms in the community. Red cells mean the dense concentration of firms, whereas blue cell mean sparse, compared to the random distribution. (b, c, d) Distribution of ‘core’ and ‘peripheral’ firms in each C-P pair, respectively. The figures show that core firms are scattered throughout the country, whereas peripheral firms are concentrated in some areas.

References
The Struggle for Existence in the World Market Ecosystem

Viviana Viña-Cervantes\textsuperscript{1}, Michele Coscia\textsuperscript{1,2}, and Renaud Lambiotte\textsuperscript{1,3}

\textsuperscript{1} Naxys Department of Mathematics, University of Namur, Namur, Belgium,\textsuperscript{2} Computer Science, University of Copenhagen, Copenhagen, Denmark,\textsuperscript{3} Mathematical Institute, University of Oxford, Oxford, England.

1 Introduction

Global trade can be considered as a complex system, whose sophisticated behavior emerges from its many interacting parts countries exporting products in different importing markets. Different economic complexity indexes have proven to be incredibly successful in predicting future economic growth [2–5]. Here, we enrich the literature on complexity and economic development by further investigating its relationship with ecology. We draw relations among countries by inferring potential competition among them across time. We see a pair of importer-product, for instance the car market in the US, as an evolving trade “niche”, with exporters appearing and disappearing like fit and unfit organisms in an ecosystem.

![Fig. 1](image)

In our analysis, the fitness of an economy in a niche correlates with its ability to displace (out-compete) unfit economies. We test this theory by creating a competition network, connecting country a to country b if a’s appearance in a market preceded b’s disappearance, as illustrated in Figure 1. Since we have different products and different years in which these relationships can be established, we use a multilayer network model.

1.1 Inferring Competition Relationships

These is the procedure of how a single edge in our multilayer directed network is established.

1. Detect whether there is an anti-correlation between the export patterns of the two countries ($\delta$);
2. Detect whether one of the two countries appeared from the market, while the other disappeared ($\kappa$);
3. Detect whether the disappearing country did not reappear in the market immediately after the event ($\lambda$).

1.2 Detecting Roles

We now turn to the detection of node roles in the multilayer competition network. We follow closely the methodology delineated in [1]. In that paper, Cooper and Brahona propose to group nodes according to their role in the network, defined in terms of the overall pattern of incoming and outgoing flows. According to this, we expect to find three categories of countries: out-performing, displaced and transitioning. According to this, we expect to find three categories of countries: \textit{out-performing countries}, countries with high values for the out-degree roles and low ones for in-degree, \textit{displaced countries}, with low values for the out-degree roles and high ones for in-degree roles and \textit{transitioning countries} with comparable values for both roles.

2 Results

2.1 Competition Network Statistical Analysis

The fundamental assumption of this paper is that the competition network that will allow us to predict an exporter’s future performance in the global market. If a country can out-compete many other countries in a product, then it is expected to export more of that product. The first question one might ask is: why do we need to calculate node roles? The number of times an exporter out-competes its rivals is simply its out-degree. One of the reasons why this is not the case is that out- and in-degree in the competition network are highly correlated. Figure 2 shows the out- and in-degree correlation. We can see that both distributions are very similar.

![Fig. 2: (Left) Out- and in-degree distributions. (Right) The relationship between out-degree (x-axis) and in-degree (y-axis). Each observation is a country.](image-url)
2.2 Role Clusters

Many topological properties of the multilayer networks are dependent on our choice of parameters. We investigate the direct effect on clustering quality of the three parameters $\delta$, $\kappa$ and $\lambda$. For each combination of parameter we calculate the average cosine distance between a country and the cluster template to which it is the most similar.

2.3 Prediction

Once we fix $\delta$, $\kappa$ and $\lambda$ such that we obtain the lowest residual energy (i.e. average cosine distance), we can perform a simple predictive task. We calculate the clusters using exclusively data from a given decade, say 1971 to 1980. Then, we look at the exports of each country in that product in the next decade from 1981 to 1990. We calculate the slope of the decade trend, in this way, we have for each country its competition network cluster for a decade and its corresponding export growth in the following decade. We then calculate the mean export growth rate for each country cluster.

In our best case, the out-competing cluster was able to correctly capture all the eleven fastest growing countries in the manufacturing sector in the 80s. The fact that we cannot predict the growth in natural resources is not crucial, as it makes little sense to plan a development strategy by aiming at discovering oil. Countries are more interested in developing capabilities for sustainable growth. Given that success in exporting a product can be a telltale sign of other societal indicators such as income inequality and poverty traps, the reach of our methodology can span multiple potential applications.

References

Influence of network correlation structure in multivariate time series forecasting

Luis Ospina-Forero\textsuperscript{1,2} Mihai Cucuringu\textsuperscript{2,3} and Gesine Reinert\textsuperscript{2,3}

\textsuperscript{1} University of Manchester, Manchester Business School, Manchester M13 9SS, UK, luis.ospina-forero@manchester.ac.uk.
\textsuperscript{2} The Alan Turing Institute, 96 Euston Rd, Kings Cross, London NW1 2DB, UK
\textsuperscript{3} University of Oxford, Department of Statistics, 24-29 St Giles’, Oxford OX1 3LB,UK

1 Introduction

Multivariate time series have recently been addressed from the perspective of network analysis, by creating networks through thresholding correlations between individual time series. In these networks, time series are the nodes, and two time series are connected by an edge if the correlation between them exceeds a given threshold. In contrast to the field of Bayesian networks, causal analysis and classical statistical time series forecasting (e.g. \cite{1, 3}), the network analysis methods used for analysis of financial time series have mainly focused on simple network summary statistics \cite{5}, a recent exception being \cite{2}. In this work, we explore the potential of network comparison methods in multivariate time series forecasting.

Using a sliding window for calculating correlations, multivariate financial time series give rise to a time series of networks. To forecast the behaviour at time $T + h$, we compare the network at time $T$ to previously observed networks and use the closest networks from the past to predict the next time step via a weighted average of the response associated to these past closest networks. The network comparison methods which we employ (GDDA, GCD, and Netdis) are non-parametric and based on the occurrence of small connected subgraphs, as shown in Figure 1 (left); see \cite{4} for a review.

2 Data and Method

Data. We consider two data sets of multivariate time series. The first one, S&P 500, is based on the \textit{The Standard & Poor’s 500} stock market index. The second, BoE, is composed of multiple macro-economic variables from the UK. Time series with missing values are omitted, and perfectly correlated variables are collapsed. The BoE and the S&P 500 datasets were obtained from www.bankofengland.co.uk, and Yahoo Finance respectively. Table 1 shows a summary of each filtered data set.

<table>
<thead>
<tr>
<th>Time points</th>
<th>Resolution</th>
<th>N. of Variables</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>S&amp;P 500</td>
<td>3092</td>
<td>Daily</td>
<td>477</td>
<td>2003/01/02 2015/04/15</td>
</tr>
<tr>
<td>BoE</td>
<td>92</td>
<td>Yearly</td>
<td>76</td>
<td>1925 2016</td>
</tr>
</tbody>
</table>

Table 1. Number of time points, variables, and time range of the two data sets S&P 500 and BoE.

Method. To combine different variations of the overall method, we divide our forecasting framework into four stages: (1) Construction of correlation matrices, $C_i^t, t = 1, \ldots, T$; (2) transformation of the correlation matrices $C_i^t$ into networks $G_i^t$; (3) creation of a structure dissimilarity vector between pairs $(G_i^t, G_j^t)$, for all $1 \leq t < T$ via network comparison methods; (4) creation of the forecast for the multivariate vector at $T + h$. 

\begin{center}
\textsuperscript{7th} International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
\end{center}
(1) For a given centered multivariate series \( R = [R_1, R_2, \ldots, R_d] \in \mathbb{R}^{T \times d} \), of \( d \) univariate time series \( R_i = [R_{i1}, R_{i2}, \ldots, R_{it}] \), \( i \in \{1, 2, \ldots, d\} \), and \( T \) the number of time points, we find the observed correlation structure between the individual time series at time \( t \) based on the previous \( l \) observations, where \( l \) is the window size used for estimating the correlation. Here we explore \{Pearson, Kendall, Spearman, partial\} correlations, with window sizes \{25,50,100\} for S&P 500, and \{7,10,25\} for BoE.

(2) We threshold the correlations to obtain a network; the threshold is set so that the average degree \( d \) of the corresponding undirected networks is fixed; we explore average degrees \{3,5,10,15,20\}. This procedure gives a time series of networks \( \{G_t\} \) for each choice of these parameter settings. We also explore the number of past observations used in the weighted average forecast to be \{3,5,10\}.

(3) We use the network comparison methods \{GCD, GDDA, Netdis\} to: (i) select previous values \( R' = (R'_1, R'_2, \ldots, R'_d) \), which have a similar network dependence structure \( (G') \) to the one observed at time \( T \); (ii) to provide weights for the previously selected values, which are later used in a weighted average forecast. Figure 1 (right) provides an example of the type of information that can be extracted from comparing the dependence structure. Crisis periods such as the 2007-2008 global economic crisis and the UK’s economic recession of the 70s are clearly highlighted. Thus, this type of information could have the potential to produce warning signs of economic changes.

![Graphs](image-url)

**Fig. 1.** Left: subgraphs of size 2-4 used by GCD, GDDA, and Netdis (from [6]). Right: network comparison of the dependence structure of the BoE macroeconomic variables across time.

(4) We create the forecast as follows. For S&P 500, \( \tilde{R}^i_t = \log P^i_t - \log P_t^{i-1} \) denotes the 1-day log return of the price of stock \( i \), where \( P_t^i \) is the official market close price on day \( t \). The log-returns give rise to the market excess returns, \( R'_i = R'_i - R_{SPY}^{\prime} \), \( i = 1, \ldots, d \), where \( SPY \) denotes the S&P 500 index ETF. We benchmark our method against future returns \( F_{i,t+h}^h = \sum_{k=t+1}^{t+h} R_k^i \), \( h \in \{1,3,5,10,15,20,30\} \) denotes the length in days of for the return. Our forecasts are a weighted average \( \tilde{F}_{i,t+h}^h = \sum_{j \in N(t)} F_{i,j+t}^h w_{i,j} \), where \( N(t) \) is the set of \( k \)-nearest-neighbour historical correlation networks closest to the one at time \( t \), and \( w_{i,j} \) the corresponding weights. To evaluate the resulting out-of-sample forecast of the future returns \( F_{i,t+h}^h \) for a horizon \( h \), we use the average Profit and Loss statistic, \( \langle PNL \rangle \), and a \( z \)-score based on the deviation of the \( PNL \) from its mean (i.e., the Sharpe Ratio)

\[
PNL' = \frac{1}{d} \sum_{i=1}^{d} \text{sign}(\tilde{F}_{i,t+h}^h) F_{i,t+h}^h
\]

\[
z = \frac{\text{PNL}}{\sqrt{\frac{1}{T-b-1} \sum_{l = b}^{T-1} (\text{PNL}_l - \text{PNL})^2}}.
\]

When there is no signal in the data, the PNL and the \( z \)-score should be symmetric around 0. We also consider the top \( x\% \) strongest forecasts (i.e., percentile portfolios).
3 Results
We applied our forecasting method to the S&P 500 and BoE at horizon $h = 1$, using market excess returns for the S&P 500 and raw returns for BoE. Table 2 shows the $z$ scores of the top 25% performing variables (Z-score 25%); and of all variables (Z-score), for the best-performing combinations of methods and parameter settings. As a baseline, we also show the performance of a naive Up strategy that sets all forecasts to positive values at each step (a "Buy and Hold" portfolio). Table 2 illustrates that our method can produce better out of sample forecasts than the simple "Buy and Hold strategy. We also note that different network comparison statistics may need to be used for different data sets, as different network comparison methods capture different aspects of the subgraph configuration of networks [4]. Figure 2 shows the cumulative PNL across time, for the S&P 500, at various horizons.

### Table 2. Summary of the top 5 forecasting variations for S&P 500 (2009-2015) and BoE (1925-2016). We show Z-scores of the top 25% performing variables (Z-score 25%), and of the whole data set (Z-score). The Up strategy also shows quantiles 5% and 95%, of 25% of randomly selected variables for Z-score 25%. The network comparison methods are shown in bold.

<table>
<thead>
<tr>
<th>S&amp;P 500 dataset</th>
<th>Methods</th>
<th>Z-score 25%</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up strategy</td>
<td>Avgd15 - Spearman - 25 step w.</td>
<td>0.81</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Netdis - Neighb. 10</td>
<td>0.55</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Avgd 20 - P. cor - 100 step w.</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Netdis - Neighb. 3</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>Avgd 3 - Pearson - 50 step w.</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>Netdis - Neighb. 3</td>
<td>0.47</td>
<td>0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bank of England dataset</th>
<th>Methods</th>
<th>Z-score 25%</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up strategy</td>
<td>Avgd 3 - Pearson - 25 step w.</td>
<td>2.52</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>Netdis - Neighb. 5</td>
<td>2.45</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>Avgd 20 - Pearson - 25 step w.</td>
<td>2.45</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>GDDA - Neighb 5</td>
<td>2.35</td>
<td>2.35</td>
</tr>
<tr>
<td></td>
<td>Avgd 10 - Spearman - 25 step w.</td>
<td>2.35</td>
<td>2.35</td>
</tr>
<tr>
<td></td>
<td>GDDA - Neighb 10</td>
<td>2.35</td>
<td>2.35</td>
</tr>
</tbody>
</table>

Fig. 2. Cumulative PNL (S&P 500) for forecasting at various horizons (days), most of which exhibit a positive slope before the 2008 crisis, and a significant decrease in performance afterwards.

**Summary.** We have provided a forecasting method based on network comparison which has the potential to improve multivariate time series forecasting.

**References**

Finding the most central countries in the international exchange multiplex using the MultiRank

Giovanni Bonaccorsi\textsuperscript{1}, Massimo Riccaboni\textsuperscript{1}, Giorgio Fagiolo\textsuperscript{2}, Gianluca Santoni\textsuperscript{3}

\textsuperscript{1} IMT School of Advanced Studies, Lucca 55100, Italy.
\textsuperscript{2} Sant’Anna School of Advanced Studies, Pisa 56122, Italy
\textsuperscript{3} Centre d’Etudes Prospectives et d’Informations Internationales, 75334 Paris, France
giovanni.bonaccorsi@imtlucca.com

1 Introduction

Single layer measures of node centrality have been used in the past to enrich economic models based on exchanges among countries. These works considered a limited number of single layer networks in isolation and extracted node properties without considering the interplay of different layers. The recent generalization of single layer algorithms to multilayer networks allows us to test how the centrality of nodes may be affected when more than one layer is taken into consideration. In this work we apply the MultiRank algorithm \cite{2} to a newly collected dataset consisting in 19 layers of exchanges among world countries, sampled at time 2003 and 2010. We analyze the resulting ranking and find a good match with commonly used indicators of development, such as GDP and the Human Development Index, signaling that our measure can be used in further economic analysis.

2 Multirank algorithm

The approach we are following is the one of \cite{2}: we summarize the multiplex from two perspectives, first as a colored network with links of different colors having different influence, second as a bipartite network of nodes and layers; and then use both dimension to obtain a generalized multiplex PageRank, the MultiRank.

From the first perspective we obtain matrix $G$ as the sum of adjacency matrices across $M$ layers weighted by their respective influence $z^{\alpha}$. So for layers $\alpha = 1, 2, \ldots M$ the element of $G$ are given by $G_{ij} = \sum_{\alpha=1}^{M} A_{ij}^{\alpha} z^{\alpha}$. From the bipartite perspective we obtain the $M \times N$ incidence matrices $B^{in}$ and $B^{out}$ representing the normalized in-degree and out-degree of each node $i$ in each layer $\alpha$:

\begin{align*}
B_{\alpha_i}^{in} &= \frac{\sum_{i=1}^{N} A_{ji}^{\alpha}}{\sum_{j=1}^{N} A_{ji}^{\alpha}} = \frac{\sum_{i=1}^{N} A_{ji}^{\alpha}}{W^{\alpha}} \\
B_{\alpha_i}^{out} &= \frac{\sum_{j=1}^{N} A_{ji}^{\alpha}}{\sum_{i=1}^{N} A_{ji}^{\alpha}} = \frac{\sum_{j=1}^{N} A_{ji}^{\alpha}}{W^{\alpha}}
\end{align*}

(1)

Where $W^{\alpha}$ is the number of link in each layer. The PageRank equation for each node now becomes: $X_i = d \sum_{j=1}^{N} \frac{G_{ji} X_j}{k_j} + \beta v_i$ where $d$ is the damping factor and $k_j = \max(1, \sum_{i=1}^{N} G_{ji})$, $v_i = \theta \left( \sum_{j=1}^{N} (G_{ij} + G_{ji}) \right)$ and $\beta = \frac{\sum_{i=1}^{N} \theta(\sum_{j=1}^{N} G_{ji})}{\sum_{i=1}^{N} v_i}$. Here $\theta(\cdot)$ is the Heaviside step function.
The equation for node centrality must be jointly estimated with the equation for layer influence:

\[ z^\alpha = \left( W^\alpha \right)^a \left( \sum_{i=1}^N B_{ai}^{in}(X_i)^s \right)^s \]  

where \( N \) is a rescaling constant.

The layer influence depends on several parameters. The \( a \) parameter regulates the effect of total weight of layers \( W^\alpha \) on the influence: with \( a = 1 \) it is proportional to it, while with \( a = 0 \) becomes independent from the number of edges present in the layer. The \( s \) parameters indicates if layers more influential are those with fewer (\( s = -1 \)) or more central nodes (\( s = 1 \)). Once \( s \) is settled the parameter \( \gamma \) allows us to suppress or enhance the contribution of low centrality nodes: with \( s = 1 \) values of \( \gamma > 1 \) (\( \gamma < 1 \)) suppress (enhance) their contribution. Viceversa with \( s = -1 \) values of \( \gamma > 1 \) (\( \gamma < 1 \)) enhance (suppress) the contribution of less central nodes.

In our base estimation we will assume \( a = 0, s = 1, \gamma = 1 \).

3 Dataset

We have collected data on 19 layers with edges mapping different kind of relation among countries. There are five categories of edges, roughly speaking: trade, finance, migration, citations and a reminder consisting of infrastructure and diplomatic relationship layers. We have selected a common set of 112 nodes and two cross sections, for years 2003 and 2010, to inspect the evolution of the MultiRank across time on the same sample of countries. We have applied the MultiRank algorithm on the weighted version of the multiplex filtering the data with an hypergeometric filter, following [1]. Then for robustness we have repeated the analysis using different data treatment.

4 Results

Figure 1 shows our results. In subfigure (a) we are plotting the MultiRank of the first cross section (year 2003) against the second (year 2010), we have highlighted in red the top 30 countries by the HDI index in 2010 and we can see that almost all belong to top half distribution of countries by the MultiRank and that their ranking is very stable over time. This reflects the fact that the HDI of developed countries, once they reach a certain level, is very stable and improvements are rare. To highlight the nodes where ranking changes the most we have identified with different marker the top 10 countries gaining (cross marker) and losing (triangle marker) positions over time, according to the MultiRank. We can see that both categories belong to the lower part of the distribution by HDI, except for few cases (they’re almost all blue points). Hence the evolution of MultiRank mirrors that of countries development: rate of growths are greater for less developed countries, while are very stable for developed ones.

In the second subplot we have plotted the joint distribution of MultiRank and GDP per capita for year 2003 over five different treatment of the data to check the robustness of the relationship.\(^1\)

\(^1\)Given the same set of weighted adjacency matrices for the multiplex we have either applied the hypergeometric filter (\( \text{filtered and unfiltered} \) in the legend) or some kind of normalization to the data (log transformation and/or rescaling by the maximum) or nothing (\( \text{unfiltered, not rescaled} \)). Finally we have tested the use of the output of hypergeometric filter (a probability matrix) as adjacency matrix (\( \text{filtered normalized to 1} \)).
We can see that there is a positive correlation among GDP and MultiRank values, even though not very strong.

Fig. 1. How country MultiRank behaves with respect to development indicators.

5 Conclusions

In this work we have applied the MultiRank algorithm to a newly collected set of heterogeneous layers regarding exchanges among countries. We have shown that the resulting rank is consistent with commonly used development indicators (GDP per capita and HDI) and that using an aggregated measure of multilayer node centrality is a promising strategy to be included in further economic analysis.

References

How far does a shock event spreads on a network?
Detecting causality on the salmon trade network

Juan Carlos Rocha*
Jessica Gephart†

Abstract

Stochastic events can affect the ability of a country to produce and export natural resources such as food. When exposed to such events, an importing country will face a deficit on the expected trade of a product and turn to another trading partner to supply its demand. This simple mechanism can therefore couple oscillations on resource exploitation and consumption in far away places of the planet, a phenomenon theorized as telecouplings. Here we use monthly time series of salmon trade to reconstruct a network of potential telecoupling effects, this is when the export dynamics of a country $c_i$ to $c_j$ causally affects the dynamics of exports from $c_k$ to $c_j$. Our results show that non-linear causal effects do not spread more than few degrees of separation in the trade network; and in fact less than 20 time series out of 402 mapped links in the network have a forecasting skill higher than 20% on secondary links. We offer empirical evidence of telecoupling in the salmon trade network as well as an innovative methodological approach to assess causality of non-linear dynamics in networked systems.

Networked complex systems are susceptible to the spreading of dynamic failures such as blackouts in power grids or the collapse of industrial production chains. Scholars have studied this failure phenomena when the structure of the network is well known. However, detecting causal signals on a network is a challenging task because the independence assumption can be violated by the network structure itself. This is, the dynamics of a node or a link over time depends on the adjacent nodes or links and vice versa. Here we use convergent cross mapping, a nonlinear causality detection technique, to identify causal relationships between the dynamics of links in the salmon trade network as a case study. By 2016 salmon trade accounted for 11.6B dollars globally and 93 countries traded salmon monthly through 406 bilateral trade relationships. In 2016 a red tide event (toxic algae bloom) off the coast of Chile, a major exporter of salmon, reduced Chile’s exports on the order of tens of thousands tons. The question we address is to what extent the export dynamics of a country like Chile have a causal effect on the trade patterns of other trade partners? More generally, which countries’ trade relationships contain causal information about the dynamics of far away resource dynamics?

To study these questions, we used UN Comtrade monthly records for salmon commodities. We aggregated all commodities by traded weight in tonnes and created a network where two countries are linked together if they traded more than 2.7 tonnes (or 1 in log-scale). We used time series from 2014 to the present discarding months were data has not been completely reported (typically 2-3 months closest to the present). All time series were normalized to zero mean and unity variance, and each represents the dynamics of a link in the network. Convergent cross mapping algorithms were used to unravel whether the trade time series from country $c_j$ to country $c_k$ has information on the dynamics of trade from countries $c_i$ to $c_j$. In particular, we measured $\rho$ (the forecasting skill of one time series over another) for all pair-wise combinations of links in the network. If $\rho$ is positive and significantly different from zero we can conclude that the trade from $c_i$ to $c_j$ has a causal effect on the trade from $c_j$ to $c_k$. Our results show that only few pairwise combinations of links have positive forecasting skill [$\rho$]. In other words, dynamics of exports from few countries have a causal effect on the resource exploitation dynamics of other countries far away. Forecasting skill of a link is weakly correlated to the average weight traded and the indegree of the salmon importing node (Fig 1).

Our findings point to important governance implications for the management of marine resources such as salmon. If two countries far apart can synchronize their resource exploitation dynamics given they trade and respond to the trading needs of common partners, it means they can co-evolve managerial frameworks that take into account teleconnections. In other words, our results shed light on which trading communities can

---

*a Stockholm Resilience Centre, Stockholm University
†National Socio-Environmental Synthesis Center
Figure 1: Global salmon network. Correlations between forecasting skill and the degree ratio (a) and average traded tons (b). The complete salmon trade network (c) shows that not all links have positive forecasting skill, only few countries can affect other’s trading dynamics (d for rho > 0.1). Europe contains a strongly coupled cluster with positive forecasting skill.

avoid governance misfit\textsuperscript{4} by designing managerial options together. In particular, the EU forms a strong cluster of trading countries that influence each other. Further work is needed to explore the instabilities created by the cyclical structure within this hub, as well as the role of trade in alternative commodities.


Part XIV

Quantifying Success
Using Network Science to Discover the Grand Masters of the Florentine Renaissance

Cameron J. DeChurch and Noshir S. Contractor

Department of Industrial Engineering & Management Sciences, Department of Communication Studies, & Department of Management & Organizations, Northwestern University
nosh@northwestern.edu

1 Introduction & Method

Many are familiar with the great masters of Florence. The household names of Michelangelo, Botticelli, Raphael, and Da Vinci have enamored art consumers for over 500 years [1, 2]. However, less is known about the grand masters of Florence – those who trained the great masters or trained those who trained the great masters. Their pupilage ties, where one famous painter trains another, represent a key mechanism through which expertise and techniques were transmitted through generations. In order to better understand the antecedents and outcomes of these pupilage networks in the Florentine Renaissance we pose two questions: (1) what organizing principles explain the emergence of the pupilage network? and (2) which Florentine painters, through their position in the network, had the greatest impact on the Renaissance?

We used historical records [1, 2, 8] to construct the pupilage networks among notable Florentine Renaissance painters from the birth of Giotto (1276) to the death of Bronzino (1572), the period art historians mark as initiating the revival of classical principles. A set of “nodes” consisting of all notable painters [1, 2] living and working in Florence during this period was identified ($N = 50$). Great masters were identified as those whose paintings received the greatest critical acclaim [4]. A directed network was constructed where links indicate pupilage ties from apprentice to master (see Figure 1).

![Fig. 1. Master-Apprentice Painter Network of the Florentine Renaissance (1276-1572)](image-url)

Two sets of analyses were conducted: (1) ERGM [3, 6] was used to determine the ordering principles, and (2) a Renaissance index ($R$-index) was constructed in order to...
identity the painters most influential in training, directly or indirectly via the pupilage network, the great masters of Florence. The $R$-index, $R_i$, was computed for each node, estimating each node’s generativity based on their closeness (or geodesic path distance), through directed pupilage ties, from all other painters in the network, and then weighting those pupils based on the relative success of their paintings [4]. For each painter $i$, the index $R_i$ is defined as:

$$R_i = \sum_{k \neq i}^{N} \left( \frac{1}{d_{ki}} \right)^{\alpha} S_k$$

(1)

where $i =$ master, $k =$ pupil, $N =$ number of nodes, $d =$ distance, and $S_k =$ success of the pupil $k$’s paintings. $R_i$ was computed at $\alpha = 1, 2, \text{and } 3$.

2 Results & Conclusion

Which network signatures explain the emergence of pupilage ties? ERGM [3, 6] identifies four factors (see Table 1): (1) the master’s birth year (being born 10 years later during the Renaissance period (1276-1572) increases the likelihood of a master training a famous apprentice about ten-fold), (2) the master’s lifespan (for each additional 5 years the master lived, the chance of training a famous apprentice is reduced nearly 80%), (3) the presence of master hubs – the propensity for a few painters to attract a disproportionate number of apprentices [7], and (4) the absence of pupil hubs – the lack of propensity for pupils to seek apprenticeship under multiple painters [7]. This model demonstrates excellent MCMC diagnostics and GOF. Two controls, edges and an undirected edgecov term reflecting whether or not the pupil was alive and at least 12 years old at the time of the master’s death, are included.

Table 1. ERGM of Master-Apprentice Network of the Florentine Renaissance

<table>
<thead>
<tr>
<th>ERGM Parameter</th>
<th>Estimate</th>
<th>SE</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control: edges</td>
<td>-11.0300</td>
<td>0.1409</td>
<td>0.0000</td>
</tr>
<tr>
<td>Control: &quot;Alive Together&quot; (edgecov)</td>
<td>4.5580***</td>
<td>0.7361</td>
<td>94.4433</td>
</tr>
<tr>
<td>1. &quot;Master’s Birth Year&quot; (nodeicov)</td>
<td>0.0045***</td>
<td>0.0008</td>
<td>10.0450</td>
</tr>
<tr>
<td>2. &quot;Master’s Age&quot; (nodeicov)</td>
<td>-0.0157†</td>
<td>0.0695</td>
<td>0.1969</td>
</tr>
<tr>
<td>3. &quot;Master Hubs&quot; (idegree1.5)</td>
<td>0.4514†</td>
<td>0.2171</td>
<td>1.5705</td>
</tr>
<tr>
<td>4. &quot;Apprentice Hubs&quot; (odegree1.5)</td>
<td>-1.4380**</td>
<td>0.5092</td>
<td>0.2374</td>
</tr>
</tbody>
</table>

Note: SE=Standard Error, OR=Odds Ratio. For interpretation, ORs for birth year converted to log odds of being born 10 years later, & master’s age of living 5 years longer. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Who were the grand masters? In order to discover which painters had the greatest overall influence on the transmission of renaissance artistic ideals through their direct and indirect pupilage of famous apprentices, we calculated a Renaissance ($R$-) index for each painter. Table 2 compares the top 6 painters identified by the $R$-index to the top 6 based on the critical acclaim of their paintings. The top 6 in each do not overlap. In short, the great masters were not the grand masters. Of the great masters, only 2 had a famous pupil (Botticelli & Andrea del Sarto) and those successes did not propel them to the list of top 6 grand masters. The grand masters who were uniquely influential in the production of great art and artists were: Verrocchio, Perugino, Ghirlandaio,

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Baldovinetti, Piero di Cosimo, and Cosimo Rosselli. If an artist’s level of success is determined by the extent to which they trained others that either went on to produce great art and/or pass their ideas down to other artists who did the same, we have a whole new perspective on what makes the “ideal Renaissance Man”.

The metrics developed here are also relevant to contemporary teams. For instance, the field of sports analytics has been fascinated by players who are not the great masters of their sport but whose presence improves others. In basketball the “Battiere effect” is named in honor of Shane Battiere whose own statistics are not record setting, but who makes everyone else’s better [5].

### Table 2. Great and grand master painters of the Florentine Renaissance

<table>
<thead>
<tr>
<th>Painter</th>
<th>R-Index (α = 1)</th>
<th>R-Index (α = 2)</th>
<th>R-Index (α = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Great masters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leonardo da Vinci</td>
<td>0.2513</td>
<td>0.0939</td>
<td>0.0470</td>
</tr>
<tr>
<td>Raphael</td>
<td>0.2293</td>
<td>0.0782</td>
<td>0.0386</td>
</tr>
<tr>
<td>Botticelli</td>
<td>0.1044</td>
<td>0.0705</td>
<td>0.0650</td>
</tr>
<tr>
<td>Michelangelo</td>
<td>0.2416</td>
<td>0.0829</td>
<td>0.0412</td>
</tr>
<tr>
<td>Masaccio</td>
<td>0.0466</td>
<td>0.0205</td>
<td>0.0200</td>
</tr>
<tr>
<td>Andrea del Sarto</td>
<td>0.2727</td>
<td>0.1206</td>
<td>0.0812</td>
</tr>
<tr>
<td><strong>Grand masters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andrea Verrocchio</td>
<td>0.5010</td>
<td>0.3472</td>
<td>0.2901</td>
</tr>
<tr>
<td>Perugino</td>
<td>0.4168</td>
<td>0.2555</td>
<td>0.2003</td>
</tr>
<tr>
<td>Ghirlandaio</td>
<td>0.3898</td>
<td>0.2299</td>
<td>0.1800</td>
</tr>
<tr>
<td>Baldovinetti</td>
<td>0.3893</td>
<td>0.2041</td>
<td>0.1349</td>
</tr>
<tr>
<td>Piero di Cosimo</td>
<td>0.3566</td>
<td>0.1959</td>
<td>0.1480</td>
</tr>
<tr>
<td>Cosimo Rosselli</td>
<td>0.3456</td>
<td>0.1783</td>
<td>0.1282</td>
</tr>
</tbody>
</table>

**Acknowledgments:** This work was supported by the National Institutes of Health [R01GM112938]. We thank Kyosuke Tanaka for his assistance with this work.

**References**

Lifelong Career Success: Joseon Dynasty’s Bureaucrats

Donghyeok Choi¹ and Juyong Park¹,²

¹ Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
² Sainsbury Laboratory, University of Cambridge, Cambridge CB4 1YE, United Kingdom

juyongp@kaist.ac.kr

1 Introduction

The Kingdom of Joseon (briefly The Empire of Korea, 1897–1910 CE) administered the Korean people and its territory in 1392–1910 CE. In accordance with the governing Confucian ideal, its people considered holding a public office by passing state-administered examinations (gwageo) and leaving their names in the historical records the most honourable achievement. For this reason, charting the dynamics of the public careers of the people could be helpful in understanding the workings of the Joseon Dynasty. Fortunately, we possess copious amounts of historical records of their personal careers. Among them, The Annals of the Joseon Dynasty [1] (조선왕조실록; pronounced Joseon-wang-jo-sillok) contains accounts of major historical events and career records of nearly 43,000 bureaucrats including 248 prime ministers, covering a total of 189,218 days of the kingdom’s existence between its foundation and demise.

These records recount dramatic events of political intrigue such as a coup d’état where a faction of bureaucrats attempted to overthrow the reigning king, which would bring power and wealth to the instigators when successful. A famous example is the Revolt of 1453 as a result of which King Danjong, Joseon’s 6th monarch, was dethroned by his uncle Prince Suyang who simultaneously defeated his brother Prince Anpyong’s revolt to become King Sejo. Mining the Annals and other official records for data on the decoration and disciplinary actions on the bureaucrats, one can construct networks that show the correlation between the social interaction networks and their personal success, an example of which is shown in Figure 1. The figure shows the bureaucrats who closely interacted with Danjong, Suyang, and Anpyong during a two-month period surrounding the revolt. It shows those socially close to Suyang likelier to be decorated and honoured, while those close to Anpyong were purged or executed afterwards.

2 Gross Lifelong Career Success Index (GLCSI)

While filled with drama and intrigue, a coup d’état that dethrones the King is an extremely rare event, having occurred only three times during the entire 519-year history of Joseon. Therefore it is more generally informative to find patterns of bureaucrats’ career rise (and fall) during more stable times. Given the strict hierarchy of Joseon’s Confucianism-based bureaucracy, the definition of career success as ‘the real or perceived achievements individuals have accumulated as a result of their work experience’ [2], and using a bureaucrat’s official rank (one to eighteen) as career success
appear appropriate for our purpose. This means that we can consider a person who has spent more time at a higher officer rank as more successful, leading us to define Gross Lifelong Career Success Index (GLCSI) as

$$GLCSI_n = \sum_{i=1}^{L} r_i \times (t_{i+1} - t_i),$$

(1)

where $t_i$ and $r_i$ are the date and the rank at the $i$-th mentioning of the individual $n$ in the Annals. This is equal to the area under the career trajectory line connecting ranks at a given time shown in inset Figure 2.

3 Results and discussion

Figure 2 shows the cumulative distribution of GLCSI. We observe that GLCSI is unequally distributed, with a few bureaucrats having outsized success measured in this way. To understand what explanatory variables drive this phenomenon, we applied the regression method to the relationship between GLCSI and various metadata variables of individual bureaucrats. So far we have identified three variables—‘The number of appearances in the Annals’, ‘Highest official rank achieved by one’s ancestors’ and ‘The
age of passing the exam’—with explanatory power. This result suggests that both individual’s ability and family background can influence one’s bureaucratic career. We are looking to find other factors affecting success, including geographical ones. The quality and the quantity of the information left behind by the Joseon dynasty is expected to shed light on many long-standing questions of history, such as whether it was a bureaucratic or an aristocratic society since titles and positions were not inherited.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF-20100004910, NRF-2013S1A3A2055285 and NRF-2016S1A2A2911945) and the BK21 Plus Postgraduate Organisation for Content Science.

References

Reciprocity and success in academic careers

Weihua Li\textsuperscript{1,2}, Tomaso Aste\textsuperscript{1,2}, Fabio Caccioli\textsuperscript{1,2}, and Giacomo Livan\textsuperscript{1,2}

\textsuperscript{1} Department of Computer Science, University College London, 66-72 Gower Street, London WC1E 6EA, UK
\textsuperscript{2} Systemic Risk Centre, London School of Economics and Political Sciences, Houghton Street, London WC2A 2AE, UK

1 Introduction

Ranging from citation counts to the well known \textit{h}-index, bibliometric citation-based indicators are increasingly relied upon to inform all aspects of academic decision-making, including faculty recruiting, and grant attribution. The onset of such a trend dates back to the 1950s, when the earliest citation-based indices to assess academic impact were put forward \cite{1}. The following decades saw the ever-increasing adoption of such indicators in all fields of science, which eventually led to the systematic analysis of academic citations and to the emergence of \textit{bibliometrics} as a research field \cite{2}.

Given the importance that citations play nowadays in shaping the prospects of an academic career, it is certainly not surprising to see increasing attention to the study of publication strategies that can attract a larger number of citations. Self-citations and their role in inflating bibliometric indicators have been studied extensively \cite{3}. Yet, very little attention has been paid to the scientific community’s collective response to the increasing adoption of bibliometric indicators, and whether this ultimately resulted in more sophisticated citation patterns involving collaborators and colleagues. Indeed, the quest for higher citation counts generates an obvious incentive for scientists to exchange citations with their closest circle of coauthors and collaborators.

