Gender Disparities in Computing

Between nursing informatics and Fachtagung prozessrechner

Thomas M. Meijers
10647023

Bachelor thesis
Credits: 18 EC

Bachelor of Science Artificial Intelligence

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Dr. M.J. Marx

ILPS, IvI
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

June 24th, 2016

Acknowledgements

First and foremost, I would like to thank my supervisor, Maarten Marx, for his insight, feedback and constantly good mood. Discussing our progress outside while smoking a cigarette together always provided me with new energy and ideas. I would also like to express my gratitude towards Casper Strømgren of Genderize.io for letting me use their API. Without it, my analysis would have been performed on a much smaller sample. Another contribution to my evaluation results is the use of SightCorp's F.A.C.E. API, for this I thank Theo Gevers and Roberto Valentini.

Furthermore, special thanks to Eva Nijsten for sometimes keeping me from my work and, indirectly, contributing to the quality of my thesis. Without Eva I would not have enjoyed this period as much. Also, many thanks to my Mother, Yvonne Sueters, for her insight in the English language and implicitely raising me to become a feminist. Continuing this line, I would also like to thank my father, Joost Meijers, for making it able for me to study Artificial Intelligence in the first place and Nina Läuger for sometimes reminding me why it is important to be, and why everyone should be, a feminist. I would like to thank Joline van der Pal, Bart Nijsten and Maud van Lier as well for their energy and support. Without them I could not possible have spent this many days at libraries.

Lastly, in no specific order, I would like to thank some of Maarten Marx his students for their efforts: Merel Stammes, Iris Kailola, Karlijn Rozestraten, Muhammad Shuduyev, Iliass Allach, Joeri Primowees, Thomas Dijkman, Amber Stoete, Hidde de Haan, Rick Bakker, de Dennis de Vries, Rens Vendrig, Tom Krommenhoek, Rob Dekker, Bart Witting, Gavin Schipper, Deveney Brehler, Jurian van Zanten, Jeffrey Ng, Marty Star, van Sander Wickeren, Robin Schouten, Alrian Kamdhi, Arnout van Dael, Joep Straatman, Saskia Woortman, Kennet Botan, Edward Gubler, Coen Welling, Rosan van der Werf, Tom Dekker, Laura Geerars, Luna de Veer, Wout Singerling, Nathalie Sinnema, Menno Lont, Fleur Koentjes, Yaleesa Borgman, Britt Ruigrok, Ferdi Zwamborn, Bob Scheer, Ossip Kupperman, Yoeri Leijdekker, Marit Beerepoot, Andrea Pineda Calderon, Maryse van Dalen, Hidaya Boujamaa, Jessy Bosman, Barend van Rooij, Jasper van derr Heide, Vincent Damen, Daryl Zandvliet, Mathijs Parmentier, Brechje Boeklagen, Tijmen Venneker, Kees Stammes, Bozana Miletic, Jimi Duiveman, Meile Houtsma, Abe Sweep, Siyen Bindels, Nadine Allewijn, Laura Hilhorst and Rens Gingnagel.

Abstract

Gender disparities have always been present in all of academia. Insight with regard to these disparities is crucial for scientific disciplines to become more egalitarian. In order to shed light on gender disparities within computing, the DBLP bibliography was analysed. As no gender was available for authors in the DBLP, the gender of these authors was inferred. Thus, a combination of gender classifiers was analysed. Classification was done based on name and images of persons. Validation of combined classification on two separate data sets resulted in $\boldsymbol{F_2}$-scores of 0.898 and 0.907. We were able to assign gender to 75.8% of the authors. Female computer scientists have less authorships on average than their male counterparts. The type of a publication or size of a conference does not affect this. The only aspect where the gender distribution favoured women is first authorship positions.

Contents

1 Introduction

Gender disparities have persisted in all of academia and male scientists dominate scientific output in most countries, as can be seen in figure 1 (Sugimoto et al., 2013). Through a bibliometric analysis, using a data set comprising nearly 5.5 million papers published between 2008 and 2012, Sugimoto et al. shed light on gender disparities between scientific disciplines and countries. Many disciplines tend to be dominated by male scientists, amongst which philosophy, mathematics and also computer science. In 2008, women earned 57% of all Bachelor's degrees in the United States, but only 18% in the area of computer science (Ashcraft and Blithe, 2009). These numbers decrease even further for Master's degrees and PhDs and can only increase gender disparities in computer science.

Gender diversity matters: Herring (2009) show that gender-diverse teams positively influence sales revenue, customer numbers and relative profits compared to gender-heterogeneous team. Gender-diverse teams are also more efficient, creative and its members more self-confident (Gratton et al., 2007). To be able to improve gender diversity, and thus decrease gender disparities, insight into these matters is crucial.

As mentioned, computer science is one of the fields that tends to be led predominantly by male researchers according to Sugimoto et al. (2013). From 1966 to 2008, women authorship of workshop and conference papers increased by a large margin but still makes up only 27% of the total of these papers (Cohoon et al., 2011).

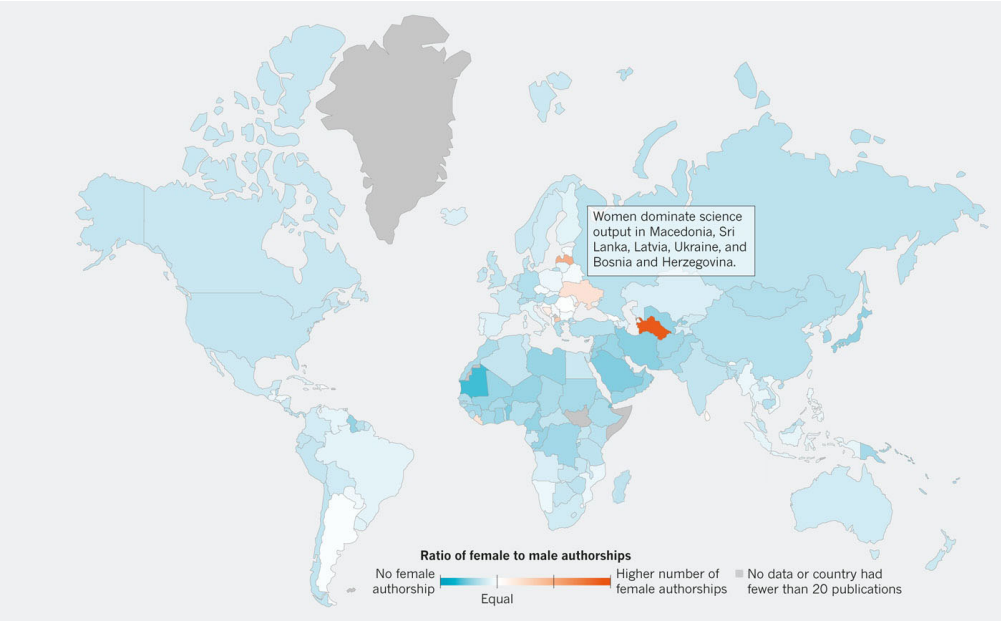

Figure 1: Leading gender authorship per country (Sugimoto et al., 2013)

1.1 Research Questions

This research aims to provide better quantitative knowledge of gender disparities in computer science. Therefore the question we want to answer is: *What gender disparities become apparent through a bibliometric analysis of the DBLP computer science bibliography?* This bibliography is a freely available bibliography tool (Ley, 2005). It is designed to provide researchers and scientists with high quality meta-data on computer science publications. It contains over 3.3 million publications, most of them being papers and articles.

To analyse gender disparities in the DBLP bibliography, we have to infer gender for all of the included authors. Most authors are known under their full name. Unfortunately, gender is not listed in the bibliography nor anything else that could directly help to obtain the author's gender. In order to answer the research question, an author's gender will be inferred through her or his name. In order to obtain a higher precision and recall, profile pictures on an author's homepage will be used to enhance classification.

When gender can be predicted with high precision and recall, the disparities within DBLP can be analyzed. The bibliometric of this analysis will focus on the difference of scientific output between female and male computer scientists. Due to the fact that DBLP contains publications with publication dates, ranging from 1936 to 2016, the trend of this difference will also be looked at. In order to successfully answer these questions, the following sub-questions will have to be addressed.

1. **With what recall and precision can the gender of a DBLP author be determined?**

   To gain quantitative knowledge on gender disparities, we need to be able to infer the gender of the authors with high precision. This will be done by assigning gender through name lists. To obtain a higher accuracy, the unclassified person's homepage will be crawled for images if the author has provided an URL. These images will then be used for image classification. This method will require us to answer the following sub-questions:

   (a) **What grade of accuracy can be achieved by inferring gender from name only?**

   Multiple countries have amassed name lists with associated gender. Good examples of these lists are the US Census Data[1] and name data by the UK Office of National Statistics[2]. There are multiple classifiers available that make use of these name lists by simply looking up a name and assigning a gender when it is statistically likely to be correct. Since authors in DBLP nearly never register their affiliation and their ethnicity is wholly unknown, we will have to validate both individual and combined classification and conclude which classifier or combination of classifiers obtains the highest accuracy

   (b) **Does inferring gender from author pictures increase our accuracy?**

[1] Since 1880, records name with associated gender if more than 5 births took place.
[2] Records name and gender since 1996 for any name with 3 or more births.

As homepages often include a profile picture, the images on these pages can be used for gender classification. Training a neural network for image recognition is a challenging task and requires a lot of data. It is also well beyond the scope of this research. There are, however, several APIs available which perform facial recognition and infer features, including gender from a person's face. Answering this questions requires testing its performance in terms of speed and effect on accuracy.

2. **What is the difference between scientific output of men and women and what aspects of DBLP influence this?**

   By assigning authorships to authors and gender to authorships per publication, scientific output can be analyzed. Through this method, general disparities within the DBLP bibliography can be discovered by comparing female and male authorships. To find out which factors influence authorship gender distribution, the following aspects will be further examined:

   (a) **Do male computer scientists occupy dominant positions more often, and is this decreasing over time?**

   The dominant positions are first and last authorship position. By assigning gender to these positions, we will be able to find the gender distribution of this aspect. Due to the DBLP containing papers, ranging from 1936 to 2016, the change of this distribution can be analyzed as well.

   (b) **Does publication type affect the gender distribution of authorships?**

   The relative output of conference and workshop papers of female computer scientist increased from 7% in 1967 to 27% in 2009 (Cohoon et al., 2011), while the number of female authors did not increase that much. Analyzing publications per type will show if there is a difference between articles and conference papers, the two main publication types in DBLP.

   (c) **Does conference size influence gender inequality regarding authorships?**

   By dividing the conferences and workshops in three groups, based on total authorships, it can be determined whether this has any effect on the composition of female and male authors. Finally we will also take the two largest artificial intelligence (AI) conferences in the DBLP bibliography, AAAI and IJCAI, and determine whether the field of AI is less diverse than the average.

1.2 Thesis outline

In section 2, related work, the literature research will be discussed, describing contemporary research in relation to the research questions. Thereafter, a description of the data, including validation data, is given in section 3. The methodology will be described in section 4, which the method of gender classification will be outlined. Further, the validation and evaluation of

the classifiers will be justified in section 5, after which an elaboration on the interpretation of disparities will follow. Concluding this section, we will describe the aspects of DBLP, which will be analyzed in regard to gender disparities. Finally, a conclusion is given in section 6, followed by a discussion.

2 Related Work

In line with the posed research questions, this section will also be divided into two parts. papers on gender disparities within computer science will be reviewed. Secondly, contemporary literature on the area of gender classification will be discussed. More specifically, the second part contains literature on gender inference from (first) names and images.

2.1 Gender disparities within computer science

As stated in the section 1, gender disparities persist in nearly all of science. Previous bibliometric analysis has shown that male and female scientists have equal scientific output in only 6% of all countries world-wide (Sugimoto et al., 2013). Furthermore, male scientists occupy dominant positions, first or last author mentioned, more often than female scientists. In the field of computer science (CS), female scientists publish relatively more conference and workshop papers each year. Absolute publications, however, are still lower than those of male scientists (Cohoon et al., 2011). Part of this can be attributed to, as Vardi (2015) states, to the 'narrow pipeline' of female computer scientists. Women earn less than 18% of all computer science degrees in North America (Ashcraft and Blithe, 2009). In addition, low ranking CS departments have relatively more female PhDs than high-ranked departments do (Baumann et al., 2011).

Gender disparities in academia are, amongst other, measured by skewed gender distributions of average publications, citations and dominant author position. While Sugimoto et al. (2013) showed that females produce and publish far fewer papers than their male colleagues, women seem to be more productive when it comes to workshop and conference papers (Cohoon et al., 2011). The percentage of female authorships of ACM conference papers was lower than 10% in 1967. In 2008 that number was 23% while accounting for 27% of the conference papers. Note that gender disparities also exist between conferences; the percentage of female authorships ranges from 10% to 44%.

2.2 Gender classification

In order to give an analysis on gender disparities, the gender of authors in the DBLP needs to be inferred. Gender classification is a common problem and has been applied to data containing scientists, social platform users and other social groups (Cohoon et al., 2011; Lin and Serebrenik, 2016; Karimi et al., 2016). Using a person's name for gender classification is a common method. The dominant gender associated with a name is dependant on country of origin and age (Blevins

and Mullen, 2015; Cohoon et al., 2011). The gender classifiers make use of name-lists with associated gender to statistically determine gender. Many of these lists are readily available, albeit mostly for western names (Sugimoto et al., 2013, appendix)

The methods used in the mentioned papers obtain accuracy rates between 75% and 92%. If the data set is highly unbalanced, as is the case with the DBLP data, containing many more men than women, different accuracy measures are required (Lin and Serebrenik, 2016). The F-measure will be used to evaluate the accuracy of the classifiers, since recall is as important as precision. Using raw accuracy as the evaluation metric would encourage assigning a classification to all entries, where the F-measure does not. Karimi et al. (2016) managed to obtain an F_1-score of 92-93% but used a balanced data set, which included country and affiliation per entry. Ambiguous names were also removed from this data set. To obtain this score, a combination of name and image was used to classify gender.

3 Data

The DBLP computer science bibliography is set up to provide free high-quality bibliographic meta-data on major computer science publications. The raw data is freely accessible in the form of an XML dump.[3] [4] Currently, the DBLP bibliography contains over 3.3 million publications which are written by over 1.7 million authors. The first articles were added in 1995 and over the last ten years, the number of records has nearly quadrupled. Due to its popularity and vast amount of publications covered, it has been used within multiple experiments and proves to be a good framework for bibliographic analysis (Elmacioglu and Lee, 2005; Schifanella et al., 2012).[5]

3.1 DBLP publications

The DBLP bibliography contains different types of publications. Most records are either journal articles or conference and workshop papers, comprising 40.1% and 53.6% respectively. Other types of publications include, among others, publications in books and PhD theses. See figure 4 for the actual distribution of publication types. Other relevant categories of meta-data that are included per publication are authors, publication date, journal, title and publisher, refer to figure 3 for a sample.

The oldest records added to the bibliography were published in 1936 , however, as the discipline of Computer Science expanded ernomously in the last decades, it naturally contains more recent records (see figure 5). More than 81% of all DBLP records have been published in the year 2000 or later.

[3] Available at `http://dblp.uni-trier.de/xml/`.

[4] The dump used for this research was downloaded April 19th, 2016.

[5] The graphs in this section have been downloaded from `http://dblp.uni-trier.de/statistics/`.

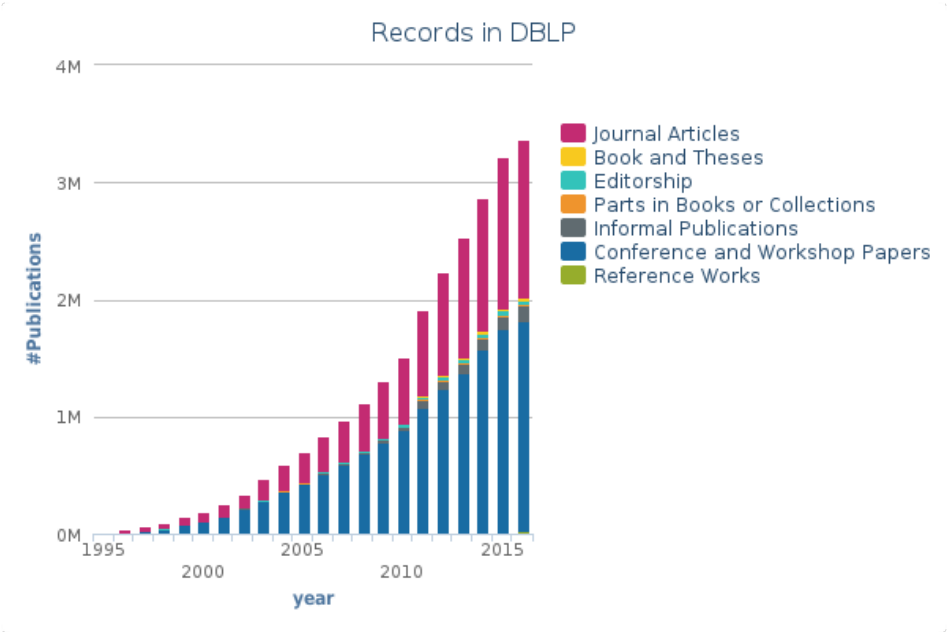

Figure 2: Total records in the DBLP bibliography per year

```
<article mdate="2011-05-25" key="journals/ngc/HasegawaSH00">
   <author>Osamu Hasegawa</author>
   <author>Katsuhiko Sakaue</author>
   <author>Satoru Hayamizu</author>
   <title>Interactive Learing and Management of Visual Inforamtion
        via Human-like Software Robot.</title>
   <pages>103-116</pages>
   <year>2000</year>
   <volume>18</volume>
   <journal>New Generation Comput.</journal>
   <number>2</number>
   <url>db/journals/ngc/ngc18.html#HasegawaSH00</url>
   <ee>http://dx.doi.org/10.1007/BF03037589</ee>
</article>
```

Figure 3: Xml sample from the DBLP publications file.