Following the broad stream of literature that has analyzed citation patterns from a network perspective (see, e.g., \cite{4}), we investigate reciprocity in an author citation network constructed from the citation history of papers in the Physical Review corpus of journals published by the American Physical Society (APS) between 1950 and 2010. We address two main research questions: 1) how much reciprocated citations contribute to a scientist’s academic reputation; 2) how prevalent is the tendency to reciprocate citations in the scientific community at large, and how it has evolved over time.

2 Methodology

We tackle the above questions by measuring the \textit{excess} reciprocity (see \cite{5} and the caption of Fig. 1) in the data with respect to a variety of ensemble of null networks obtained by rewiring links (i.e., citations) in the empirical author citation networks. The rewiring yields network configurations that are only partially randomized with respect to the empirical networks, as we preserve the number of citations received and the \textit{h}-index (i.e., ultimately, the level of success) of each individual author, and we preserve

The 7\textsuperscript{th} International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
the inherent time ordering of citations (i.e., older papers can cite more recent papers, not the other way around). Furthermore, we constrain link rewiring to only take place between authors belonging to the same scientific community, which we do by considering null models where the rewiring is constrained in terms of the network geodesic distance between authors and in terms of the network’s modularity structure.

3 Results

Our main results, published in [6], are outlined in the following.

1. We find excess reciprocity to be negatively correlated with long-term career success, i.e. authors who manage to attract larger number of citations do so by reciprocating at a lower rate than low-impact authors (see Fig. 1). This result suggests that even potentially sophisticated shortcuts to artificially boost bibliometric indicators based on the exchange of citations, rather than on mere self-citations, cannot significantly improve an author’s impact.

![Graph A](image1.png)

**Fig. 1.** Relationship between excess reciprocity and long-term career success. (a) Centroids of author clusters (identified via k-means) based on the cumulative number of citations received by authors who started their career between 1970 and 1990. (b) 95% confidence level interval for the excess reciprocity of the corresponding clusters of authors in the first 20 career years. The excess reciprocity of an author $i$ is computed as $\rho^{(i)} = (\rho_0^{(i)} - \langle \rho_{NM}^{(i)} \rangle) / (1 - \langle \rho_{NM}^{(i)} \rangle)$, where $\rho_0^{(i)}$ is the number of citations reciprocated by author $i$ in the empirical citation network, while $\langle \rho_{NM}^{(i)} \rangle$ is the corresponding average quantity in the null model ensemble.

2. At the level of the whole network, excess reciprocity steadily increased from 1950 to 1990 and then stabilized around a roughly constant value (see the left panel in Fig. 2). We interpret such increase as the echo of the academic community’s collective response to the increasing importance of bibliometric indicators, whose adoption started...
3. The above increase in reciprocity is exceedingly fuelled by the exchange of citations between coauthors. Indeed, not only such trend has increased in absolute terms in recent years, to the point that roughly half of all reciprocated citations are exchanged between coauthors, but its difference with respect to expected rate of citations between coauthors in the null models has also increased significantly over the years (see the right panel in Fig. 2).

Fig. 2. (a) Average network-wide excess reciprocity from 1950 to 2009. (b) Fraction of reciprocity contributed by past coauthors in the empirical networks (dark blue) and in the null model ensembles (light blue).

References

Ties and Event Flow Network in Basketball

Gyuhyeon Jeon1 and Juyong Park1,2

1 Graduate School of Culture Technology, Korea Advanced Institute of Science & Technology, Daejeon, Republic of Korea, 34141
2 Sainsbury Laboratory, University of Cambridge, Cambridge, United Kingdom, CB4 1YE
juyongp@kaist.ac.kr

1 Introduction and Methods

The unpredictability of the result of a sports game is one of the factors that make it exciting to watch for many people [1, 2]. At any given point in time, the unpredictability is at a maximum when the score difference is zero. Since each team’s goal is to pull the game away from such unpredictable state in their favor, the characteristics of ties and lead reversals (changes) can show interesting patterns that help us understand the dynamics of the game. Basketball leagues such as the NBA provide rich data for such understanding, since a basketball game consists of many scoring events, resulting more frequent ties and lead changes than other sports in general [3]. In this work we present the scoring dynamics in basketball games determined from patterns of occurrence of ties. We introduce multiple variables to characterize ties and provide results measured from an entire season of NBA.

We extracted all the tie and lead change events from the play-by-play data of games played during the 2013-2014 NBA regular season. When a lead change occurs without going through an explicit tie (such as when a 2-point shot by a team trailing by one point is successful), we considered that a virtual tie has occurred. During the season, 1145 games have produced 11,273 ties (including the virtual ones). A tie and the situation surrounding it can be visualized via a graph that shows the score difference between two teams during a game, for instance Fig. 1 (a). Writing $L_n$ as the $n$-th tie observed in a game, we measure $P_{n,t}$, the peak (largest) point difference, and $P_{n,s}$, the time it happens, during the interval between $L_{n-1}$ and $L_n$. We also measured the number of events (such as shooting, turnovers, and rebounds) in the interval. Finally, we construct a flow graph between events to visualize the dynamics of the game.

2 Results

Characteristics of the Ties. The games showed an average of 9.84 ± 6.93 ties ($max = 46$ events). Approximately 70% of the ties were followed by lead changes. Ties occurred most frequently and had the shortest duration in the first quarter; Compared with the first quarter, the last quarter had two thirds as many ties, each lasted 3.25 times longer. Nevertheless, the percentage that were followed by lead changes did not change.

Duration of a lead by either team lasted on average 167.67 ± 318.75 seconds ($max = 2855$ seconds), and the mean peak score difference was 3.21 ± 3.10 points ($max = 30$...
points. The score peak occurred roughly in the middle of the interval, i.e. at $0.48 \pm 0.19$ with the interval length normalized to 1.

Previous studies have found that events in sports games follow a Poisson or negative binomial distribution [3, 5]. We also found that the number of ties followed a negative binomial.

**Intra-Interval Dynamics.** We compared the number of events inside an interval, before and after the peak score. We find that there were 50% more events post-peak. Also, before the peak, there was little difference between the number of events taken by the two teams; however, after the peak, the losing (trailing) team generated 30% more events than the leading team. Regarding shooting in particular, the losing team took 56% more shootings. The quality of the scoring was also different. Before the peak, the leading team makes more 3-points shots than the losing team. After the peak, however, the losing team makes more 3-point shots than the leading team.

Sequences of events towards a scoring event do not show wide variety because the pathway to a scoring event is limited [4]. We drew the path of an event needed to reach a shooting event. Fig. 1 (b) and (c) are weighted graphs that present the event flow towards a shooting event. The node represents an event, and the thickness of an edge represents the number of occurrences between the nodes. From the results, the leading team was found to have little difference in the flow of events before and after reaching the peak. However, the flow of the losing team changed noticeably after the peak: The losing team wastes fewer scoring opportunities, and generates more scoring opportunities from opponents’ fouls. We believe that these changes indicate how the losing team adapts and begins to outscore the opponent.

3 Discussion

We proposed a method to observe tied events in NBA games to understand the dynamics of scoring. Our approach shows the general characteristics of ties and provides a picture of what happens between ties or lead changes. Future work will include investigation on further sophisticated aspects of the problem, such as what happens between the last tie and the end of the game to see if the nature of adjustments changes as the game draws to a close, the season-to-season difference, and the individual team characteristics. It will also include more analysis of the event flow network.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF-20100004910, NRF-2013S1A3A2055285 and NRF-2016S1A2A2911945) and the BK21 Plus Postgraduate Organisation for Content Science.

References

Fig. 1. (a) Ties, lead changes, and peak points (b, c) Event flow path of the trailing team represented as a weighted directed graph. (b) Before the peak score. (c) After the peak score.


Part XV

Resilience and Control
Trade-off between time and energy cost for controlling complex networks

Yongzheng Sun\textsuperscript{1,2}, Siyang Leng\textsuperscript{2}, Ying-Cheng Lai\textsuperscript{3}, Celso Grebogi\textsuperscript{4}, and Wei Lin\textsuperscript{2}

\textsuperscript{1} School of Mathematics, China University of Mining and Technology, Xuzhou 221008, PR China, yzsung@gmail.com, 
\textsuperscript{2} School of Mathematical Science, Fudan University, Shanghai, 200433, PR China 
\textsuperscript{3} School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, Arizona 85287-5706, USA 
\textsuperscript{4} Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom

1 Introduction

In the past two decades, it has shown that the structure of many real systems can be described by complex networks. It is highly desirable to be able to apply proper control to guide the network dynamics toward a desired state. When controlling a real network, a fundamental question is how to design a controller which can be constructed in practice. There are two important factors in designing the control input. One is the cost energy, the other is the cost of time. Even if a network is controllable in principle, it may not be controllable in practice if it costs an infinite amount of energy or if it require too much time to reach the desired state. Existing works on network controlling focus either on open-loop control strategies and their energy consumptions or on closed-loop control schemes with an infinite-time duration [1,2]. A universal framework for exploring the time and energy cost of controlling complex networks is lacking.

The goal of the present study thus is to investigate the time and energy cost for controlling of complex nonlinear networks. We articulate a finite-time, closed-loop controller with an eye toward the physical and mathematical underpinnings of the trade-off between the control time and energy as well as their dependence on the network parameters and structure. To ensure that our framework is physically significant, we focus on the control energy and the time required to achieve control and investigate their tradeoff. We study how network parameters and structure affect the control time and energy and test the control framework using a variety of real biophysical systems including stem cell differentiation, food webs, random ecosystems, and neuronal networks. We herein present and extend our work recently published [3].

2 Results

We consider nonlinear dynamical networks described by

\[ \dot{x}_i = f(x_i) + \sum_{j=1}^{N} \varepsilon_{ij} \Gamma x_j(t) + u[x(t)]B_i, \quad 1 \leq i \leq N, \]  

(1)
where \( N \) is the network size, \( x_i = [x_{i1}, \cdots, x_{id}]^\top \in \mathbb{R}^d \) denotes the \( d \)-dimensional state variable of the \( i \)-th node, \( \mathbf{x} \) represents the state variables of the whole network, \( f : \mathbb{R}^d \to \mathbb{R}^d \) is a nonlinear velocity field governing the nodal dynamics, \( C = (c_{ij}) \in \mathbb{R}^{N \times N} \) is the coupling matrix determined by the network structure, \( \Gamma \in \mathbb{R}^{d \times d} \) describes the internal coupling configuration of each node, \( u_x(t) = [u_i(t)]_{1 \leq i \leq M} \in \mathbb{R}^{d \times M} \) is the closed-loop control protocol to be designed, and \( B_i = [b_{i1}, \cdots, b_{iM}]^\top \in \mathbb{R}^M \) \((b_{im} = 0, 1)\) characterizes the driving of the controller \( u \) to the \( i \)-th node.

In principle, the linear feedback controller \( u_i = -k \text{sign}(x_i)^\alpha = u_i^L \) can be used to drive the system to \( x_i = 0 \), but the time can diverge. The feedback controller \( u_i = -k \text{sign}(x_i)^\alpha = u_i^L \) can drive the system to \( x_i = 0 \) for all \( t \geq T_j^f \) with \( T_j^f < \infty \), where \( \text{sign}(x_i)^\alpha = [\text{sign}(x_{i1})|x_{i1}|^\alpha, \cdots, \text{sign}(x_{id})|x_{id}|^\alpha]^\top \), \( \text{sign}(\cdot) \) is a sign function, \( k \) is the control strength, and \( \alpha \in (0, 1) \) is the steepness exponent.

\[
E_{\text{potential}} = \int_0^\infty u_i^L \ dx_i.
\]

To gain physical insights into the control process, we consider the potential function \( E_{\text{potential}}^{LF} \) (dashed curves) underlying closed-loop feedback controllers \( u_i^L \). The controlled system trajectory in the phase space is shown in Fig. 1(a).

![Fig. 1. (Color online) Physical underpinning of our closed-loop feedback controller. (a) System moving according to the potential function \( E_{\text{potential}}^{LF} \) (dashed curves) underlying closed-loop feedback controllers \( u_i^L \). (b) Controlled system trajectory in the phase space by \( u_i^L \), where a control switch occurs when the system crosses the unit sphere \(|\mathbf{x}| = 1\).](image)

The basic principle is then to design two controllers in complementary regions of the phase space. This consideration leads us to propose the following global, compound controller: \( u_i = u_i^L 1_{\mathcal{U}} + u_i^F 1_{\mathcal{S}^c} = u_i^\mathcal{F} \), where \( 1 \leq i \leq M \), the unit ball is defined by \( \mathcal{U} = \{|\mathbf{x}| < 1\} \), \( \mathbf{x} = [x_{i1}, \cdots, x_{id}]^\top \), \( |\cdot| \) denotes an appropriate norm of the underlying vector, \( \mathcal{S}^c \) is the complement of \( \mathcal{S} \), and \( \mathcal{F} \) is the indication function for a given subscript set. As shown in Fig. 1(b), the compound controller \( u_i^\mathcal{F} \) switches from \( u_i^L \) to \( u_i^F \) when the system enters the unit sphere.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
We show that the global controller $u^S = [u^S_i]_{1 \leq i \leq M}$ enables finite-time control, and provide an estimate of $T^S_f$, the time required to achieve control. Assume that the nodal dynamics satisfy $\| f(x) \| \leq l \| x \|$ ($\forall x \in \mathbb{R}^d$). Let $\rho \equiv k - l - \eta_{\text{max}}, \; \eta_{\text{max}}$ is the maximum eigenvalue of the matrix $H \equiv \left( [C \otimes \Gamma]^T + C \otimes \Gamma \right)/2$. The upper bound for $T^S_f$ is given by

$$T^S_f = \frac{1}{\rho} \left[ \ln \| x(0) \| + \frac{l}{1-\alpha} \right], \; x(0) \not\in \mathcal{W}, \; x(0) \in \mathcal{W}.$$  

For our controller $u^c$, the required energy cost is $E^S_c = \int_0^{T^S_f} \sum_{i=1}^N \| u^c_i(t) \|^2 dt$. A lengthy calculation leads to the following upper bound for the energy cost

$$E^S_c = \begin{cases} k^2 \frac{1}{2 \rho} \left[ 1 - \| x(0) \|^{-2} + \frac{\zeta}{\alpha} \right], & x(0) \not\in \mathcal{W}, \; x(0) \in \mathcal{W}, \\
\frac{1}{\rho} \| x(0) \|^{1+\alpha}, & x(0) \not\in \mathcal{W}, \; x(0) \in \mathcal{W}, \end{cases}$$

where $\zeta = (Nd)^{1-\alpha}$.

We see that, for given values of $\alpha$ and $x(0)$ as well as specific network dynamics with $l$, $C$ and $\Gamma$, the estimation (2) is on the order $O(1/k)$. Accordingly, $u^S$ with a larger value of $k$ can expedite control, and $E^S_c$ is bounded from above by a quantity on the order of $O(k)$. This indicates that, for a given network and given values of $\alpha$ and $x(0)$, increasing $k$ will raise the energy cost. In addition, for fixed values of $\alpha$ and $x(0)$, if $k$ is sufficiently large, increasing $l$ or $\eta_{\text{max}}$ will lead to larger upper bounds for both control time and energy. It is easy to see that $T^S_f$ is an increasing function of $\alpha$, implying that control can be expedited by using a smaller value of the steepness exponent $\alpha$. In addition, the condition $\partial_\alpha (E^S_c) < 0$ implies that smaller values of $\alpha$ lead to higher energy costs. These results reveal a trade-off between the control time and energy cost for our controller $u^S$ with respect to variations of $\alpha$ or $k$.

We demonstrate that our closed-loop controller can drive two different cell fates to the critical expression level to enable stem cells to remaster their cell fate for cellular differentiation. Our control method can also successfully restore the nonlinear ecosystems out of extinction to a sustainable state. These results suggest that to develop closed-loop control with optimized control time and energy not only is fundamental to the network control field but also has applied values.

Acknowledgments This work is supported by the National Natural Science Foundation of China (Grant Nos. 61403393, and 11711530203).

References

Chimera states in complex networks: from nonlocal to fractal modular and multilayer topologies

Iryna Omelchenko, Anna Zakharova, and Eckehard Schöll
Institute of Theoretical Physics, Technische Universität Berlin, Germany,
omelchenko@itp.tu-berlin.de

1 Introduction

The investigation of coupled systems is bridging between nonlinear dynamics and network science, with a plethora of applications in physics, biology, technology, and giving rise to an abundance of new dynamical scenarios. Among these a very peculiar type of dynamics, when a network exhibits a hybrid nature combining both coherent and incoherent parts, called chimera states. The most surprising aspect was the discovery of these states in a system of identical oscillators coupled with a symmetric interaction function. Chimera states have recently received much attention because they arise spontaneously as symmetry-breaking dynamical patterns in networks, they have been associated with unihemispheric sleep and epilepsy in neuroscience. We discuss properties of chimera states and dynamical mechanisms leading to their appearance in the variety of network topologies: from regular rings with nonlocal coupling to complex networks with fractal connectivities and multiple layers.

2 Results

Chimera states are complex spatio-temporal patterns which consist of coexisting domains of spatially coherent and incoherent dynamics in systems of coupled oscillators. Initially observed in the systems of phase oscillators [1, 2], chimera states can also be found in a large variety of different systems [3, 4], and higher spatial dimensions. The habitat of chimera states extends to more complex oscillatory units characterized by phase-amplitude dynamics, which include neural models. Moreover, next to the classical chimera state, which exhibits one coherent phase-locked and one incoherent region, patterns with multiple domains of incoherence have been observed.

We consider networks of $N$ coupled units:

$$\dot{x}_i(t) = F(x_i(t)) + \frac{\sigma}{g} \sum_{j=1}^{N} G_{ij} H(x_j(t) - x_i(t))$$

with $i \in \{1,\ldots,N\}$, where $x_k = (u_k,v_k)^T \in \mathbb{R}^2$ are dynamical variables, and $F$ defines dynamics of each individual oscillator, which we choose to be FitzHugh-Nagumo, or Van der Pol oscillator. Coupling topology is given by adjacency matrix $G$. The parameter $\sigma$ denotes the coupling strength, and $g = \sum_{j=1}^{N} G_{ij}$ is the number of links for each node (corresponding to the row sum of $G$). Fig. 1 shows an example of chimera state...
in the nonlocally coupled network of FitzHugh-Nagumo oscillators. One can clearly
distinguish coherent and incoherent parts, a characteristic signature of chimera states.
Elements that belong to the incoherent part are scattered along the limit cycle, as shown
in Fig. 1(b). The systems elements perform a nonuniform rotational motion, but neigh-
bouring oscillators are not phase locked, Fig. 1(c) shows mean phase velocities for each
oscillator calculated as \( \bar{\omega}_k = \frac{2\pi M_k}{\Delta T} \), \( k = 1, \ldots, N \), where \( M_k \) is the number of com-
plete rotations around the origin performed by the \( k \)th unit during the time interval \( \Delta T \).
The interval of constant \( \bar{\omega}_k \) corresponds to the coherent domain. This phase velocity
profile is a prominent feature of such patterns, and is usually used as a clear indica-
tion of chimera states. In our talk, we will present an overview of our recent studies

on chimera states in the networks with different topologies. Starting with nonlocally
coupled rings [5–7], Fig. 2(a), where each network element is coupled to a fixed range
of its neighbours, we will move towards more sophisticated coupling schemes, such as semi-fractal (or hierarchical) connectivities, Fig. 2(b), which can be generated using Cantor construction algorithm for a fractal set. We will demonstrate that symmetric topologies with large clustering coefficients promote the emergence of chimera states, and discuss the increasing role of amplitude dynamics of the oscillators [8–10]. As a
further step, motivated by potential application to epileptology and epilepsy surgery, we
compare two topologies: an empirical structural neural connectivity and a mathematically constructed network with modular-fractal connectivity, Fig. 2(c). We analyse the properties of chimeras and partially synchronized states in both networks, and qualitatively simulate the dynamics of epileptic seizures [11]. Finally, we will discuss possible mechanisms of controlling chimera patterns by feedback [12, 13], and by multiplexing (Fig. 2(d)).

Summary. We analyse chimera states, which consist of coexisting spatial domains of coherent (synchronized) and incoherent (desynchronized) dynamics in networks of coupled nonlinear oscillators. We show that a plethora of novel chimera patterns arise if one goes beyond phase oscillator model and regular ring topologies. We test the robustness of chimeras with respect to changes in the structure of the network, and inhomogeneity of individual oscillators. Our results account for a wide range of applications in biology, physics, and technology.

This work was supported by Deutsche Forschungsgemeinschaft in the framework of Collaborative Research Center SFB 910.

References

Robustness of hierarchical network organization in models of functional connectivity

Ali Safari1, Miguel Ángel Muñoz2, and Paolo Moretti1

1 Institute for Materials Simulation, Friedrich-Alexander-University Erlangen-Nuremberg, Dr-Mack-Str 77 90762 Fürth, Germany, ali.s.safari@fau.de, http://www.matsim.techfak.uni-erlangen.de/staff/ali-safari.shtml
2 Departamento de Electromagnetismo y Física de la Materia e Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, Granada E-18071, Spain

1 Introduction

Human brain networks [1, 2] are well known examples of biological systems, which exhibit a hierarchical modular structure. Hierarchical modular organization is often described by simple mathematical models of synthetic hierarchical modular networks (HMNs). As different types of algorithms are used in the literature to generate HMNs, in the following we refer the model proposed in [3] without loss of generality.

Recent convergence of neuroscience and network science opens up new opportunities for approaching the study of brain function [4–6]. A fundamental issue in this context is how structure and function are related. Provided that structural brain networks are hierarchical in nature, to which extent can we say the same regarding functional brain networks? Functional networks arise from activity correlations between different brain regions. Assessing the similarity between structural and functional connectivity represent a crucial aspect of the current structure-function debate. Similarity measures were proposed in the past (see for instance [8]), highlighting how similarity itself may be non-trivially dependent on the choice of thresholds used to extract functional networks. A promising route to address the threshold problem is provided by persistent homology techniques, which leverage concepts of computational topology (see e.g. [9]). The open question remains however, as to how spectral network properties may be used to assess the similarity of structural and functional connectivity.

In this work we address the problem of structural-functional similarity from the spectral perspective. A now well-known spectral feature of structural HMN is the fact that they possess small spectral gaps (defined e.g. as the difference between largest and second largest eigenvalue of the adjacency matrix \( \Lambda_1 - \Lambda_2 \)) [3]. Does this property extend to functional networks? How does it depend on the choice of thresholds? How does it depend on the nature of the functional process hosted by the network (e.g. sub-critical vs. supercritical)?

2 Results

Our starting point is the HMN model used in Refs. [3, 10]. Its hierarchical modular nature is reflected in the fact that \( \Lambda_1 - \Lambda_2 \) is small. In order to verify to which extent this
property extend to functional connectivity, we generate functional networks by computing coactivation matrices, as suggested in [11]. The functional connectivity of a network of size $N$ will be then given by a dense matrix, whose element $ij$ is the probability of nodes $i$ and $j$ to be simultaneously active at the same time in a simple activity spreading simulation. As a model for activity spreading we use simple types of SIS dynamics as the ones used in [3, 10]. While these models do not capture the complexity of a system like the brain, they provide a reliable probing tool to study functional patterns under minimal assumptions, and with the introduction of a single tuning parameter, i.e. the spreading rate $\lambda$. It is then well known that a critical $\lambda_c$ arises. For $\lambda < \lambda_c$ the system is in a subcritical phase (or, more precisely, a Griffiths phase [3]). For $\lambda > \lambda_c$ the system is supercritical, and activity is unbounded as in a simple model for pathological conditions such as epilepsy. By generating a dense coactivation matrix for this type of dynamics, we can then apply a finite and positive threshold $T$ and extract the sparse adjacency (or weight) matrix of the functional network. Upon increasing $T$, the network becomes sparser and sparser, up to the point $T_c$ when it undergoes a percolation-like transition in which the largest connected component size $s_1$ vanishes. Obviously, the spectral gap too should undergo a similar transition: for a fully connected dense network it should be large, for a fragmented network it should be vanishing. Do the two transitions coincide? To answer this question we can thus act on two parameters, the spreading rate $\lambda$ and the threshold $T$. Figure 1 shows our main result. In the subcritical case, the spectral gap drops by more than two orders of magnitude well before the network fragments. The functional network indeed inherits the small-gap property of the structural one that generated it. In other words, it is hierarchical too. In the supercritical case, instead, the two transitions become closer and, more importantly, as soon as the spectral gap starts decreasing, the giant connected components shrinks too. As a consequence, in the supercritical case the functional network never exhibits a small spectral gap except for when it is fragmented: the functional network is not hierarchical. Figure 1 also shows how this clear-cut separation between subcritical and supercritical is reflected by the degree distributions of the ensuing functional networks, which exhibit exponential tails in the subcritical case (mimicking the degree distribution of the underlying structural network), and heavy power-law tails in the supercritical regime.

**Summary** In this work we addressed the question, as to which extent we can consider functional (coactivation) networks to be hierarchical as their structural counterparts. We have shown that this is the case as long as the dynamic process generating the functional patterns is not supercritical. Below the critical point, functional networks indeed inherit essential properties of the underlying structural HMNs, namely the small spectral gaps and the exponential degree distributions. This correspondence is lost for supercritical types of dynamics, where we can safely conclude that the resulting functional networks are not hierarchical. Our work of course relies on essential simplifications of the original neuroscience problem. Namely our simulated dynamics is a simplification of neural activation processes, and our definition of the functional network, constructed here for convenience from a coactivation matrix, differs slightly from the methods used in experiments. Nevertheless, we believe that our results provide useful insights regarding the connection between structure and function in brain networks, and a viable technique to assess the persistence of the hierarchical organization in functional networks.
Fig. 1. Comparisons between spectral gaps and giant connected component sizes (right) and degree distribution (left) for functional coactivation networks generated from HMNs of size $N = 1024$. The top row refers to the subcritical case, the bottom row to the supercritical case.

References

Resilience of Hierarchical Networks and Interdependent Hierarchical Networks

Louis M. Shekhtman and Shlomo Havlin
Department of Physics, Bar-Ilan University, Ramat Gan, Israel 52900,
lsheks@gmail.com

1 Introduction

Much work has focused on the resilience of networks and interdependent networks under varying levels of failures. The resilience of the network has often been measured based on the largest fraction of connected nodes, i.e., the giant component $P_\infty$, remaining after a fraction $1-p$ nodes of the network are removed (percolation). Researchers have considered many different realistic network structures, such as various degree distributions, spatial networks, modular networks, and others. Despite this significant research, none so far have incorporated the hierarchical structure present in many real-world networks [1, 2].

For example, infrastructure across cities will likely be distributed such that nodes are tightly connected within small neighborhoods, somewhat less connected across the whole city, and have even fewer connections to other cities. One can describe such a network using a model where nodes are each assigned to modules and then these modules merge together as one moves up the hierarchy (see Fig. 1). For our model, we use a stochastic block model with Erdős-Rényi blocks. The vector $k$ describes the average degree at each level of the hierarchy and the vector $m$ gives the number of modules in the hierarchy at each level.

Furthermore, while previous work identified interconnected nodes, those nodes with links outside their neighborhood, to be more likely to fail due to having higher betweenness, longer distance connections, and increased load [3–5], here we have various levels of interconnections (between neighborhoods, between cities, etc). For our model, we first consider the failure of the nodes with interconnections at the highest level, followed by those with interconnections at the next level, etc. This follows from the idea that those with interconnections at the highest level are also likely to have the highest betweenness since they connect larger groups of nodes, represent longer distance connections, and have greater load passing through them.

2 Results

We develop an analytic solution for the size of $P_\infty$, during a targeted attack that removes a total fraction of $1-p$ nodes. We use the generating functions framework from
Callaway et al. [6]. Following the derivation leads to the following solution

\[ m_j P_m = p_{coj-1} \left[ e^{-k_j} (1 - r_j) \left( 1 - e^{-\left( \sum_{i=1}^{j-1} k_i \right) m_j P_m} \right) \right. \\
\left. + r_j \left( 1 - e^{-\left( \sum_{i=1}^{j} k_i \right) m_j P_m} \right) \right], \quad p_{coj} < p < p_{coj-1}, \tag{1} \]

where \( p_{coj} \) represents the cutoff point at which all nodes interconnected at level \( j \) are removed, more explicitly it is given by \( p_{coj} = \exp \left( - \sum_{k=1}^{j} k_i \right) \), \( r_j \) is the fraction of nodes that survive at level \( j \) (which can be converted to an overall survival probability by \( r_j = (p - p_{coj})/(p_{coj-1} - p_{coj}) \)), and \( m_j \) \( (k_j) \) represent the respective entries for level \( j \) in vectors \( m \) \( (k) \). Analyzing the terms of Eq. (1) can provide some intuition as follows: once level \( j \) is reached the size of \( P_m \) will be reduced by a factor \( m_j \) since the network is already segregated into \( m_j \) components. Also \( 1 - p_{coj-1} \) fraction of nodes were already removed prior to reaching this layer which can be recognized by the initial factor of \( p_{coj-1} \). The terms in the large brackets relate to the likelihood of a node to be in the giant component at this point which involves the two cases that the node is either interconnected at this layer or not. Eq. (1) is transcendental and thus must be solved numerically for a given set of parameters.

We compare our theory to extensive simulations of large model networks and confirm its validity in Fig. 1.

We find that depending on the number of levels in the hierarchy and the extent of the connections at each level, the giant component can experience multiple transitions as the network separates at the various levels. The transitions before the final one are all abrupt and are the result of entire groups of communities becoming segregated.
The final transition however is continuous. Through analytic considerations we also determined the number of abrupt transitions that will occur in a hierarchical network and the locations (values of $p$) for all of the transitions.

We also consider the case where the lowest layer has a scale-free distribution and find that multiple transitions exist in this case as well. We further show that trade-offs exist between optimizing the hierarchy at one layer as opposed to another. We suggest extensions to our model where costs are considered for links at each layer of the hierarchy and the minimal cost for a certain resilience is defined as an optimal point.

We next extended our framework to networks of interdependent hierarchical networks (NoNs) [7, 8]. We considered two types of NoNs, namely treelike NoNs where the dependencies among the networks form a tree and NoNs with loops such as random-regular (RR) NoNs where each network depends on a fixed number of other networks. We developed analytic theories for these cases as well and confirmed them through extensive simulations. In the case of treelike NoNs we find that the final transition is also abrupt as in the case of interdependent Erdős-Rényi networks. For RR NoNs the final transition may be either continuous or abrupt depending on the specific parameters. Our results show that the hierarchical structure of networks has important effects on resilience and could be useful for designing robust infrastructure.

References

Temporal network-based analysis of turbulent mixing

Giovanni Iacobello¹, Stefania Scarsoglio¹, J.G.M. Kuerten², and Luca Ridolfi³

¹ Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin, Italy
giovanni.iacobello@polito.it
² Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
³ Department of Environmental, Land and Infrastructure Engineering, Politecnico di Torino, Turin, Italy

1 Introduction

Complex network-based tools have been recently attracting a lot of interest in the analysis of fluid flows. So far, most of the research has focused on geophysical flows [1], two phase flows [2], as well as turbulent flows [3, 4]. In particular, the investigation of fluid flows from a Lagrangian point of view – i.e., by tracing particle trajectories during the motion – has recently been explored [5, 6]. However, particle trajectories are typically exploited to build static networks, thus losing transient information on the flow dynamics. In this work, we propose a temporal network-based approach [7–9] for the study of turbulent mixing, that fully captures the temporal evolution of particle dynamics.

2 Data description and network building

A turbulent channel flow was simulated through a direct numerical simulation at $Re_t = Hu_t/ν = 950$ and performed for a time $T^+ = Tu_t/ν = 15200$ with a time step $Δt^+ = 4.75$, where $H$ is the half-channel height, $u_t$ is the friction velocity and $ν$ is the kinematic viscosity [10]. A set of $100 × 100$ fluid particles were initially arranged at $x^+ = 0$ on a uniformly spaced grid in the plane $(y^+, z^+)$ (see Fig. 1(a)), where $(x^+, y^+, z^+)$ are the streamwise, wall-normal and spanwise coordinates, respectively. Particles were grouped into $N = 100$ levels, $l_i$, according to their $y^+$ value at the initial time $t^+ = 0$ (see Fig. 1(a)). Trajectories of the particles are then computed over time.

If two particles are sufficiently close to each other at a given time, then a connection is established between them. Specifically, a particle $A$ is connected to a particle $B$ if $A$ lies inside a reference ellipsoid centred in $B$, and vice versa (by symmetry). For example, in Fig. 1(b) the trajectories of two particles $A$ and $B$ and their corresponding reference ellipsoids are shown. The corresponding time-sequence of connections between $A$ and $B$ is illustrated in Fig. 1(c). The ellipsoid was geometrically anchored and centred to each particle location and it was chosen as reference geometry to take into account the anisotropy of the flow. Since the average Euclidean distance between particles increases in time (mainly due to the streamwise dispersion), the semi-axes of the ellipsoid were set to grow proportionally to the average distance in each direction.

As particles follow the turbulent motion, their relative distance changes in time, thus particles belonging to different levels generate different time-sequences of connections.
Therefore, we modelled the turbulent particle dynamics by means of a weighted time-varying network \[ W_{ij}(t^+) \], in which nodes correspond to levels, \( l_i \) (i.e., groups of particles initially at the same \( y^+ \)). At each time, the weight of a link between two nodes \((i, j)\) – i.e., two levels – is equal to the total number of connections shared by particles belonging to levels \( l_i \) and \( l_j \). By doing so, a time-varying network of \( N_T = 3200 \) undirected weight matrices (each one of size \( 100 \times 100 \)) is obtained, that is able to capture the temporal effects of turbulent mixing and flow advection on particle dynamics.

### 3 Results

Since \( W_{ij}(t^+) \) embeds the information of the modelled particle dynamics, the weight matrices at three representative times are shown in Fig. 2(a), where colors indicate the intensity of the weight, i.e., the number of connections between each pair of levels. At \( t^+ = 0 \), the weight matrix displays a band structure corresponding to the initial arrangement, where nodes close in space are strongly linked each other. As time increases, the weight matrix changes structure: a three-square pattern of high \( W_{ij}(t^+) \) values emerges, as a consequence of turbulent dispersion, while turbulent mixing enables the activation of links between initially distant levels (e.g., levels 20 and 90). Finally, for large times \((t^+ = 15200)\) the turbulent mixing dominates the particle dynamics, approaching an asymptotic behaviour, where particles are well-mixed independently from their starting level. Consequently, the resulting weight matrix does not show any clear pattern.

The change in the particle dynamics is also highlighted by a different network topology of \( W_{ij}(t^+) \). This is shown in Fig. 2(b), where the change in the network topology as function of time is evident. In particular, at \( t^+ = 900 \) the nodes tend to cluster into three groups in analogy with the three-square pattern in Fig.2(a). This pattern is lost at \( t^+ = 15200 \), in which the network topology appears as a random layout. From a centrality perspective, the analysis of the average strength over time is able to highlight and discern characteristic advection-mixing regimes in a straightforward way. Hence, the main advantage of the proposed approach is its ability to fully incorporate the information about particle dynamics into the weight matrices, at any time.
Summary. In this work, the turbulent mixing in a channel flow is investigated through a Lagrangian approach. A set of fluid particles is traced in time, where each particle belongs to a level corresponding to its initial wall-normal coordinate. By exploiting the pairwise distance between particles, a weighted time-varying network is built in which nodes represent levels, and the link weight is equal to the number of connections between each pair of levels. Different particle dynamics are found to result into different network topologies and link weight patterns. In this way, it is possible to entirely capture the features of the turbulent mixing and geometrize the evolution of the particle swarm.

References

10. The use of supercomputer facilities was provided by NWO, Physical Sciences.
Using non-normality for control energy reduction in network controllability problems

G. Lindmark∗and C. Altafini∗

∗Linköping University, SE-58183, Linköping, Sweden, gustav.lindmark@liu.se, claudio.altafini@liu.se

1 Introduction

Controlling a network refers to the possibility of steering the state variables associated with its nodes in order to accomplish a desired task. Several of the classical notions of controllability developed in the last 50 years in Control theory have been recently adapted to the context of complex networks [4, 6]. For instance the notion of structural controllability has proven useful when the problem is to determine a minimal set of driver nodes (i.e. nodes on which an external control input is acting) that render a given network controllable [3]. However, a network might be controllable in theory but not in practice when unreasonable amounts of control energy are required to steer it in some direction [8, 1, 9]. To handle such problems, several approaches to formulate a more “practical” degree of controllability with limited energy have been proposed [1, 7, 9]. For linear dynamics, the measures of control energy are normally formulated in terms of the eigenvalues of the controllability Gramian, e.g. the smallest eigenvalue of the Gramian or the trace of the Gramian (the sum of its eigenvalues).