3.2 DBLP authors

Over 1.7 million authors are listed in the DBLP bibliography. Additional information includes a link to the author's DBLP page and optionally an URL to her or his personal homepage,

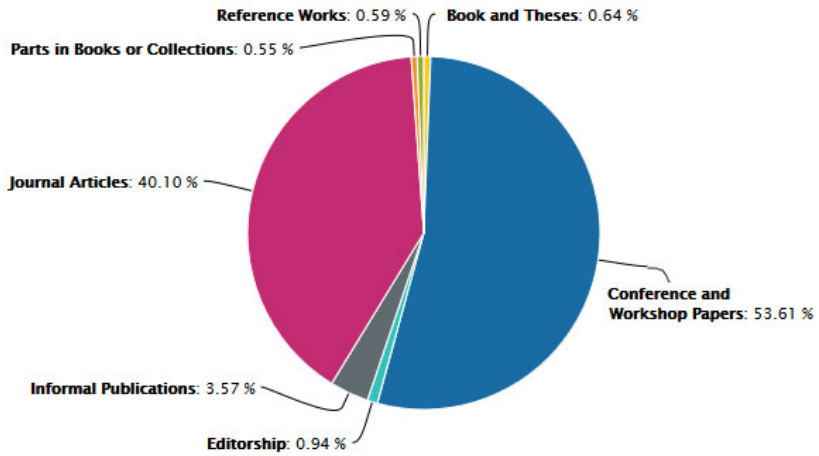

Figure 4: Types of publications included in the DBLP bibliography

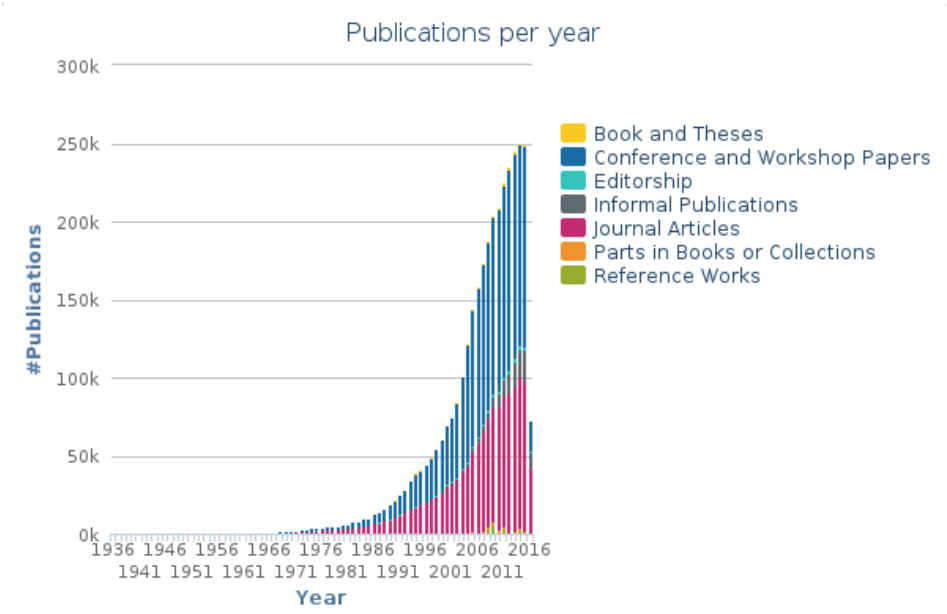

Figure 5: Publications per year in the DBLP bibliography

affiliation and one or several synonyms[6] (refer to figure 6 for a sample). The distribution of papers, written per author, is a power law distribution. Most authors have only one publication, while there is one author with 1297 publications (see figure 7). Most authors, nearly 270,000, have 2 co-authors. While there are a total of 280,000 authors who have 0 or 1 co-authors, the distribution follows a power law again after 2 co-authors or more (figure 8 for details).

[6]If an author has published under more than one name.

```
<www mdate="2012-01-16" key="homepages/96/10754">
   <author>Junnosuke Shino</author>
   <title>Home Page</title>
</www>
<www mdate="2015-01-04" key="homepages/96/3827">
   <author>Peter A. Henning</author>
   <author>Peter Henning</author>
   <title>Home Page</title>
<url>http://de.wikipedia.org/wiki/Peter_Henning_%28Physiker%29</url>
</www>
```

Figure 6: Xml sample from the DBLP author file.

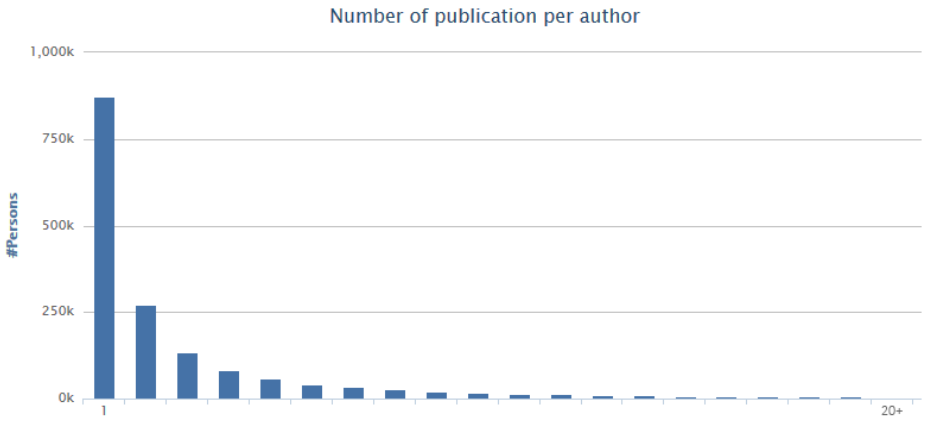

Figure 7: Publications per author

The difficulties in the data set are, of course, sparseness; some articles do not have any author or other meta-data listed, nor synonymy and homonymy[7]. To solve the issue of synonymy, DBLP tries to add a cross-reference to the respective author's entry in DBLP. There, 24,435 authors have one synonym listed and only 8 authors have listed 4.[8] Homonymy is solved by appending a four-digit number to the author's name.[9] These issues have to be reported to DBLP and are then fixed by hand, therefore the data is not without faults (Ley, 2009).

[7] Different authors publishing under the same name.

[8] For example: 'Patricia Flatley Brennan' also published under the names 'Patricia F. Brennan', 'Patricia Brennan', 'Patti Brennan' and 'Patti Flately Brennan'.

[9] There are 57 authors known as 'Wei Li', making it the most ambiguous name.

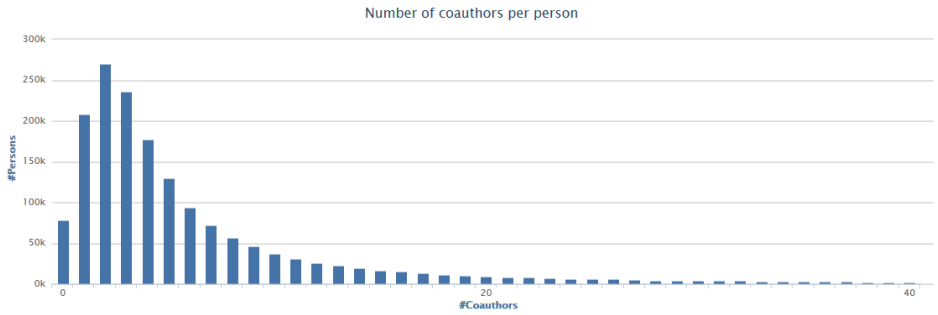

Figure 8: Coauthors per author

3.3 Validation Data

The accuracy of gender classification will be measured by two validation sets. The main validation set is a subset of the authors in the DBLP bibliography. For purposes of gender inference through images, the sample was taken from the set of authors who have at least one URL to their homepage. Of the 28,537 authors in this set a sample of authors was taken and annotated by hand. Each author is assigned a gender and a direct URL to his or her profile picture. The resulting validation set includes 634 authors of which 631 have a direct link to a profile picture. The number of males and females is respectively 503 and 131. There are no duplicate first names in this validation set.

The second validation set is taken from Wikipedia. The entries in this set are persons who have a personal Wikipedia page. The information listed per entry may be a direct URL the to thumbnail taken from the corresponding Wikipedia page. Most entries have either a normal first name listed or initials as within the DBLP data. The entries with non-real names, mostly artist names, were mostly removed. The Wikipedia set contains 849,051 entries, of which 721,501 males and 127,443 females. In this set, 264,676 entries have a direct URL to a thumbnail. While this validation set has many more entries, it has only 71,719 unique names, since many first names occur multiple times.

4 Methodology

This section will be split up into two parts. The main subject of the first part regards for obtaining a high accuracy for gender inference. This is crucial for the second part, which is the analysis of differences between female and male computer scientists.

4.1 Gender inference

4.1.1 Data parsing

The DBLP data set, as is mentioned in subsection 3, contains certain flaws and difficulties. Firstly, the data is incomplete. Not all publications have an authors listed and some publications have been assigned to the wrong author due to synonymy. These flaws will be addressed in subsection 4.2 as they have no direct impact on classification accuracy. The case of homonomy and authors who have initials listed, is a problem for accuracy. Special characters pose another challenge. The data set includes many Scandinavian and Turkish names, as well as names from other countries which use special characters. These characters cannot be encoded by the 'standard' Latin-1 character set, therefore all data needs to be parsed to Unicode. Secondly, the size of the data set is significant, for the XML dump has an uncompressed size of 1.68 GB. In order to cope with the mentioned difficulties, the data set has been parsed in the following way.

Decrease the size of the data as much as possible: The importance of this is due to memory issues and the speed of operations on the data set. To achieve a lower size, the data set was first split into a set of authors and a set of publications. For both sets, only the relevant information was saved:

Authors:
- Homonyms
- DBLP person ID
- Homepage URL

Publications:
- Publication type
- Publication title
- Book title
- Authors
- Publication year
- Journal
- Publisher

Solve the issue of homonymy: Since some authors have published under multiple names, it would distort the analysis if we regard these different names as different persons. For each author who uses homonyms, the best name will be determined. This will be a name with no initials (when possible), and is most often used by the author in terms of publications. By replacing all homonyms with a single name, the publications can be counted per author and not per name. The 'best' name will also be the one used to infer the author's gender.

4.1.2 Accuracy measures

To infer gender per author, three classifiers will be used. Two of these will infer gender from an author's first name, while the other classifier will infer gender from an image. To determine which classifier, or combination of classifiers predicts gender best, the weighted harmonic mean (F-measure) of recall and precision will be used.

Recall and Precision: Recall and precision will be calculated per gender. For completeness, recall is defined in equation 1 and precision in equation 2. Recall and precision are important, since the classifiers that are used may return 'unknown' when gender cannot be determined. Since female scientists are severely underrepresented in the data set, majority vote would result in an accuracy between 80% and 90%. Since we are especially interested in the actual distribution of female and male scientists, we are not interested in raw accuracy.

$$\boldsymbol{R_g} = \frac{TruePositives_g}{TruePositives_g + FalseNegatives_g} \quad | \quad g \in \{male, female\} \tag{1}$$

$$\boldsymbol{P_g} = \frac{TruePositives_g}{TruePositives_g + FalsePositives_g} \quad | \quad g \in \{male, female\} \tag{2}$$

***F*-measure**: The final accuracy measure will be the weighted harmonic mean of recall and precision. The formula for the F-measure is listed in equation 3. The weighting factor used, β, is 2. The resulting score, F_{2g}, prioritizes recall over precision. The reason for this is that the analysis of gender disparities within the DBLP bibliography depends heavily on recall.[10]. Since we have two classes, female and male, we will take the average of F_{2f} and F_{2m}. The final equation, used to determine the effectiveness of the (combined) classifiers, is listed under equation 4.

$$\boldsymbol{F_{\beta g}} = (1 + \beta^2) * \frac{P_g * R_g}{(\beta^2 * P_g) + R_g} \quad | \quad g \in \{male, female\} \tag{3}$$

$$\boldsymbol{F_{2f+m}} = \frac{1}{2} * (\frac{5 * P_f * R_f}{(4 * P_f) + R_f} + \frac{5 * P_m * R_m}{(4 * P_m) + R_m}) \tag{4}$$

Accuracy: While not used as the main accuracy measure, accuracy will also be included in the classifier evaluation results. The reason for including is, that it directly shows for how many cases gender is inferred correctly. Again for completeness, accuracy is listed in equation 5.

$$\boldsymbol{A} = \frac{TotalCorrect_f + TotalCorrect_m}{TotalActual_f + TotalActual_m} \tag{5}$$

4.1.3 Gender inference by name

The classifiers used for gender classification by name are SexMachine[11] and Genderize.io[12]. The names used should be first names. These are parsed from the author's full name, by splitting the full name on spaces and returning the first string. Since name notation in DBLP follows the

[10] For we cannot say anything about authors whose gender is unknown

[11] https://github.com/ferhatelmas/sexmachine/

[12] https://store.genderize.io/

pattern '[first name] ¡middle names¿ [last name]', this will return either a first name or initials. Two classifiers are used to be able to perform combined classifiers as well as a comparison between them. Both classifiers work by collecting data on names combined with gender and country. If a name has been observed a factor more often for one gender than another, it will be classified as that gender. Since associated gender per name can vary per country, the classifiers used the option to specify a country together with a name. This research will not make use of this option, for the DBLP lists country nor language.

SexMachine: The classifier is open-source and based of a C program programmed by Jörg Michael. The Python implementation can be installed through pip and imported in your local project, thus benefiting performance. Its data source includes over 40,000 unique names with associated gender and sometimes country. According to the original description, parts of the data have been classified by native speakers. For example, Turkish names were classified by a native Turkish speaker. The classification output by SexMachine is 'female', 'mostly_female', 'mostly_male', 'male' and an unknown category. For the purpose of this research the classification will be binary, which means that the 'mostly_' prefix will be parsed. The result, being one of 'female', 'male' or 'unknown' will be assigned to the respective author.

Genderize.io: The data used by this classifier, comes mainly from scraping social network sites. It includes over 216,286 unique names from 79 countries. The classifier is accessed through an API, accepting up to 10 names, optionally with specified country, per request. The classification returned is either 'Male', 'Female' or an unknown category. The classification assigned to the respective author will be parsed to equal the classification output of SexMachine.

Both classifiers will be used to classify the first names in all data sets independently. As the classification is binary, per classifier each author will be assigned either 'female', 'male' or 'unknown'.

4.1.4 Gender inference by image

To infer gender from images FaceCorp's F.A.C.E. API[13] will be used. A request will be answered with a response, containing how many persons are in the image and for each person multiple attributes, including gender. Gender classification is coupled to a confidence score and either 'male' or 'female'. The first classified gender will be taken, since manual inspection of a set of pictures showed that most of them contain just one person.

As both validation sets have direct URLs to a picture, these pictures will be downloaded in batches, after which they are sent to the F.A.C.E. API in batches as well. For each entry with an URL available, gender will be inferred independently through this method. The image classifier will also be used for classifier comparison and combination.

The URLs of the entries in the actual DBLP data set point to the respective author's homepage. The HTML of these homepages will be downloaded in batches, after which the BeautifulSoup library is used to collect all the links in the ¡IMG¿ tags. Per author, the set of image URLs from

[13] http://face.sightcorp.com/

his or her homepage are then requested in a batch again. The headers of these responses are used to sort the responses on size (large to small) and check their image type. For performance reasons, the images are written to disk and sent in a request to the F.A.C.E. API one by one, ordered by descending file size. As soon as a classification request returns either 'male' or 'female' this classification is assigned to the corresponding author. If no gender is assigned from any image or images are found 'unknown', it will be assigned to the respective person. This means that it is not the classifiers overall performance which is evaluated, but its classification accuracy with regard to a sparse data set, such as the DBLP data set.

4.1.5 Final classification

The final classification, performed on the authors in the DBLP data set, will be carried out through a combination of all three classifiers. The combined classification will be carried out hierarchically. If the main classifier can assign a classification, either 'male' or 'female', then this output will be used. For all unclassified authors, the secondary classifiers output will be assigned and the third classifier will assign its classification to all remaining unclassified authors.