In this work we define two network centrality measures meant to quantify two important contributions to the control energy: The first centrality measure quantifies the influence that each node has on the rest of the network, and the second the ability to control a node indirectly from other driver nodes. The two centrality measures reflect the fact that nodes in directed networks can be important for different reasons. By combining the centralities into node rankings for driver node placement, the control energy is significantly reduced w.r.t. random driver node placement, hence the practical degree of controllability is improved. The proposed node rankings can be interpreted on the basis of the non-normality of the weighted adjacency matrix of the network, and it turns out that the control energy improvements are particularly large for networks with a high degree of non-normality. We herein present our work [5].

2 Results

We consider the discrete-time linear network model

\[ x(t+1) = Ax(t) + B_w u(t), \]

where \( x(t) \in \mathbb{R}^n \) is the state at time \( t \in \mathbb{N}_0 \), \( A \in \mathbb{R}^{n \times n} \), \( B_w = [e_{k_1}, \ldots, e_{k_m}] \in \mathbb{R}^{n \times m} \) and \( u(t) \in \mathbb{R}^m \). The vector \( e_k, k \in 1, \ldots, n \), denotes the \( k \)-th vector of the canonical
basis of $\mathbb{R}^n$. We represent the network with the directed graph $G(A)$ with node set $V = \{v_1, \ldots, v_n\}$ and edge set $E = \{(v_i, v_j), i, j \text{ s.t. } A_{ij} \neq 0\}$. Each control input is assumed to act on only one node (driver node), and the set of driver nodes is $K = \{v_1, \ldots, v_m\} \subseteq V$. When the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the controllability Gramian

$$W = \sum_{t=0}^{T-1} A^t B_\infty^T (A^T)^t$$  \hspace{1cm} (2)$$

are large, the amount of energy that is required to steer the network from the origin in the corresponding eigendirections is low. Hence, $\lambda_{\min}(W)$ is a metric for the worst case control energy and $\text{Tr}(W)/n = \sum_{i=1}^{n} \lambda_i/n$ relates to the average control energy. The Gramian $W$ depends on the chosen driver nodes according to

$$W = \sum_{i=k_1, \ldots, k_m} W^{(i)},$$

where $W^{(i)}$ is the controllability Gramian when $v_i$ is the only driver node. The centrality $p_i = \text{Tr}(W^{(i)}), i \in 1, \ldots, n$, was proposed in [9] for the maximization of $\text{Tr}(W)$. It can be interpreted as the excitation energy of the network from a unit pulse input applied to $v_i$ and it is a measure of the network impact of $v_i$. Driver node placement according to $p_i$ reduces the average energy needed to steer the network but tends to perform badly when for instance the metric $\lambda_{\min}(W)$ is considered [8].

The ability to control a node indirectly from other driver nodes is another characteristic that is essential in this context. Nodes that are hard to control indirectly need to be driver nodes themselves. Since this is not captured by the quantity $p_i$, we introduce a new centrality measure based on the fictitious output equation $y(t) = e_i^T x(t)$ and the corresponding observability Gramian

$$M^{(i)} = \sum_{t=0}^{T-1} (A^T)^t e_i e_i^T.$$  \hspace{1cm}

The centrality $q_i = \text{Tr}(M^{(i)})$ is low for nodes with few and weak incoming edges. It has a dual interpretation as compared to $p_i$; $q_i$ is the total energy that reaches the node $v_i$ from all other nodes. Choosing nodes with low $q$-centrality as driver nodes limits the energy that is required for steering the network in the most difficult directions.

Rankings of the nodes for driver node placement can be constructed by combining the network centralities $p$ and $q$. When the matrix $A$ is normal, i.e. $A^T A = A A^T$, then $p_i = q_i \forall i$, meaning that the nodes that are most suitable as driver nodes considering the $p$ centrality are the worst nodes considering the $q$ centrality. This is the case with for instance undirected networks. For directed networks we propose the rankings $r_{\text{diff},i} = p_i - q_i$ and $r_{\text{quot},i} = p_i/q_i$ because they exploit the non-normality of the network. Driver node placement based on these rankings results in reduced energy requirements for controlling the network, i.e. both the metrics $\text{Tr}(W)$ and $\lambda_{\min}(W)$ are simultaneously improved w.r.t. e.g. random driver node placement, see Figure 1. In this Figure we use the directed scale-free network model suggested in [2] to generate random networks with 500 nodes, indegree distribution $P_{\text{in}}(k_{\text{in}}) \propto k_{\text{in}}^{-3.14}$, and outdegree distribution $P_{\text{out}}(k_{\text{out}}) \propto k_{\text{out}}^{-2.88}$. Furthermore, problems with minimal controllability are avoided by the addition of self-loops and edges that guarantee strong connectivity.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. Control energy metrics of random directed scale-free networks. The displayed values show the averages over 1000 realizations of networks with 500 nodes. Driver nodes are placed based on the proposed rankings and compared with the set of driver nodes that maximize $\text{Tr}(W)$ (denoted $\text{max Tr}(W)$) and random placements. Both the metrics $\lambda_{\text{min}}(W)$ and $\text{Tr}(W)/n$ improve as the number of placed driver nodes increases from $m = 100$ to 250, but they are always better for the proposed rankings (red, blue) as compared to random placements (green).

References

Alternative routing maps with smaller search information

Hengda Yin, Richard G. Clegg and Raúl J. Mondragón
Queen Mary University of London, School of Electronic Engineering and Computer Science, Mile End, London E1 4SR, United Kingdom, h.yin@qmul.ac.uk

1 Introduction

In a communications network, path diversity is related to its robustness as alternative routes can maintain the information flow if for some reason a route is not available. The existence of alternative routes is related to the existence of cycles in the network, if a branch of a cycle is not available, the traffic can be delivered via the other branch of the cycle. However it could be that many different cycles contain the source and destination nodes creating a large amount of possible alternative paths. Our aim here is to generate networks that have the same path diversity as a given network but are smaller in size and the information needed to describe the alternative routes is also small. The applications involve the simplifications of large-scale networks in different areas (i.e. Internet network and transport network).

The deforestation algorithm [1] is a good candidate to obtain a simpler description of networks alternative paths as it generates smaller networks with the same path diversity. The algorithm is based on link–contraction, that is the agglomeration of nodes that do not contribute to alternative paths into a super–node with the restriction that the agglomeration should not introduce multilinks. For example in Fig 1(a), node 1, 2, 3, 4 can be aggregated into a super–node A. The aggregation stops with the network shown in Fig. 1(b) as it is not possible to aggregate more nodes without introducing multilinks. However, the deforestation algorithm can produce simplified networks with the same path diversity but different connectivity [4](Figs. 1(c)–(d)). This is because the connectivity of the simplified networks depends on the order that the nodes are aggregated, as a consequence, whether the simplified network has a simpler routing map is still unknown.

Rosvall et al. [4] introduced the Search Information to judge whether a network is difficult to navigate. This information measures the average information needed to send a signal between source and destination via the shortest path. A network is easier to navigate if its search information is low. For each shortest path \( \ell(s,t) \), the probability that node \( s \) can locate node \( t \) is: \( P[\ell(s,t)] = \frac{1}{k_s} \prod_{j \in \ell(s,t)/st} \frac{1}{k_j} \), where \( j \) denotes the nodes in the shortest path \( \ell(s,t) \) excluding the source \( s \) and destination \( t \) nodes and \( k_j \) is the degree of the node \( j \). The search information needed to locate a node \( t \) from a node \( s \) is: \( S(s \rightarrow t) = -\log_2 P(s \rightarrow t) \). Based on the assumption that a routing path is easier to describe if its search information is small, we distinguish different simplified networks via the average of \( S(s \rightarrow t) \) for all shortest paths for all possible source \( s \) and destination \( t \) pairs.
Fig. 1: An example of deforestation. (a) The original network and (b) its simplification. (c) A network with many alternative routes (above) and a possible start for its simplification (below). (d) Two different simplified networks of the network in (c) obtained using different starting nodes.

Since the topology of the simplified networks depends on the aggregation sequence, we tested different aggregation strategies: (a) Random sequence. A node and one of its neighbors both are chosen at random. (b) High degree sequence, aggregating the neighbors with high degree first. (c) Low degree sequence, aggregating the neighbors with low degree first. The random strategy represents the situation where there is no information about the neighbors degree. A motivation behind the degree strategies is that high degree nodes have a large contribution to the search information, i.e. node 7 in Fig 1(a). Similarly, a low degree node has a low contribution to the search information. We assume that in the simplified network all the alternative paths have the same weight, and that the information needed to travel inside the super-node is known.

2 Results and Future work

We evaluated the search information of different real networks and its related simplified networks. The results (Table 1) show that, firstly, in most networks the search information of the simplified networks is smaller than the original networks. Also, there is less variability in the value of the search information and in the structure of the simplified networks. However, the simplified Football network has larger average search information. This might be explained by its specific structure as this network has many highly connected clusters. Secondly, different strategies do create simplified networks with different topology which influence their search information. Still there is not a clear cut which strategy is the best. For example in the Dolphins network, applying the low and high degree strategies creates simplified networks with smaller search information. In Polbooks and C. elegans networks, the search information under the different strategies looks similar and in the NetSci network the results are the same.

In the future, this research will be extended to include other aggregation strategies based on biased random walks [5, 2] and weighted links. We will also exploit that the simplified networks tend to be highly clustered so they can be divided into communities using Radicchi et al. method [3]. This method is attractive because it gives a hierarchical
communities structure which in our case can be used to create a hierarchy of routes in different routes domains.

References


Table 1: Number of nodes, edges and search information of several real network and their simplified network. The number inside the brackets is the standard deviation (SD) obtained from 10 simplified networks.
Percolation is progressively shifting from being merely a theoretical concept to becoming a tool for modelling and analysis: many processes in various applied fields can be recast as essentially a percolation problem [1–4]. Yet, with this recasting comes the realisation that complex networks are more then a pattern of connections, they include an additional layer of information that may arguably be even more decisive then the connectivity between the nodes.

Here we introduce a generalisation of bond percolation on complex networks. The conventional bond percolation can be viewed as a continuous time process defined on an interval $p(t) \in [0, 1]$, where $p$ is the probability that an arbitrary edge is present (or alternatively not removed), all edges are present from the start $p(0) = 1$, and none at the end of the process $p(1) = 0$. Suppose the edges are supplied with an additional layer of discrete information. This information can either be coupled to the network topology, as for instance, degrees of incident to the edge nodes, community association, etc., or irrelevant, as for instance, weights, labels, and vector features. In either case, we refer to the set of all possible features as colours $i = 1, 2, \ldots, N$, and to the network as the edge-coloured network. Note, if the features are vectors, their discrete states have to be projected on a line for indexing.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Conventional bond percolation</th>
<th>Colour-dependant percolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupancy probability</td>
<td>scalar, $p \in [0, 1]$</td>
<td>vector, $p \in [0, 1]^N$</td>
</tr>
<tr>
<td>Configuration space</td>
<td>interval, $[0, 1]$</td>
<td>hypercube, $[0, 1]^N$</td>
</tr>
<tr>
<td>Critical point</td>
<td>scalar, $p_c \in [0, 1]$</td>
<td>manifold, $P_c \subset [0, 1]^N$</td>
</tr>
<tr>
<td>Time process</td>
<td>$p(t), p'(t) &lt; 0$</td>
<td>$p(t), p'(t) &lt; 0$, $p(0) = 1$, $p(1) = 0$</td>
</tr>
</tbody>
</table>

Table 1. Correspondence between the concepts in conventional and colour-dependent percolations.
process state rather then a single scalar number. Needless to say, when $p = p_1 = p_2 = \cdots = p_N$ the problem degenerates to the conventional bond percolation. Gradually removing edges form the network in time reflects in the vector being time-dependant $p(t)$, where $t \in [0, 1]$ and $p(0) = 1$, $p(1) = 0$. In this setting, we have a very rich problem: both topology of the network and the colour-labelling scheme contribute to the percolation properties. A few better studied types of percolation can be shown to be partial cases of colour-dependant percolation.

Many concepts from conventional percolation obtain new intriguing properties in the colour-dependant framework, see Table 1 for comparison. All configurations of percolation probability vector $p$ form an $N$-dimensional hypercube and all critical probability vectors $p_c$ form a manifold in this hypercube. Furthermore, a gradual removing of all edges corresponds to a continuous path that joins the vertices $\bar{1}$ and $\bar{0}$ in the hypercube.

![Fig. 1. Critical manifolds for a few examples of networks coloured with $N = 3$ colours. Blue colour indicates the interior of the manifold, red colour – exterior. The yellow lines give examples of evolution paths for the percolation vector throughout time: a. trivial path, $p_1 = p_2 = p_3$; b.c. paths that evolve inside the critical manifold.](image)

**Results**

Consider a coloured network that is defined by a *light-tailed* multidegree distribution. We say that a configuration of the probability vector is critical $p_c$, if it corresponds to a *heavy tailed* component size distribution in the network. By extending the results of our previous works, Ref. [1, 5, 6], we derived a concise criterion that identifies the critical vectors [7]: edge-coloured network features the critical behaviour at $0 < p_c < 1$, if and only if,

$$ v \in \ker[\text{diag}\{p_c\}M - I], \text{ and } \frac{v}{\sum_i v_i} \geq 0, $$

(1)

where

$$ M_{i,j} = \frac{E[k_i k_j]}{E[k_j]} - \delta_{i,j}, \ i, j = 1, \ldots, N, $$

(2)
is a matrix composed of expectation values of the multi-degree distribution, and \( \delta_{i,j} \) is the Kronecker’s delta. The latter criterion can be viewed as a parameter equation for a surface placed in the \( N \)-dimensional space. One may also view this equation as a generalisation of the Molloy and Reed criterion\[8\] to probability vectors. The precise description of how percolation happens in this system is given by the combination of the critical manifold and the path along which \( p(t) \) evolves in time. In Figure 1 we illustrate a few possibilities. What is perhaps most remarkable, is that this path might pierce the critical manifold multiple times (multiple criticalities, Fig.2a.) or even evolve inside the critical manifold (wide criticality, Fig.2b,c). In the latter case, the network maintains infinite expected component size for a long period of time (if we are the thermodynamic limit), whereas below the thermodynamic limit, mean size of the non-giant components features macroscopic fluctuations.

![Graphical Illustration](image-url)

**Fig. 2.** The average size of non-giant components as a function of time. The panels correspond to manifolds depicted in Fig 1. Panel a. depicts multiple critical points, whereas Panels b,c depict an intricate situation when the network becomes critical for a long, continuous period of time.

**References**

Rigidity percolation in random rod packings
Samuel Heroy¹, Dane Taylor², Feng Shi³, M. Gregory Forest⁴, and Peter J. Mucha⁴

¹ Oxford University, Mathematical Institute, Oxford OX2 6GG, UK, heroy@maths.ox.ac.uk
² University at Buffalo, Department of Mathematics, Buffalo, NY 14260, USA,
³ University of North Carolina, Odum Institute for Research in Social Science, Chapel Hill, NC 27599, USA,
⁴ University of North Carolina, Department of Mathematics, Chapel Hill, NC 27599, USA

1 Introduction

In composite materials research, the widely studied bond percolation problem has been applied in numerous settings, both experimental and theoretical [1, 7–10]. Typically, these composites consist of dielectric polymer and conducting particles—when these particles become packed with sufficient density to form a giant connected component, the composite experiences a conductivity phase transition as its bulk conductance increases by several orders of magnitude. Interestingly, in many composites a similar phase transition in mechanical properties is observed at higher particle loading densities [1, 7]. In a recent experimental study of carbon nanotube (CNT)/polymer composites [7], Penu et al. conjecture that mechanical (or ‘rheological’) percolation occurs simultaneously with the formation of a ‘rigid CNT network.’ In the current modeling study, we explore this hypothesis computationally by simulating rod dispersions and using a graph compression algorithm to identify rigid networks within such dispersions. Furthermore, we identify a phase transition associated with the formation of a rigid network, and show that this transition coincides with a simple mean field prediction.

2 Methodology

Our modeling system is rather simple—we model CNTs as cylindrical rods, and contacts between them as idealized hinges, such that two interacting rods stay fixed at but may rotate about their ‘point’ of contact. Then we identify the presence of a giant rigid component (rigidity percolation), wherein the rods forming this component are not only connected but furthermore the constraints resulting from hinge-like contacts eliminate any nontrivial degrees of freedom within the component. We study the evolution of such a component in systems of randomly placed and oriented rods, using Monte Carlo sampling to determine the dependence of the rigidity percolation threshold on rod aspect ratio, system size, and number density. While previous modeling studies have characterized elasticity in 3D fiber systems [6], these studies rely on specific constructions (e.g. perturbed lattices) distinct from the random rods setting. The method we introduce below has the advantage of being general to any construction with hinge-like contacts, and moreover our study is the first application of rigidity analysis to random rod systems.

Our rigidity analysis is based on the identification of certain topological ‘rigid motifs,’ rules for how rigid bodies connect to form larger rigid bodies—e.g. three rods
intersecting pairwise form a single rigid body (in either 2- or 3-dimensional space). We prove these hierarchical rigid motifs analytically using rigidity matroid theory [2], and integrate them into an iterative algorithm—Rigid Graph Compression (RGC)—which we use to decompose large ensembles into mutually rigid sets of rods (see Fig. 1a).

We first verify our method in 2D, finding that three rigidity motifs are sufficient to accurately estimate the rigidity percolation threshold and correlation length parameter associated with the formation of a spanning rigid component [4]. Then, we apply RGC to 3D rod networks, using seven motifs, and robustly estimate the rigidity percolation threshold and associated correlation length exponent for three different rod aspect ratios. Being that this method rests on an incomplete set of provably rigid motifs, an output label from RGC of ‘rigid’ is true (i.e. there are no false positives), but the method may fail to detect some rigid samples. Drawing from the verifiable success of RGC in two dimensions, we provide several pieces of evidence that our analysis captures rigidity percolation in 3D rod systems with sufficient accuracy. Specifically, we use a graphlet-based analysis to show that 3D-RGC only fails to identify relatively unlikely motifs, and then a reshuffling approach to ensure consistent results in large systems.

3 Results and Discussion

Whereas the critical volume fraction required for rigidity percolation in 3D rod systems depends strongly on aspect ratio, this threshold apparently has constant dependence on the system’s mean degree (equal to \(2N_c/N_r\), where \(N_c\) denotes the total number of contacts endured by \(N_r\) rods) across differing aspect ratios (Fig. 1b). Furthermore, this dependence agrees strongly with a mean field prediction in Maxwell’s isostatic condition, which is given by equating the number of constraints in a \(D\)-dimensional system \((D \times N_c)\) with the number of degrees of freedom in that system \((N_r \times (2D - 1))\) for axisymmetric rods, or, in three dimensions:

\[
3N_c = 5N_r
\]  

Solving for the mean degree, we find \(2N_c/N_r = 10/3\), which agrees remarkably well with our experimental results (Fig. 1b). Interestingly, this mean field approach, which assumes erroneously that all constraints are independent, grossly underpredicts the rigidity percolation threshold in analogous 2D rod systems [4, 5]. Apparently, this effect is less significant here. Moreover, Maxwell’s condition requires the entire system having \(\geq 10/3\) contacts per rod—however, rigidity percolation only requires this to be true for rods within the giant rigid component, and so (1) might even slightly overpredict the rigidity percolation threshold.

This result has interesting theoretical and practical implications for materials scientists and engineers, suggesting a number of experiments that could be used to further investigate the connection between this rigidity mechanism and the observed mechanical percolation phase transition. However, real composite materials differ from our idealized simulation systems—in particular, if the filler phase consist of CNTs, these will tend to bundle and cluster [3], presenting a picture quite unlike our uniformly distributed systems. We begin to explore this complicated parameter space, showing that certain experimental results are reasonable given our theoretical model.
This work is supported by the U.S. Army Research Office (#W911NF-13-1-0013, #W911NF-16-1-0356).

Fig. 1. (a) In our simulation scheme, rods are scattered randomly in a cubic domain with periodic boundary conditions. We infer a network representation wherein nodes represent rods and edges represent pairwise contacts. Within this network, RGC searches for rigid motifs (right)—provably rigid subgraphs, and then hierarchically compresses these motifs into single nodes. As RGC proceeds, many rods may emerge into one node (here, node ‘1’), which represents a giant rigid component. (b) Early results suggest that rigidity percolation, here represented by a phase transition in the mean size of the largest rigid component, is associated with a threshold mean degree of $\approx 10/3$ across varying aspect ratios.

References

Redundancy in the Structure and Dynamics of Complex Networks

Rion Brattig-Correia\textsuperscript{1,2,3}, Alexander J. Gates\textsuperscript{4}, Nathan Ratkiewicz\textsuperscript{1}, Alain Barrat\textsuperscript{5,6}, Luis M. Rocha\textsuperscript{1,3*}

1 School of Informatics, Computing, and Engineering, Indiana University, Bloomington IN, USA
2 CAPES Foundation, Ministry of Education of Brazil, Brasília, DF, Brazil
3 Instituto Gulbenkian de Ciência, Portugal
4 Center for Complex Networks Research, Northeastern University, Boston, MA, USA
5 Aix-Marseille Université, CNRS, CPT, Marseille, France
6 Data Science Laboratory, Institute for Scientific Interchange (ISI) Foundation, Torino, Italy

* correspondence: rocha@indiana.edu

Understanding complex networked systems is key to solving some of the most vexing problems confronting humankind, from discovering how thoughts and behaviors arise from dynamic brain connections, to preventing the spread of disease. However, a critical gap remains in understanding how the structure of networks affects the dynamics of complex systems, which we have been addressing with methods to compute and remove redundancy from them. Whereas patterns of connectivity (structure) have been amply studied in complex networks & systems (CNS), and patterns of dynamics have been studied in dynamical systems theory, much remains to be done to characterize redundancy and its importance for understanding and controlling the dynamics of complex systems—which, ultimately, defines their function. In this talk we will present our recent studies of redundancy in both the structure and dynamics of complex systems.

The first concept we present is the invariant sub-graph that is revealed by the computation of all shortest paths (metric closure) of a weighted graph \cite{1,2}. We refer to this subgraph as the metric backbone of a complex network \cite{4}. The size of the backbone subgraph, in relation to the size of the original graph, defines the amount of redundancy in the network: edges not on this backbone are superfluous in the computation of shortest paths. We demonstrate the utility of the metric backbone with an analysis of the SocioPatterns datasets \cite{5}, which yield duration of contact between pairs of individuals (via wearable sensors), and have been used in models of epidemic spread to evaluate containment policies \cite{6,7}. We show that the SocioPatterns contact networks are very redundant, with the proportion of semi-metric edges ranging from 50-91%—four of the six networks with larger than 80% redundancy. This means that all shortest paths can be computed with fewer than 50 to 9% of the edges (which comprise the metric backbone). The figure shows the primary school contact network \cite{3} and its metric backbone, which contains only 9% of the edges in the original network. Importantly, the social structure of the contact network is preserved in the backbone subnetwork—the preservation of community structure in the metric backbone is observed in all the six SocioPatterns networks. Finally, we compare epidemic spread simulations, using the original and backbone networks, for different spreading rates to demonstrate the utility of the metric backbone in dynamical processes on networks.

The second concept stems from a dynamics perspective. Based on recent work \cite{9,10} we show that the control of complex networks crucially depends on redundancy that exists at the level of variable dynamics. To understand the effect of such redundancy, we study automata networks—both systems biology models and large random ensembles of Boolean networks (BN). In these discrete dynamical systems, redundancy is conceptualized as canalization: when a subset of inputs is sufficient to determine the output of an automaton. We discuss two types of canalization: effective connectivity and input symmetry \cite{9}. First, we show that effective connectivity strongly influences the controllability of multivariate dynamics. Indeed, predictions made by structure-only methods can both undershoot and overshoot the number and which sets of variables actually control BN\cite{10}. To understand how control and information effectively propagate in such complex systems, we uncover the effective graph that results after computation of effective connectivity \cite{11}. To study the effect of input symmetry, we further develop our dynamics canalization map, a parsimonious dynamical system representation of the original BN obtained after removal of all redundancy \cite{2}. Mapping canalization in BN via these representations allows us to understand how control pathways operate, aiding the discovery of dynamical modularity and robustness present in such systems. Finally, we demonstrate the utility of the approach with the analysis of a battery 50+ systems biology automata networks, including systems biology models of cancer.

The 7\textsuperscript{th} International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Figure 1: The primary school contact network [3] (left) and its metric backbone (right). The network was built from 125,773 records of contact between 242 individuals (232 students and 10 teachers. Data collection lasted 2 days, between October 1st and 2nd in 2009. Nodes are colored according to classroom membership (10 classrooms), and the 5 grades are shown with lighter and darker shades of the same color; teachers are colored in black. Metric backbone contains 9% of the edges of original network. Rendering of graph community structure with ForceAtlas2 [8] performed independently for the network and its backbone, showing preservation of community structure: main student communities (10 classrooms), and community pairs (5 grades) did not change from those in original network, neither did most of the teacher nodes.

References
Percolation transition on scale-free networks with assortative degree correlation

Toshihiro Tanizawa
National Institute of Technology, Kochi College,
200-1 Monobe-Otsu, Nankoku, Kochi, JAPAN
tanizawa@ee.kochi-ct.ac.jp

1 Introduction

Though the main trend in research on network science at present seems to have shifted to a complex system of networks, such as multiplex or multilayered networks, fundamental and important questions regarding to the physical properties of a single network still remain, some of which can be found in percolation transition on networks with degree correlation between node connection.

In the work in 2007, Noh reported the possibility that the universality class of the percolation transition of networks with tendency in connection between nodes with similar degrees (assortative degree correlation) is different from the universality class of the percolation transition of networks with no specific tendency in node connection (neutral degree correlation) or networks with tendency in connection between nodes with different degrees (disassortative degree correlation) [1]. Noh drew this conclusion from the results of his numerical simulations, in which the peak in the finite size cluster at the transition does not appear for ER networks with assortative degree correlation.

In contrast, Tanizawa et al. analytically calculated several properties of the limiting case of scale-free networks with assortative degree correlation, which is hierarchically connected random regular graphs and found that the critical exponent $\beta$ defined by $S \sim |p - p_c|^\beta$ for the giant component fraction $S$ against targeted node removal satisfies $\beta = 1$, which is the “mean field” value [2].

This work intends to clarify the discrepancy.

2 Model

In order to tackle this problem by numerical simulations, I first generate scale-free networks with the power-law degree distribution,

$$P(k) \propto k^{-1}, \quad (1)$$

by the configuration model algorithm. In this work, the total node number is 10000, the minimum degree is one, and the maximum degree is 50.

Next, I introduce positive degree correlation between nearest neighbor nodes into these neutral networks by edge swapping according to the joint degree probability between the degree of node $i$ ($k_i$) and the degree of node $j$ ($k_j$),

$$P(k_i, k_j) \propto \frac{1}{1 + a |k_i - k_j|}, \quad (2)$$
with a controlling parameter, \( a(\geq 0) \). This joint degree probability is the same type as the one adopted in the work of Wu and Holme [4]. For \( a = 0 \), no degree correlation is induced and the network remains neutral. As the value of \( a \) increases, the tendency of making connection between nodes of almost same degrees \((k_i \approx k_j)\) increases and the network becomes more assortative.

In this work, the assortativity is measured by the index \( r \) introduced by Newman [3]:

\[
r = \frac{\sum_i e(k_i, k_i) - \sum_i a_i b_i}{1 - \sum_i a_i b_i},
\]

where \( e(k_i, k_j) \) is the fraction of the edges from a node with degree \( k_i \) to a node with degree \( k_j \), \( a_i = \sum_j e(k_i, k_j) \) and \( b_i = \sum_j e(k_j, k_i) \). In this work, networks are undirected. Hence, \( e(k_i, k_j) = e(k_j, k_i) \) and \( a_i = b_i \).

3 Results

Figure 1 shows the results of numerical simulations obtained from the average over 500 sample runs. Figure 1(a) shows that the assortativity increases monotonically as the increase of \( a \) and saturates at about 0.8 for \( a = 5 \). I verified from the histogram of the numbers of edges between nodes of degree \( k \) and \( k' \) that the topological structure at \( a = 5 \) is almost onion-like.

In Fig. 1(b), no peculiar behavior can be seen in the giant component collapse during uniform node removal apart from the abrupt initial downturn for assortative networks \((a > 0)\).

Figure 1(c) shows, however, interesting behaviors in the second largest connected cluster. For neutral networks \((a = 0)\), the size of the second largest cluster has a single peak at the transition where the giant component vanishes. The peak disappears, however, for assortative networks \((a > 0)\). Instead, another peak grows at the initial stage of node removal followed by a growing dip at about 70% of remaining node fraction as the value of \( a \) increases. The plateau between 10% and 60% of remaining node fraction for \( a = 3 \) begins to develop other fine structures for \( a \geq 4 \) suggesting the appearance of the second peak at about 55% of remaining node fraction.

4 Discussion

As shown in Fig. 1(c), the peaks in the profiles of the finite size cluster at transition disappear for scale-free networks with assortative degree correlation, which is similar to the results of Noh [1] dealing with correlated ER networks. However, there are also differences. Contrary to the case of Noh, other peaks seem to grow according to the increase of assortativity. This suggests that networks with assortative degree correlation behave more likely as an assembly of hierarchically connected random regular graphs than a single entirely connected object. On my opinion, this viewpoint seems to be lacking in Noh’s discussion.

It is true that we cannot draw any decisive conclusion solely from numerical simulations. I am also taking an analytical approach to clarify this problem by extending a system of equations used in the works of Tanizawa et al. [2] and Mizutaka et al. [5].
Fig. 1. Results obtained from numerical simulations of site percolation on scale-free networks with total nodes $N = 10000$, degree exponent $\lambda = 1$, the minimum degree $m = 1$, and the maximum degree $K = 50$ averaged over 500 runs: (a) The assortativity index defined by Newman as the function of controlling parameter, $\alpha$, in the joint degree probability, Eq. (2). As the value of $\alpha$ increases, the assortativity increases from zero and almost saturates at about 0.8 for $\alpha = 5$. (b) The collapse of the largest connected component fraction against uniform node removal. (c) The node number of the second largest connected component due to uniform node removal.

Summary. Percolation transition on scale-free networks with assortative degree correlation is studied by numerical simulations. As in the previous work of Noh [1], the peak in the finite size cluster at transition disappears. However, other structures also appear as the assortativity increases, which suggests that the hierarchical (onionlike) structure of networks with highly assortative degree correlation might play a vital role in this percolation transition.

Acknowledgment

This work is supported by Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (JSPS KAKENHI), Grant Number 17K05590.

References

Heterogeneity in Percolation on Complex Networks

Reimer Kühn\textsuperscript{1} and Tim Rogers\textsuperscript{2}

\textsuperscript{1} Mathematics Department, King’s College London, Strand, London WC2R 2LS, UK,
\textsuperscript{2} Centre for Networks and Collective Behaviour, Department of Mathematical Sciences,
University of Bath, Bath, BA2 7AY, UK
reimer.kuehn@kcl.ac.uk, t.c.rogers@bath.ac.uk

In this contribution we report results we have recently obtained which reveal that there is a considerable degree of heterogeneity in the problem of percolation on complex networks, which has so far largely escaped attention of the community working in this field.

To be specific, we consider bond percolation, but remark at the outset that analogous results can easily be obtained for site percolation. For fixed networks, bond percolation is defined as follows: starting from an arbitrary large (connected) network, we consider each edge independently, keeping it with probability $\rho$ and deleting it with probability $1-\rho$. The largest connected component remaining after this random edge removal process is referred to as the percolating cluster or giant component; write $S$ for its size measured as a fraction of the total number $N$ of nodes in the network. For large sparse networks it was shown in \cite{1,2} that this quantity can be computed to close approximation using a message-passing protocol. Knowledge of $S$ gives global information about the robustness of a network to attack or infection; in particular, there is a critical value of $\rho$ below which no percolating cluster survives in the thermodynamic limit, and $S = 0$. One of the main results of \cite{2,3} was to identify the percolation threshold $\rho_c$ as the reciprocal of the largest eigenvalue of the non-backtracking (or Hashimoto) matrix that encodes the relationship between variables in the message-passing equations.

Here we take a much more detailed look and consider for each node $i$ of the network the probability $\langle \sigma_i \rangle$ that it remains part of the giant component when averaged over many instances of the percolation process. The heterogeneity in the responses of individual nodes to percolation is captured by the empirical distribution of the $\langle \sigma_i \rangle$ defined as

$$\varphi(s) = \frac{1}{N} \sum_i \delta(s - \langle \sigma_i \rangle).$$  \hspace{1cm} (1)

In a similar vein, for nodes that do not remain part of the giant component, we look at the average size $\langle n_i \rangle$ of the finite cluster to which node $i$ belongs, as well as its empirical distribution

$$\psi(n) = \frac{1}{N} \sum_i \delta(n - \langle n_i \rangle).$$  \hspace{1cm} (2)

We evaluate these distributions for single large instances of complex networks as well as in the thermodynamic limit $N \to \infty$ for the configuration model class. Figure 1 displays two examples.

We also perform a weakly non-linear analysis of the message passing equations that allow one to obtain the site dependent percolation probabilities $\langle \sigma_i \rangle$ and the average cluster sizes $\langle n_i \rangle$ in the vicinity of the percolation transition at $\rho_c$. Figure 2 compares...
results of full solutions of the message passing equations with those of the weakly non-linear analysis carried to second order in $|\rho - \rho_c|$, for a selected subset of vertices for a single large instance of a network with power-law degree distribution.

![Graph showing distribution of probabilities to be part of the giant cluster for the percolation problem on a graph with power-law degree distribution, $p(k) \propto k^{-3}$ with $k_{\text{min}} = 2$, at $\rho = 0.5$ (black). Results for the thermodynamic limit shown together with an unfolding according to degree for $k = 2, 3, \ldots, 9$, and $\{k \geq 10\}$ (blue, red, green, ... curves showing dominant contrib from shifting to higher $s$ with increasing $k$). Right panel: Distribution $\psi(n)$ of average cluster sizes for percolation on an Erdős-Rényi network of mean degree $c = 4$ at $\rho = 0.3$. Results for the thermodynamic limit, shown together with an unfolding according to degree for $k = 0, 1, 2, \ldots, 9$, and $\{k \geq 10\}$ (blue, red, green, ... with peak positions shifting to higher $n$ with increasing $k$). From [4].](image1)

![Graph showing probability $\langle \sigma_i \rangle$ to appear in the percolating cluster (solid lines), compared to the second order expansion (dashed lines). Right panel: expected size $\langle n_i \rangle$ of finite clusters containing given nodes (solid lines), compared to the second order expansion (dashed lines). Both panels show results for a selected number of vertices from a graph with power-law degree distribution, $p(k) \propto k^{-3}$ with $k_{\text{min}} = 2$. From [4].](image2)

We note that our results could be of interest in the context of analyzing cascading failures, epidemic spreading, or probabilistic information spreading in complex networks. The considerable detail which our methods allow to unearth might be useful for instance in the design of optimal vaccination strategies that exploit information beyond degree. Indeed the fact that percolation probabilities conditioned on degrees are themselves broadly distributed is for instance a clear indicator of the fact that the so-called degree based approximation (which postulates that heterogeneity is fully characterized by degree), on which the majority of attempts to design optimal vaccination strategies...
has been based, misses important information which could be exploited to improve upon such strategies.

References

Dynamic competitive opinion control: theory, simulations, and experiments

Valerio Restocchi¹, Markus Brede¹ Sebastian Stein¹, Lewis Hill¹, and Soheil Eshghi²

¹ University of Southampton, SO17 3SX, Southampton, UK
valerio.res@gmail.com,

² Yale University, 06510, New Haven, CT, USA

1 Introduction

Opinion control, i.e., the study of strategically influencing agents on social networks with the aim to align their opinions, behaviours, or choices with certain targets, has been extensively studied in competitive and non-competitive scenarios, mostly via variants of models based on the seminal independent cascade model [1]. However, these models may not be appropriate in situations in which agents are subject to various sources of social influence, and decisions can be changed over time. The voter model, by allowing agents to repeatedly change their opinion, provides a more accurate description of the underlying opinion dynamics mechanism, but, although recent work has started to address the problem of opinion control in the voter model [2], to the best of our knowledge these studies focussed on static control and have rarely addressed competition among multiple parties. In this paper, we discuss our most recent results on dynamic and competitive opinion control. Our approach brings together theoretical results, agent-based simulations, and experimental work with the goal of gaining a more comprehensive understanding of this topic.

2 Results

In our simulations and theoretical work, we consider the case of a strategic influencer (SI) who is competing with one opponent, and we focus on the dependency of optimal strategies on time horizons and network degree heterogeneity. Our results suggest that low-degree nodes represent better control targets when the time horizon is short, whereas hubs are ideal targets for long time horizons. Moreover, our results suggest that optimal strategies vary depending on the initial state of the network. Importantly, we find that, when many nodes are aligned with the SI’s opinion, its best strategy is to neutralise its opponent’s influence. Conversely, when most voters hold a different opinion than that of the SI, the best strategy is to directly target opposing low-degree nodes and avoid the opponent’s influence. These results are displayed in Fig. ??.