In order to obtain the highest accuracy, all possible orders of combinations will be evaluated by both data sets. The combined classifier, resulting in the highest F_{2f+m} score, will be used to infer gender on the DBLP data set. Note that only a very small fraction of the DBLP authors has a URL listed. The importance of image classification is the evaluation of its added value in a combined classifier.

4.1.6 Assigning output and gender

The main analysis will be carried out through a comparison of the differences in scientific output between female and male computer scientists. The final combined classifier will be used to assign each author in DBLP a gender.

DBLP authors: The authors will be assigned both absolute and relative authorships. Absolute authorships are the number of publications they have authored. Relative output is a fraction of a publication. For example, if the total number of authors for an article is 3, the author will get assigned one-third of an authorship. In case of a publication where some authors were not assigned a gender, these authors' 'part' of the publication is not distributed between the other authors. Thus, the sum of all relative output of authors will not equal the number of publications in the DBLP bibliography.

DBLP publications: Each publication will be assigned the number of female and male authorships for which the corresponding authors' genders are inferred. In case an author's gender is unknown, it will not be adjoined to the number of female or male scientists. Some authors in the DBLP publication list are not in the DBLP author list. The gender of these authors will be separately inferred with the same combined classifier. The resulting output will be assigned

accordingly. For each publication with more than two authors, the associated gender of the dominant positions, first and last author, will also be saved.

Since not all authors will be classified as either female or male, the sum of female and male authors will be lower than the total number of authors. For analyzing gender distribution of publications, the authorships which were not assigned a gender, will be left out of the analysis. Be aware that for publications it is authorship that is counted and not unique authors. If one author co-authored 20 publications and the author is assigned a female gender, one female authorship will be added to each publication she worked on. The result is an addition of 20 female authorships to the total set of publications.

Due to the DBLP bibliography being incomplete, there may be authors with no assigned authorships, publications without authors or publications with authors of which no gender was inferred. Papers with an unknown gender distribution and authors with no assigned scientific output will be, naturally, excluded from the analysis.

4.2 Analyzing gender disparities

The main interest of the analysis will be scientific output, which equals authorship. Average scientific output per author, both relative and absolute output, will be compared. For authorships we will look at the publications and analyse the average distribution of female and male scientists. Since publications include publication year, the trend in this distribution will be looked at as well. A factor, possibly denoting disparities as well, is the gender distribution of dominant positions. Thus the average distribution of first and last authorship will be included in the analysis, as well as its trend over time.

Finally, it will be researched if there are disparities between different types of publications. Since most publications in DBLP are either journal articles or workshop and conference papers, these two types will mainly be used for the analysis. This line of research will be continued even more specifically, by inspecting differences in average scientific output per gender for conferences. We will examine whether the total authorships of a conference affects its gender distribution; a larger number means that either a conference is larger, that it has existed for some time or both. Concluding, these averages will be compared to the average of the two larger Artificial Intelligence conferences listed in the DBLP, namely AAAI and IJCAI.

5 Evaluation

As with previous sections, the evaluation of the results will also be conducted in two parts. The first part will be the accuracy of the individual and combined gender classifiers. The final analysis of gender disparities will be done in the second part of this section.

5.1 Gender classification accuracy

The evaluation of gender classifiers will initially be done individually, firstly comparing the name classifiers after which all three are compared with one another. Secondly, multiple combinations of classifiers will be evaluated. The performance of the combined classification is the most significant, for this classification method will be applied to the actual DBLP bibliography. All evaluations will be performed on both data sets.

5.1.1 DBLP validation - Name classifiers

As described in section 3, this validation set counts 634 entries, out of which 503 males and 131 females. Both name classifiers have a higher recall for women than for men, while precision is higher for men than for women (For numeric results on single classifiers refer to table 1 and 2). Comparing the classifiers, Genderize.io scores higher on recall while SexMachine scores higher on precision. The higher accuracy of Genderize.io can be attributed to its higher recall. Finally, Genderize.io performs better, measured by our main accuracy measure, having a mean $\boldsymbol{F_2}$-score of .864 compared to SexMachine's score of .807.

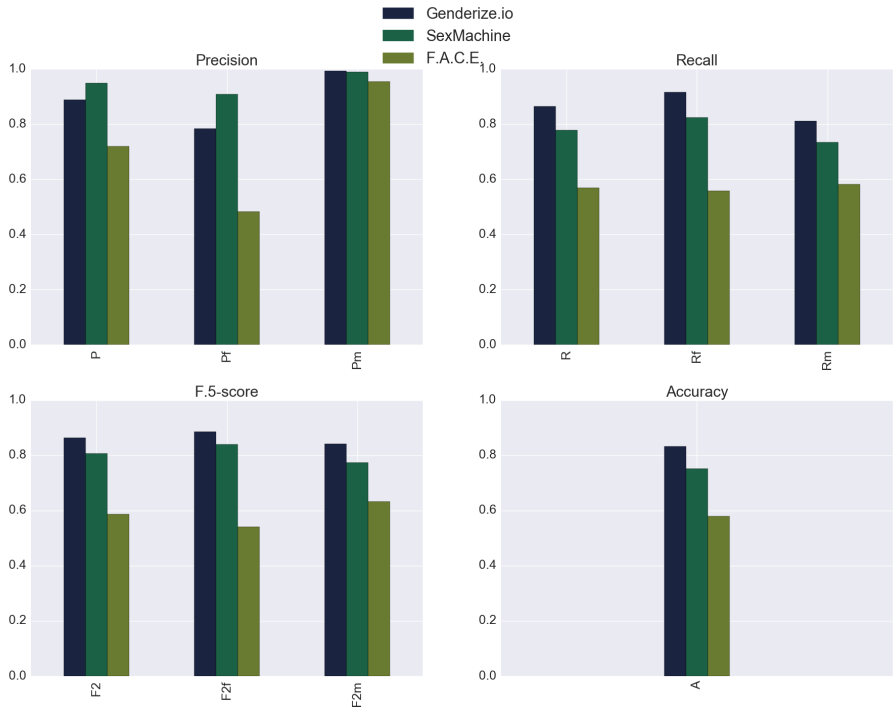

Figure 9: DBLP validation - Individual classification.

	P	P_f	P_m	R	R_f	R_m
Genderize.io	0.889	0.784	**0.993**	**0.864**	**0.916**	**0.811**
SexMachine	**0.948**	**0.908**	0.989	0.779	0.824	0.734
F.A.C.E.	0.719	0.483	0.954	0.570	0.557	0.583

Table 1: DBLP validation - Precision and recall (single classifiers).

	A	F_2	F_{2f}	F_{2m}
Genderize.io	**0.832**	**0.864**	**0.886**	**0.842**
SexMachine	0.752	0.807	0.840	0.774
F.A.C.E.	0.580	0.586	0.541	0.632

Table 2: DBLP validation - Accuracy and F_β-scores (single classifiers).

5.1.2 Wikipedia validation - Name classifiers

The gender distribution for the Wikipedia set is 721,501 males and 127,443 females. The validation results of the Wikipedia validation set classification are comparable to those of the DBLP validation set classification. Precision is nearly identical, while recall is approximately 5% higher, resulting in a slight increase of the F_2-score. One reason for the higher recall could be the relatively lower amount of unique names combined with it including more western names.[14] The relatively higher accuracy for the Wikipedia validation set can also be attributed to this. The low performance of the image classifier is due to the fact that less than a third of the entries include a URL to a thumbnail.[15]

	P	P_f	P_m	R	R_f	R_m
Genderize.io	0.894	0.798	0.991	**0.894**	**0.892**	**0.895**
SexMachine	**0.942**	**0.891**	**0.993**	0.840	0.821	0.859
F.A.C.E.	0.698	0.444	0.951	0.140	0.153	0.127

Table 3: Wikipedia validation - Precision and recall (single classifiers).

Of the classifiers used, Genderize.io performs best overall, mainly due to its higher recall. Its F_2-score lies between 86% and 90% which is approximately 5% percent higher than the score of SexMachine. Its results are quite similar to the results of the DBLP validation set classification. The slightly higher results on this set can presumptively be attributed to the its relative homogeneity.[16]

[14] As discussed in section 2, Western names are found more often in name lists.

[15] For a discussion on the purpose of the image classifiers, refer to the end of section 4.1.4.

[16] The wikipedia set has relatively less unique names and, as manual inspection suggests, less non-Western names.

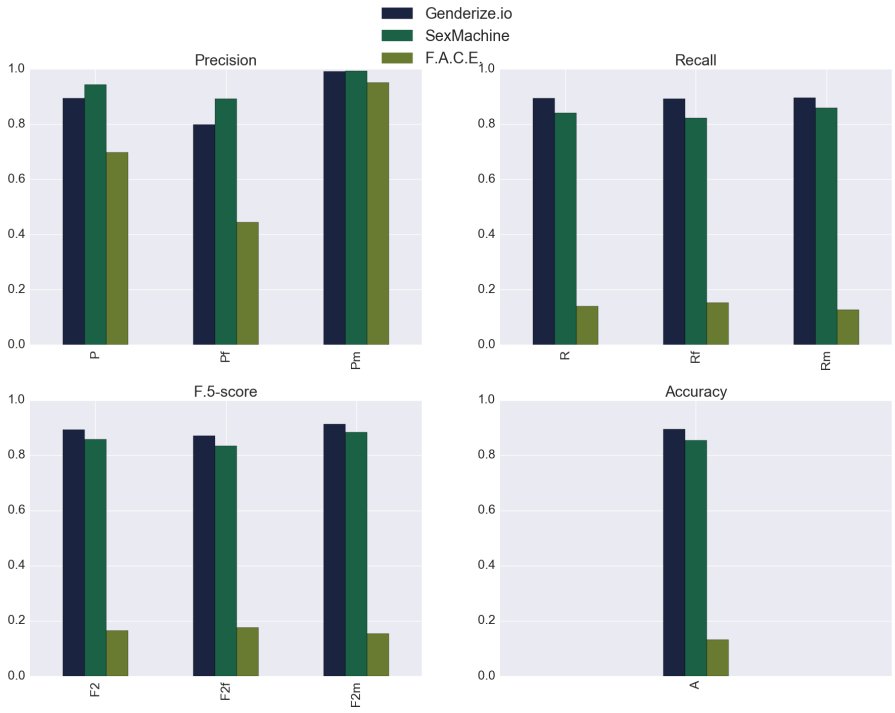

Figure 10: Wikipedia validation - Individual classification.

	A	F_2	F_{2f}	F_{2m}
Genderize.io	**0.895**	**0.892**	**0.871**	**0.913**
SexMachine	0.853	0.859	0.834	0.883
F.A.C.E.	0.131	0.165	0.177	0.154

Table 4: Wikipedia validation - Accuracy and F_β-scores (single classifiers).

5.1.3 Image classification

Note: Image classification results can be found in the same figures and tables mentioned in the previous section. These are figures 9 and 10 and tables 1, 2, 3 and 4.

Image classification scores lower on all accuracy measures compared to the name classifiers. Part of this can be attributed to broken URLs, but mainly gender inference through images results more often in 'unknown' results by not being able to detect a face in the images. Interestingly, the only measure the image classifier scores almost as good on as the name classifiers, is its precision of male classification, while it scores extremely low on precision of female classification. The data does not give any apparent clarification for this behavior, as recall is not many times higher for females than for males. It is possible that the classification model is trained on far more

images containing male faces and thus its facial recognition for females is much worse. Image classification, used individually, performs significantly worse compared to name classification.

The Wikipedia validation results can be discussed rather quick, as it is not a good data set to test the performance of the image classifier individually. Its precision is the only point of interest and is quite similar to the precision on the DBLP set.

5.2 Combined classification

Combined classification analysis will be done for most combinations.[17] The order of the capitals shows in which order classifications were assigned. For example, **G F** means that Genderize.io's classification was first assigned, after which F.A.C.E.'s classification was assigned to all remaining unknown cases.

5.2.1 DBLP validation - Combined classification

As expected, combining classifiers results in overall higher results for all accuracy measures (results displayed in figures 11 and 12 and tables 5 and 6). Recall is up to 5.2% higher compared to the highest recall with single classification, going up from 86.4% to 91.6%. This without severely impacting precision, resulting in an overall increase of $\boldsymbol{F_2}$-score.

Classification with Genderize.io as the main classifier, combined with the remaining two classifiers, regardless of order, yields the best results, although almost all other classifiers are on par. A point of interest is by how much the results increase when using the image classifier on top of the name classifiers. Combining either **G S** or S G with F.A.C.E. results in an increase in recall by approximately 4%, a decrease in precision around 3% resulting in an $\boldsymbol{F_2}$-score which is around 3% higher.

	$\boldsymbol{P}$	$\boldsymbol{P_f}$	$\boldsymbol{P_m}$	$\boldsymbol{R}$	$\boldsymbol{R_f}$	$\boldsymbol{R_m}$
G S	**0.889**	**0.784**	**0.993**	0.875	0.916	0.833
G F	0.858	0.727	0.989	**0.914**	**0.954**	**0.875**
S G	0.887	0.787	0.988	0.868	0.901	0.835
S F	0.866	0.745	0.986	0.888	0.916	0.861
G S F	**0.858**	0.727	**0.989**	**0.916**	**0.954**	0.879
G F S	**0.858**	0.727	**0.989**	**0.916**	**0.954**	0.879
S G F	0.856	0.728	0.984	0.910	0.939	**0.881**
S F G	**0.858**	**0.729**	0.987	0.914	0.947	**0.881**

Table 5: DBLP validation - Precison and recall for combined clasification.

[17] Combinations with F.A.C.E. as the main classifier has been excluded. See appendix A for these results.

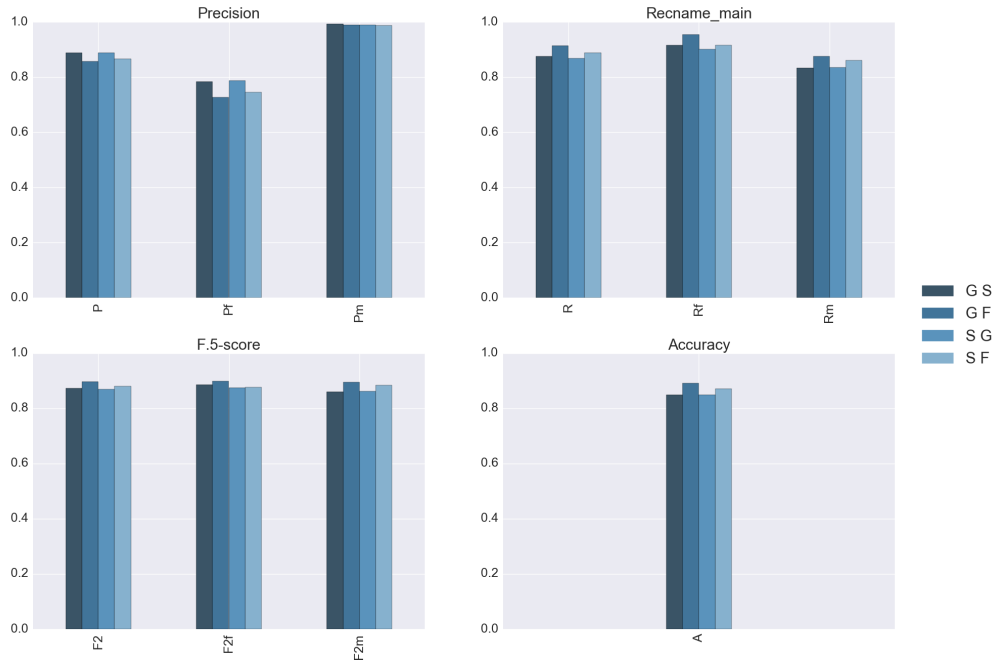

Figure 11: DBLP validation - Two classifier combination.

	A	F_2	F_{2f}	F_{2m}
G S	0.850	0.873	0.886	0.861
G F	**0.890**	**0.897**	**0.898**	**0.895**
S G	0.848	0.869	0.875	0.862
S F	0.871	0.880	0.876	0.883
G S F	**0.893**	**0.898**	**0.898**	0.899
G F S	**0.893**	**0.898**	**0.898**	0.899
S G F	0.892	0.894	0.887	**0.900**
S F G	**0.893**	0.897	0.893	**0.900**

Table 6: DBLP validation - Accuracy and F_2-score for combined classification.