We also find that network heterogeneity plays an important role in finding the SI’s best response to its opponent. In particular, we find that optimal control strategies vary markedly depending on in-degree heterogeneity with optimal control being the more focused on out-degree hubs the less in-degree heterogeneous the network. Moreover, by analysing the dependence of the resulting equilibria on resource endowments of...
Fig. 1. Dependence of optimal influence strategies on time horizons $T$ of the SI (A) against an opponent that uses a random allocation strategy (B): figure (a) represents the average degree of nodes influenced by A, and (b) the relative standard deviation of distribution of influenced degrees. Bottom row: (c) fraction of opposing initial opinions, (d) fraction of neighbors of influenced nodes which have opposing initial conditions, and (e) fraction of mutually influenced votes. Both A and B have the same resources and initial states are equally distributed between A and B. The lines in (c) and (d) and (e) indicate a range of one standard deviation around the expectation for a random allocation.

Fig. 2. Screenshot of the game. The players can try to influence nodes by deploying 'influence tokens', of which they have a limited amount. One influence token counts as an extra node of that colour being linked to the node. The bars above the nodes represent their probability of being in either state (red or blue) at the beginning of the next round.
the competing influencing parties, we show that resource-poor opponents gain a dis propor tionate advantage by exploiting out-degree heterogeneity, whilst resource-richer opponents prefer homogeneous network architectures.

For our experimental work, we build an opinion control game (Fig. ?? in which human subjects play either against other human subjects or against intelligent agents. The goals of this experiment are twofold: first, we aim at understanding people’s decision-making process in a dynamic opinion control scenario. Second, we intend to find dominant strategies that emerge when the participants play repeated games against multiple strategies. In this poster, we will discuss preliminary results from the experiments, as well as an outline of future research directions and challenges that can be addressed with this approach.

References

1 Introduction

Networks provide a useful paradigm to incorporate contact patterns and various heterogeneities within a population. The basic ingredients of such models are nodes and links, usually representing individuals and the contacts between them, but they may represent also groups of individuals (such as the population at some geographic location), and the connectedness of these groups (such as transportation routes). In simple disease outbreak models, the status of an individual can be susceptible \((S)\), infected \((I)\) or recovered \((R)\). A key parameter associated with most epidemic models is the basic reproduction number (denoted by \(R_0\)), which denotes the expected number of secondary infections generated by a typical infected individual introduced into a fully susceptible population. The reproduction number is also a threshold quantity: if \(R_0 < 1\) the epidemic will die out, while if \(R_0 > 1\) the disease will spread. Another important measure of epidemic severity is the final epidemic size, which is the total number of individuals who become infected during the time course of the epidemic. These two quantities are often connected via the so-called final size relation.

Pairwise models have been successfully used to approximate stochastic epidemics on networks and represent an improvement on compartmental models. The former are formulated in terms of the expected values for the number of susceptible \(\langle S \rangle\), infected \(\langle I \rangle\) and recovered \(\langle R \rangle\) nodes, which depend on the expected values of \(\langle SS \rangle\) pairs \(\langle SS \rangle\) and \(\langle SI \rangle\) pairs \(\langle SI \rangle\). Upon using a closure approximation formula for the triplets of nodes, the pairwise SIR model reads as

\[
\begin{align*}
\dot{\langle S \rangle}(t) &= -\tau \langle SI \rangle(t), \\
\dot{\langle I \rangle}(t) &= \tau \langle SI \rangle(t) - \gamma \langle I \rangle(t), \\
\dot{\langle SS \rangle}(t) &= -2\tau \frac{n-1}{n} \frac{\langle SS \rangle(t) \langle SI \rangle(t)}{\langle S \rangle(t)}, \\
\dot{\langle SI \rangle}(t) &= \tau \frac{n-1}{n} \left( \frac{\langle SS \rangle(t) \langle SI \rangle(t)}{\langle S \rangle(t)} - \frac{\langle SI \rangle(t) \langle SI \rangle(t)}{\langle S \rangle(t)} \right) - (\tau + \gamma) \langle SI \rangle(t).
\end{align*}
\]
where $\tau$ is the per contact infection rate, $\gamma$ is the recovery rate, $n$ is the average degree, $[S] + [I] + [R] = N$ is the total number of nodes in the network, and only those equations are listed which are necessary to derive a complete self-consistent system.

2 Non-Markovian Recovery

The Markovianity of the recovery process is a strong simplifying assumption. For many epidemics, the infectious period has great importance and it can be measured empirically. Recently, pairwise approximations of the SIR dynamics with non-Markovian recovery have been derived. In the special case of fixed recovery time $\sigma$, the pairwise model turned out to be

\[
[S](t) = -\tau [SI](t),
\]

\[
[SS](t) = -2\tau \frac{n-1}{n} \frac{[SS](t)[SI](t)}{[S](t)},
\]

\[
[I](t) = \tau [SI](t) - \tau [SI](t - \sigma),
\]

\[
[SI](t) = \gamma \frac{n-1}{n} \frac{[SI](t)[SI](t)}{[S](t)} - \tau \frac{n-1}{n} \frac{[SI](t)[SI](t)}{[S](t)} - \tau [SI](t)
\]

\[-\tau \frac{n-1}{n} \frac{[SI](t - \sigma)[SI](t - \sigma)}{[S](t - \sigma)} e^{-\int_{0}^{\sigma} \frac{n-1}{n} \frac{[SI](u)}{[S](u)} du}. \tag{2}\]

Considering a general distribution for the recovery period, the pairwise model can be formulated as a system of integro-differential equations, which we omit here.

3 Results

For arbitrary infectious periods, the basic reproduction number of the pairwise model is

\[
R_0^p = \frac{n-1}{N} (1 - \mathcal{L}[f_\sigma](\tau)) [S]_0, \tag{3}\]

where $\mathcal{L}[\cdot]$ is the Laplace transform and $f_\sigma$ is the probability density function of the recovery process given by the random variable $\mathcal{S}$. Numerical tests and analytical results have both confirmed that, in general, the following implicit relation for the final epidemic size holds

\[
\frac{n-1}{\frac{1}{n} - 1} = R_0^p \left( \frac{\frac{n-1}{s_{\text{un}}} - 1}{\frac{n-1}{s_{\text{un}}} - 1} \right) = \frac{n-1}{N} (1 - \mathcal{L}[f_\sigma](\tau)) [S]_0 \left( \frac{\frac{n-1}{s_{\text{un}}} - 1}{\frac{n-1}{s_{\text{un}}} - 1} \right), \tag{4}\]

where $s_{\text{un}}$ is the fraction of nodes unaffected by the epidemics. We shall analyze how the basic reproduction number (3), which is not only an epidemic threshold but also determines the final size via (4), depends on the variance of the recovery time distribution. Using gamma, lognormal and uniform distributions we show that once the mean infectious period is fixed, smaller variance in the infectious period gives a higher reproduction number and consequently a more severe epidemic. Next we generalize this result without restricting ourselves to special distributions.
4 Summary

Pairwise models have been proven to be a flexible framework for analytical approximations of stochastic epidemic processes on networks that are in many situations much more accurate than mean field compartmental models. The non-Markovian aspects of disease transmission are undoubtedly important, but very challenging to incorporate them into both numerical stochastic simulations and analytical investigations. Here we present a generalization of pairwise models to non-Markovian epidemics on networks. For the case of infectious periods of fixed length, the resulting pairwise model is a system of delay differential equations, which shows excellent agreement with results based on the explicit stochastic simulations. For more general distribution classes (uniform, gamma, lognormal etc.) the resulting models are PDEs that can be transformed into systems of integro-differential equations. We derive pairwise reproduction numbers and relations for the final epidemic size, and initiate a systematic study of the impact of the shape of the particular distributions of recovery times on how the time evolution of the disease dynamics play out.

Joint work with István Z. Kiss, Zsolt Vizi, and Joel C. Miller.

References

5. Vizi Zs, Kiss IZ, Miller JC, Röst G Smaller variance in the infectious periods leads to larger reproduction number and more severe outbreak for network epidemics, arxiv preprint, arXiv:1712.06026
Modularity and stability in bipartite networks with extinctions and plasticity

Somaye Sheykhali, Juan Fernández-Gracia, and Víctor M. Eguíluz

Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), E07122 Palma de Mallorca, Spain

1 Introduction

The tolerance of complex systems to the spreading impacts of perturbations has been the subject of numerous theoretical and experimental studies [1]. Biological systems, and in general any complex system, are expected to withstand the loss of elements, either by random failure or driven by directed perturbation (e.g., environmental change or targeted attacks) [2–4]. In the context of ecology, loss of biodiversity as a consequence of environmental perturbations disrupts ecosystems and their functioning. Ecological adaptation is often expected to mitigate the consequences of species extinction by reducing the risk of cascading extinction and allowing species persistence.

Recent studies of ecosystems have revealed key features of the topological structure of the interactions of the species that compose them: modularity, that captures the block structure [5, 6], stability, which can be measured as the largest eigenvalue of the appropriate matrix [7], and robustness, defined in terms of the largest connected component [8, 2]. Theory predicts that networks with a clustered or modular structure of species interactions benefit a dynamical stability as a consequence of localized dispersion of extinction [9].

Here we investigate the robustness, modularity and stability of mutualistic bipartite ecological networks under a process of species extinction - adaptation. For the extinction we test two conditions, random extinction and directed extinction. As for the adaptation, we consider rewiring of links with two variants: random rewiring and co-occurrence-driven rewiring. To be specific, by co-occurrence rewiring with consider that rewiring only occurs to other species which share neighbors with the extinct species. Altogether we have 4 extinction-rewiring scenarios: random extinction–random rewiring (RR), random extinction–co-occurrence rewiring (RC), directed extinction–random rewiring (DR) and directed extinction–co-occurrence rewiring (DC).

2 Results

We study the evolution of the bipartite networks in the context of extinction and adaptation dynamics. We analyze empirical mutualistic networks extracted from real data [10]. For the sake of clarity we show results only for the three largest networks of different mutualistic interaction types: plant–pollinator, seed–dispersal and plant–ant. See Table 1 for their basic network characteristics. The results hold in general for all the networks of our study.
Table 1. Sample mutualistic ecological bipartite networks. The table shows basic quantities of the three bipartite networks exemplifying three different types of mutualistic relations.

<table>
<thead>
<tr>
<th>Ecosystem</th>
<th>Network ID</th>
<th>Type of interactions</th>
<th>N</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pollinators</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Atlantic Forest [12]</td>
<td>Net2</td>
<td>Seed dispersal</td>
<td>207</td>
<td>1121</td>
</tr>
<tr>
<td>Rain forest [13]</td>
<td>Net3</td>
<td>Plants</td>
<td>51</td>
<td>923</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ants</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

The robustness of the system, measured as its ability to keep its nodes connected in a single component, is strongly affected by the extinction scenario, being less robust against preferential extinction. For random extinction the rewiring scenario does not impact the robustness. For directed extinction, co-occurrence rewiring makes it even less robust. Modularity increases as species become extinct and the network size is reduced. Comparing scenarios, modularity also increases with directed extinction and even more with co-occurrence rewiring, similarly to what happens with stability.

Fig. 1. Modularity of the mutualistic networks in the different extinction-adaptation scenarios. Rewiring improves modularity for increasing fraction of extinct nodes. Each column illustrates the modularity of one data set (Table 1) as a function of the probability of rewiring and the fraction of extinct species. Colors shows the $P_r$ (the probability of rewiring).

3 Summary.

Taken together, the analysis of changes of modularity, stability and robustness against species extinction and adaptation shows how three relevant structural properties of bipartite networks, modularity, robustness and stability, evolve as the number of extinctions-adaptation events increase. Our preliminary results suggest that extinction driven by highly connected species followed by co-occurrence rewiring might be an evolutionary attractive mechanism, as it enhances modularity and stability at expenses of topological robustness, compared to the case of random extinctions and random rewiring. The
high modularity in ecological networks makes the network effectively resistant against perturbations, as localized dispersion naturally limits the spread of extinction. Longitudinal data on the structure of bipartite mutualistic ecological networks could confirm our results by showing a preference for co-occurrence rewiring.

References

Optimal attacks on random graphs

Nicole Balashov\textsuperscript{1}, Reuven Cohen\textsuperscript{1}, Simi Haber\textsuperscript{1}, and Michael Krivelevich\textsuperscript{2}

\textsuperscript{1} Department of Mathematics, Bar-Ilan University, Ramat-Gan 5290002, Israel
\textsuperscript{2} Department of Mathematics, Tel-Aviv University, Tel-Aviv 6997801, Israel

1 Summary

We consider optimal attacks or immunization schemes on different models of random graphs. We derive bounds for the number of nodes needed to be removed from a network with \( n \) nodes such that all remaining components are fragments of size \( o(n) \).

We obtain bounds for different regimes of random regular graphs and Erdős-Rényi random graphs, some of which are tight. It can be shown that the performance of adaptive attacks by degree is bounded away from optimality. We also give a randomized polynomial time attack and prove optimal performance in certain cases.

2 Bounds on optimal attacks

One of the most studied questions in complex networks is the resilience of networks under different failure models and attack strategies [1–3]. In particular, one wishes to know the optimal attack strategy that will lead to fragmentation by removal of a minimal fraction of the nodes. This information is important for estimating the vulnerability of network infrastructures, and also for devising optimal immunization strategies for populations and computer networks.

The main methods that have been proposed for targeted attacks on networks via node removal have been based on attack by highest degree [1, 3, 4], and attack by highest betweenness centrality [5]. Some methods based on more advanced algorithms for graph partitioning have also been proposed [6], and led to improved upper bounds on the minimal fraction of nodes that should be removed to fragment a networks.

We define \( c_f \) as the minimal fraction of nodes that are to be removed before the network becomes fragmented into sublinear components. It is clear that for any network \( c_f \leq q_c = 1 - p_c \) where \( p_c \) is the percolation threshold for random removal of nodes.

For a \( d \)-regular graph (where each node has degree \( d \)) this leads to \( c_f < 1 - \frac{1}{d+1} = \frac{d-2}{d+1} \).

Improved bounds were obtained in [7, 8]:

\[
\frac{d-2}{2d-2} \leq c_f \leq \frac{d-2}{d+1}.
\]

When \( d \) is a large constant a lower bound tending to one as a function of \( d \) was given in [12].

In this extended abstract we establish better lower and upper bounds on \( c_f \), focusing on sparse random graphs. In particular for random regular graphs we show

\[
1 - 2\frac{\alpha(G)}{n} \leq c_f \leq 1 - \frac{\alpha(G)}{n},
\]

(1)
where \( \alpha(G) \) is the independence number of \( G \). When \( d \) is large enough the independence number is known \cite{9} to satisfy \( \alpha(G) \approx 2n \ln d/d \). In this case we obtain
\[
c_f \approx 1 - \frac{\alpha(G)}{n}.
\]

We provide matching results for Erdős-Rényi random graphs. For the latter model we have obtained tight results also when the average degree is small.

3 Results

For the analytical results, we mainly exploit structural graph properties such as expansion, domination number and independence and obtain deterministic connections to the shattering number. We then apply known estimations of these parameters for random graphs either directly or via contiguity arguments.

As an example we establish a lower bound on the minimal fraction of nodes needed to be removed in order to shatter a regular network into small components of size \( o(n) \). Consider shattering a graph to \( m \) disjoint clusters of sizes \( C_i \), \( i = 1, \ldots, m \) by deleting a set of \( |S| = c_f n \) nodes. \( c_f \) is therefore the fraction of removed nodes. Notice that \( \sum_{i=1}^{m} C_i = S = n \). Thus, \( \frac{1}{d} \sum_i |C_i| + c_f = 1 \). Denote by \( B_i \) the number of nodes on the boundary of \( C_i \), i.e., the number of neighbors of nodes of cluster \( i \). Since the clusters are disconnected for all \( i, j \), all boundary nodes for any cluster \( i \) must be removed. Therefore, for every \( i \), all nodes \( B_i \) are deleted. Now, in a random regular graph each cluster is locally tree like. Therefore, \( B_i = (d - 1)C_i - C_i = (d - 2)|C_i| \). Since every node has exactly \( d \) neighbors, a node can not participate in more than \( d \) of the \( B_i \)'s. Thus \( \frac{1}{d} \sum_i B_i \leq D \). Therefore,
\[
N \geq \sum_i C_i + \frac{d - 2}{d} \sum_i C_i = \frac{2d - 2}{d} \sum_i C_i.
\]  
(2)

Summarizing the above we get:
\[
c_f = 1 - \frac{1}{N} \sum_i C_i \geq \frac{d - 2}{2d - 2}.
\]  
(3)

This approach can be shown to give asymptotically tight solution for \( d = 3 \) (i.e., it matches the upper bound shown in \cite{7}). However, for large values of \( d \) it deviates considerably from the exact value. Indeed, for \( d \to \infty \), Eq. (3) leads to \( c_f \geq \frac{1}{2} \), where as will be shown below \( c_f \to 1 \).

In order to give a better lower bound on \( c_f \) for random regular graphs with large constant degree we observe the following: Random regular graphs are locally tree like, having a bounded number of short cycles. Therefore after the network is shattered we expect the remaining components to be trees. A tree can be shattered into isolated vertices by removing at most half of its nodes. Therefore, it can be deduced that the number of nodes remaining after the attack is at most twice the size of the largest independent set. Since removing all but an independent set clearly shatters the graph, we obtain Equation (1). With some further effort we can improve the lower bound to close the
gap and match the upper bound. The same approach may be applied to the Erdős-Rényi model $G(N, p = c/N)$ where we get

$$c_f \approx 1 - \frac{\log c}{c}. \quad (4)$$

When $c = 1 + \varepsilon$ for small $\varepsilon > 0$, a different approach is needed. We can show that $c_f$ equals a quarter of the kernel size.

### 4 Algorithmic aspects

Finding an optimal shattering set is NP-hard, but when the input is a random graph the problem becomes tractable, that is, there exists a polynomial time algorithm succeeding with probability tending to one as the number of vertices approaches infinity. We propose the following algorithm for finding a shattering set in a graph. We describe the algorithm and demonstrate its asymptotic optimality for random cubic graphs. Let $G \sim G_{n,3}$ be a random cubic graph.

1. Find a Hamilton cycle $H$ in $G$
2. Start from an arbitrary vertex $v_0$ and advance along $H$ creating a segment.
3. When visiting a vertex $v$, it is incident with two edges on the cycle and a third edge $e$. If $e$ is the second edge in the segment going backward (in $H$), delete $v$.

Each edge is seen once going forward and once going backward, and we delete vertices for half of the edges seen going backward. Hence, we remove exactly $1/4$ of the vertices. This is optimal as can be seen in Eq. (3).

Random cubic graphs with $n$ vertices are known to be Hamiltonian [10] with probability tending to one as $n \to \infty$. Using a variant of ideas from the proof of [11] we could get an algorithm finding a Hamilton cycle in a random cubic graph with asymptotic probability one in time $O(n^{7/2})$. The rest of the algorithm runs in linear time. These results are supported by simulations.

### References

Controlling synchronization in bow-tie architectures

Edgar A. P. Wright¹, Sooyeon Yoon¹ José F. F. Mendes¹, and Alexander V. Goltsev¹,²

¹ Departamento de Física & I3N, Universidade de Aveiro, Aveiro, Portugal, wrighteap@ua.pt
² A. F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia

1 Introduction
Sparse directed networks can be partitioned into topologically distinct components connected together in a bow-tie: the CORE, the largest set of nodes reachable from each other along any sequence of directed links, and the IN and OUT components, all nodes in sequences of directed links leading into and out of the CORE respectively. The bow-tie architecture is a feature of many real directed networks, such as gene regulatory [1], metabolic [2] and neuronal networks [3, 4], and synchronization has been proposed as a key mechanism for communication among the interacting elements of these systems [5, 6]. Bow-tie architectures have an important structural feature: feedforward connectivity from IN nodes without incoming links – sources – to OUT nodes without outgoing links – sinks. Source nodes behave independently from and affect the dynamics of all downstream nodes, presenting a topologically-defined subset of nodes to target for control over synchronization dynamics on bow-tie architectures. We explore the effect of source node dynamics on synchronization in bow-tie architectures, revealing how they can disrupt or enhance synchronization in the CORE.

2 Toy-model
We consider a toy-model based on a directed Erdos-Renyi network with \( N = 10^5 \), where each node has an incoming link with probability \( \frac{2\langle q_{in} \rangle}{(N-1)} \), and the mean in-degree \( \langle q_{in} \rangle \) is the topological control parameter: \( \langle q_{in} \rangle \) determines the relative number of nodes in each bow-tie component [7]. Synchronization is modelled with Kuramoto dynamics:

\[
\dot{\theta}_n = \omega_n + K \sum_{m=1}^{N} A_{mn} \sin(\theta_m - \theta_n),
\]

where \( \omega_n \) is the natural frequency, \( K \) is the coupling strength, and \( A_{mn} \) is an element of the network’s assymetric adjacency matrix. Synchronization among a subset \( X \) of oscillators is given by the order parameter

\[
r_X = \frac{1}{N_X} \left| \sum_{n \in X} e^{i \left( \theta_n - \psi \right)} \right|,
\]

where \( \psi = \Omega t + \psi (t = 0) \) is the phase corresponding to the group velocity \( \Omega \). The pair-correlation function

\[
C_X = N_X \left( \langle r_X^2 \rangle_t - \langle r_X e^{i\psi} \rangle_t \langle r_X e^{-i\psi} \rangle_t \right),
\]
 introduced in [8], was used to study the emergence of synchronization, where \( \langle r_X \rangle_t = \frac{1}{T} \int_0^T dt \, r_X \) for some observation time \( T \gg 1 \).

An ensemble of oscillator networks with normally distributed natural frequencies and uniformly distributed initial phases was generated, and quantities of interest – e.g. those in equations (2) and (3) – were averaged both over the ensemble and the steady state.

### 3 Results

Synchronization emerges above the network’s percolation threshold. This is in contrast to synchronization on undirected networks, where both transitions occur at the same mean degree, and consistent with similar findings in the Ising model on directed ER networks [9].

Collectively, source nodes with random initial phases and natural frequencies suppress the emergence of synchronization among CORE nodes, as shown through the random removal of a fraction of IN nodes \( f_{IN} \) (cf. 1).

![Figure 1](image-url)  
*Fig. 1.* Order parameter \( r_{CORE} \) and pair-correlation function \( C_{CORE} \) as a function of the fraction of randomly removed IN nodes \( f_{IN} \). Subplots (a), (b) and (c) correspond to networks with \( q_{IN} \) 1.1, 1.4 and 1.7 respectively.

Equivalently, feedback links established between non-repeated pairs of randomly selected sink and source nodes were shown to enhance or suppress synchronization depending on whether the coupling strength is positive or negative. Establishing the maximum number of positive feedback links is as effective in enhancing synchronization as randomly destroying all paths from IN to CORE nodes.

The collective effect of the out of phase oscillation of source nodes at the same natural frequency is also explored, and shown to enhance synchronization in the CORE and determine the group velocity. As shown in 2, the group velocity can (i) vary non-monotonically with \( q_{IN} \) and (ii) differ significantly from the mean natural frequency of CORE nodes.

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 2. (a) Order parameter $r_{\text{CORE}}$ and (b) average instantaneous velocity (group velocity) in the CORE $\langle \dot{\theta} \rangle_{\text{CORE}}$ as a function of $\langle q_{in} \rangle$. Different markers/guidelines represent the natural frequencies of the source nodes $\omega_s$ – relative to the average of the natural frequency distribution $\langle \omega \rangle$: $\omega_s - \langle \omega \rangle$.

References


European hinterland container transport as a multiplex network - How is robustness affected by the multi-mode and multi-operator structure?

Camill Harter, Rob Zuidwijk, and Otto Koppius
Erasmus University Rotterdam, 3062PA Rotterdam, The Netherlands, harter@rsm.nl, rzuidwijk@rsm.nl, okoppius@rsm.nl

1 Motivation

Multiplex networks find applications in a growing number of complex biological and man-made systems. Most of these applications have shown that robustness of the multiplex representations differs substantially from their uniplex counterparts due to inter-layer dependencies and restrictions [1]. For instance, robustness induced by multiplexity in transport networks is often linked to restrictions regarding the usage of multiple layers along a chosen path. Prior work finds for instance that the robustness of the air transport network against failure of connections deteriorates if airlines are considered separate layers and only one layer can be used in a trip [3].

Network layers defined by transport operators are only one out of many types of layers that can form a multiplex transport network. Among those, networks formed by multiple transport modes deserve attention since fostering multi-modal transport is a popular approach to relieve transport networks from increasing congestion [2,7]. Transport modes can be considered separate independent layers that provide backup options if transport with the planned mode becomes infeasible. Beyond that, so-called inter-modal transport enables mode changes, i.e. routes comprising consecutive alternative transport modes.

Even though it can be assumed that complementary modes enhance robustness due to the sole fact that there is more capacity in the network, the actual effects of the different multi-modality setups are yet unclear. In this project, we show how robustness of the different shapes of complex multi-modal networks varies and how it compares to unimodal networks at the example of the European hinterland container transport network (EHCTN). Hinterland networks perform the landside transport of overseas container haulage. The EHCTN is distinctive as being both a multi-mode and a multi-operator network. Containers in the hinterland can be transported via truck, barge, rail, or combinations of these. Robustness of this system against disruption is critical given the crucial role of containerized transport on global supply chain performance [6]. Increasing container volumes lead to congestion and emission problems in the hinterland, which calls for a robust and flexible intermodal system [2]. However, intermodal transport is still far from unleashing its full potential as switches between modes and operators still come with considerable technical, process- and trust-related barriers that cause time and monetary effort, thereby making it less attractive. A goal of this project is to quantify the potential of intermodal transport to increase robustness if these barriers are reduced.
Table 1: Key figures of EHCTN (as of 2018)

<table>
<thead>
<tr>
<th>Rail connections</th>
<th>Barge connections</th>
<th>Number of countries</th>
<th>Number of cities</th>
<th>Number of terminals</th>
<th>Number of operators</th>
<th>Unique directional links between cities</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,111</td>
<td>808</td>
<td>25</td>
<td>538</td>
<td>484</td>
<td>1,101</td>
<td>1,079</td>
</tr>
<tr>
<td>Average departures per week (rail)</td>
<td>Average departures per week (barge)</td>
<td>Average travel time in days (rail)</td>
<td>Average travel time in days (barge)</td>
<td>Average distance in km (rail)</td>
<td>Average distance in km (barge)</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>2.8</td>
<td>1.9</td>
<td>2.5</td>
<td>651</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

2 Dataset: European hinterland container transport network

For our analysis we make use of a unique dataset containing all rail and barge services scheduled in the European hinterland. Truck transport is not included in the dataset as trucks can move between any two points and can be booked flexibly. We create a directed network with nodes representing cities that have at least one container terminal and edges representing transport connections. Edges have two labels indicating transport mode and operator as well as two weights indicating travel time and weekly frequency. Mode and operator layers consist of all edges with the respective mode or operator label and their adjacent nodes. There can be parallel edges for services on the same connections provided by different operators or a different mode.

An initial analysis of the dataset revealed that rail and barge layers are complementary, i.e. the number of connections served by both rail and barge is limited [4]. Moreover, rail services mostly operate on long-haul connections whereas barge services tend to serve shorter distance links. Figure 1 shows an illustration of the structure of the different mode layers and table 1 shows key figures of the dataset.

Fig. 1: Visualization of full network (a), rail (b) and barge layers (c). Edge thickness by number of services and node thickness by degree
3 Methodology and simulation setup

We measure robustness of the EHCTN by the ability to efficiently re-schedule container deliveries after failure of transport services. The methodology derives from [3], but is adjusted to the hinterland setting. A set of OD pairs is created representing the delivery assignments for containers. Based on shortest paths, the capacity for each edge is determined as the initial load plus a proportional tolerance. A re-scheduling algorithm comes into place when a proportion of the edges is removed from the network searching for alternative feasible paths with a maximum tolerated extra length of $n$. Robustness is determined based on the share of containers that need to be re-scheduled or cannot be delivered. Unlike the comparison in [3], we do not exclude the possibility to create routes that go across layers in the multiplex case. Instead we increase the travel time on an edge if a layer is being crossed, i.e. we assume that switching modes or operators comes with a time effort representing the additional challenges coming with such a switch. This has implications on the shortest path algorithm. A simple adaptation of the Dijkstra algorithm inspired by the time-dependent Dijkstra algorithm by [5] allows for edge costs that vary depending on the previously taken edge. Thus, instances can be created by varying the costs for mode and/or operator switches. The instances in [3] can be reproduced by setting the switching costs to 0 in the uniplex case and to $\infty$ in the multiplex case.

4 Results

Preliminary simulations suggest that the isolated mode and operator layers are less robust than any instance of the full network. Moreover, robustness can be enhanced if barriers for mode and operator switches are reduced.

References

Precursor Theory of Adaptive Networks

Leonhard Horstmeyer\textsuperscript{1,2}, Christian Kühn\textsuperscript{2,3} Tuan Minh Pham\textsuperscript{1,2}, Jan Korbel\textsuperscript{1,2}, and Stefan Thurner\textsuperscript{1,2,4}

\textsuperscript{1} Section for the Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090, Vienna, Austria
leonhard.horstmeyer@meduniwien.ac.at.
\textsuperscript{2} Complexity Science Hub Vienna, Josefstädterstrasse 39, A-1090 Vienna, Austria
\textsuperscript{3} Faculty of Mathematics, TU Munich, Boltzmannstr. 3, 85748 Garching München, Germany
\textsuperscript{4} Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

1 Introduction

Adaptive co-evolving dynamic networks play a key role in ecological, epidemiological, social or financial systems. In adaptive network models the node states and the network topology are dynamically coupled, i.e., they co-evolve. These models have in common that they can collapse, e.g., in the form of mass extinctions, disease outbreaks, financial crises or drastic decline of network connectivity. Our aim is to understand these collapse transitions and their precursors in adaptive network models. Structural information about the network is of crucial importance for this task. These models differ strongly with respect to the level of aggregation at which this information is required. On one side there are systems whose collapse depends solely on a very particular network configuration and on the other side there are systems whose collapse depends on aggregated structural variables, such as the overall density of certain network motifs.

In this contribution we contrast two representative models that differ in this respect: The Jain-Krishna model and the adaptive SIS model. In the first model the collapse occurs when cycles are broken. In fact we could show that the collapse is always preceded by a state in which the network just contains a single cycle. This is a very particular network configuration and one might expect that detailed information about links is required to detect this state. We could show, however, that this state can surprisingly be detected purely macroscopically by a quantisation phenomenon, in which the node states can only take values from a set of equidistant levels. In the second model one might expect that aggregated information is sufficient, because the model is tractable by a mean field description. We could, however, show that the prediction of the collapse just via aggregated motif densities may systematically fail.

2 Results

The Jain-Krishna model was introduced in [1] as a co-evolving network model of species evolution. The basic ingredients of this model are species that form catalytic ties with other species. The species abundances are the node states and the catalytic relations give rise to a directed network. The abundances are governed by a linear reaction dynamics. On a much slower time scale species are chosen for extinction when
their abundances are low in relative terms. New species are then introduced that form new catalytic ties. It turns out that the presence of catalytic cycles is crucial for the survival of the entire ecosystem as such. Feeding off these cycles is profitable for new species, but may weaken that very cycle in the long run. The ecosystem collapses when a critical species is removed from the network, namely one that is a member of the last catalytic cycle. We show that this vulnerable state of the network, i.e. the presence of a single cycle, leads to a quantisation of the relative species abundances. Figure 1 depicts a pre-critical network and a critical network with a single cycle. The histograms clearly show the non-quantised and quantised population levels, respectively. Thus, the bare knowledge of the species populations is sufficient to detect this topological state of the network. Without interference the network collapses and we could further show that it is predictable with good accuracy.

In the adaptive SIS model [2] the nodes can be either infected (I) or susceptible state (S). The contacts amongst the nodes constitute an undirected network. Infected nodes recover at a rate \( r \) and infect a neighbouring susceptible node at a rate \( \lambda \). Susceptible nodes can cut their contact to an infected neighbour and establish a new contact with a susceptible node at a rate \( w \). If the infection rate becomes lower than a critical rate \( \lambda_c \), the disease cannot sustain itself anymore and all SI-links vanish. We studied the critical behaviour of the network as the system is gradually moved towards this critical threshold and into the disease-free state. We found in [3] that the critical curves of network quantities that contain topological information lead to mis-predictions of the true location of the transition. In Figure 2 we depict two examples. The effective branching ratio, which is a proxy for the risk of a secondary infection for any potential first infection does not feel the presence of the threshold at all. Fitting a power law would yield \( \lambda_c = 0 \)
as a prediction. Likewise for the standard deviation of the degree distribution. Thus the classical methods from early-warning-sign theory, which revolve around estimating the location of the threshold via critical curves systematically fail.

![Fig. 2](image)

**Fig. 2.** (a) Effective Branching ratio $\langleSSI\rangle/\langleSI\rangle$ and (b) standard deviation of the degree distribution. Neither quantity senses the critical threshold. The black dashed line indicates the critical transition and the dotted line the asymptotic value. $N = 400, w = 0.01, r = 0.002, \langle k \rangle = 20$

**Summary.** We have investigated and contrasted network collapses of two adaptive network models: the Jain-Krishna model and the adaptive SIS model. We have shown that the collapse can be predicted accurately in the first model purely from information of the nodes, even though it relies crucially on a particular network topology. In the second model we have shown that a prediction may be systematically off, even in the presence of aggregated knowledge of the network topology. These results have three ramification. First, classical scaling methods might lead to wrong predictions in adaptive network collapses. Secondly, when the rupture of cycles are the decisive factor of collapse, then the quantisation phenomenon may be a viable signal. Lastly, these results may be the first step towards a classification of adaptive network collapses, via their predictability in the presence or absence of network topological data. In a next step we would like to investigate real world systems where the rupture of cycles can be observed.

**References**

Part XVI

Social Networks
Social structure in towns and corruption risk at city hall

Johanes Wachs\textsuperscript{1}, Taha Yasseri\textsuperscript{2} Balázs Lengyel\textsuperscript{3}, and János Kertész\textsuperscript{1}

\textsuperscript{1} Department of Network and Data Science, Central European University, Budapest, Hungary, H-1051
\texttt{johanneswachs@gmail.com, KerteszJ@ceu.edu}
\textsuperscript{2} Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, UK
\textsuperscript{3} Agglomeration and Social Networks Lendület Research Group, Hungarian Academy of Sciences, Budapest, Hungary, H-1097

1 Introduction

Corruption is a plague to society: gains accrue to small groups, while the multiple costs are borne by everyone. Significant variation in its level between and even within countries suggests a relationship between social structure and the prevalence of corruption, yet, large scale empirical studies thereof have been missing due to lack of data. We relate the structure of the social capital of cities and the exposure to corruption in their local governments. Using datasets from Hungary, we quantify corruption risk by suppressed competition and transparency in the town’s awarded public contracts and we characterize social capital using social network data from a popular online platform. We find that settlements with fragmented social networks, indicating an excess of \textit{bonding social capital} have higher corruption risk and towns with more diverse external connectivity, suggesting a surplus of \textit{bridging social capital} are less exposed to corruption. We interpret fragmentation as fostering in-group favoritism and conformity, which increase corruption, while diversity facilitates impartiality in public life and stifles corruption.

2 Data and network measures

Public procurement contracts constitute a major channel of public funds to private hands and are highly vulnerable to corruption [1]. A set of corruption risk indicators have been derived based on public contract data, e.g., if the contract attracted a single bidder. We examine 20,524 municipal government contracts from the period of 2006-2014 issued by Hungarian settlements awarding at least five contracts a year on average. 169 settlements in Hungary meet this criterion, excluding Budapest\textsuperscript{4}.

We quantify the structure of a city’s social capital using complete data from “iWiW”, a now defunct online social network once used by approximately 40\% of the adult Hungarian population [2]. Data from iWiW includes, besides friendship ties, information on each user’s location, their gender, their time-stamped connections to other users, and

\textsuperscript{4} Budapest, the capital of Hungary, is an order of magnitude larger (\approx 2 million inhabitants) than the second largest city. Separate regulations are valid for Budapest, e.g., there is a central city government and 23 district governments making comparison with other settlements difficult.
education. We use data from iWiW to quantify bonding and bridging social capital in Hungarian settlements.

We measure fragmentation as an indicator of excess bounding social capital of the settlement’s internal social network using network community detection methods and modularity [3, 4]. If $Q(S)$ is the modularity of a settlement’s social network considering only within-settlement links and $Q_{\text{max}}(S)$ is the maximum possible modularity for the same partition and the same number of links then fragmentation $F_S$ of a settlement $S$ is defined as the quotient

$$F_S = \frac{Q(S)}{Q_{\text{max}}(S)}.$$

A highly fragmented settlement consists of tightly-knit groups that are weakly connected thus it indicates excess bonding social capital (see Fig. 1).