5.2.2 Wiki validation - Combined classification

Results of the combined classification of the Wiki validation set follow the same pattern as with individual classification (Refer to figures 13 and 14 and tables 7 and 8 for graphical and numeric results). Overall, recall for dual classification is around 2%-3% higher, but with three classifiers it stalls at 91%, similar to the results for the DBLP validation set. The best combination for the

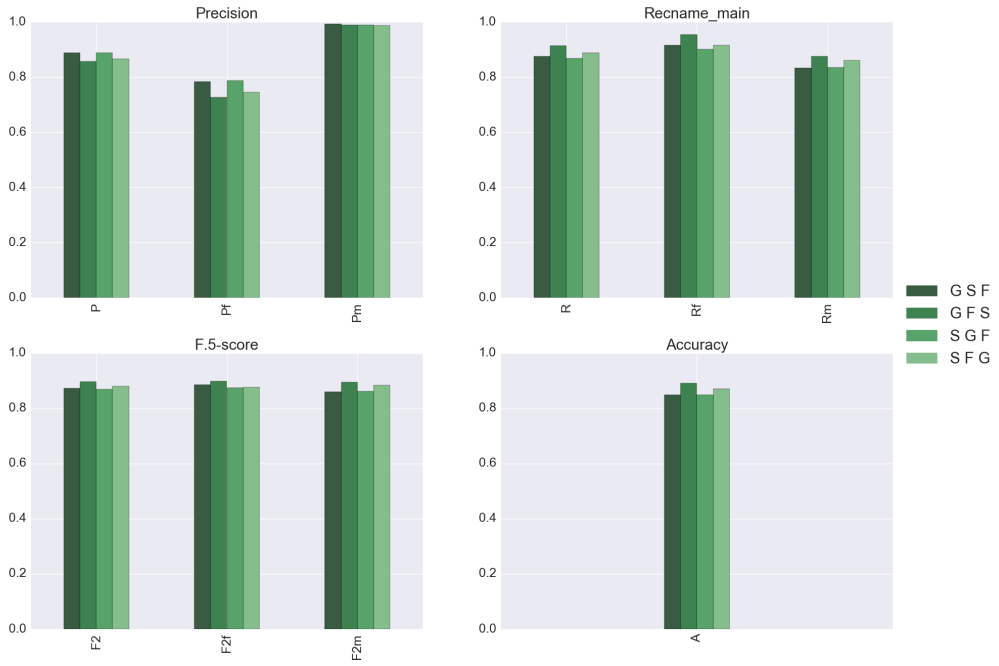

Figure 12: DBLP validation - Three classifier combination.

Wikipedia validation set is **S G F**, again with other combinations being almost on par. The main difference for this validation set is that results do not in- or decrease THAT much when adding a third classifier. This can, once more, mainly be attributed to relatively less unique names and more Western names.

	P	P_f	P_m	R	R_f	R_m
G S	0.895	0.798	0.991	**0.903**	0.900	**0.906**
G F	0.890	0.790	0.991	0.902	**0.902**	0.903
S G	0.907	0.825	0.989	0.902	0.891	0.913
S F	**0.931**	**0.869**	**0.993**	0.858	0.844	0.873
G S F	0.891	0.791	**0.991**	**0.910**	**0.908**	0.912
G F S	0.890	0.790	**0.991**	**0.910**	**0.908**	0.912
S G F	**0.904**	**0.818**	0.989	0.909	0.899	**0.919**
S F G	0.903	0.816	0.989	**0.910**	0.900	**0.919**

Table 7: Wiki validation - Precison and recall for combined clasification.

To conclude, a combination of classifiers with main classification through name yields the best

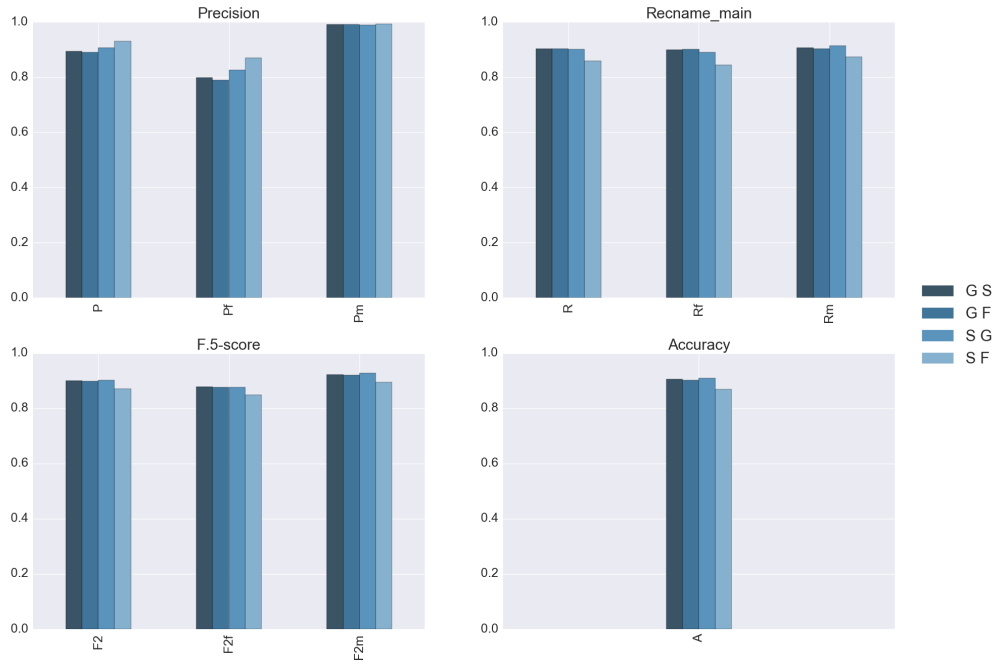

Figure 13: Wiki validation - Two classifier combination.

results. Comparing the evaluation on both sets, image classification mainly increases accuracy when data contains more or less frequent occurring (non-Western) names.

	A	F_2	F_{2f}	F_{2m}
G S	0.905	0.900	**0.877**	0.922
G F	0.903	0.898	**0.877**	0.919
S G	**0.910**	**0.902**	**0.877**	**0.927**
S F	0.869	0.872	0.849	0.895
G S F	0.912	0.905	**0.882**	0.927
G F S	0.911	0.904	**0.882**	0.927
S G F	**0.916**	**0.907**	**0.882**	**0.933**
S F G	**0.916**	**0.907**	**0.882**	0.932

Table 8: Wiki validation - Accuracy and F_2-score for combined classification.

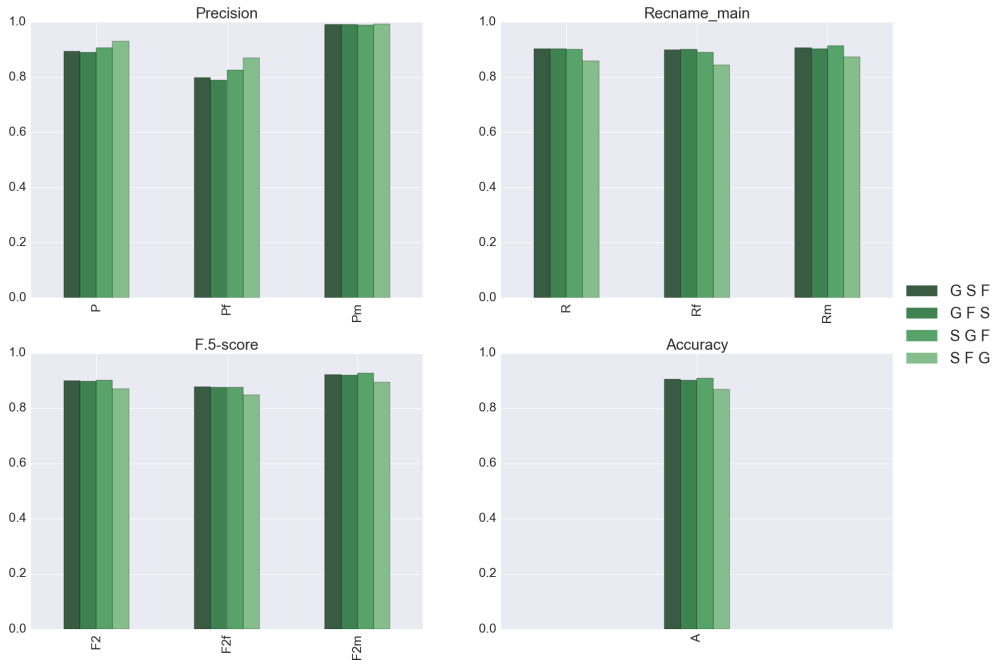

Figure 14: Wiki validation - Three classifier combination.

5.2.3 Final classification

The classification for the DBLP bibliography data will take place through a combined classifier. Due to the data not containing direct image links, the images had to be gathered through the method described in section 4.1.4. Due to performance reasons, extremely low recall (below 5%) and only a fraction of the set containing images, the image classification will not be applied to the actual data set. The final classification will therefore be done through a combination of Genderize.io and SexMachine. The performance of these two combinations is nearly equal. Combined classification with the order of Genderize.io and SexMachine will be used due to performing better on the DBLP validation set (see figure 15).