Fig. 1. Sampled social networks and adjacency matrices of high (A) and low (B) fragmentation settlements. Node colors indicate membership in communities. In the adjacency matrices, percentages indicate the share edges staying within each community. In the fragmented settlement, communities have a significantly fewer connections with other communities.

We introduce the notion of diversity as an indicator of surplus in bridging social capital for settlements as follows. We define the subgraph $\{\text{alters}_i\}$ of alters of node $i$ with all links between them. We define the settlement diversity as

$$D_S = \frac{1}{|S|} \sum_{i \in S} Q(\{\text{alters}_i\}),$$

where $|S|$ is the number of nodes in the settlement $S$. This measure captures the typical diversity of social perspectives that the members of the town access. At the settlement level this measure captures bridging social capital.
Fig. 2. Plots of marginal effects of the key social capital variables and their predicted impact on a settlement’s rate of closed procedure or single bidder contract awards; shaded regions represent 90% confidence intervals are indicated. As the variables are standardized, unit changes on either axis can be interpreted as standard deviation changes.

3 Results

Our aim is to relate bonding and bridging social capital in settlements to corruption risk in their public contracts. We predict contract corruption risk using multiple regressions:

\[ C_S = \beta_1 F_S + \beta_2 D_S + \theta X_S + \epsilon_S, \]

where \( C_S \) is one of two corruption risk indicators (closed or single bidder rate and average corruption risk index), averaged at the settlement level, \( X_S \) is a matrix of control variables, and \( \epsilon_S \) is an error term. The control variables include average income, education level, unemployment rate, the presence of a university in the town, total population, rate of population above 60, victory rate in the last three elections of the major, and distance from the capital.

We find a significant relationship between social network structure and the dependent variables measuring corruption. More fragmentation consistently predicts more corruption, while more ego diversity consistently predicts less corruption (see Fig. 2). In both cases adding the network features significantly improves the adjusted \( R^2 \) of the model. Moreover, comparing the coefficients, we see that the social network features have effect sizes similar to that of any social, political, or economic control. All models pass a variance inflation factor (VIF) test for feature collinearity.

References

The Contagion Effects of Repeated Activation in Social Networks

Pablo Piedrahita, Javier Borge-Holthoefer, Yamir Moreno, and Sandra González-Bailón

Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Catalonia

1 Introduction

Demonstrations, protests, riots, and shifts in public opinion respond to the coordinating potential of communication networks. Digital technologies have turned interpersonal networks into massive, pervasive structures that constantly pulsate with information. In this context, threshold models have become the standard for how we think about interdependence and the collective effects of social influence [1, 2]. In the latest developments, the activation of individual thresholds responds to their group of reference, as determined by their connectivity in the network, thus changing from actor to actor [3, 4]. The threshold model shows two important elements: first, activation is modelled as a step function that goes from 0 to 1 when thresholds are reached; and second, thresholds can only be reached once, that is, activation is assumed to be a one-off event.

Here, we propose a model that aims to analyze the contagion dynamics that emerge in networks when repeated activation is allowed, that is, when actors can engage recurrently in a collective effort. We have theoretical and empirical reasons to allow repeated activation to be the driving force of contagion dynamics. The empirical reason is that most instances of diffusion do not involve a single activation but many activations building up momentum in time. Before a hashtag becomes a trending topic, a period of buzz is first required; prior to a protest day, calls announcing the mobilization are distributed in waves. Our model aims to capture this temporal dimension. We also have a theoretical reason to relax the assumption of single activation. The intermittent dripping of information that social networks facilitate often leads to bursts of activity [5], as when news suddenly become trending topics [6, 7]. Coordination dynamics underlie these bursts of activity: sudden peaks in communication require the adjustment of individual actions, that is, the alignment of many individual decisions so that everybody uses the same trending hashtag or talks about the same news at the same time. We analyze how the structure of communication networks impacts on the ability to coordinate actors, and we identify the conditions under which large-scale coordination is more likely to emerge.

2 Model

Our model of contagion relaxes the assumption that actors can only transition once from an inactive to an active state. We also allow the effects of each activation to vary over time to the extent that they coevolve with the contagion dynamics taking place in the
rest of the network. These modelling choices make sense if we think about how online networks facilitate contagion dynamics: users are constantly exposed to signals that might shift their inclination to act. To bring these empirical intuitions into a tractable framework, we follow classic integrate-and-fire models of synchronized coordination [8–10], see Figure 1A. These models have been used extensively to study coordination dynamics in biological and physical settings.

3 Results

Besides the combination of values for the two parameters ($\omega$ and $\varepsilon$, see Figure 1A) of the integrate-and-fire model, the other crucial element that determines how fast coordination will emerge—that is, how long it takes for all nodes to start pulsating, or activating, concurrently—is topology on which interactions take place. To determine the impact that different network structures have on those dynamics, we run experiments on four network topologies: Erdős-Rényi, regular, small world, and scale-free networks (Figure 1B) [11]. Panels C and D of Figure 1, and their corresponding caption, summarize the results of such numerical exploration of parameters and topologies.

References

The model, adapted from [8], assumes that actors reach their activation threshold at different speeds. The speed of activation is a function of two parameters: \(\omega\), which determines how quickly the actor reaches the threshold zone (i.e. it defines the concavity of the curve that maps progression towards activation); and \(\varepsilon\), or the strength of the signal received from the neighbors in the network when they activate, a pulse that shifts the state of the focal actor closer towards the threshold (the timing of which varies over time). The lower panels in the figure illustrate how actor \(i\) advances towards activation. When a node activates, as node \(i\) does in \(t_2\), she shifts the state of her neighbors with the \(\varepsilon\) signal and resets her state back to the beginning of her phase.

**B: Network topologies used in the simulation experiments.** We use four topologies to run the explained dynamics. All networks were generated using the configuration model [12], with the exception of the small world network, for which we used the Watts-Strogatz model [13].

**C and D panels.** In both cases, results are shown for 100 realizations of the simulation. The color scheme indicates how long it takes, for each combination, to reach large-scale coordination (which here we define as at least 75% of the nodes activating simultaneously). Lighter colors indicate earlier coordination, darker colors indicate later coordination; black signals that no large-scale coordination was possible.

**C: Impact of social influence (parametrized as \(\varepsilon\)) on large-scale coordination across \(\omega\) values.** Subpanels 1–4 summarize coordination dynamics for different values of \(\omega\) (the intrinsic motivation parameter) and \(\varepsilon\) (social influence strength) across the four network topologies. Every dot in the plots corresponds to a combination of parameters \(\varepsilon\) and \(\omega\); the distribution of \(\omega\) and \(\varepsilon\) is homogenous across nodes and edges, respectively. The findings suggest that random, homogenous networks are more conducive to large-scale coordination. Heterogeneous networks characterized by the presence of hubs (i.e. scale free networks) do not allow large-scale coordination soon after the social influence signal weakens. Small world networks are also more restrictive in the emergence of coordination than regular networks, in spite of the global shortcuts created by random rewiring (or because of them).

**D: Impact of actor heterogeneity on large-scale coordination across \(\varepsilon\) values.** Subpanels 1–4 summarize coordination dynamics for different distributions of \(\omega\) and \(\varepsilon\) values (social influence strength). The distribution of \(\omega\) depends on the standard deviation (vertical axis); \(\varepsilon\) is homogenous across edges. The results show that, once more, all networks are less efficient in allowing large-scale coordination than the random benchmark provided by the Erdős-Rényi topology. Overall, low to mild heterogeneity increases the probability of global coordination, whereas high heterogeneity hinders it.
Homophily explains perception biases in social networks

Fariba Karimi*, Eun Lee*, Claudia Wagner, Hang-Hyun Jo, Markus Strohmaier, Mirta Galesic1

*These authors contributed equally to this work.

Abstract

Individuals’ perceptions about the prevalence of attributes in their social networks are commonly biased by the limited information available to them: False consensus indicates the tendency to consider one’s own opinion as more common than it really is, and false uniqueness is the tendency to think that one’s own view is less frequent than it really is. In this paper, we show how these two perception biases can be explained in a general framework by means of homophily in social networks. Using a generative network model with adjustable homophily—symmetric and asymmetric homophily between groups—and group size, we demonstrate analytically and numerically under which conditions and to what extent the two perception biases emerge. Our model is validated by empirical evidences in cross-cultural survey study. Moreover, we investigate the potential for perception biases in six real-world networks with various levels of homophily and minority group’s size. Our results show that (i) homophily explains perception biases in social networks, (ii) these biases are related to the asymmetric nature of homophily in networks, and (iii) extremely disproportionate group size can enlarge the difference of possible perception bias between majority and minority groups. Finally, we explore a possibility of reducing individual perception biases by collecting more information about the perception of their neighbors. These results advance our understanding of the impact of network structure on perception biases and offer a quantitative approach to address related issues in society.

Introduction

In this work we study how two important structural properties of social networks, i.e., homophily and disproportionate group size, can facilitate perception biases on the prevalence of attributes. Homophily, the tendency to be connected to similar others[4], and disproportionate group sizes, e.g., between minority and majority, can be observed in a variety of situations, such as gender imbalance in STEM fields, income inequality, and imbalanced population of smokers.

We show empirically, analytically, and numerically how a simple model of social networks can explain two well-documented and apparently contradictory perception biases, i.e., false consensus and false uniqueness. First, we conduct a cross-cultural survey asking about the frequency of different attributes in people’s social circles and the population. The survey results confirm an importance of the homophily and disproportionate group size on people’s perception biases. Next, in order to gain analytical insights on the extent to which homophily and disproportionate group size impact individual samples and consequently people’s perceptions, we propose a generative network model with tunable homophily and disproportionate group size. We show that the model predictions align well with observed false-consensus and false-uniqueness effects drawn from the survey. Finally, in six empirical networks, we numerically demonstrate that our predictions are consistent with the observed sample biases, and show how the perception biases can be reduced by aggregating individual perceptions with neighborhood information.

Results

We define the perception bias by assuming that each node relies on the prevalence of an attribute in their direct neighbourhood to perceive the overall prevalence of the attribute. We focus on binary attributes where one attribute value is more uncommon than the other in a population (for example, smoking or having a theft experience). Then we measure to what extent the fraction of minority nodes in the local neighbourhood of a node $i$ deviates from the true fraction of minority nodes, $f_a$. For node $i$, we define perception bias, denoted by $P_{\text{indv.}}$, as follows:

$$P_{\text{indv.}} = \frac{1}{f_a} \sum_{j \in \Lambda_i} \frac{x_j}{k_i},$$

(1)

where $k_i$ is the degree of node $i$, $\Lambda_i$ is a set of neighbors of node $i$, $x_j$ is the attribute of node $i$’s neighbor $j$ (1 for minority and 0 for majority) and $f_a$ is the fraction of minority nodes in the entire network.

To systematically study the relation between perception bias, homophily, and group size, we develop a network model that allows us to create scale-free networks with tunable homophily and group sizes[2]. This network model is a variation of the Barabasi-Albert preferential attachment model with the addition of homophily parameter (we call this model BA-homophily). BA-homophily model has a simple addition of homophily parameter $h$ on the preferential attachment model in which a probability of an attachment by a newly arriving
Fig. 1. Measuring the perception bias. Nodes belong to one of the two groups (blue and orange); Blue nodes are the majority and orange nodes are the minority. The minority fraction is 1/3 in both networks \( f_a \approx 0.33 \). Subfigure (a) depicts a homophilic network, while subfigure (b) shows a heterophilic network. Perception bias is measured on an individual-level (grey, dashed circles) and a group-level (grey, shaded area). On an individual-level, in the homophilic network, the central node \( i \) perceives the size of the minority as approximately 1/6 \( \approx 16\% \), while in the heterophilic network it perceives the size of the minority approximately 4/6 \( \approx 67\% \). Therefore, in homophilic networks, node \( i \) underestimates the minority size by factor of 0.5 and in heterophilic network it overestimates the minority size by the factor 2.

The generative network model offers a very simple representation of real-world networks. To examine the possibility of perception biases in real networks and validate the plausibility of the simple model, we study six empirical networks with various ranges of homophily and minority sizes. The network characteristics of the datasets are presented in the paper. These empirical networks have different structural characteristics and show different levels of homophily or heterophily with respect to one specific attribute.

To examine perception biases in these empirical networks, we need to estimate homophily. In the first approximation, we assume that homophily is symmetric in all networks. The symmetric homophily is equivalent to Newman’s assortativity measure that measures the Pearson coefficient of the edges that exist between nodes with the same attribute values compared to what we would expect from their degree. The Newman assortativity measure corresponds directly to the homophily parameter in our model when adjusted for the scale. In reality, however, the tendency of groups to connect to other groups can be asymmetrical. For example, it has been observed that in scientific collaboration homophily among women is stronger than homophily among men [1]. Given the relation between number of edges that run between nodes of the same group and homophily, we can estimate the asymmetric homophily such as homophily between minorities \( (h_{aa}) \) and between majorities \( (h_{bb}) \) [2]. As the figure 2 shows, it turns out that asymmetric homophily has an important impact on the predictability of perception bias in empirical networks.

References
Fig. 2. Group-level perception biases in six empirical social networks. The figure shows how accurately do the (a) minority group and (b) majority group estimate the size of the minority in social networks with different levels of homophily. Higher deviation from 1 indicates larger perception biases. Brazil is a sexual contact network, POK is a Swedish online dating network, USF51 is a Facebook friendship network, DBLP is a scientific collaboration network, Github is a coding collaboration network and APS is a citation network. All these networks exhibit binary attributes (male/female for Brazil, POK, USF51, DBLP and Github) and (two branches of physics QSM/CSM for APS). We use the measured homophily and minority size in the empirical networks to generate synthetic homophilic networks with similar characteristics [2]. This will enable us to compare the perception bias in the empirical networks with perception bias in synthetic networks and evaluate the goodness of the model. The perception bias increases as heterophily or homophily increases in networks. The results of the generative network model with asymmetric homophily are in excellent agreement with the empirical results highlighting the importance of considering the asymmetric homophily. The inset shows the perception gap (differences between the perception of minority and majority) for each network. For more details see [3].
Recruiting Mechanisms on Online Labour Markets: 
Agent-Based Models of Bipartite Social Networks

Martin Lukac
Centre for Sociological Research, KU Leuven, 3000 Leuven, Belgium, 
martin.lukac@kuleuven.be, 
ORCID: orcid.org/0000-0002-1747-3168

1 Introduction

Online labour markets (OLMs) are platforms that match buyers of electronically deliverable services with workers. Freelancing has been growing steadily over the last decade and was repeatedly called “an unprecedented development” [9] in globalising labour. OLMs have the power to connect geographically dispersed workers and employers as contracting for work becomes possible online. OLMs are already spurring large shifts towards part-time employment and have immediate implications for global inequality. Considering the enormous scale of OLMs, hiring decisions can have significant social and economic consequences and influence wellbeing of millions of workers worldwide.

An important aspect that distinguishes OLMs from traditional labour markets is assessment of job applicants. Employers and employees online face information asymmetry problem. Employees can easily communicate low bandwidth information (i.e. education, experience, credentials, etc.) via their online profiles, however high bandwidth information (such as quality, trustworthiness, or motivation) are difficult to demonstrate [1]. Trying to alleviate this problem, platforms implement various reputation systems that aim at aggregating information about freelancers’ talents and previous conduct on the platform [7][5][18]. Especially newly joining freelancers are often forced to accept jobs that do not reflect their talents just to obtain initial reputation to be considered for more complex tasks [6][13]. As such, reputation systems are unlikely to provide unbiased information about freelancer’s capabilities.

Previous work is available on reputational signals [14][18][11], geographical differences [12], bidding strategies [16], and gender biases [2]. No work, however, has investigated the effect of reputation and recruiting systems of OLMs on their social and economic outcomes.

With particular focus on reputation, in this study I investigate whether recruiting mechanisms at OLMs can directly influence platform’s social and economic outcomes. In other words, the aim of this study is to examine the link between micro-mechanisms of recruiting and macro-outcomes by conducting a series of simulation experiments (see [3][17][10]). Earnings and reputation are mutually reinforcing factors that are directly related to future employability of workers [13][6]. Recruiting online is, however, done with limited information about worker’s skills and employers often choose only from restricted pools of candidates that apply for a specific job opening.
2 Methodology

The paper makes use of agent-based modelling [15][8], projecting online labour markets as bipartite networks with employees and employers as agents with different sets of possible behaviours (see Fig. 2). OLMs constitute relatively well tractable systems where interactions between employers and employees are well structured by the possible actions that either can take. Each employee is assumed to have a set of characteristics, such as their true talent score (unobserved by the employer), number of previously finished jobs, and earnings obtained on the platform. All of these are aggregated together into a reputation score based on the previous research [6][18].

3 Preliminary Results

Preliminary results show that even despite various efforts to minimise the information asymmetry problem with reputation systems, availability of limited information during the recruiting process leads to winner-take-all outcomes. Restricted pools of applicants often result in favouring freelancers with best communication of the high bandwidth information—their reputation, which is directly dependent on their previous success on the platform (see Fig. 1). This dynamic hinders entry of new workers to the platform without first undertaking a number of jobs that might be unrelated to their skills or severely underpaid just to build reputation to be considered for tasks that fit their skill-sets. Workers with high reputation land disproportionately higher share of jobs even though they might not be the best pick from the local pool.

4 Conclusion

I develop an agent-based model of recruiting behaviour to understand the effect of reputation and recruitment on social and economic outcomes of OLMs. OLMs were initially considered ‘‘‘borderless’ virtual business platforms’’ [4] that were to globally equalise access to labour. However, recruiting of freelancers online gave rise to new barriers for
equal outcomes. Discrimination based on artificially created reputation scores starts a
discussion about fundamental inequalities originating from algorithmically calculated
scores. From a methodological perspective, it is the first paper to theorise OLMs as bi-
partite networks, which opens doors to future application of network models on OLM
data. The results underline the usefulness of computational experimentation to probe
potential outcomes of various socio-technical systems to ensure the most equitable so-
cial outcomes.

References

Farrar, Straus & Giroux.
*Procedia Computer Science*, 130, 172-179.
International Workshop, Proceedings*. Berlin: Springer. Available at http://john-joseph-
horton.com/papers/online_labor_markets.pdf
are Social Networks? *Am. Econ. Rev.*, 97(3), 890-915.
ing the playing field for international service markets? *IPP2014: Crowdsourcing for Politics
motivate sellers to seek third-party certifications in an online labor market. Working paper,
University of Arizona, Tucson.
Heng Chen, Mak Kaboudan, & Ye-Rong Du (Eds.) *The Oxford Handbook of Computational
Sci.*, 49(1), 1504-1520.
393(6684), 440-442.
*Mark. Sci.*, 32(6), 860-891.
On the clustered structure of a decentralized social network and its dynamics

Matteo Zignani, Christian Quadri, Sabrina Gaito, and Gian Paolo Rossi

Università degli Studi di Milano, Department of Computer Science

While in the last decade we witnessed the growing of online social networks like Facebook, Instagram, or Twitter which are based on a centralized infrastructure and are making their data less and less accessible; various forms of micro-blogging platforms are sprining up, moving towards a decentralized architecture and becoming new sinks of data for social computing. Furthermore, these types of social platform put the social relationships and the users’ contents at the heart of their actions, not providing any advertisement and recommendation algorithms. Among these alternatives, Mastodon is the newest, the most growing and presents some interesting characteristics for the study of the clustered structure of online social networks and its relation with the group metadata [1]. Despite many common features, Mastodon is not centralized like Twitter, but is made by interconnected communities being located in different servers, named instances; and each instance is independently owned, operated, and moderated and supports specific interests, languages, and needs. So, the role of instances as metadata is twofold: they express a common interest and they may impose physical constraints, since each instance is hosted on a physical server and users must select which server to use.

Dataset: To support our analysis we gathered the social network made by the relationships between Mastodon’s users, and the information (interests, geographical position, supported languages) of the instances they belong to. To establish a relationship between two users, Mastodon offers the same Twitter’s mechanism: users may follow other users; so the resulting social network is directed. To gather the social structure we developed an ad-hoc crawler which bypasses the Mastodon API, since the latter does not provide the follower/following relationships. To build the initial seed, we gathered the users whose posts have been published on the timeline of the different instances. The resulting social network is made by more than 500 thousands users and more than 6 millions links [3]. Further, we collected the growth of the social network along a two-month period by gathering the new relationships people have established1.

In this work we focus on three main questions about the link creation and the clustered structure of Mastodon.

a) Are users bound by instances when they establish relationships? To evaluate the impact of the community-based architecture on the creation of links in the Mastodon network, we measured for each node the portion of neighbors who are hosted in different instances, then we computed the average ratio for each instance. We named it the extra-boundary index. An extra-boundary index close to 1 indicates that people in an instance are likely to establish relationships with users outside the instance. In this case

1The dataset is available at https://dataverse.mpi-sws.org/dataset.xhtml?persistentId=doi:10.5072/FK2/AMYZGS
the users are scarcely bound by the instance. In Fig. 1a we report the extra-boundary index as a function of the size of the instance. In the figure we observe that users belonging to small instances are more likely to search for friendships outside their instance, since most of the small instances are characterized by an extra-boundary index greater than 0.5, i.e. half of their neighbors lie outside their home-instance. This result on small instances is quite expected as well as the tendency of users in large instance to establish relationships with members of the same instance, since they have more opportunities to find people with common interests. So, the boundary effect is carried out only by large instances.

Through the data about the growth of the Mastodon network we are able to investigate if the previous behavior is stable during the evolution of the network or periods when people change how they search for new friends exist. To this aim, we computed the extra-boundary index for five snapshots of the network. In Fig. 1d we show the trend of the index in the instances with more than one thousand users. We observe that a common and well-defined trend does not exist, rather in a few instances the boundary effect is more evident, i.e. people keep establishing new relationships within the same instance; while in other large instances, e.g. mastodon.xyz or mastodon.cloud, people interchange opposite behaviors, i.e. sometimes they explore new instances, sometimes they create new links within their instance. A similar trait characterizes small communities but for the sake of readability of the figure we do not report it. Generally, the stability of the boundary effect is not a feature common to most of the instances, but is instance-dependent.

b) Does the decentralized architecture based on instances impact on the local clustering coefficient? We analyze the differences in the local clustering between the main Mastodon instances, as shown in Fig. 1b. In the figure we report the average local clustering coefficient versus the mutual degree in the instances: pawoo.net, mstdn.jp, mastodon.social and mastodon.xyz. Instances are very different from one another: i) mstdn.jp, the second largest instance, has a higher average clustering coefficient (0.35) than the other instances (pawoo.net - 0.26, mastodon.social - 0.13 and mastodon.xyz - 0.08), and it is also higher than the clustering coefficient of the entire network; ii) in the mstdn.jp subnetwork the clustering coefficient increases up to a peak (\( cc = 0.46 \)) at degree around 30, then slows down. That indicates the presence of clustered regions around nodes with a small-medium connectivity. The same behavior, at a different magnitude order, has been observed in the Twitter Japanese subgraph [2], where there are quasi-clique subgraphs centered around high degree nodes. The above results highlight that the clustered structures around nodes strongly depends on the instances, as also indicated by Fig. 1c. In the figure we report the cumulative distribution function of the variation of the clustering coefficient on the instance subgraph w.r.t. the whole network. The distribution is concentrated around the interval \((-0.1, 0.1)\) and reveals that the tendency of neighborhood’s nodes of being clustered is limited within the instance boundary.

What is the impact of instances on the clustered structure of the Mastodon network? The previous results on the clustering coefficient and the extra-boundary index may suggest that, despite people search for friendships out of their instance, it is unlikely that triangles among users in different instances close. So, we might hypothesize that clustered regions of the network unlikely overcome the boundary of the instances.
To investigate this question, we compare the clustered structures detected by a popular community detection algorithm (the Louvain algorithm) with the group metadata, i.e. the instances of the Mastodon network. Unlike other similar approaches, here we do not use the concept of ground truth when we refer to instances, since we want to avoid the idea of a true partition of the nodes. Moreover, the usage of the community detection algorithm is limited to the identification of clustered regions in the graph. The comparison has been conducted at the level of the whole partitions of the set of nodes belonging to both the instances and the detected communities. Among the methods to evaluate the similarity of partitions we adopt the normalized mutual information (NMI). We obtain a NMI equal to 0.26, thus the clustered regions returned by the Louvain method do not align with how nodes are grouped into instances. So, instances scarcely impact on the clustered structure of the network. The reasons of this result should be further investigated, however it may be due to the presence of many clustered groups within large instances or the union of small instances physically separated by infrastructural constraints but supporting a single interest-driven community.

References

Temporal evolution of mentions and retweets networks in the Spanish General Elections

S. Martin-Gutierrez, J. C. Losada, and R. M. Benito

Grupo de Sistemas Complejos, ETS Ingeniería Agronómica, Alimentaria y de Biosistemas,
Universidad Politécnica de Madrid, Avda. Complutense s/n 28040 Madrid, Spain,
http://www.gsc.upm.es

samuel.martin@upm.es https://orcid.org/0000-0002-5685-7834
juancarlos.losada@upm.es https://orcid.org/0000-0002-4373-603X
rosamaria.benito@upm.es https://orcid.org/0000-0003-3949-8232

1 Introduction

Twitter is one of the most used media by journalists, politicians and politics enthusiasts to share their views and news. Furthermore, politicians consider Twitter one of the more consequential social media for their job [5]. These facts have inspired an ever growing number of academic works that use Twitter data to analyze different aspects of political campaigns [2].

Whereas in other data-driven researches the main objective usually is to characterize different aspects of a specific electoral campaign [1], in this work we take advantage of the availability of Twitter data gathered during two electoral campaigns [4] that are very close in time (the Spanish general elections held on the 20th of December 2015 and the repetition of the elections on the 26th of July 2016) to compare the collective user behavior manifested in two analogous social systems with a similar context. Our objective is to study and characterize emergent behaviors that are recurrently manifested in political contexts.

To this end, we have computed and compared the time series of the number of tweets per day, or daily activity, for both elections. We have also studied the relationship between the number of tweets, retweets and mentions with respect to the number of unique users that take part in the conversation. In order to explore the evolution of the interactions among users, we have built networks of mentions and retweets and studied their temporal evolution. This has enabled us to verify that they show similar topological properties in both electoral periods. Besides, we have studied the mention and retweet subgraphs induced by political users and obtained results that imply a lack of communication among different parties. Furthermore, we have found that an intensification of the interaction can be detected between parties after the formation of a coalition.

2 Results

The results that we have obtained from the computation and analysis of the daily user activity time series for both elections indicate that they present a significant linear correlation (see figure 1). Additionally, by studying the distribution of user activity we have
found that in both elections its exponent fluctuates in the same tight interval. The value of the exponent obtained in a previous work [1] also lies within this interval. These facts suggest the existence of recurrent activity patterns in different political campaigns.

We have shown that the daily rate of tweets, retweets and mentions follow a power law with respect to the number of unique users that participated in the conversation each day. However, whereas in 2015 the growth for the three quantities was slightly super-linear with respect to the number of users, in 2016 we observe an approximately linear behavior. Hence, in 2015, when more users join the conversation, the activity experiences a proportionally higher increment than in 2016. To understand this behavior we have followed the work by Leskovec et al. [3], where it is shown that when real-world networks evolve through time, the number of links $E(t)$ scale with the number of nodes $N(t)$ as $E(t) \propto N(t)^\alpha$.

We have assessed the consistency of the topology of the mentions and retweets networks from one election to the other by computing the degree distribution and the degree correlations of the aggregated networks. The variation of the power law exponent of the degree distributions from one electoral period to the other is of 1% at most, whereas the degree correlations are shifted less than 10% from one year to the other. It is worth to mention that the values of these properties are also comparable to the results obtained in a previous work for a different political context [1]. This indicates that the underlying interaction dynamics are robust in the face of a change in social context; in this case, the transition that took place between the political system of the 2011 Spanish elections (two-party system) to the 2015 Spanish elections (multi-party system). On the other hand, the analysis of the daily evolution of the degree assortativity of the networks.
Fig. 2. Power law relationships of the total number of tweets, retweets and mentions per day as a function of the number of unique users that participated in the conversation each day for the 2015 campaign (top) and the 2016 campaign (bottom). Note that the data corresponding to the day of the elections (marked with a circle) were not included in the fit.

has enabled us to detect fluctuations that seem to be caused by disruptive events, both exogenous and endogenous, that have a relevant impact on the political conversation.

References

The homophily shows the Dunbar’s number: modeling and data

János Török¹,², Yohsuke Murase³ Hang-Hyun Jo⁴,⁵, Jnos Kertész²,⁶, and Kimmo Kaski⁷

¹ MTA-BME Morphodynamics Research Group, Budapest University of Technology and Economics, H-1111 Budapest, Hungary torok@phy.bme.hu,
² Department of Theoretical Physics, Budapest University of Technology and Economics, H-1111 Budapest, Hungary
³ RIKEN Center for Computational Science, 7-1-26, Minatojima-minami-machi, Chuo-ku, Kobe, Hyogo, 650-0047, Japan
⁴ Asia Pacific Center for Theoretical Physics, Pohang 37673, Republic of Korea
⁵ Department of Physics, Pohang University of Science and Technology, Pohang 37673, Republic of Korea
⁶ Department of Network and Data Science, Central European University, Nádor u. 9, H-1051 Budapest, Hungary
⁷ Department of Computer Science, Aalto University, Espoo FI-00076, Finland

In human societies homophily, the tendency of similar individuals getting associated and bonded with each other, is known to be a prime tie formation factor between a pair of individuals [1]. This association and bonding can be related to one or more features including gender, race, age, education level, economic and social status, and many more. Consequently homophily has a large impact on a number of fundamentally important social phenomena like segregation, inequality, perception biases, and the transmission of information between groups of individuals [2–5].

Homophily also ensures that our acquaintances have much in common with us. An average person has a complex social network based on his activities and social preferences. This network is, however, not unstructured but almost always groups of different size can be observed (see Fig. 1 (a)). These groups are often homogeneous in a social trait (e.g. city in the above example) but very often in more traits. This is natural since e.g. among those who like to play hobby football I will join the group with which I have more in common. This brings us to the main concept of our Abstract, the feature overlap. It was introduced in an earlier paper [6] and is defined as follows. Let us consider two persons \((i, j)\) with many social features (e.g. gender, location, education level, etc.) for which they can have different traits (e.g. male/female). We define the feature overlap \(o_{ij}\) as the fraction of matching features. We consider only features for which we have data for both individuals.

We can also measure how an ego is similar compared to its acquaintances. To this end for each ego we average the feature overlaps between the ego and its acquaintances \(o_{ego, i} = \langle o_{ij} \rangle_j\). This is an important quantity because this can be easily measured in social data and in models using traits and homophily. In Fig. 2 we show the evolution of the overlap as function of ego degree for two datasets. The first one is the iWiW[7] which was the Hungarian facebook, and a mobile phone dataset[8], for the latter only talkative users are selected. In the first one we had four features (age, gender, location,
education level), in the second three features (age, gender, location). Both figures show similar features, a decreasing part then a maximum around $k=100-200$.

It is important to note that the degree seen in a social network site is not related to the actual social degree of an individual [9]. The majority of the links on a communication channel are a subset of the real social ties. On the other hand some links do not exist in real life. They are often sings of followers of important persons (OSN), or non-social work relations (phone). So the degree in Fig. 2 reflects only the number of links seen in the data.

If one makes connection to only a few persons on a communication channel they will be very close ones and naturally very similar to the ego itself. As more and more acquaintances are added the similarity decreases naturally. Surprisingly it starts to increase again and has a peak at around the Dunbar’s number of 150[10]. The reason could be that egos having 100–200 acquaintances connect to their real friends. Those ones with more connections have to connect unrelated persons with little or no similarity (sharp decrease for large degrees). The ones with intermediate (10–20) acquaintances seem to choose one or two acquaintances per egocentric group and thus the low similarity between them.

The above idea is verified on the data by analyzing more in detail the structure of egocentric networks, and the results are reproduced using the combined version of the social network model using homophily [6] and the model of social link sampling of communication channels[9].

---

**Fig. 1.** Egocentric networks. Left: iWiW data, nodes are colored according to home city. Right: model quarters are colored if the trait is same as for the ego.
Fig. 2. The average overlap as function of the degree.

References

Topological study of the convergence in the Voter Model

Inés Caridi\textsuperscript{1}, Sergio Manterola\textsuperscript{2}, and Pablo Balenzuela\textsuperscript{2}

\textsuperscript{1} Calculus Institute, University of Buenos Aires and CONICET, Argentina, ines@df.uba.ar
\textsuperscript{2} Departamento de Física UBA-CONICET, Universidad de Buenos Aires, Argentina.

1 Introduction

The Voter Model has been extensively studied due to its simple formulation and theoretical capabilities. The study of the "active links", those edges which link nodes in a different state, has been a key element in the analysis of the model. Typically, the density of active links, $\rho$, is used to characterize the approach to equilibrium. In this work, we study the process of convergence to equilibrium via the analysis of the topological properties of what we call the "Active Link Network" (ALN).

We found that the ALN goes from a state similar to the underlying random network in the initial state to one extremely disassortative when the dynamics approaches to equilibrium. In this state, the ALN is dominated by "star-like" motifs, where different opinions play different topological roles on the network.

2 Results

The Voter Model is one of the most well-known models of opinion formation in which a population of agents evolves toward a consensus through a dynamics of imitation [1]. The original rules of the model establish that at each step $t$ of time, one of the $N$ agents (agent $i$) and one of his neighbors on the contact network (agent $j$) are selected at random. If the states of agents are different, agent $i$ takes the state of agent $j$. The model has been studied following different techniques [2], including mean-field-like approach [3]. The set of active links covers those edges which connect agents in a different state, and then, are susceptible to change their states. The typical observable is the density of active links with respect to the number of links of the contact network, $\rho$. For an ensemble of finite networks, $\rho$ follows an exponential decay to zero in the process of convergence to equilibrium [4]: high values of $\rho$ would indicate a system close to its initial state and low values, a system near consensus. Nevertheless, the study of the active links in terms of a network has not been carried out yet. In this work, we focus on the Active Links Network (ALN), which is defined by the induced network from the set of active links. We consider the evolution of a Voter Model of $N$ agents immersed in a contact network, $G$, of the type of Erdos-Renyi with mean degree $k = 12$. Let us call $s_i$ the opinion state of the $i$-agent, which can take two possible values: 1 and $-1$. At the beginning, the states of the agents are randomly chosen with the same probability (0.5) for each one. The degree of the $i$-agent on the contact network is $k_i$, and the mean degree of his neighbors on the ALN is $kk_i$. The set of active links, changing along the
dynamics, is a subset of the set of links of $G$. The Active Links Network, $G_a$, is defined as the subgraph of $G$ induced from the set of active links.

Panel a) and d) of Figure 1 show snapshots of the ALN at two time steps of one realization, panel b) and e) show the degree correlation of the ALN (mean of the degree of the neighbors of the nodes vs mean of the degree for each one of the agents of the ALN), and panel c) and f) show the degree of the contact network vs degree on the ALN for each node(s) which belongs to the ALN. In all the cases, white circles correspond to agents in state 1 and black ones, to agents in state $-1$.

![Image](image_url)

**Fig. 1.** Realization of $N = 1000$ agents, whose underlying network of contacts is an Erdos Renyi of $k = 12$, a), b) and c) (up and from left to right) correspond to a time step when $\rho = 0.026$ and d), e) and f) (down and from left to right) correspond to other time step with $\rho = 0.246$.

The particular realization represented in the Figure achieves the consensus of the dynamics with all the agents in the opinion state $-1$. We can observe that in the step corresponding to $\rho = 0.026$, it is possible to note that points are separated in accordance with their opinion state. White circles take values of small mean degree while the mean degree of their neighbors is high. The opposite occurs for black circles. The snapshot of the corresponding ALN (panel a) revels that white nodes are playing the role of *leaves* of stars and black nodes, the role of *centres*. The same situation occurs another time at the dynamics of this realization (not shown here), although interchanging states 1 by $-1$ and vice versa: in this case, those agents in state opinion 1 are playing the role of centres of the starts, and those in state opinion $-1$ are the leaves.

In Figure 2 (left), we plot the mean degree vs $\rho$ for each one of the populations (states 1 and $-1$) for each time step of 90 realizations of the dynamic. We use different colors associated with different ranges of $\rho$ values.
Let us call the \textit{activity} of an agent $a_i$, which takes the value 1 when agent belongs to the ALN, and 0 in another case. The \textit{role} to agent $i$, $r_i$, takes three possible values: $r_i = 1$ if $kk_i > k_i$ (the agent is a leave), $r_i = 0$ if $kk_i < k_i$ (the agent is a star), and $r_i = 2$ if $kk_i = k_i$ (other case). In order to quantify the fact that the two populations play different topological roles on the ALN along the dynamics, we calculated the mutual information between the \textit{role} and the \textit{state} of the agents for the set of agents belonging to the ALN, $I(r,s|a=1)$. Figure 2 (right) shows the normalized mutual information between the \textit{role} and the \textit{state} vs $\rho$, where the normalized mutual information is defined by $I(r,s|a=1)/H(r|a=1)$, being $H(r|a=1)$ the Entropy of $r$, given that the nodes belongs to the ALN, and thus, the maximum value that $I(r,s|a=1)$ can take. Then, the normalized mutual information ranges from 0 to 1. It is clear that for high values of $\rho$, data about the state of the opinion of the agents of the ALN does not bring information about the topological role of the agent on the ALN and vice versa. On the other hand, when $\rho$ decreases, the fact of knowing one of the variables brings information about the other one. The extreme scenario is reached when the normalized mutual information takes the maximum value 1: all the agents of one of the populations are playing the role of leaves, and the others, the role of stars. This case is that represented in Figure 1 (first line of plots). Between these two cases (high and low values of $\rho$), normalized mutual information reflects a slow transition.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2}
\caption{Panel left: mean degree of the nodes from a state’s population vs $\rho$. Each step of each realization contributes two points to the plot: one for the mean degree of the population of agents in state 1 and one for agents of state $-1$. Different colors are used to identify ranges of values of $\rho$. Panel right: normalized mutual information vs $\rho$ for each one of the 90 realizations of a dynamic of $N=1000$ agents with an underlying network of contacts of Erdos Renyi of $k=12$.}
\end{figure}

\section*{References}

Part XVII

Structural Network Measures
Dimension as an Invariant of Street Networks

Franz-Benjamin Mocnik
Heidelberg University, 69120 Heidelberg, Germany
mocnik@uni-heidelberg.de
www.mocnik-science.net

1 Introduction

Street networks have been examined in respect to their structure [1,2,3,4]. The author of this paper has previously examined networks from various domains, thereby demonstrating that the polynomial volume law applies to many of them [5]. This paper focuses on the geographical domain only. Thereby, it examines the dimension of street networks, which turns out to be very stable.

2 Networks and Space

Things exist and happen in space and time. While different conceptualizations of space exist – space as a container, space instantiated by objects and their relations [6], etc. – they all agree on the fact that space can provide structure to things that are placed in space. Mocnik [7,8] has discussed the influence of space on networks inheriting a spatial structure. Thereby, networks with a spatial structure, or spatial networks in short, can be regarded as a prototypical example of spatial information that consists of relations.

The question of which properties are inherited from space depends on its conceptualization – Euclidean and metric space, topological space, or geographical space, to name a few. Tobler’s law is an example of how space and relations mutually influence: ‘everything is related to everything else, but near things are more related than distant things’ [9]. When being interpreted in a geographical context, the law describes geographical information as being continuous. Tobler’s law can prototypically be modelled by the Mocnik (network) model [5,7,8]. The model introduces edges between near nodes for a given set of nodes. Thereby, a node \(n_1\) being ‘near’ to another node \(n_2\) is defined as

\[
\text{dist}(n_1,n_2) \leq \rho \cdot \min_{m \neq n_1} \text{dist}(n_1,m)
\]

for some given constant \(\rho > 1\). The resulting model has been shown to resemble many spatial networks, among them, street and public transport networks [5].

3 The Polynomial Volume Law

One of the most prominent characteristics of space is the possibility to measure volumes. In particular, the volume of a ball scales as \(r^d\) in dependence of the radius \(r\), thereby encoding the dimension \(d\). A concept similar to the volume of a ball can be
introduced for thematic information represented by a network. The ball $B_n(r)$ can be defined as all nodes that are within network distance $r$ from the centre node $n$. The volume of this ball is, in turn, defined as the number of the contained nodes.

When relations are influenced by space, it can be hypothesized that the volume of a ball in the corresponding network scales in the same way as the volume of a ball in space, called the polynomial volume law [5]. This law has turned out to be valid for a large number of networks, indicating that they are, in fact, strongly influenced by space. The volumes in a network can statistically be examined (for various centre nodes and radii) and fitted by the polynomial volume law. The resulting estimation of the parameter $d$ can be regarded as the dimension of the network. This dimension by the polynomial volume law has been compared to other concepts of dimension in a network, exposing strong similarities and some differences [5]. Examples of brain, public transport, and road networks, among others, have been shown to follow this law with reasonable estimates of the dimension (Figure 1).

![Fig. 1. Polynomial volume law at the examples of real-world networks. (a) Bus network of Manhattan; (b) Subway in New York City; (c) Road network of California; and (d) Brain network. The figure has been adapted from a previous publication [5] under the CC BY 4.0 license.]

4 Dimension as an Invariant of Street Networks

While several real networks have been shown to follow the polynomial volume law, it has not yet been examined whether the estimate of the dimension is stable across many real-world networks of similar type to which they can thus serve as an invariant. Table 1 shows the dimension for a number of street networks. In fact, the dimension is very similar for these networks. The estimates are distributed around $2$, which suggests itself due to the two-dimensional surface of the Earth. This is despite the varying size of the examined cities and their different network structures. In particular, the networks differ strongly in their centralization, which measures the variance of the centrality within the network, and their assortativity, given as the Pearson correlation coefficient of the degrees.

5 Conclusion

The influence of space on networks can not only be traced but is seemingly also very stable, as is demonstrated here at the example of street networks. Future research might examine further examples of networks and test for which kind of networks the dimension can serve as an invariant.
Table 1. Street networks. The networks are extracted from OpenStreetMap (http://openstreetmap.org) using OSMnx (https://github.com/gboeing/osmnx).

<table>
<thead>
<tr>
<th>City</th>
<th>Dimension</th>
<th>Nodes</th>
<th>Edges</th>
<th>Centralization Degree</th>
<th>Centralization Betweenness</th>
<th>Assortativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heidelberg, Germany</td>
<td>1.91 ± 0.51</td>
<td>2726</td>
<td>6403</td>
<td>9.74 · 10^{-4}</td>
<td>1.43 · 10^{-1}</td>
<td>1.76 · 10^{-1}</td>
</tr>
<tr>
<td>Oxford, UK</td>
<td>1.92 ± 0.57</td>
<td>3363</td>
<td>7512</td>
<td>5.26 · 10^{-4}</td>
<td>1.77 · 10^{-1}</td>
<td>4.56 · 10^{-2}</td>
</tr>
<tr>
<td>Reykjavik, Iceland</td>
<td>1.93 ± 0.58</td>
<td>10157</td>
<td>20877</td>
<td>1.92 · 10^{-4}</td>
<td>1.96 · 10^{-1}</td>
<td>3.87 · 10^{-2}</td>
</tr>
<tr>
<td>Lund, Sweden</td>
<td>1.93 ± 0.60</td>
<td>1792</td>
<td>4298</td>
<td>8.95 · 10^{-4}</td>
<td>1.78 · 10^{-1}</td>
<td>1.88 · 10^{-1}</td>
</tr>
<tr>
<td>Greenwich, UK</td>
<td>1.94 ± 0.59</td>
<td>4750</td>
<td>10891</td>
<td>5.70 · 10^{-4}</td>
<td>2.71 · 10^{-1}</td>
<td>3.65 · 10^{-2}</td>
</tr>
<tr>
<td>Cambridge, UK</td>
<td>1.96 ± 0.60</td>
<td>3345</td>
<td>7360</td>
<td>5.39 · 10^{-4}</td>
<td>2.61 · 10^{-1}</td>
<td>1.00 · 10^{-1}</td>
</tr>
<tr>
<td>Melbourne, Australia</td>
<td>1.98 ± 0.53</td>
<td>180971</td>
<td>406940</td>
<td>2.07 · 10^{-5}</td>
<td>2.14 · 10^{-1}</td>
<td>1.59 · 10^{-1}</td>
</tr>
<tr>
<td>Uppsala, Sweden</td>
<td>1.98 ± 0.57</td>
<td>8116</td>
<td>18931</td>
<td>3.29 · 10^{-4}</td>
<td>1.31 · 10^{-1}</td>
<td>1.86 · 10^{-2}</td>
</tr>
<tr>
<td>Bonn, Germany</td>
<td>1.99 ± 0.51</td>
<td>5827</td>
<td>13871</td>
<td>4.50 · 10^{-4}</td>
<td>2.21 · 10^{-1}</td>
<td>7.37 · 10^{-2}</td>
</tr>
<tr>
<td>Bremen, Germany</td>
<td>1.99 ± 0.53</td>
<td>8903</td>
<td>20346</td>
<td>3.05 · 10^{-4}</td>
<td>2.44 · 10^{-1}</td>
<td>1.09 · 10^{-1}</td>
</tr>
<tr>
<td>Greater London, UK</td>
<td>2.03 ± 0.54</td>
<td>124665</td>
<td>295088</td>
<td>2.11 · 10^{-3}</td>
<td>1.27 · 10^{-1}</td>
<td>4.37 · 10^{-2}</td>
</tr>
<tr>
<td>Santa Fe, NM, USA</td>
<td>2.03 ± 0.57</td>
<td>4083</td>
<td>9895</td>
<td>6.32 · 10^{-4}</td>
<td>1.60 · 10^{-1}</td>
<td>8.33 · 10^{-2}</td>
</tr>
<tr>
<td>Vienna, Austria</td>
<td>2.06 ± 0.49</td>
<td>16054</td>
<td>36183</td>
<td>1.71 · 10^{-4}</td>
<td>1.62 · 10^{-1}</td>
<td>1.86 · 10^{-1}</td>
</tr>
<tr>
<td>Manhattan, NY, USA</td>
<td>2.07 ± 0.38</td>
<td>4473</td>
<td>9729</td>
<td>8.56 · 10^{-4}</td>
<td>2.00 · 10^{-1}</td>
<td>3.65 · 10^{-1}</td>
</tr>
<tr>
<td>Albuquerque, NM, USA</td>
<td>2.08 ± 0.54</td>
<td>22537</td>
<td>57020</td>
<td>1.10 · 10^{-4}</td>
<td>2.65 · 10^{-1}</td>
<td>2.05 · 10^{-1}</td>
</tr>
<tr>
<td>Münster, Germany</td>
<td>2.13 ± 0.55</td>
<td>6723</td>
<td>15878</td>
<td>3.93 · 10^{-4}</td>
<td>1.81 · 10^{-1}</td>
<td>1.83 · 10^{-2}</td>
</tr>
<tr>
<td>Paris, France</td>
<td>2.17 ± 0.51</td>
<td>10176</td>
<td>19947</td>
<td>3.97 · 10^{-4}</td>
<td>8.13 · 10^{-2}</td>
<td>2.15 · 10^{-1}</td>
</tr>
</tbody>
</table>

Software. An implementation of the Mocnik model as well as of the algorithm to estimate the dimension of a network is published as part of NetworKit (https://github.com/kit-parco/networkit), an open-source toolkit for large-scale network analysis.

Funding. The author has been funded by Deutsche Forschungsgemeinschaft as part of the project A framework for measuring the fitness for purpose of OpenStreetMap data based on intrinsic quality indicators (FA 1189/3-1).

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Interpreting and comparing the length of complex networks

Gorka Zamora-López1,2 and Romain Brasselet3

1 Dept. of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain,
gorka@zamora-lopez.xyz,
WWW home page: http://www.zamora-lopez.xyz
2 Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain,
3 Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy.

1 Introduction

The phenomenon of small-world networks has fascinated popular culture and science for decades. Discovered in the realm of social sciences it stands for the observation that two randomly selected individuals are usually connected by a short chain of social ties, even considering the large size and geographic widespread of the human population [1]. Most real networks studied so far, from natural to man-made systems, have been found to exhibit this small-world effect as well [2] [3] [4]. Usually, when we say that "a complex network is small-world" we mean that its average pathlength – the average distance between all pairs of nodes – is much smaller than the number of nodes the network is made of [3]. Despite its significance, a quantitative determination of how short or how long a network is compared to others has remained unresolved. In the physical world we define reference metric systems in order to quantify the size of objects. However, when dealing with complex networks the same procedure is not suitable because every network is a metric space of its own whose properties are strongly bounded by its number of nodes \( N \) and its number of links \( L \) (or density \( p \)). Therefore, the question of whether a network is smaller or larger than another implies the comparison of two different metric spaces with each other instead of the more familiar situation, in which objects are compared within the space they live. The usual strategy to deal with this problem in the literature is to compare the properties of real networks to those of well-known graph models, e.g., random graphs, regular lattices or scale-free networks [5] [6] [7] [8]. While these models represent a variety of null-hypotheses, useful to answer particular questions we may have about the data, they do not correspond to absolute or universal references.

2 Results

In order to solve this problem, in the present work we establish a reference framework under which the length of networks can be interpreted and compared. Instead of relying on a comparison to typical models, we will interpret the length of a network based on how it deviates from the shortest and the longest value it could possibly take. To do so,
Fig. 1. Upper and lower boundaries for the pathlength and efficiency of complex networks.

(a) Average pathlength of ring lattices (red) and random graphs (green) of \( N = 1000 \) nodes, compared with the corresponding upper and lower boundaries for ultra-long (yellow) and ultra-short (grey) networks. Shaded areas mark values of pathlength or efficiency that no graph of \( N = 1000 \) nodes can achieve, depending on density. The pathlength of the two models decay towards the ultra-short boundary at sufficiently large density. (b) Same as (a) displayed in terms of efficiency. (c) – (f): Average pathlength of sample empirical networks from three domains (neural, social and transportation). Absolute pathlength (c) and three different relative versions of the pathlength (d)–(f). Results in panel (e) show that the pathlength of all empirical networks lies close to that of equivalent random graphs. With this result alone, we would be prone to interpret that all networks are “small-world”. However, comparison to the true lower limit for a network’s pathlength, panel (f), reveals that only some networks are close to the lower limit while others are much longer than this limit.

we have first performed an exhaustive search to discover the upper and the lower limits for the average pathlength of networks, of arbitrary size and link density. We have achieved this for both directed and undirected graphs, either sparse (disconnected) or dense (connected). We study the disconnected cases in terms of the efficiency [9]. We have found that all these limits are given by families of singular network configurations we generically name as ultra-short and ultra-long networks. Among these configurations a few observations are to be highlighted: (i) All graphs (directed or indirected) with diameter \( \text{diam}(G) = 2 \) have exactly the same pathlength and they are ultra-short, regardless of the specific internal arrangements of their links. (ii) The optimal configurations are not always unique and may vary according to size or density. (iii) Undirected ultra-short and ultra-long networks can be generated adding links, one-by-one, to star and path graphs respectively but optimal directed graph generation is often non-Markovian. That is, the optimal configuration for a digraph with \( L + 1 \) arcs cannot always be achieved by adding one arc to the optimal digraph with \( L \) arcs.

In practical terms, we could show that typical models (random, scale-free and ring networks) undergo a transition as their density increases, eventually becoming ultra-short, 1(a) and (b). The convergence rate, however, depends on the properties of each model. Application to well-known empirical networks (neural, social and transportation) illustrates how this framework allows for a complete and accurate interpretation.
of the length of complex networks, Fig. 1(c)–(f). While many graphs display a path-length comparable to random graphs, when compared to the ultra-short boundary, we find that only the cortical connectomes are optimal, meaning that it is almost impossible to reduce their pathlength even by performing a selective rewiring of their connectivity. On the contrary, transportation networks are the least efficient of the three classes here illustrated. They are 2 to 10 times longer than their corresponding ultra-short limit.

Summary. Our theoretical findings solve the practical problem of assessing, comparing and interpreting how short (or how long) a complex network is. A major difficulty for the evaluation and comparison of networks is that the outcome of graph measures very much depends on the size and the number of links. In order to disentangle the influence of size and density from the purely topological contribution the common procedure has been to consider reference models and interpret the results in terms of the values the graph measures take on these null-models. Our results allow, for the first time, to assess complex networks based on how far their length deviates from the minimal and maximal possible, instead of relying on the arbitrary choice for a reference model. Future efforts shall be carried to identify the limits of other graph measures and thus contribute to a more reliable analysis and interpretation of complex networks.

References

Coexisting Mesoscale Patterns in Complex Networks

M. J. Palazzi1, J. Borge-Holthoefer1, C. J. Tessone2, and A. Solé-Ribalta1

1 Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Spain
2 URPP Social Networks, Universität Zürich, Switzerland.

1 Introduction

The detection and identification of emergent structural patterns has been one of the main focuses in modern network theory. Such interest is not surprising, because these structural patterns lie at the core of the discipline as one of the keys to the origins – which are the assembly rules that led to an observed pattern? – and dynamics – how does the structure constrain the performance of the network’s activity? – of a network.

In this context, modularity [8] has concentrated most of the efforts [6]. However, there are other architectural principles which, in some contexts, may play more important roles. In Ecology, for instance, nestedness [1] has been crucial to understand the stability and diversity of ecological systems [10, 2], but also more colourful possibilities have been explored [7]. In this scenario, the accent has been mainly placed on designing and improving algorithms to detect those patterns, on understanding the dynamical constraints that those patterns impose or in describing plausible microscopic rules that make this patterns emerge. Nonetheless, we have limited knowledge on how different structural signatures may be intertwined, how they affect and limit each other – or if they do at all. We have examples in which two or more structural features have been jointly considered [5, 4, 3]. But such consideration overlooked to what extent the inherent constraints of one pattern limit –or boost– the presence of the other.

Here, we explore the relationships between three structural patterns, one at the macroscale: nestedness $N$; and two at the mesoscale: modularity $Q$ and in-block nestedness $I$.

2 Results

We experimentally study the dependence between the aforementioned structures by means of a probabilistic network model introduced in [9]. The model asks for four parameters: the number of modules $B$, the fraction of inter modules links $\mu$, the fraction of links outside the perfect nested structure $p$, and the slimness of the nested structure $\xi$. For a detailed description of the model see [9]. Figure 1(A) shows the effect of the parameters on the generated networks. To quantify modularity, we employ the measure of Girvan and Newman [8], and to quantify nestedness and in-block nestedness we employ the definitions introduced in [9]. We obtain the different measures for $1.45 \times 10^5$ networks in a wide range of parameters with a fixed community size $N_B = 50$, so as we add communities we are also increasing the size of the network. We represent our results over different heatmap ternary plots. A ternary plot is a three-variables diagram on
which the variables sum to a constant, and displays the proportion of the three variables in an equilateral triangle. Here, the bottom-left vertex represents purely in-block nested networks, the bottom-right represents purely nested networks, and the top vertex those networks that are purely modular. This representation allows us to explore which network structural configurations map onto which region on the ternary. Panel (B) shows the distribution of networks across the ternary, the colorbar indicates the amount of networks in each bin of the ternary. The colorbar in panels (C), (D) and (E) shows the normalized values of $N$, $Q$ and $I$, respectively. Clearly, two sharp boundaries emerge, $F_1$ and $F_2$.

Boundary $F_1$, orthogonal to the center of the $N - I$ axis, corresponds to networks with $B = 1$, regardless the values of $\xi$, $p$ and $\mu$, panels (G-I) and $N = I$, panels (C-E). This boundary is induced by the definitions of $N$ and $I$, since in-block nestedness measure reduces to nestedness when the network has a single block with nested structure. Boundary $F_2$, almost orthogonal to the $I - Q$ axis, is obtained for high values of $\xi$ and low values of $p$ and $\mu$ (panels (G-H)), i.e., networks where the blocks internal organization approximates to a star network with low noise within and between the blocks. This situation resembles a ring of stars graphs. Here we have almost constant and maximal values of $I$ and complementary values for $N$ and $Q$, panels (C-E).

Finally, in order to set approximate bounds for the mesoscale patterns, we analyze the situation corresponding to the second boundary. In particular, we perform an analytical formulation for a growing ring of stars graphs –i.e. $\xi \to \infty$ and increasing number of $B$ stars--, having $N_B$ nodes on each star, and a single link connecting them through the central node. This configuration maximises the contribution of each star to the modularity and in-block nestedness. These expressions allow us to completely characterize that second boundary on the ternary plot (see left column of the table). We show that nestedness can be used to provide approximate bounds for modularity, both quantities being complementary (i.e. $Q = 1 - N$; see right column of the Table) in the limit $N_B \to \infty$, and that in-block nestedness, provides a parsimonious transition between nested to modular networks taking properties of both.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$Q$</th>
<th>$I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = BN_B^2 - BN_B^2 - 3BN_B + B + 2N_B + 2$</td>
<td>$Q = 1 - \frac{1}{N_B} - \frac{1}{B}$</td>
<td>$I = 1 - \frac{3}{BN_B} - \frac{2}{N_B}$</td>
</tr>
</tbody>
</table>

Nestened when considering limiting cases: $\lim_{N_B \to \infty} N = \frac{1}{B}$ and $\lim_{B \to \infty} N = 0$

The expressions for modularity

$\lim_{N_B \to \infty} Q = 1 - \frac{1}{B}$ and $\lim_{B \to \infty} Q = 1 - \frac{1}{N_B}$

and finally, for in-block nestedness

$\lim_{N_B \to \infty} I = 1$ and $\lim_{B \to \infty} I = 1 - \frac{2}{N_B}$

References


The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Fig. 1. (A) examples of synthetic networks generated with the graph model introduced in [9]. (B) shows the distribution of the generated networks over the ternary plot. (C), (D) and (E) show the normalized values of $N$, $Q$ and $I$, respectively. Panels (F) - (I): results with respect to the graph model parameters, number of blocks $B$, shape $\xi$, and noise parameters $p$ and $\mu$, respectively.

Mixing patterns and individual differences in networks

George T. Cantwell\textsuperscript{1} and M. E. J. Newman\textsuperscript{1,2}

\textsuperscript{1} Department of Physics, University of Michigan, Ann Arbor, Michigan, USA
\textsuperscript{2} Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan, USA

1 Introduction

We study mixing patterns in networks, meaning the propensity for nodes of different kinds to connect to one another. The phenomenon of assortative mixing, whereby nodes prefer to connect to others that are similar to themselves, has been widely studied \cite{2, 4}, but here we go further and examine how and to what extent nodes that are otherwise similar can have different preferences. Many individuals in a friendship network, for instance, may prefer friends who are roughly the same age as themselves, but some may display a preference for older or younger friends.

We introduce a network model that captures variance in assortativity along with methods for fitting it to observed network data. The fit allows us to make best estimates of the preferences of individual nodes, define metrics to quantify individual variation in assortativity, perform sensitive community detection even in the absence of traditional assortative structure, and accurately predict missing data in unlabeled or partially labeled networks.

2 Estimating preferences

Consider a network where nodes are divided into a number of discrete, non-overlapping groups, types, or categories, and where individual nodes have preferences about the types of the nodes with which they have network connections. For the moment, assume that the type of every node is explicitly specified.

Let $k_{is}$ be the number of edges from node $i$ to nodes of type $s$ and $k_i = \sum_s k_{is}$ be the total number of edges from node $i$. Then the ratio $k_{is}/k_i$, the fraction of edges from node $i$ to nodes of type $s$, gives us a simple measure of $i$’s preferences. This measure is a rather crude one. In particular, it gives us little information when the degree $k_i$ is small. If someone lists only one friend, who happens to be a man, it would not be correct to conclude that this person has a 100% preference for male friends.

To quantify preferences more accurately, we consider how the observed value of $k_{is}$ arises in probabilistic fashion from a node’s true underlying preference. We let $x_{is}$ be the underlying (latent) preference, and assume that node $i$ connects to each node in group $s$ with a probability proportional to $x_{is}$. We assume a Dirichlet prior for $x_{is}$ and use the method of maximum likelihood to find estimates for the parameters of this distribution.

The Dirichlet choice for the prior allows us to write a probability for observing a network, given the Dirichlet parameters and the groups of each node. Let $A$ be the
adjacency matrix of the network, let $g_i$ be the group of node $i$, and let each group have its own set of Dirichlet parameters, $\alpha_s$. Then we can write

$$P(A|\alpha, g) \propto \prod_i B(\alpha_{g_i} + k_i) / B(\alpha_{g_i})$$

where $B(x)$ is the multi-dimensional beta function. Maximizing this expression with respect to $\alpha$ provides us with estimates for the underlying distribution of preferences. Figure 1 shows two examples comparing the naive estimates of preference to the estimates from the method described here. Note, in both cases the naive estimates show a very different picture of the variance, in these cases significantly over-estimating it.

3 Assortativity and variance coefficients

The parameters of the fitted Dirichlet model can be used to define assortativity and variance coefficients. We define the normalized preference assortativity, $R$, as

$$R = \frac{\sum_r p_r (\alpha_r / \alpha_0 - K_r / m)}{\sum_r p_r (1 - K_r / m)},$$

where $p_r$ is the proportion of nodes in group $r$, $K_r$ is the sum of the (out-)degrees in that group, $m$ is the sum of all (out-)degrees, and $\alpha_0 = \sum_s \alpha_s$. Further, we define the
normalized variance coefficient, \( V \), as
\[
V = \sum_r \frac{p_r}{\alpha_r + 1}.
\] (3)

While it is well established that many networks are significantly more assortative than would be expected by chance (large \( R \)), we also find that many networks show significant preference variation (large \( V \)). And even when \( V \) is not large the result is still an interesting one: it tells us a lot about a population if we find that their preferences are entirely determined by, say, gender or race, with little to no variation within any given group.

4 Recovering missing data

When we are missing some (or all) of the group labels on nodes, we can use Eq. (1) to simultaneously recover the labels and find the Dirichlet parameters \( \alpha \) by employing an expectation–maximization algorithm. The process is reminiscent of community detection using the stochastic block model [1], but our approach should outperform the stochastic block model whenever there is significant within group variation. Indeed our model is capable of accurately recovering group labels in this scenario even when there is no average assortativity.

5 Summary

Nodes often show assortative (or disassortative) preferences on average but the preferences can also exhibit a large degree of variation [5]. We introduce a model of individual preferences in networks and their variation and a method for fitting it to observed network data. The fitted model parameters can be used to define coefficients that represent the average assortativity and variance of preferences. For fully labeled networks in which the characteristics of every node are known, maximum likelihood estimation of the parameters is straightforward. When networks are only partly labeled, or completely unlabeled, the model can be used to recover the missing labels and we present an expectation–maximization algorithm for doing so. Among other things, our methods can be used even when average mixing patterns are weak.

References

Centrality measures for graphons: Accounting for uncertainty in networks

Michael T. Schaub\textsuperscript{1,2}, Marco Avella-Medina\textsuperscript{3} Francesca Parise\textsuperscript{4}, and Santiago Segarra\textsuperscript{5}

\begin{itemize}
\item \textsuperscript{1} IDSS, Massachusetts Institute of Technology, Cambridge MA, USA
\item \textsuperscript{2} Department of Engineering Science, University of Oxford, Oxford, UK
\item \textsuperscript{3} Department of Statistics, Columbia University, New York, USA
\item \textsuperscript{4} LIDS, Massachusetts Institute of Technology, Cambridge MA, USA
\item \textsuperscript{5} Department of Electrical and Computer Engineering, Rice University, Houston TX, USA
\end{itemize}

1 Summary

As relational datasets keep increasing in size and their data-acquisition is permeated by uncertainty, graph-based analysis techniques can become computationally and conceptually challenging. In particular, node centrality measures rely on the assumption that the graph is perfectly known — a premise not necessarily fulfilled for large, uncertain networks. Accordingly, centrality measures may fail to faithfully extract the importance of nodes in the presence of uncertainty. To mitigate these problems, we suggest a statistical approach based on graphon theory: we introduce formal definitions of centrality measures for graphons and establish their connections to classical graph centrality measures \cite{1}. A key advantage of our approach is that centrality measures defined at the modeling level of graphons are inherently robust to stochastic variations of specific graph realizations. Using the theory of linear integral operators, we define degree, eigenvector, Katz and PageRank centrality functions for graphons and establish concentration inequalities demonstrating that graphon centrality functions arise naturally as limits of their counterparts defined on sequences of graphs of increasing size. The same concentration inequalities also provide high-probability bounds between the graphon centrality functions and the centrality measures on any sampled graph, thereby establishing a measure of uncertainty of the measured centrality score.

2 Centrality measures and uncertainty

A fundamental task in network analysis is to identify salient features in the underlying system, such as key nodes or agents in the network. To identify such important agents, researchers have developed centrality measures in various contexts, each of them capturing different aspects of node importance. Prominent examples for the utility of centrality measures include the celebrated PageRank algorithm \cite{2}, employed in the search of relevant sites on the web, or the identification of influential agents in social networks to facilitate viral marketing campaigns \cite{3}.

Most existing applications of network centrality measures follow the paradigm in Figure 1a: a specific graph — such as a social network with friendship connections —
Fig. 1. Schematic – Network centrality analysis. (a) Classical centrality analysis computes a centrality measure purely based on the observed network data. (b) If networks are subject to uncertainty, we may adopt a statistical perspective on centrality, by positing that the observed network is but one realization of a true, unobserved latent model. Inference of a model then would lead to a centrality estimate that accounts for the uncertainty in the data in a well-defined manner. (c) Illustrative example. Top panel: A network of 100 nodes is generated according to a graphon model with a well-defined increasing connectivity pattern. Bottom Panel: This graphon model defines a latent (expected) centrality for each node (blue curve). The centralities of a single realization of the model (red curve) will in general not be equivalent to the latent centrality, but deviate from it. Estimating the graphon-based centrality thus allows us to decompose the observed centrality into an expected centrality score (blue), and a fluctuation that is due to randomness.

is observed, and conclusions are then drawn about the importance of each agent based a centrality measure that ranks the nodes according to the observed network structure. A common feature among all centrality measures is that the importance of each agent derives entirely from the observed network. A crucial assumption for the applicability of these centrality measures is thus that the observation of this network captures all the data we care about, i.e., the observed data is assumed to be complete and noise free.

However, in many instances in which centrality measures are employed, this assumption is arguably not fulfilled: we usually do not observe the complete network at once. Further, even for those parts that we observe there are measurement errors, such as false positive or false negative links, and other forms of uncertainty [4]. The key question is therefore how to identify crucial nodes via network-based centrality measures without having access to an accurate depiction of the “true” latent network.

3 Graphon centralities

One answer to the above problem is to adopt a statistical inference-based viewpoint towards centrality measures, by assuming that the observed graph is a specific realiza-
tion of an underlying stochastic generative process (Figure 1b). In this work [1], we use (sparse) graphons to model such underlying generative process. Graphons provide a rich non-parametric statistical framework and it has been recently shown that they can be efficiently estimated from one (or multiple) noisy graph observations [5]. In particular, graphons encapsulate a broad class of network models including the stochastic block model [6], random dot-product graphs, the infinite relational model, and others [7].

We harness the flexibility provided by the graphon framework to suggest a statistical approach to agents’ centralities that inherently accounts for network uncertainty. Our main contribution is to show that, based on an inferred graphon (which accounts for all statistically significant patterns in the data), one can compute a latent centrality profile of the nodes that we term graphon centrality function. This graphon centrality may be interpreted as a fundamental measure of node importance, irrespective of the specific realization of the graph at hand. This leads to a robust estimate of the centrality profiles of all nodes in the network. See Figure 1b-c for an illustration of these ideas.

In fact, we provide explicit high-probability bounds between the distance of such latent graphon centrality functions and the centrality profiles in any realized network, in terms of the network size. Hence, for a network of a given size we know what size of fluctuations to expect from our centrality scores and can thus perform a robust centrality analysis, that accounts for the uncertainty in the data.

Acknowledgements This work was supported by the Swiss National Science Foundation through the grants P2EGP1_168962, P300P1_177746 (M. Avella-Medina) and P2EZP2_168812 (F. Parise); the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 702410 (M. Schaub); and the MIT IDSS Seed Fund Program (S. Segarra)

References

The distribution of shortest path lengths in random networks

Eytan Katzav\(^1\), Ofer Biham\(^1\), and Reimer Kühn\(^2\)

\(^{1}\) Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel, \(^{2}\) Mathematics Department, King’s College London, Strand, London WC2R 2LS, UK
eytan.katzav@mail.huji.ac.il

The theory of complex networks provides a useful conceptual framework for the study of a large variety of systems and processes in science, technology and society. These studies are based on network models, in which the nodes represent physical or virtual objects, while the edges represent the interactions between them. Typically, these networks exhibit random structures, which can be characterized by their statistical properties at the local and global scales. The local structure of a network is captured by the degree distribution and by certain correlations between the properties of nearby nodes. However, the large scale structure of a network is captured by the spectrum of path lengths between random pairs of nodes. The shortest path between each pair of nodes is of particular importance because it provides the strongest interaction and fastest response between these nodes. The average lengths of these paths were studied extensively, while the entire distribution of shortest path lengths (DSPL), \(P_{DSPL}(L = \ell)\), has attracted limited attention.

The DSPL provides a useful framework for structural analysis of networks, such as network hyperbolicity [1], as well as for analysis of dynamical processes on networks, such as the propagation of information, traffic navigation [2], and epidemic spreading. To give just one example, considering an epidemic which starts from a random individual, \(i\), in the limit of high infection rate, the temporal spreading of the infection is determined by the shell structure around node \(i\) (Fig. 1). Thus, the expectation value of the number of nodes infected up to time \(t\), \(N_I(t)\), can be expressed in terms of the DSPL and is given by \(N_I(t) = 1 + (N - 1)[1 - P_{DSPL}(L > t)]\), where \(N\) is the network size.

![Fig. 1. Illustration of the shell structure around a reference node in a random network. The \(\ell^{th}\) shell consists of the nodes which are at distance \(\ell\) from the reference node.](image-url)
Recently, we developed a suite of analytical approaches for the calculation of the DSPL in a wide range of random networks both in and out of equilibrium. The first approach is the Random Shells Approach (RSA), which is designed to treat Erdős-Rényi (ER) networks [3]. We later improved RSA in Ref. [4] to take into account the detailed size and micro-structure of the giant cluster. This improvement yields very good results even in the vicinity of the percolation transition of the ER ensemble, which takes place when the mean degree is $c = 1$. To obtain a complete understanding of the problem, including the subpercolating regime, we have developed a different methodology, based on a topological expansion [5], which yields an exact result, namely an exponential distribution, for the DSPL with $c < 1$, conditioned on the nodes being on the same cluster. An interesting conclusion is that the mean distance between random nodes is $E[L|L < \infty] = \frac{1}{1 - c}$, which means that in the vicinity of the percolation transition it diverges. Among other things, it means that even within the ER ensembles the common lore that distances are "small-world" is far from being the full picture.

![Fig. 2](image)

(a) The tail distribution of shortest path lengths, $P_{DSPL}(L > \ell)$, for Supercritical ER networks above the percolation threshold, using the Random Paths Approach. (b) The exact result for the DSPL conditioned on finite distances $P_{DSPL}(L = \ell|L < \infty)$ for Subcritical ER networks below the percolation threshold, using the topological expansion. In both cases $N = 10^4$. The analytical results (solid lines) agree with the results of computer simulations (symbols).

The second methodology, named the Random Paths Approach (RPA), was also developed in the context of the ER ensemble [3], but unlike RSA lent itself to generalization. We identified and formalized the relation between RPA and the cavity method, which allowed application of the RPA to configuration model networks [6], and in particular to random regular graphs (were the exact solution is a Gompertz distribution) as well as to scale-free networks - see Fig. 3. We found that except for the very dilute limit, the distance between most pairs of nodes is centered around to the typical distance (mean or mode), which is given by $\langle L \rangle \simeq \ln N / \ln \left(\langle K^2 \rangle / \langle K \rangle - 1\right)$. Also, when the 2nd moment of the degree distribution diverges (as in certain scale-free networks), the $N$-dependence enters in a more complicated way, which may lead to an “ultra-small” network [7], i.e. with a mean distance that scales like $\log \log N$ or $\log N / \log \log N$.

Recently, we have developed a methodology to study growing networks based on master equations. We successfully calculated analytically the DSPL for the Node Duplication
model, both undirected [8] and directed [9]. It turns out that although networks generated by this model are scale-free their mean distance scales like $\log N$, and therefore they are small-world networks, unlike generic scale free networks, that were shown to be ultra-small [7]. Moreover, the mean distance is even much longer than a corresponding configuration model with the same degree distribution.

The DSPL is only one member in a family of distributions of important metric properties - another one being the distribution of shortest cycle lengths (DSCL). Cycles play an important role in the study of critical phenomena on networks using high temperature expansions as well as in dynamical processes such as the first return of diffusive particles. We calculated the DSCL for the configuration model using the DSPL [10].

References

7. R. Cohen and S. Havlin, Scale-free networks are ultrasmall, PRL 90, 058701 (2003).
Logistic core–periphery detection in networks

Francesco Tudisco and Desmond J. Higham
University of Strathclyde, Department of Mathematics and Statistics, Glasgow (UK), f.tudisco@strath.ac.uk, d.j.higham@strath.ac.uk

1 Introduction

A key task in network science is to assign nodes to either the core or the periphery. Here, core nodes should be strongly connected across the whole network, whereas peripheral nodes should be strongly connected only to core nodes. The core–periphery concept emerged from the study of economic, social and scientific citation networks, and was formalized in a seminal paper of Borgatti and Everett [1]. A review of recent work on modeling, analyzing and detecting core–periphery structure can be found in [2].

In this work we summarize results from [5], where a new core–periphery detection algorithm is developed and tested, using techniques from nonlinear spectral theory.

2 Two core–periphery detection models

Let $G = (\{1, \ldots, n\}, E)$ be an undirected graph with adjacency matrix $A = (a_{ij})$. A core ranking assignment is a permutation vector $u$ which maps vertex $i$ to position $u_i$. The closer $u_i$ to $n$, the higher the rank of $i$ as a member of the core. We denote by $\text{CR}(n)$ the set of such vectors.

2.1 Functional core–periphery quality function optimization

For $\alpha \in \mathbb{R}$ and a core–ranking vector $x \in \text{CR}(n)$, the core–quality function

$$f_\alpha(x) = \sum_{i \leftrightarrow j \in E} a_{ij} \mu_\alpha(x_i, x_j), \quad \mu_\alpha(x, y) = \left(\frac{x^\alpha + y^\alpha}{2}\right)^{1/\alpha},$$

has been considered in [4, 5]. As thoroughly discussed there, in order to reveal the presence of a core–periphery structure, it is intuitively reasonable to look for the core ranking assignment $x \in \text{CR}(n)$ that maximizes $f_\alpha(x)$. However, $f_\alpha$ is neither convex nor concave, which makes the optimization of $f_\alpha$ particularly challenging.

2.2 Logistic core–periphery random model best fitting

We define the Logistic Core–Periphery Random Model (LCP) where, given $x \in \text{CR}(n)$, edges between nodes $i$ and $j$ are generated with independent probabilities given by

$$P(i \leftrightarrow j) = \sigma(\mu_\alpha(x_i, x_j)/n) = p_{ij}(x).$$

Here, $\sigma(x) = 1/(1 + e^{-10(x-1/2)})$ is a logistic sigmoid function, suitably rescaled to compensate for the fact that $0 \leq \mu(x_i, x_j)/n \leq 1$ when $x \in \text{CR}(n)$.
Note that for any $\alpha > 1$ the probability $p_{ij}(x)$ is large if at least one of the nodes $i$ and $j$ has a high core rank. Thus, the LCP can be used to artificially generate networks with a planted core–periphery structure but, unlike more classical block-based models, allows a smooth transition between the set of core nodes and the set of peripheral ones. Figure 1 shows example matrices with entries given by the probability $p_{ij}(x)$, for different choices of $\alpha > 1$. Moreover, for a given a network we may assign a core–ranking to the nodes by fitting the random model; that is, by maximizing the log-likelihood

$$L_\alpha(x) = \sum_{ij \in E} \log p_{ij}(x) + \sum_{ij \in \overline{E}} \log(1 - p_{ij}(x))$$

among all $x \in CR(n)$. Our main results show that a relaxed version of the problem has a unique solution that may be obtained from a computationally attractive iteration.

3 Main results

We first show that the quality function optimization and maximum likelihood approaches of §2.1 and §2.2 are equivalent.

**Theorem 1 ([5]).** For any $\alpha > 1$, a vector $x^* \in CR(n)$ is a solution of $\max_{x \in CR(n)} L_\alpha(x)$ if and only if it is a solution of $\max_{x \in CR(n)} f_\alpha(x)$.

Then, we show that if we relax the condition $x \in CR(n)$ and ask that $x$ has positive components with $\|x\|_p = 1$, then the quality function maximization approach can be recast in terms of the Perron eigenvector of the nonlinear operator defined entry-wise by

$$x \mapsto \Phi_\alpha(x) = n \sum_{j=1}^{n} a_{ij} |x_i|^{\alpha-2} x_i \mu_\alpha(x_i, x_j)^{1-\alpha}, \quad i = 1, \ldots, n.$$ 

**Theorem 2 ([5]).** Let $\alpha > 1$ and $p > \alpha$. Then the nonlinear eigenvalue equation $\Phi_\alpha(x) = \lambda x^{p-1}$ has a unique positive solution $u^* > 0$ such that $\|u^*\|_p = (|u_1^*|^p + \cdots + |u_n^*|^p)^{1/p} = 1$, which is also the unique solution of $\max_{x > 0, \|x\|_p = 1} f_\alpha(x)$.

Finally, the following result shows that the solution to the relaxed problem can be computed via a globally convergent Nonlinear Spectral Method (NSM).

**Corollary 1 ([5]).** Given an initial guess $u_0 > 0$ and $p > \alpha > 1$, $q = p/(p - 1)$, consider the following iterative method

$$
\begin{align*}
\{v_{k+1} &= \Phi_\alpha(u_k) \\
u_{k+1} &= \|v_{k+1}\|^{1-q} v_{k+1}|_{q}^{-2} v_{k+1}, & k = 0, 1, 2, 3, \ldots
\end{align*}
$$

Then $u_k > 0$ for all $k \geq 0$ and $\|u_k - u^*\| = O\left((\frac{p-1}{p})^k\right)$, i.e., $u_k$ converges to the unique solution $u^* > 0$ of Theorem 2.
### Core London Train Stations

We summarize some results from [5] on the London train transportation network. Figure 2 (left) shows the geographical location of the top 10 core train stations in the city of London identified by the new NSM, highlighting stations at key intersections that are not necessarily close to the city centre. Figure 2 (right) shows the names of the top ten core stations identified by (i) the adjacency matrix eigenvector (which is known to be related to rich-club structure [3]), (ii) the simulated–annealing method proposed in [4], and (iii) the NSM. The table also gives the overall number of passengers, in millions, entering or exiting each station, according to recent reports (Wikipedia, April 2018). We note that the top ten stations selected by NSM involve around 35% more passengers. Further validating experiments from [5] will presented at the conference.

### Acknowledgements

The work of F.T. is funded by the Marie Curie fellowship MAGNET no 744014. The work of D.J.H. is supported by grant EP/M00158X/1 from the EPSRC/RCUK Digital Economy Programme.

### References


---

**Fig. 2.** Left: Geographical location of top 10 core train stations in city of London identified by NSM. Right: Millions of passengers per years using the top 10 core stations identified by adjacency Perron eigenvector, simulated–annealing method and NSM.

Note that on sparse networks the method scales linearly with the number of nodes. In fact, each iteration requires $O(|E|)$ floating point operations, where $|E|$ is the number of edges in the graph. Moreover, the free parameter $p$ allows us to tune the overall number of iterations $k^* = O(\ln \varepsilon / \ln \frac{p-1}{p})$ required to achieve the precision $\|u_k - u^*\| = O(\varepsilon)$. 

4 Core London Train Stations

We summarize some results from [5] on the London train transportation network. Figure 2 (left) shows the geographical location of the top 10 core train stations in the city of London identified by the new NSM, highlighting stations at key intersections that are not necessarily close to the city centre. Figure 2 (right) shows the names of the top ten core stations identified by (i) the adjacency matrix eigenvector (which is known to be related to rich-club structure [3]), (ii) the simulated–annealing method proposed in [4], and (iii) the NSM. The table also gives the overall number of passengers, in millions, entering or exiting each station, according to recent reports (Wikipedia, April 2018). We note that the top ten stations selected by NSM involve around 35% more passengers. Further validating experiments from [5] will presented at the conference.

Acknowledgements. The work of F.T. is funded by the Marie Curie fellowship MAGNET no 744014. The work of D.J.H. is supported by grant EP/M00158X/1 from the EPSRC/RCUK Digital Economy Programme.

References

Network Constraints on the Mixing Patterns of Binary Node Metadata

Matteo Cinelli\textsuperscript{1}, Leto Peel\textsuperscript{2,3}, Antonio Iovanella\textsuperscript{1}, and Jean-Charles Delvenne\textsuperscript{2,4}

\textsuperscript{1} Department of Enterprise Engineering, University of Rome “Tor Vergata”, Via del Politecnico 1 00133 Rome, Italy
\textsuperscript{2} Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Universit Catholique de Louvain, Louvain-la-Neuve B-1348, Belgium
\textsuperscript{3} Namur Institute for Complex Systems (naXys), Universit de Namur, Namur B-5000, Belgium
\textsuperscript{4} Center for Operations Research and Econometrics (CORE), Universit Catholique de Louvain, Louvain-la-Neuve B-1348, Belgium

1 Introduction

Assortative mixing is the tendency of nodes with a common attribute (also referred to as node metadata) to be connected to each other in a network. It is typically measured using Newman’s assortativity coefficient \cite{2}, which is the network analogue of Pearson’s correlation. Just as correlation plays an important role in identifying relationships between pairs of variables, assortativity plays a fundamental role in understanding how a network is organised with respect to a given attribute of the nodes. The assortativity coefficient for a categorical node attribute measures the proportion of links in the network that connect nodes with the same attribute value relative to the expected proportion of links under a null model. Similar to Pearson’s correlation, assortativity is constrained to lie in the range \( r \in [-1, 1] \), however the structure of the network places further constraints that often make the maximum and minimum values unattainable. These constraints introduce issues when we try to interpret values of assortativity, as \( r \) is no longer the proportion of the maximum value relative to the expected value.

Here we consider the case of binary metadata, e.g. gender of actors in a social network, and we investigate how different constraints affect the range of permissible values of assortativity. These constraints are imposed by maintaining different characteristics of the network structure (specific degree sequence or specific graph topology) and node metadata (proportion of nodes per category or specific assignment of categories to nodes). Thus, by combining both types of constraints, we consider three different spaces of assortativity (omitting the fourth combination as it corresponds to a single configuration): the metadata-graph space (mgs), the graph space (gs) and the metadata space (ms). We demonstrate that these bounds can be substantially far from -1 and 1.

2 Reformulation and Bounds Applicability

The assortativity coefficient to categorical node attributes can be written as:

\[
    r = \frac{\sum_i e_{ii} - \sum_{i,j} e_{ik} e_{jk}}{1 - \sum_{i,j} e_{ik} e_{jk}} \tag{1}
\]
It reaches its minimum value \( r_{\min} \) when the number of links among nodes with the same attribute equals 0, i.e. when \( \sum e_{ii} = 0 \). In a graph with \( n \) nodes and a binary metadata assignment, we have \( n_0 \) nodes with metadata value \( c_i = 0 \) and \( n_1 \) nodes with metadata value \( c_i = 1 \) thus, \( n = n_1 + n_0 \). The binary metadata assignment implies the presence of three kinds of links in the network (1-1), (1-0) and (0-0). The three amounts of such links, whose sum equal the number of network links \( m \), can be labeled as \( m_{11}, m_{10} \) and \( m_{00} \). Therefore, assortativity can be re-written in terms of such edge counts:

\[
    r = \frac{(m_{11} + m_{00})m - (m_{11} + \frac{m_{10}}{2})^2 - (m_{00} + \frac{m_{10}}{2})^2}{1 - (m_{11} + \frac{m_{10}}{2})^2} \quad (2)
\]

In [1] it is established that, for a given degree sequence and the proportion of labeled nodes, such edge counts can be confined to a limited range. The bounds of this range are defined by partitioning the ordered degree sequence such that we maximise or minimise the edge counts. For example, to determine the upper bound of the number of edges \( m_{11} \) that connect pairs of nodes with metadata value \( c_i = 1 \), we should consider that the maximum value occurs when \( n_1 \) nodes are arranged into a complete subgraph or, if the degree sequence doesn’t allow such situation, when \( n_1 \) nodes with the highest degree only connect to each other and not to any nodes with metadata \( c_i = 0 \). Substituting these bounds into the expression of assortativity (Equation 2) provides bounds on the assortativity in the metadata-graph space (i.e. for the networks with a given degree sequence and \( n_1 \) value) and typically yields a range that is tighter than \(-1 < r < 1\). Since the bounds exploit specific partitions of the degree sequence, it is possible to extend them to the graph space by considering the observed metadata assignment as fixed. With respect to the metadata space (i.e. the ensemble of graphs with same topology and \( n_1 \) value), the bounds cannot be determined combinatorially in the same way that we define the bounds for the graph space and metadata-graph space. In the metadata space, the bounds of assortativity directly depend upon the specific network topology and set of all possible permutations of node metadata. Instead we must rely upon a complete enumeration of the \( \binom{n}{n_1} \) possible permutations (when feasible) or a heuristic algorithm to explore such metadata space.

3 Results

As an example, we investigate gender assortativity in the US college, Smith, extracted from the Facebook 100 dataset [3]. Smith displays \( r = 0.025 \) that is positive but close to 0, i.e. close to a random distribution of node metadata. In Figure 1 we compare this observed value of assortativity to the bounds we derived for the graph space and metadata-graph space. We immediately observe that the disassortative area of such network is severely bounded such that, given \( n_1 \) and the observed degree sequence, Smith could not, in practice, display very much disassortativity at all. We evaluate the assortativity range of Smith in the metadata space by computing the assortativity for random permutations of the metadata over the network nodes. This provides us with the empirical distribution of assortativity that we compare against the observed value. Given such empirical distribution, we exploit an algorithm of metadata swapping in order to obtain the boundaries of assortativity in the metadata space for which combinatorial bounds cannot be computed. The boundaries of the metadata space reveal, also in this case, a narrow range of assortativity into which evaluate the observed configuration. Considering the size of three ranges of assortativity we observe how such a measure is narrowly
constrained with respect to all the graph generated via link rewiring and to all the possible node metadata assignments obtained via label reshuffling, thus regardless of the null model that we consider more appropriate for the actual network. For what concerns the actual gender assortativity value displayed by Smith we can conclude that such a network is more assortative than random and that its value, if rescaled with respect to the upper bounds of (ms) and (gs), would provide evidence of assortative mixing.

Fig. 1. Histogram and assortativity landscape for the college Smith with \( n = 2625, m = 77259, n_{\text{females}} = n_0 = 2596, n_{\text{males}} = n_1 = 29 \). Smith has a gender assortativity value \( r = 0.02522 \) (solid line) that occurs in correspondence with \( m_{11} = 25 \) and \( m_{10} = 1404 \). The upper bound to the metadata-graph space is not represented in the Figure and its values is \( r_{\text{mgs}}^{\text{max}} = 1 \).

References

Part XVIII

Urban Systems and Networks
Insights from self-organizing maps for characterizing accessibility to healthcare networks

Jerome R. Mayaud1*, Sam Anderson2, Martino Tran1, Valentina Radic2

1School of Community and Regional Planning, University of British Columbia, 6333 Memorial Road, Vancouver, Canada *jerome.mayaud@gmail.com
2Department of Earth, Ocean and Atmospheric Sciences, 2020–2207 Main Mall, Vancouver, Canada

1 Introduction

As urban populations continue to grow worldwide, it becomes increasingly important to evaluate network accessibility (the ease with which residents can reach key places or opportunities). In an effort to become more liveable and sustainable, ‘smart cities’ are seeking to optimize accessibility to essential services such as healthcare, while ensuring equity is maintained amongst different population groups. For the first time, we apply a powerful machine learning (ML) tool, the self-organizing map (SOM), to cluster income characteristics and relate them to accessibility metrics of healthcare service networks across the rapidly growing, ‘smart’ City of Surrey, Canada. We perform this analysis for both 2016 and 2022, to examine the potential shifts in accessibility over time.

2 Methods

We use a SOM to reduce the dimensionality of our income datasets for 2016 and 2022 and draw out patterns in the underlying data to relate them to accessibility metrics. First, we used open-source data to evaluate accessibility across Surrey. We divided the surface area of the city into a grid of 1,480 equally sized hexagons (500 m diagonal diameter), following the method of [1]. We then calculated travel-time estimates (optimally combining public transportation and walking) between every pair of grid cells – an ‘origin’ and a ‘destination’ (O-D) – using OpenTripPlanner (OTP). The spatial layout of road networks and pedestrian infrastructure used in OTP was acquired from OpenStreetMap, and public transport routing was derived from geolocated timetables for a typical commuting day in September 2017.

We combined our O-D matrices with high-resolution (Dissemination Area) census population data from 2016 and population projections for 2022 [2]. The census data were spatially reorganized into the aforementioned hexagonal grid. Since census income data are reported in bins, we used MATLAB’s built-in ‘ksdensity’ method to perform a kernel density estimation (KDE) that provided us with a smoothed, normalized continuous probability distribution for income in each grid.
cell. These distributions were used as the input for our SOM algorithm. The population counts for each grid cell were used to perform catchment area analysis using the O-D matrices. Accessibility for each origin grid cell (for a total of \( n \) grid cells) was calculated as:

\[ F_o,T = \sum_{d=1}^{D} F_d f(t_{odr}), \]

where \( F_o,T \) is the number of facilities \( F \) that can be reached from origin \( o \) within time threshold \( T \) (30 mins in this study), \( F_d \) is the number of facilities in destination cell \( d \), and \( f(t_{odr}) \) is a time threshold function whose value (0 or 1) depends on whether \( t_{odr} \) is greater or smaller than \( T \). We also calculated the travel-time to the nearest facility for each grid cell.

The essential concept behind SOM analysis is to cluster observations onto a two-dimensional topology of nodes (patterns) presented in a regular ‘map’ [3]. We created a SOM through an iterative ‘training’ process that compares input data samples to each SOM pattern, and computes the minimum Euclidean distance to determine the closest match for each sample [4]. We quantified the magnitude of income distribution change between 2016 and 2022 by performing principal component analysis (PCA) to determine the first two modes of each SOM pattern distribution in both study years. A ‘cluster distance’ metric was then derived from the Euclidean distance in PC-space between the 2016 and 2022 income cluster patterns.

### 3 Results

Our method yielded an 8-cluster (4 rows x 2 columns) SOM (Fig. 1). We found that income homogenization is projected to occur over the study period: higher-income clusters (e.g. cluster 1 in red) will become more prevalent by ‘absorbing’ members from other clusters. This will result in decreased average accessibility for the absorbing clusters, and vice versa. By 2022, SOM cluster transitions and demographic shifts will result in \( \sim 12,000 \) more seniors residing in areas with no access to a hospital and \( \sim 1,500 \) with no access to a walk-in clinic.

![Fig. 1: Spatially mapped SOM topology, coloured according to clustering, for (a) 2016, and (b) 2022; (c) Frequency distributions for each SOM cluster, showing in bold the frequency that they occur in the 2016 and 2022 maps.](image)

Cluster distance analysis for both hospitals and walk-in clinics (Fig. 2) suggests that ageing populations (i.e. redder colours) will be concentrated in areas with relatively poor hospital and clinic accessibility, although income distributions in these...
areas will remain relatively stable. In contrast, median age will decrease (bluer colours) in areas experiencing large changes in income distribution and/or with access to numerous healthcare facilities.

![Fig. 2: Scatterplots showing relationships between Euclidean cluster distance and access to healthcare facilities (in terms of the number of facilities and travel-time to the nearest facility) for each cell. A third coloured variable shows the difference in median age between 2016 and 2022.](image)

4 Conclusion

Our results suggest that a dual accessibility problem may soon arise in Surrey. First, large senior populations will reside in areas with access to numerous, and close-by, clinics, which will put pressure on existing facilities for specialized services. Second, lower-income seniors will increasingly reside in areas poorly connected to healthcare service networks; these populations are likely to be reliant on public transportation, so accessibility equity may suffer. As fertility rates decline and life expectancies increase around the world, shifts to elderly, less mobile populations will pose major challenges for healthcare accessibility in ageing cities like Surrey.

Our study confirms there is much potential in applying SOM and other ML methods more widely in fields related to accessibility and urban networks. SOMs in particular help to uncover and visualise data patterns in ways that are communicable and relevant to policymakers.

References

Simple Spatial Scaling Rules behind Complex Cities

Ruiqi Li¹, H. Eugene Stanley²

¹ College of Information Science and Technology, Beijing University of Chemical Technology, North Third Ring Road 15, Chaoyang District, Beijing 100029, China
² Center for Polymer Studies and Physics Department, Boston University, Boston, MA 02215, U.S.A.

1 Introduction

Although most of wealth and innovation have been the result of human interaction and cooperation, we are not yet able to quantitatively predict the spatial distributions of three main elements of cities: population, roads, and socioeconomic interactions. By a simple model mainly based on spatial attraction and matching growth mechanisms, we reveal that the spatial scaling rules of these three elements are in a consistent framework, which allows us to use any single observation to infer the others. All numerical and theoretical results are consistent with empirical data from ten representative cities. In addition, our model can also provide a general explanation of the origins of the universal super- and sub-linear aggregate scaling laws and accurately predict kilometre-level socioeconomic activity. And the theoretical analysis method is original which is based on growth instead of mean-field assumptions. The active population (AP) concept proposed by us is another contribution, which is a mixture of residential and working populations according to the duration of their activities in the region. AP is a more appropriate proxy than simply residential population for estimating socioeconomic activities. The density distribution of AP is \( \rho(r) \propto r^{-\beta} \left( R_1^{1+\beta} - r^{1+\beta} \right)^{-r^{-\beta}} \) which can also reconcile the conflict between area-size allometry and the exponential decay of population from city centre to urban fringe found in the literature. Our work opens a new avenue for uncovering the evolution of cities in terms of the interplay among urban elements, and it has a broad range of applications.

2 Results

![Figure 1](https://example.com/figure1.png)

**Figure 1**: The spatial scaling results for two cities. Results for **a** London and **b** Beijing. The exponents of AP in London and Beijing are 1.70 and 1.91, respectively, which gives us \( \beta = 0.30 \) for London and 0.09 for Beijing. Then according to our theory, the exponents of road length and socioeconomic interactions captured by nighttime lights can be predicted to be \( 2 - \beta /2 = 1.85 \) (1.95) and \( 2 - 3 \beta /2 = 1.55 \) (1.86) for London (Beijing), which are very close to the empirical results -- by linear regression, the exponents and standard deviations for road length and socioeconomic interactions of London are 1.85 (sd 0.008) and 1.66 (sd 0.0118). For Beijing, they are 1.95 (sd 0.0164) and 1.74 (sd 0.0105), respectively.
References


On the Statistics of Urban Street Networks

Jérôme Benoit and Saif Eddin Jabari
New York University Abu Dhabi, Abu Dhabi, United Arab Emirates
jerome.benoit@nyu.edu

1 Introduction

We seek to understand the statistics of urban street networks. Such an understanding will improve urban policies in general and urban transportation in particular. In our work here we investigate urban street networks as a whole within the frameworks of information physics [9] and statistical physics [7].

Although the number of times that a natural road crosses another one has been widely observed to follow a discrete Pareto probability distribution [3] among self-organized cities [1, 4, 8], very few efforts have focused on deriving the statistics of urban street networks from fundamental principles. Here a natural road (or road) denotes an accepted substitute for a “named” street [8].

Our approach explicitly emphasizes the road-junction hierarchy of the initial urban street network rather than implicitly splitting it accordingly into two dual but distinct networks. Most of the investigations indeed seek to cast the initial urban street network into a road-road network [8] and to describe its valence probability distribution.

This holistic viewpoint adopted by the urban community [1, 2] also fits with the mindset of information physics [9], which is built upon partial order relations [5, 9]. Here the partial order relation derives from the road-junction incidence relation. Applying then information physics enables us to envisage urban street networks as evolving social systems subject to an entropic equilibrium similar to the Paretian one effectively observed among cities of a same country [6].

2 Method

The relation that ties natural roads and junctions is bijectively reduced into an algebraic structure known as Galois lattice [5]. Then, by imposing natural consistency constraints, information physics [9] enables us not only to evaluate urban street networks but also to assess a probability distribution and, ultimately, an information measure. It appears that urban street networks effortlessly reduce to Galois lattices with two nontrivial layers: the natural roads form the lower layer and the junctions form the upper one, while the partial ordering relation is “passing through.” This causes urban street networks to become a toy model for the emerging paradigm.

The passage from Galoisean hierarchy to Paretian coherence is then achieved by invoking Jaynes’s Maximum Entropy principle [7] with the first logarithmic moment as the sole characterizing constraint and our complete ignorance as initial knowledge. The corresponding most plausible probability distribution expresses for each natural road...
or junction, indiscriminately, the likeliness of possessing a given number of equally likely states; it is a discrete Pareto probability distribution whose entropy is the imposed first logarithmic moment, as desired. This probability distribution must ultimately be decomposed with respect to the structure of the Galois lattice and the algebraic rules imposed by information physics theory. In other words, a physical meaning remains to be given to the evaluation.

Hypothesizing a crude asymptotic binomial paired-agent model with the spirit of the city model [6] allows us finally to predict the statistics of urban street networks. Here each natural road or junction is envisioned as an intranetwork whose very survival relies on the ability of each of its agent to preserve a crucial number of intraconnections. Thereby each urban street network becomes characterized by two generalized binomial combination numbers which asymptotically rise up as two characterizing exponents beside the Paretian exponent \( \lambda \): the numbers of vital connections for natural roads and for junctions, respectively \( v_r \) and \( v_j \).

Fig. 1. Relative Frequency Distributions (RFD) for the urban street network of London: circles represent relative frequencies for the valences of the road-road topological network; crosses represent relative frequencies for the valences of the junction-junction topological network. The red fitted curve for the natural road statistics describes the Maximum Likelihood Estimate (MLE) for the discrete Pareto probability distribution (1a) estimated according to the state of the art [3] \( (n_r = 4, 2\lambda v_r = 2.610(0.065), n = 5000 \text{ samples, } p\text{-value} = 0.929) \). The green fitted curve for the junction statistics shows the best Nonlinear Least-Squares Fitting (NLSF) for the nonstandard discrete probability distribution (1b) with \( n_r \) and \( 2\lambda v_r \) fixed to their respective MLE value \( (2\lambda v_j = -1.3) \); since fast evaluation of the normalizing function \( W \) has yet to be found, no MLE approach can be used for now. Having a number of vital connections \( v_j \) negative means that the associated generalized binomial combination number is smaller than one, i.e., that the number of agent intraconnections for junctions is relatively much smaller than the one for natural roads.
3 Results and Discussion

Our approach recovers the discrete Pareto probability distribution widely observed for natural roads evolving in self-organized cities, and foresees a nonstandard bell-shaped distribution with a Paretian tail for their joining junctions. The probability for a natural road to cross \( n_r \) natural roads is

\[
\Pr(n_r) = \frac{n_r^{-2\lambda \upsilon_r}}{\zeta(2\lambda \upsilon_r; \gamma_j)}
\]

where \( \zeta(\alpha, \beta) = \sum_{m=0}^{\infty} m^{-\alpha} \) is the generalized zeta function, and the probability for a junction to see \( n_j \) junctions through its joining natural roads reads

\[
\Pr(n_j) = \frac{\sum_{r=0}^{n_j} \beta_j \lambda_j \gamma_j \upsilon_j}{\lambda_j \upsilon_j \gamma_j \upsilon_j} \frac{(n_j - n)^{2\lambda \upsilon_j}}{}
\]

where \( \beta(\alpha, \beta, \gamma) = \sum_{m,n} m^{-\alpha} n^{-\beta} (m + n)^{-\gamma} \) is the two-dimensional generalized Mordell-Tornheim-Witten zeta function; the number of junctions per natural road \( n_r \) is assumed to span from some minimal value \( \eta_j \) for practical reasons [3].

Figure 1 exhibits the urban street network of London as a case study. The probability distribution for natural roads \( \Pr(n_r) \) (1a) is highly plausible, as expected for any recognized self-organized city [1, 8]. The validation of the probability distribution for junctions \( \Pr(n_j) \) (1b) appears more delicate for the time being. Meanwhile a crude data analysis is not conclusive enough. Interestingly, this case study reveals that the number of intraconnections for junctions might be relatively much smaller than the one for natural roads in self-organized cities.

Thus the statistical model for urban street networks (1) appears fine enough to study urban macro behaviours with the exponents \( \lambda, \upsilon_j, \) and \( \upsilon_j \) as complexity parameters. Future work includes (i) finding patterns via the ratio \( \upsilon_j / \upsilon_j \) among self-organized cities, (ii) extending the model to designed cities, (iii) applying the paradigm to more intricate systems, and (iv) full investigation of the resulting Paretian statistical physics.

References

Decentralized Optimal Route Navigation

Trivik Verma1,2, Marc Folini1 Yi Wang2, Hans R. Heinimann2, and Nuno A. M. Araújo3

1 ETH Zurich, Environmental Systems Science, Universitätstrasse 16, CH-8092 Zurich
   vtrivik@ethz.ch.
2 ETH Zurich, Future Resilient Systems, Singapore-ETH Centre, 138602, Singapore
3 Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, 1749-016, Portugal

1 Introduction

Urban traffic planning is based on two broad assignment principles, user equilibrium (UE) and system optimum (SO) [12]. In the former, in the resulting system state, no user benefits from changing their path and in the latter, the whole system yields a theoretical optimum performance at the expense of some users who are not taking their available shortest paths. The two types of assignments (UE and SO) are used to replicate the route choice behaviour of travellers, rather than providing routing guidance. Hence, they are mostly adopted as the lower level models for planning, not navigation.

Both models require a lot of information. The UE solution is likely to be achieved in the situation where traffic information is perfect (shortest path is the true one) and users choose the shortest path. In particular, one needs to know where the users originate their trips, where they want to go, compute user equilibrium and convey back to the users; in real time for a navigation system, and statically, for future urban planning. Such navigation systems would help with the first condition of suggesting a shortest path but have no control over what the users do in reality. The SO solution may only be achieved with specific tolling implemented in order to alter the cost some users pay for their shortest paths for the good of the system. Together, these approaches provide a basis for centralized planning and routing schemes.

Even if we assume to have all the necessary information a priori, unpredictable user behaviour or events like rain, accidents, or any other disturbance, could disrupt the equilibrium or optimum conditions [4]. Istantaneous duress coupled with the use of GPS based apps, have given rise to cut-through traffic [2]. Long-term trend in rising demand has also added a lot of stress on transportation networks. As a result, new congestion patterns are emerging [10]. Despite claims from companies that promote Mobility as a Service [7], they do not provide solutions to traffic congestion. As proved in economic theory, these algorithms do not converge to solutions that promote system optimality [9]. The degree of real-time communication and the infrastructural resources required for relaying all the computed information to all users in the system and incentivising them to change their paths, are both huge overheads.

We hypothesise that the situation wherein all the users choose a path that deviates exponentially from the shortest path in an equilibrium setting with their degree of altruism, diverts the selfish user equilibrium towards altruistic system optimum without imposing infrastructural costs. In this work, we study the favouring (and counter-intuitive)
effect of uncertainty of supply, demand and user behaviour on system efficiency and elucidate how the system can push towards global optimum behaviour with minimal information.

2 Results

Let us consider a real road network and two points, origin $A$ and destination $B$ as shown in Fig. 1(a). Under user equilibrium, there is one (or multiple paths with the same cost) shortest path(s) on which all users are assigned such that no user gains from switching to another path. We compute the $k$-shortest paths ($k$ paths in decreasing order of their free-flow time) with limited overlap [3] between $A$ and $B$ and assign every user on an alternative among the $k$-shortest paths using a beta-PERT distributed variable $\Xi$ which is a continuous distribution function defined by the minimum, most likely and maximum values and transformed on a standard beta distribution [5]. We use it to model the riders’ degree of altruism,

$$P(\Xi < \xi) \approx f(\lambda, \xi_{\text{mode}}),$$

where $\lambda$ defines the shape of the distribution and $\xi_{\text{mode}}$ is the most likely choice of path assigned in the distribution. The selection of paths is devoid of any centralised optimisation (such as in the standard Frank Wolfe formulation [6] used for solving the Nash Equilibrium of flow [8] and System Optimal [11]). Our method solves traffic flow based on this path assignment and plot the total travel time against $\xi$ (for fixed $k$).

We quantify the efficiency of decentralized uncertainty in the road networks using the level of altruism as a means of social behaviour,

$$\eta = \frac{<t_{ge}> - <t_{ue}>}{<t_{so}> - <t_{ue}>},$$

Fig. 1. (a) $k$-shortest paths: We model the system with the city of Zurich and run our algorithm for $k = 5$ as shown in the figure. (b) pert-BETA Distribution: The distribution shows the variation of the cumulative probability density function for different values of $\lambda$ (shape). As $\lambda$ increases, the distribution becomes less uniform and peaks toward the most likely path choice of 1 in this case. The vertical line is the mean $E(\Xi)$. 

The $7^{th}$ International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
where $t_{\xi,ue,so}$ are system-wide measures for travel time, average commuting time, degree of congestion, etc. for our decentralized altruism scheme ($\xi$), user equilibrium ($ue$) and system optimum ($so$) formulations, respectively.

The different variations of user altruism are modelled by the pert-BETA distribution and illustrated in Fig. 1(b). Lower values of $\lambda$ associate with a uniformly distributed user set on the $k$-shortest paths, while a larger value spreads the users unequally over the paths and closer to the mode of the distribution (a likely path that most users would go for as opposed to the shortest path from $A$ to $B$). Notice in Fig. 2(a), how lower values of $\lambda$ account for randomness in the system where users are taking random paths and there is no effective savings in cumulative time for the system as a whole. However, as the shape of the distribution increases, the system time of travel becomes lesser and the savings in time are greater, especially for larger systems (more possibilities for congestion).

![Fig. 2](image_url)

**Fig. 2.** (a) System-time vs Altruism for different shapes: The cumulative system time savings are more prominent with an increase in the shape parameter ($\lambda$) and the maximum savings occur at $\xi \approx 2$ (longer than the shortest path available). The inset of the figure shows similar behaviour in savings and a completely random scenario at $\lambda = 0$ (uniformly distributed on all paths). (b) System-time vs Altruism for different trip sizes: Larger systems exhibit larger time savings and the inset shows a linear increase in system time for smaller sizes.

Interestingly, Fig. 2(b) shows that as the system size increases and more users go from $A$ to $B$, the savings in cumulative travel time becomes prominent after a certain number of users enter the system and more so for larger system sizes, especially for the most likely choice of path much larger than the shortest path ($\xi = 2$). In fact, for a smaller system size, there is negative savings in time which can be taken for a lack of congestion in the system, and hence longer paths means longer times of travel.

### 3 Dataset

We use real city street drivable networks extracted from Open Street Maps using the OSMnx package in python [1]. The set forms the basis for solving standard methods used in traffic research ($ue$ and $so$) on a set of randomly selected node pairs $A$ and $B$. 

---

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
We acknowledge financial support from the Portuguese Foundation for Science and Technology (FCT) under Contract no. UID/FIS/00618/2013 and the National Research Foundation of Singapore (NRF) under its Campus for Research Excellence and Technological Enterprise (CREATE) programme (FI 370074011) for the Future Resilient Systems project at the Singapore-ETH Centre (SEC).

References

7. Geha, J.: Fremont: City takes steps to keep commuters on freeways and off its streets. East Bay Times (nov 2016)
10. Rodriguez, J.F.: 45,000 Uber and Lyft drivers may now operate in SF. SF Examiner (2016)
Application of Complex Network Analysis for Urban Drainage Systems

Julian David Reyes-Silva¹, Jonatan Zischg², Robert Sitzenfrei², and Peter Krebs¹

¹ Technische Universität Dresden, Institute of Urban Water Management, 01062 Dresden, Germany, julian_david.reyes_silva@tu-dresden.de
² Unit of Environmental Engineering, Department for Infrastructure, University of Innsbruck, 6020 Innsbruck, Austria

1 Introduction

Over the past years, there has been an increasing trend to apply graph theory for analyzing various urban infrastructure systems in order to analyze their structural characteristics, possible similarities, spatial and/or temporal topological patterns and scaling properties. Nevertheless, applications on urban water networks, i.e. drinking water supply and drainage, have been limited, due to lack or difficulty of accessing real information. Studies in this field are focused mainly on the evolution and common topological properties of such systems [1][2]. In this context, the present study aims for the characterization of structural and functional properties of urban drainage networks (UDNs) in terms of topological metrics. Since the main objective of these systems is to collect and transport sanitary sewage (e.g. from households) and stormwater runoff to a fixed outlet, e.g. a wastewater treatment plant, they are considered and analyzed as directed acyclic networks. The main goal of this study is to assess the advantages, limitations and insights on the UDN’s structure and function that can be derived from the application of complex network analyses. In order to do this, graph metrics are evaluated as potential surrogate measures of hydraulic simulations.

2 Methods

Due to the lack of information regarding real UDNs, virtual cases are used. These artificial systems are obtained based on an enhanced version of the stochastic urban drainage systems generator by Möderl et al. [3], in which the layout and size of the networks are determined based on an adapted Galton-Watson branching process, while design is done according to standard procedures. Validation of the tool was carried based on 2 real case studies [3]. Added features are a routine to include and quantify meshness in the system, and an automatic calculation of several topological metrics. Meshness is calculated here based on the average total degree (ATD) of inner nodes (k > 1), i.e. excluding source and outlet nodes. Under the assumption that in UDNs a node can have maximum 4 connections, it was determined that completely meshed systems will have an ATD close to 4, i.e. 100% meshness, while in branched networks it will be around 2.5, i.e. 0%, following a linear relationship, with slope = 61.71 and intercept = –146.9.
Determining the meshness degree of a real UDN can be challenging and furthermore it can be a key parameter to characterize the systems structural vulnerability, see Zhang et al.[4] for an example.

Usage of virtual cases has the advantage of controlling physical factors that determine the layout and dimensions of UDNs (e.g. terrain slope or percentage of impervious areas), thus uncertainties and/or influences of such factors on topological properties can be more easily analyzed. Furthermore, extreme cases such as completely branched (resembling river networks) or fully meshed (in a grid-like pattern) systems, for which real examples are difficult to find, can be modelled and analyzed. An additional advantage of the selected tool is its option of generating input files for hydrologic-hydraulic simulations in EPA Stormwater Management Model (SWMM) [5]. This allows to validate outcomes from the complex network analysis in terms of actual behavior of the drainage network. Nevertheless, analyses on the specific location of network failures, e.g. node flooding, cannot be done with the proposed tool due to its simplifications when generating such systems.

Initial results were obtained from a total of 300 virtual networks of different sizes (from 153 and 1278 nodes) and different structural configurations, covering the entire range of meshness. In order to evaluate functional properties, the daily mean wastewater flow (Qdw) for each pipe was calculated, based on results from hydrodynamic simulations. For every virtual system, assessment of the relationship between topological metrics and network function was done based on the Spearman correlation between Qdw and a specific local measure, e.g. directed Edge Betweenness Centrality (EBC). In order to evaluate the validity of these results, similar analyses were performed for 8 real UDNs, subnetworks from a german city, which size and degree of meshness varied within the ranges of the virtual cases (between 327 and 1166 nodes and from 37 to 52 % meshness). Information about calibration and validation of the SWMM models corresponding to these real networks can be found in Kaeseberg et al. [6].

3 Results

Although several metrics were analyzed for each network, the main results are:

1. Similarity between the frequency distributions of Shortest Path Lengths (SPL) and of distances along the network to the outlet. Previous studies have suggested that travel time distribution, i.e. the hydrological response of the network, can be derived from the frequency distribution of distances to the outlet [7]. In this context, analysis of the SPL histogram could be used to assess the response of the network, with advantage that only connectivity information is needed, and furthermore it requires considerably less computational time. Nonetheless, this has yet to be tested. An example of spatial distribution of SPL in a virtual UDN, frequency distributions of SPL and distances to the outlet, and travel time distribution, obtained from SWMM simulations, can be seen in Fig 1. SPL is measured here as the amount of pipe sections in between a node and the outlet of the system.

2. Correlation between EBC and Qdw decreases with higher Meshness, see Fig 2. Identification of crucial pipe sections in terms of flow quantity based only on EBC can yield appropriate results mainly for networks with low meshness, i.e. branched
systems, due to the lack of multiple pathways connecting pairs of nodes. Similar results were obtained for the real UDNs (orange diamonds in Fig 2), leading to the conclusion that, despite its simplifications, the selected tool can be used to analyze the relationship between topological metrics and network function. However, additional research on topological metrics that account for the different pathways of flow in meshed networks needs to be conducted.

Fig. 1.

a) SPL in a fully branched virtual UDN, outlet is red triangle. b) Histogram of SPL. c) Distribution of distances to outlet. d) Travel time distribution.

Fig. 2.

Correlation Coefficient ($\rho$) between EBC and Qdw as a function of Meshness for virtual (blue dots) and real (orange diamonds) UDNs.
References

Mapping the discourse on Smart Cities by means of NLP (Natural Language Processing) and complex network analysis

E. Blumenfeld-Lieberthal1, N. Serok1, O. Levy2, E.L. Milner3, S. Havlin2. (1) Tel Aviv University, Ramat Aviv, Tel Aviv, Israel, efratl@post.tau.ac.il (2) Bar Ilan University, Ramat Gan, Israel, (3) Ben Gurion University, Be’er Sheva, Israel

Introduction

Since 2008, more than half of the world’s population lives in cities, while in the western world this percentage is considerably higher (above 80% in most of the Americas and Europe). At the same time, new technologies allow collecting data from the bottom up customers and consumers and disseminating it back, using advanced technologies. The conjunction between this rapid urbanism and technological advances gave rise to a new domain in the academic and technological world known as ”Smart Cities” (SC). The current work on SC is extensive and carried in many disciplines, while different studies focus on different aspects of this term. These range from studies that regard the SC as a technology-oriented city [1] through others that emphasize green, sustainable properties of cities, and to works that stress the importance of human capital and high education [2]. The multi-disciplinary discourse is changing in time and evolving in several directions, therefore, it is difficult to understand the general trends of the discourse and its dynamics. Thus, the aim of this work is to map the discourses on Smart Cities in terms of topics, dynamics, and interdisciplinary trends, by means of network analysis.

Method

We developed a new framework, which integrates two theoretical bodies: natural language processing (NLP) and complex networks theory. We identified the research fields that study smart cities and extracted 639 papers, published in the top-ranked journals, in the following research fields: Technology, Urban Studies and Planning, Political Science, Law, Economics, and building and construction.

We analyzed these papers by means of Open Calais™ an NLP software that is general enough on the one hand (to identify topics from various fields) and on the other hand, accurate in extracting metadata from unstructured text. To construct the networks, we used a co-word methodology where the nodes represent the topics, the links represent the co-occurrence of two topics in the same paper, and their weight represents the number of
time two topics appeared in different texts. To study the dynamics of the discourse we divided the studies period, based on the distribution of the number of papers published per year, into two sub-periods (figure 1) and constructed a different network for each of them.

![Figure 1: Number of papers on SC at different years. After 2012, there is a significant growth in the number of published papers on SC, and in 2016, this number presents a major jump. The two periods of the discourse are 2012-2015 and 2016-2017 (223, and 416 papers correspondingly)](image)

We used the Louvain algorithm [3] to identify communities in the networks, and calculated the strength between the significant communities (with a threshold of the number of nodes = 40). Next, we performed a K-shell decomposition [4], then analyzed and compared the K-core of each network.

**Results**

We found that while the number of papers increased over time by 86%, the number of significant communities decreased from 8 to 6 (figure 2) and the strength between the communities doubled itself. The intra-strength (between nodes in the same community), however, increased only by 167%. This suggests that the discourse on SC became more focused (fewer communities that are better connected to one another). In terms of the discussed topics – while at the first period there were separated communities for Urban Planning, Sustainability, and Politics – the three were united into one community in the second period, and new communities that relate to System Theory, Emerging Technologies, and Logic emerged. The topics that constructed the Transportation community in the first period were disseminated into different communities in the second one, which sug-
gests this topic became more relevant and thus, was discussed in different fields. Additionally, the topics found in the cores of the networks (normalized by their weight and the number of papers in each period) suggest that issues concerning data, big-data, transportation, and technologies became more dominant in the discourse, while other issues, concerning human rights, social justice, and quality of life were pushed out from the discourse core. To summarize, the discourse on SC is becoming more focused on technologies and the use of available data while the effect of these technologies and the used of data on our lives occupy a significantly smaller volume of the discourse. Based on our analysis, human equity and rights are absent from the discourse or remain within its boundaries.

References
Decongestion of urban areas with hotspot-pricing

Albert Solé-Ribalta\textsuperscript{1,2}, Sergio Gómez\textsuperscript{2}, and Alex Arenas\textsuperscript{2}

\textsuperscript{1} Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Spain
\textsuperscript{2} Dept. Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Spain

Urban life is characterized by a huge mobility, mainly motorized. Amidst the complex urban management problems there is a prevalent one: traffic congestion. US, Canada and most of the European countries are in the top 15. The amount of hours wasted in congestion is significative ranging from 14 to 50 hours per year and with maximums up to 100 hours in Los Angeles. The corresponding socio-economic and environmental impacts are evident. An striking, as well as controversial, strategy to overcome the problem is congestion pricing. It consists in taxing vehicles for accessing a road/area, at certain times. Since the supply quantity is fixed (no more lanes or roads are usually added to the transportation network) the access to demanded areas is taxed. Assuming that congestion is an inevitable consequence of urban motorized areas, the challenge is to develop strategies towards a sustainable congestion regime at which delays and pollution are under control. Here, we rely on the Microscopic Congestion Model [3, 4] (MCM) to identify urban traffic hotspots in real scenarios and devise a mechanism to palliate its congestion.

The MCM describes the full state of system for any amount of congested junctions and it is based on assuming that the growth of vehicles observed in each congested node of the network is constant, which corresponds to the stationary state. This assumption allows us to describe, with a set of balance equations (one for each node), the increment of vehicles at the junction queues’. Mathematically, the increment of the vehicles per unit time at every junction \(i\) of the city, \(\Delta q_i\), satisfies:

\[
\Delta q_i = g_i + \sigma_i - d_i, 
\]

where \(g_i\) is the average number of vehicles entering junction \(i\) from the area surrounding \(i\), \(\sigma_i\) is the average number of vehicles that arrive to junction \(i\) from the adjacent links of that junction, and \(d_i\) corresponds to the average of vehicles that actually finish in junction \(i\) or traverse towards other junctions. Clearly, the equations in (1) are not independent since the vehicles that go out from a junction are the incoming ones into the next towards the destination. Specifically, the system is coupled through the incoming flux variables \(\sigma_i\), that can be expressed as

\[
\sigma_i = \sum_{j=1}^{S} P_{ji} p_j d_j, 
\]

where \(P_{ji}\) accounts for the routing strategy of the vehicles (probability of going from \(j\) to \(i\)), \(p_j\) stands for the probability of traversing junction \(j\) but not finishing the path at \(j\) (i.e. \(1 - p_j\) is the probability for a vehicle that reach \(j\) that \(j\) is his destination) and \(S\) is the number of nodes in the network. Probabilities \(P\) and \(p\), can be obtained in terms of node and edge betweenness for any routing strategy. See [3] for more details.
The taxing mechanism is an scheme that charges directly vehicles crossing congested spots (junctions) considering the overall topological structure and traffic functionality of the network. The aim is to eliminate the congestion hotspots using a network topology pay-per-use scheme. Specifically, we build up a flow model based on two steps: (1) detection of the hotspots using the Microscopic Congestion Model, and (2) prediction of the required tax to be applied to every congested junction to encourage drivers to divert the excess flow to neighboring and less congested regions. To estimate the required tax for each junction, we use the economic concept of elasticity [2]. The elasticity measures the response of the demand of a good in terms of an increase of its price and it is formally obtained as the ratio between the relative increase of the demand of a good and the relative increase of its price. Specifically, the predicted fraction of flow of vehicles after a tax $c$ is applied (i.e. those vehicles that decide to pay instead of diverting their paths) is given by:

$$
\phi = \phi_0 \left( \frac{c}{c_0} \right)^\mu \tag{3}
$$

where $\phi_0$ is the observed fraction of flow after applying the tax $c_0$, and $\mu$ is the elasticity value (see [1] for a detailed description of the technological implementation of taxes).

The overall taxing approach we propose follows a similar idea to the one proposed by Vickrey back in 1963, with the main difference that we now can analytically predict the model behaviour considering real data. Results show that our approach applied on real cities is able to reduce congestion and transit times further than the state-of-the-art, see fig. 1. We refer the reader to [5] for a complete description of the work.

References

Fig. 1. (A) Distribution of the congestion after applying the different taxing schemes in the city of Milan. The value of the congestion is given by the accumulation rate of vehicles, in vehicles per minute, for the different congested junctions. The distributions are computed considering all data from Monday to Friday and are shown grouped by hour of the day. (B) Comparison of travel times (in minutes) between no taxing, cordon pricing and hotspot pricing, with fixed total income, for the city of Milan. Below, maps of Milan showing the expected ratios between incoming and outgoing vehicles of each junction after the establishment of the cordon pricing tax (C) and the hotspot pricing scheme (D). Junctions with a ratio greater than 1 are congested since they receive more cars that the ones they can route.
# Organization and Committees

## General Chairs
- Hocine Cherifi, University of Burgundy, France
- Renaud Lambiotte, University of Oxford, UK
- Pietro Lió, University of Cambridge, UK

## Advisory Board
- Jon Crowcroft, University of Cambridge, UK
- Raissa D’Souza, Univ. of California, Davis, USA
- Eugene Stanley, Boston University, USA
- Ben Y. Zhao, University of Chicago, USA

## Program Co-Chairs
- Luca Maria Aiello, Nokia-Bell Labs, UK
- Chantal Cherifi, University of Lyon 2, France
- Luis Mateus Rocha, Indiana University, USA

## Poster Chairs
- Elsa Arcaute, UCL, UK
- Márton Karsai, ENS/INRIA Lyon, France
- Jat Singh, University of Cambridge, UK

## Lightning Chairs
- Sebastian Anhert, University of Cambridge, UK
- Jean-Charles Delvenne, UC Louvain, Belgium
- Huijuan Wang, TU Delft, Netherlands

## Media and Publicity Chairs
- Bruno Gonçalves, New York University, USA
- Feng Xia, Dalian University of Technology, China
- Carlos Gershenson, UNA de Mexico, Mexico
- Leto Peel, UC Louvain, Belgium

## Tutorial Chairs
- Jinhu Lü, Chinese Academy of Sciences, China
- Sarah Morgan, University of Cambridge, UK
- Eiko Yoneki, University of Cambridge, UK

## Local Chairs
- Jat Singh, University of Cambridge, UK
- Petra Vertes, University of Cambridge, UK

## Local Committee
- William Grant, University of Cambridge, UK
- Giovanna Maria Dimitri, Univ. of Cambridge, UK
- Sandra Servia Rodriguez, Univ. of Cambridge, UK
- Helena Andres, University of Cambridge, UK

## Publication Chair
- Sabrina Gaito, University of Milan, Italy

## Submission Chair
- Christian Quadri, University of Milan, Italy

## Web Chair
- Matteo Zignani, University of Milan, Italy

---

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Program Committee

Filip Agneessens University of Surrey, UK
Nesreen Ahmed Intel Labs, USA
Sebastian Ahnert University of Cambridge, UK
Masaki Aida Tokyo Metropolitan University, Japan
Marco Aiello University of Stuttgart, Germany
Luca Maria Aiello Nokia Bell Labs, UK
Deepak Ajwani Nokia Bell Labs, UK
Tatsuya Akutsu Kyoto University, Japan
Reka Albert The Pennsylvania State University, USA
Antoine Allard Laval University, Canada
Claudio Altafini Linköping University, Sweden
Lucila Alvarez-Zuzek IFIMAR-UNMdP, Argentina
Fred Amblard IRIT - University Toulouse 1 Capitole, France
Marco Tulio Angulo National Autonomous Univ. of Mexico, Mexico
Alberto Antonioni Carlos III University of Madrid, Spain
Nino Antulov-Fantulin ETH Zurich, Switzerland
Nuno Araujo Universidade de Lisboa, Portugal
Elsa Arcaute University College London, UK
Alex Arenas URV, Spain
Panos Argyrakis Aristotle University of Thessaloniki, Greece
Tomaso Aste University College London, UK
Martin Atzmueller Tilburg University, Netherlands
Rodolfo Baggio Bocconi University, Italy
James Bagrow University of Vermont, USA
Sven Banisch Max Planck Institute, Germany
Yaneer Bar-Yam New England Complex Systems Institute, USA
Paolo Barucca University of Zurich, Switzerland
Anastasia Baryshnikova Calico Life Sciences, USA
Nikita Basov St. Petersburg State University, Russia
Gareth Baxter University of Aveiro, Portugal
Mariano Beguerisse Diaz Spotify Limited, Sweden
Rosa M. Benito Universidad Politécnica de Madrid, Spain
Jacob Biamonte University of Malta, Malta
Ginestra Bianconi Queen Mary University of London, UK
Jeremy Blackburn University of Alabama at Birmingham, USA
Anthony Bonato Ryerson University, Canada
Javier Borge-Holthoefer Internet Interdisciplinary Institute, Spain
Pierre Borgnat ENS de Lyon, France
Stefan Bornholdt Universität Bremen, Germany
Dan Braha NECSI, USA
Ulrik Brandes ETH Zurich, Switzerland
Lidia A. Braunstein UNMdP, Argentina
Markus Brede University of Southampton, UK
Marco Bressan Sapienza University of Rome, Italy

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Dirk Brockmann  Humboldt University, Germany
Piotr Bródka  Wroclaw Univ. of Science and Technology, Poland
Raffaella Burioni  Università di Parma, Italy
Kanat Camlibel  University of Groningen, Germany
Paolo Campana  University of Cambridge, UK
Carlo V. Cannistraci  TU Dresden, Germany
Vincenza Carchiolo  Università di Catania, Italy
Alessio Cardillo  University of Bristol, UK
Rui Carvalho  Durham University, UK
Giona Casiraghi  ETH Zurich, Switzerland
Ciro Cattuto  ISI Foundation, Italy
Remy Cazabet  University of Lyon, France
L. Elisa Celis  EPFL, Switzerland
Tanmoy Chakraborty  IIIT-Delhi, India
David Chavalarias  CNRS, France
Kwang-Cheng Chen  University of South Florida, USA
Hocine Cherifi  University of Burgundy, France
Chantal Cherifi  University of Lyon, France
Peter Chin  Boston University, USA
Fu Lai Chung  Hong Kong Polytechnic University, Hong Kong
Matteo Cinelli  University of Rome Tor Vergata, Italy
Richard Clegg  Queen Mary University of London, UK
Giacomo Como  Politecnico di Torino, Italy
Luciano Costa  Universidade de SaPaulo, Brazil
Emanuele Cozzo  BiFi, Spain
Regino Criado  Universidad Rey Juan Carlos, Spain
Matthieu Cristelli  CNR, Italy
Mihai Cucuringu  University of California, Los Angeles, USA
Bhaskar Dasgupta  University of Illinois at Chicago, USA
Joern Davidsen  University of Calgary, Canada
Fabrizio De Vico Fallani  Inria - ICM, France
Mathias Dehmer  UMIT, Austria
Charo I. Del Genio  The University of Warwick, UK
Michela Del Vicario  IMT Lucca, Italy
Pietro Delellis  University of Naples Federico II, Italy
Jean-Charles Delvenne  University of Louvain, Belgium
José Devezas  INESC TEC and DEI-FEUP, Portugal
Matías A. Di Muro  IFIMAR-CONICET, Argentina
Wenbo Du  Beihang University, China
Louis J Dubé  Laval University, Canada
Jordi Duch  Universitat Rovira i Virgili, Spain
Marten During  C2DH, Luxembourg
Mohammed El Hassouni  Mohammed V University in Rabat, Morocco
Michael T.M. Emmerich  Leiden University, Netherlands
Frank Emmert-Streib  Tampere University of Technology, Finland

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gunes Ercal</td>
<td>SIUE, USA</td>
</tr>
<tr>
<td>Pau Erola</td>
<td>University of Bristol, UK</td>
</tr>
<tr>
<td>Ernesto Estrada</td>
<td>University of Strathclyde, UK</td>
</tr>
<tr>
<td>Tim Evans</td>
<td>Imperial College London, UK</td>
</tr>
<tr>
<td>Mauro Faccin</td>
<td>Université Catholique de Louvain, Belgium</td>
</tr>
<tr>
<td>Giorgio Fagiolo</td>
<td>Sant'Anna School of Advanced Studies, Italy</td>
</tr>
<tr>
<td>Shobeir Fakhraei</td>
<td>University of Southern California, USA</td>
</tr>
<tr>
<td>Kate Farrahi</td>
<td>University of Southampton, UK</td>
</tr>
<tr>
<td>Alessandro Flammini</td>
<td>Indiana University Bloomington, USA</td>
</tr>
<tr>
<td>Manuel Foerster</td>
<td>University of Hamburg, Germany</td>
</tr>
<tr>
<td>Mattia Frasca</td>
<td>University of Catania, Italy</td>
</tr>
<tr>
<td>Xinchu Fu</td>
<td>Shanghai University, China</td>
</tr>
<tr>
<td>Angelo Furno</td>
<td>University of Lyon, France</td>
</tr>
<tr>
<td>Sabrina Gaito</td>
<td>University of Milan, Italy</td>
</tr>
<tr>
<td>Lazaros Gallos</td>
<td>Rutgers University, USA</td>
</tr>
<tr>
<td>José Manuel Galán</td>
<td>Universidad de Burgos, Spain</td>
</tr>
<tr>
<td>Yerali Gandica</td>
<td>Université Catholique de Louvain, Belgium</td>
</tr>
<tr>
<td>Jianxi Gao</td>
<td>Rensselaer Polytechnic Institute, USA</td>
</tr>
<tr>
<td>David Garcia</td>
<td>Medical University of Vienna, Austria</td>
</tr>
<tr>
<td>Álvaro García-Recuero</td>
<td>Atos Research &amp; Innovation, France</td>
</tr>
<tr>
<td>Silvia Giordano</td>
<td>SUPSI, Switzerland</td>
</tr>
<tr>
<td>James Gleeson</td>
<td>University of Limerick, UK</td>
</tr>
<tr>
<td>Kwang-II Goh</td>
<td>Korea University, Korea</td>
</tr>
<tr>
<td>Jesús Gomez-Gardeñes</td>
<td>Universidad de Zaragoza, Spain</td>
</tr>
<tr>
<td>Bruno Gonçalves</td>
<td>New York University, USA</td>
</tr>
<tr>
<td>Przemyslaw Grabowicz</td>
<td>Max Planck Institute, Germany</td>
</tr>
<tr>
<td>Jelena Grujic</td>
<td>Vrije Universiteit Brussel, Belgium</td>
</tr>
<tr>
<td>Jean-Loup Guillaume</td>
<td>L3i - Université de la Rochelle, France</td>
</tr>
<tr>
<td>Mehmet Gunes</td>
<td>University of Nevada, USA</td>
</tr>
<tr>
<td>Emre Guney</td>
<td>Pompeu Fabra University, Spain</td>
</tr>
<tr>
<td>Sergio Gómez</td>
<td>Universitat Rovira i Virgili, Spain</td>
</tr>
<tr>
<td>Aric Hagberg</td>
<td>Los Alamos National Laboratory, USA</td>
</tr>
<tr>
<td>Edwin Hancock</td>
<td>University of York, UK</td>
</tr>
<tr>
<td>Chris Hankin</td>
<td>Imperial College London, UK</td>
</tr>
<tr>
<td>Jin-Kao Hao</td>
<td>University of Angers, France</td>
</tr>
<tr>
<td>Yukio Hayashi</td>
<td>JAIST, Japan</td>
</tr>
<tr>
<td>Denis Helic</td>
<td>Graz University of Technology, Austria</td>
</tr>
<tr>
<td>Shaun Hendy</td>
<td>The University of Auckland, Australia</td>
</tr>
<tr>
<td>Laura Hernandez</td>
<td>CNRS-Université de Cergy-Pontoise, France</td>
</tr>
<tr>
<td>Henry Hexmoor</td>
<td>Southern Illinois University, USA</td>
</tr>
<tr>
<td>Babak Heydari</td>
<td>Stevens Institute of Technology, USA</td>
</tr>
<tr>
<td>Philipp Hoevel</td>
<td>University College Cork, Australia</td>
</tr>
<tr>
<td>Seok-Hee Hong</td>
<td>University of Sydney, Australia</td>
</tr>
<tr>
<td>Ulrich Hoppe</td>
<td>University Duisburg-Essen, Germany</td>
</tr>
<tr>
<td>Agnes Horvat</td>
<td>Northwestern University, USA</td>
</tr>
</tbody>
</table>
Mark Humphries  The University of Manchester, UK
Laurent Hébert Dufresne  University of Vermont, USA
Yuichi Ikeda  Kyoto University, Japan
Roberto Interdonato  CIRAD - UMR TETIS, France
Giulia Iori  City, University of London, UK
Francesco Iorio  Wellcome Sanger Institute, UK
Gerardo Íñiguez  Next Games & Aalto University, Finland
Sarika Jalan  IIT Indore, India
Mahdi Jalili  RMIT University, Australia
Jaroslaw Jankowski  West Pomeranian Univ. of Technology, Poland
Marco Alberto Javarone  Coventry University, UK
Hawoong Jeong  KAIST, South Korea
Tao Jia  Southwest University, Chongqing, China
Di Jin  Tianjin University, China
Hang-Hyun Jo  APCTP, South Korea
Nick Jones  Imperial College, UK
Bertrand Jouve  CNRS, France
Byungnam Kahng  Seoul National University, South Korea
Andreas Kaltenbrunner  NTENT, USA
Rushed Kanawati  Université Paris 13, France
Márton Karsai  ENS de Lyon, France
Mehmet Kaya  Firat University, Italy
Przemysław Kazienko  Wroclaw Univ. of Science and Technology, Poland
Dror Kenett  Johns Hopkins University, USA
Yoed Kenett  University of Pennsylvania, USA
Khaldoun Khashanah  Stevens Institute of Technology, USA
Hamamache Khedouci  University of Lyon 1, France
Hyoungshick Kim  Sungkyunkwan University, South Korea
Maksim Kitsak  Northeastern University, USA
Mikko Kivelä  Aalto University, Finland
Konstantin Klemm  IFISC (CSIC-UIB), Spain
Peter Klimek  Medical University of Vienna, Austria
Xiaojie Kong  Dalian University of Technology, China
Johan Koskinen  The University of Manchester, UK
Nicolas Kourtellis  Telefonica Research, Spain
Danai Koutra  University of Michigan, USA
Jérôme Kunegis  University of Namur, Belgium
Valentina Kuskova  National Research University, Moscow, Russia
Ryszard Kutner  Faculty of Physics, University of Warsaw, Poland
Haewoon Kwak  Qatar Computing Research Institute, Qatar
Lucas Lacasa  Queen Mary University of London, UK
Renaud Lambiotte  University of Oxford, UK
Christine Largeron  Université de Lyon, France
Vito Latora  Queen Mary University of London, UK
Stephen Law  University College London, UK

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Anna T. Lawniczak  University of Guelph, Canada
Benedicte Le Grand  Universite Paris 1 Panthéon - Sorbonne, France
Eric Leclercq  University of Burgundy, France
Deok-Sun Lee  Inha University, South Korea
Sune Lehmann  Technical University of Denmark, Denmark
Elizaveta Levina  University of Michigan, USA
Fabrizio Lillo  University of Bologna, Italy
Gustav Lindmark  Linköping University, Italy
Yang-Yu Liu  Harvard University, USA
Jianguo Liu  Michigan State University, USA
Giacomo Livan  University College London, UK
Alessandro Lomi  University of Lugano, Switzerland
Alessandro Longheu  DIEEI - University of Catania, Italy
Jinhu Lü  Chinese Academy of Sciences, China
Claudio Lucchese  Ca’ Foscari University of Venice, Italy
John C.S. Lui  Chinese University of Hong Kong, Hong Kong
Matteo Magnani  Uppsala University, Sweden
Clemence Magnien  LIP6 (CNRS - Sorbonne Université), France
Fragkiskos Malliaros  University of Paris-Saclay, France
Eric Malmi  Google, USA
Giuseppe Mangioni  University of Catania, Italy
Madhav Marathe  Dept. of Computer Science, Virginia Tech, USA
Radek Mark  Czech Technical Univ. in Prague, Czech Republic
Andrea Marino  University of Pisa, Italy
Antonio Marques  King Juan Carlos University, Spain
Christoph Martin  Leuphana University of Lüneburg, Germany
Cristina Masoller  Universitat Politècnica de Catalunya, Spain
Rossana Mastrandrea  IMT Institute of Advanced Studies, Italy
Naoki Masuda  University of Bristol, UK
Michael Mathioudakis  University of Helsinki, Finland
Matúš Medo  UESTC, China
Natarajan Meghanathan  Jackson State University, USA
Jörg Menche  CeMM, Austria
Jose Fernando Mendes  University of Aveiro, Portugal
Ronaldo Menezes  University of Exeter, UK
Anke Meyer-Baese  FSU, USA
Radoslaw Michalski  Wroclaw Univ. of Science and Technology, Poland
Bivas Mitra  Indian Institute of Technology Kharagpur, India
Marija Mitrovic Dankulov  Institute of physics Belgrade, Serbia
Raul Mondragon  Queen Mary University of London, UK
Misael Mongiovi  Università di Catania, Italy
Yamir Moreno  University of Zaragoza, Spain
Esteban Moro  Universidad Carlos III de Madrid, Spain
Sotiris Moschouyiannis  University of Surrey, UK
Igor Mozetič  Jozef Stefan Institute, Slovenia

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Animesh Mukherjee, Indian Institute of Technology, Kharagpur, India
Tsuyoshi Murata, Tokyo Institute of Technology, Japan
Katarzyna Musial, Bournemouth University, UK
Alessandro Muscoloni, Technische Universität Dresden, Germany
Mirco Musolesi, University College London, UK
Michael Mäs, ETH Zurich, Switzerland
Muaz Niazi, COMSATS Institute of IT, Pakistan
Andrea Omicini, Alma Mater Studiorum–Università di Bologna, Italy
Gergely Palla, MTA-ELTE, Hungary
Pietro Panzarasa, Queen Mary University of London, UCL
Fragkiskos Papadopoulos, Cyprus University of Technology, Cyprus
Symeon Papadopoulos, Information Technologies Institute, Greece
Evangelos Papalexakis, University of California Riverside, UK
Michela Papandrea, SUPSI, Switzerland
Noseong Park, UNC Charlotte, USA
Juyong Park, KAIST, South Korea
Han Woo Park, YeungNam University, South Korea
Andrea Passarella, IIT-CNR, Italy
Leto Peel, Université Catholique de Louvain, Belgium
Tiago Peixoto, University of Bath, UK
Matjaz Perc, University of Maribor, Australia
Giovanni Petri, ISI Foundation, Italy
Juergen Pfeffer, Technical University of Munich, Germany
Carlo Piccardi, Politecnico di Milano, Italy
Clara Pizzuti, National Research Council of Italy (CNR), Italy
Chiara Poletto, INSERM, Sorbonne Université, France
Nataša Pržulj, UCL, UK
Christian Quadri, University of Milan, Italy
Marco Quaggiotto, ISI Foundation, Italy
Filippo Radicchi, Northwestern University, USA
Jose J. Ramasco, IFISC (CSIC-UIB), Spain
Asha Rao, RMIT University, Australia
Gesine Reinert, University of Oxford, UK
Benjamin Renoust, Osaka University, Japan
Pedro Ribeiro, University of Porto, Portugal
Massimo Riccaboni, IMT Institute for Advanced Studies, Lucca, Italy
Laura Ricci, Università di Pisa, Italy
Luis M. Rocha, Indiana University Bloomington, USA
Luis E. C. Rocha, University of Greenwich, UK
Francisco Rodrigues, University of São Paulo, Brazil
Luca Rossi, IT University of Copenhagen, Denmark
Camille Roth, Sciences Po, France
Tarik Roukny, Massachusetts Institute of Technology, USA
Amir Rubin, Ben Gurion University of the Negev, Israel
Meead Saberi, Monash University, Australia
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution/Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ali Safari</td>
<td>Universität Erlangen-Nürnberg, Germany</td>
</tr>
<tr>
<td>Alessandra Sala</td>
<td>Bell Labs, Ireland</td>
</tr>
<tr>
<td>Marc Santolini</td>
<td>Northeastern University, USA</td>
</tr>
<tr>
<td>Francisco C. Santos</td>
<td>Universidade de Lisboa, Portugal</td>
</tr>
<tr>
<td>Jari Saramäki</td>
<td>Aalto University, Finland</td>
</tr>
<tr>
<td>Antonio Scala</td>
<td>Italian National Research Council, Italy</td>
</tr>
<tr>
<td>Michael Schaub</td>
<td>Massachusetts Institute of Technology, USA</td>
</tr>
<tr>
<td>Maximilian Schich</td>
<td>The University of Texas at Dallas, USA</td>
</tr>
<tr>
<td>Rossano Schifanella</td>
<td>University of Turin, Italy</td>
</tr>
<tr>
<td>Ingo Scholtes</td>
<td>ETH Zurich, Switzerland</td>
</tr>
<tr>
<td>Frank Schweitzer</td>
<td>ETH Zurich, Switzerland</td>
</tr>
<tr>
<td>Caterina Scoglio</td>
<td>Kansas State University, USA</td>
</tr>
<tr>
<td>Simone Severini</td>
<td>University College London, UK</td>
</tr>
<tr>
<td>Rajesh Sharma</td>
<td>University of Tartu, Estonia</td>
</tr>
<tr>
<td>Amitabh Sharma</td>
<td>Harvard University, USA</td>
</tr>
<tr>
<td>Cynthia Siew</td>
<td>University of Warwick, UK</td>
</tr>
<tr>
<td>Tiago Simas</td>
<td>Telefonica Innovation Alpha, Spain</td>
</tr>
<tr>
<td>Anurag Singh</td>
<td>NIT Delhi, India</td>
</tr>
<tr>
<td>Per Sebastian Skardal</td>
<td>Trinity College, UK</td>
</tr>
<tr>
<td>Michael Small</td>
<td>The University of Western Australia, Australia</td>
</tr>
<tr>
<td>Tom Snijders</td>
<td>University of Groningen, Germany</td>
</tr>
<tr>
<td>Chaoming Song</td>
<td>University of Miami, USA</td>
</tr>
<tr>
<td>Massimo Stella</td>
<td>Institute for Complex Systems Simulation, UK</td>
</tr>
<tr>
<td>Markus Strohmaier</td>
<td>RWTH Aachen University &amp; GESIS, Germany</td>
</tr>
<tr>
<td>Blair D. Sullivan</td>
<td>North Carolina State University, USA</td>
</tr>
<tr>
<td>Jie Sun</td>
<td>Clarkson University, USA</td>
</tr>
<tr>
<td>Pál Sundsøy</td>
<td>NBIM, Norway</td>
</tr>
<tr>
<td>Michael Szell</td>
<td>Central European University, Hungary</td>
</tr>
<tr>
<td>Boleslaw Szymanski</td>
<td>Rensselaer Polytechnic Institute, USA</td>
</tr>
<tr>
<td>Bosiljka Tadic</td>
<td>Jozef Stefan Institute, Slovenia</td>
</tr>
<tr>
<td>Andrea Tagarelli</td>
<td>DIMES, University of Calabria, Italy</td>
</tr>
<tr>
<td>Lucia Tajoli</td>
<td>Politecnico di Milano, Italy</td>
</tr>
<tr>
<td>Kazuhiro Takemoto</td>
<td>Kyushu Institute of Technology, Japan</td>
</tr>
<tr>
<td>Frank Takes</td>
<td>Leiden University, Netherlands</td>
</tr>
<tr>
<td>Fabien Tarissan</td>
<td>CNRS - ENS Paris-Saclay (ISP), France</td>
</tr>
<tr>
<td>Dane Taylor</td>
<td>University at Buffalo, SUNY, USA</td>
</tr>
<tr>
<td>Claudio Juan Tessone</td>
<td>Universität Zürich, Switzerland</td>
</tr>
<tr>
<td>Claudia Test</td>
<td>GESIS, Germany</td>
</tr>
<tr>
<td>My Thai</td>
<td>University of Florida, USA</td>
</tr>
<tr>
<td>I-Hsien Ting</td>
<td>National University of Kaohsiung,</td>
</tr>
<tr>
<td>Olivier Togni</td>
<td>Lab. LE2I, Burgundy University, France</td>
</tr>
<tr>
<td>Vincent Antonio Traag</td>
<td>CWTS, Leiden University, Netherlands</td>
</tr>
<tr>
<td>Ljiljana Trajkovic</td>
<td>Simon Fraser University, Canada</td>
</tr>
<tr>
<td>Jan Treur</td>
<td>Vrije Universiteit Amsterdam, Netherlands</td>
</tr>
<tr>
<td>Chi K. Tse</td>
<td>Hong Kong Polytechnic University, Hong Kong</td>
</tr>
</tbody>
</table>

The 7th International Conference on Complex Networks and Their Applications. 11 - 13 Dec., 2018, Cambridge (UK)
Liubov Tupikina  
Ecole Polytechnique, France

János Török  
Budapest Univ. of Tech. and Economics, Hungary

Stephen Uzzo  
New York Hall of Science, USA

Lucas D. Valdez  
Boston University, USA

Sergi Valverde  
University Pompeu Fabra, Spain

Piet Van Mieghem  
Delft University of Technology, Netherlands

Balazs Vedres  
CEU, Hungary

Huijuan Wang  
Delft University of Technology, Netherlands

Xiaofan Wang  
Shanghai Jiao Tong University, China

Guanghui Wen  
Southeast University, China

Richard Wilson  
University of York, UK

Ernst Wit  
University of Groningen, Germany

Jinshan Wu  
Beijing Normal University, China

Bin Wu  
Beijing University, China

Feng Xia  
Dalian University of Technology, China

Bo Xu  
Dalian University of Technology, China

Gang Yan  
Tongji University, Shanghai, China

Taha Yasseri  
University of Oxford, UK

Wenwu Yu  
Southeast University, China

Zi-Ke Zhang  
Hangzhou Normal University, China

Yi-Cheng Zhang  
University of Fribourg, Germany

Junfei Zhao  
Columbia University, USA

Matteo Zignani  
University of Milan, Italy

Antonio G. Zippo  
Consiglio Nazionale delle Ricerche, Italy

Vinko Zlatic  
Institute Rudjer Boskovic, Croatia

Arkaitz Zubiaga  
The University of Warwick, UK

Katharina Anna Zweig  
University of Kaiserslautern, Germany

Yves-Alexandre de Montjoye  
Imperial College London, UK

Marco van der Leij  
University of Amsterdam, Netherlands