The result of assigning gender through the chosen combined classifier can be seen in figure 16. 75.8% of the authors were assigned a gender. Finally, nearly all authors in the DBLP bibliography were assigned one authorship. Only 1,461 authors have zero authorships within the DBLP publications. Furthermore, every publication has at least one female or male authorship assigned to it, meaning there are no publications where no gender could be assigned to any of the authors. These results allow that nearly all authors, for whom gender was inferred, and all publications are to be used in the analysis. An important note is that all analysis in the next section is done on the set of authors and publications for which gender was assigned (see

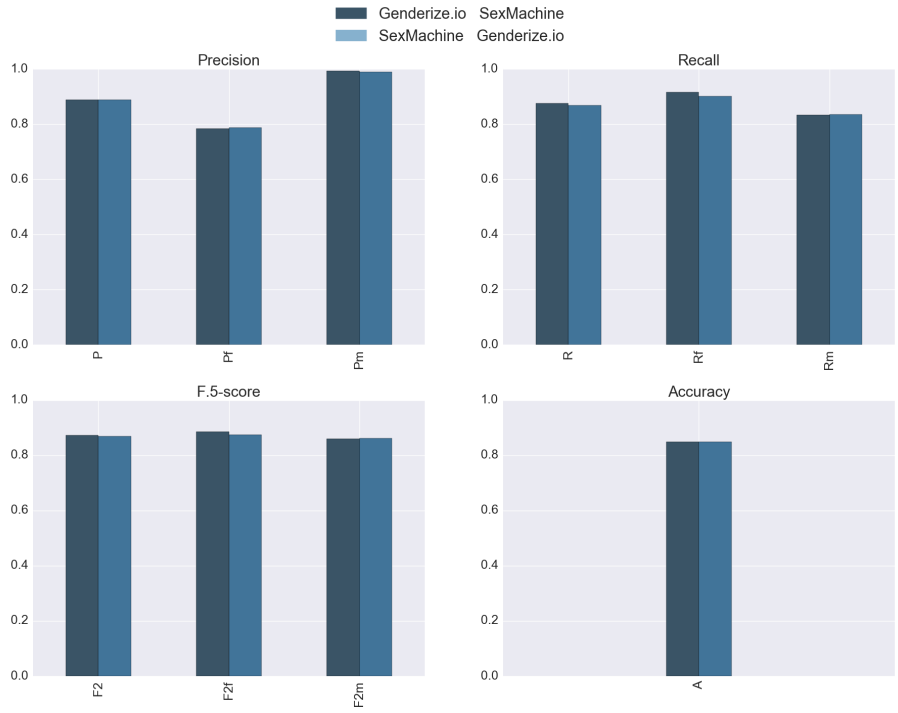

Figure 15: DBLP validation - Combination of name classifiers.

figure 17). This means that our conclusions could be incorrect, as the gender distribution of the unclassified authors and publication authorships could differ a lot.

5.3 Gender disparities within DBLP

Analysing gender disparities within DBLP will be done through scientific output. Thus authorships, and most importantly gender distribution of authorships, will be the main factor in finding inequalities. Aspects of the DBLP publications to be analysed are general authorships, average authorships per author, dominant position and differences in distribution between types of publications. Most of these factors can also be analysed historically, meaning the change in gender distribution over time.

5.3.1 General authorships

Authorship in general for DBLP shows the first disparity. While figure 18 is included for completeness, figure 19 shows the gender distribution of authorships for the classified data. While females make up 26.1% of all authors, they only account for 20.9% of all authorships.

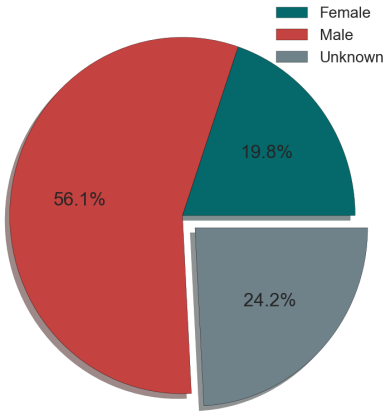

Figure 16: Result gender classification of authors.

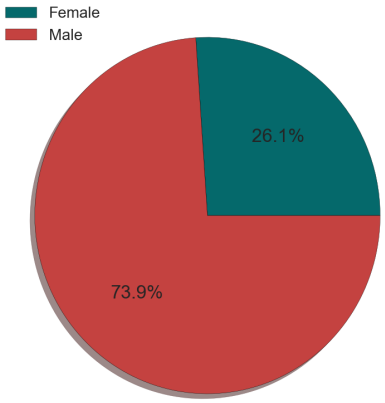

Figure 17: Final gender distribution of classified authors.

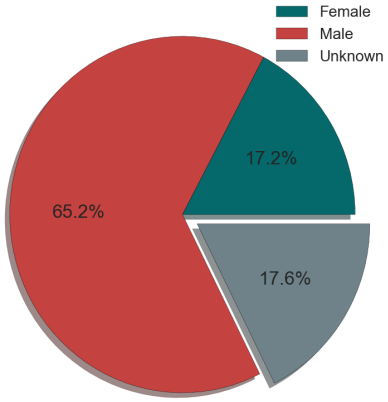

Figure 18: Result gender classification on authorships.

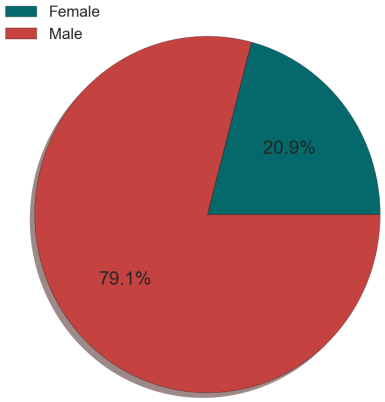

Figure 19: Final gender distribution of classified authorships.

As female scientists have a relatively lower total scientific output, they have 1.57 less average authorships than male scientists. Thus relative output is lower as well, lying 0.66 authorships below the average relative output of males. A point of interest is the average authorships for authors with unknown gender, see figure 20. Expectation was that it would be similar to the average. It is, however, lower than THE males' and females'. An explanatory reason could be

that, taking into account that the unknown group most likely includes more non-Western and thus less frequent occurring names, are affiliated with smaller universities that publish less on an international level. Finally, by taking the quotient of absolute authorships divided by the relative authorships minus one, we get the average number of co-authors. On average, women have 0.16 co-authors more, which equals 9%.

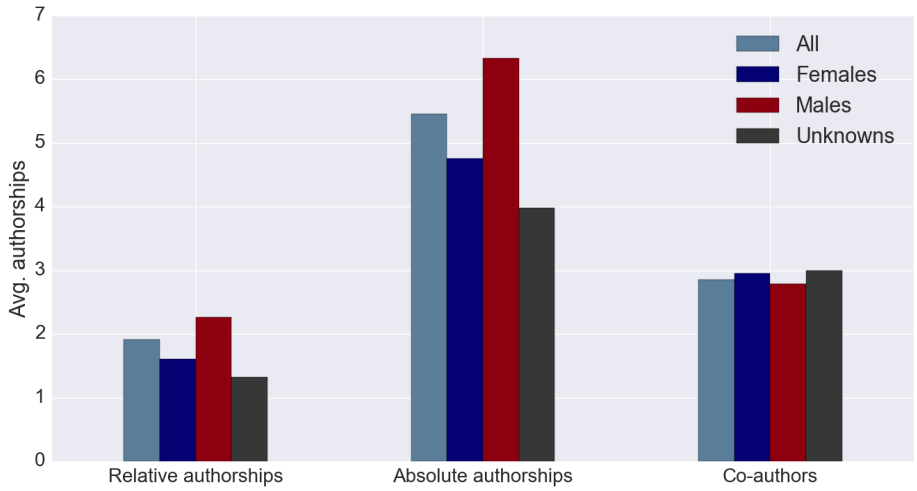

Figure 20: Average authorships and co-authors.

Finally, if we look at the authorships gender distribution per year (figure 21 we see that female authorships, related to male authorships, increases each year. However, this does not say anything about the trend in average authorships, since the data does not incorporate the increase of female and male scientists.

5.3.2 Dominant positions

First and last authorship position seem to follow the same pattern as the general authorship graph does (refer to figure 22 and 23). As relative female authorship increases, so do first and last female authorships. Interesting is that first authorships are, on average, accounted to women 30% of the time, while women concist of only 26.1% of the total amount of authors. This means that women, on average, occupy first authorship relatively more often than males do. The same is not true for last authorship as it is, on average, occupied 78% of the time by male scientists. The reason for this pattern could be that a last authorship position is often associated with senior researchers, which are males more often.

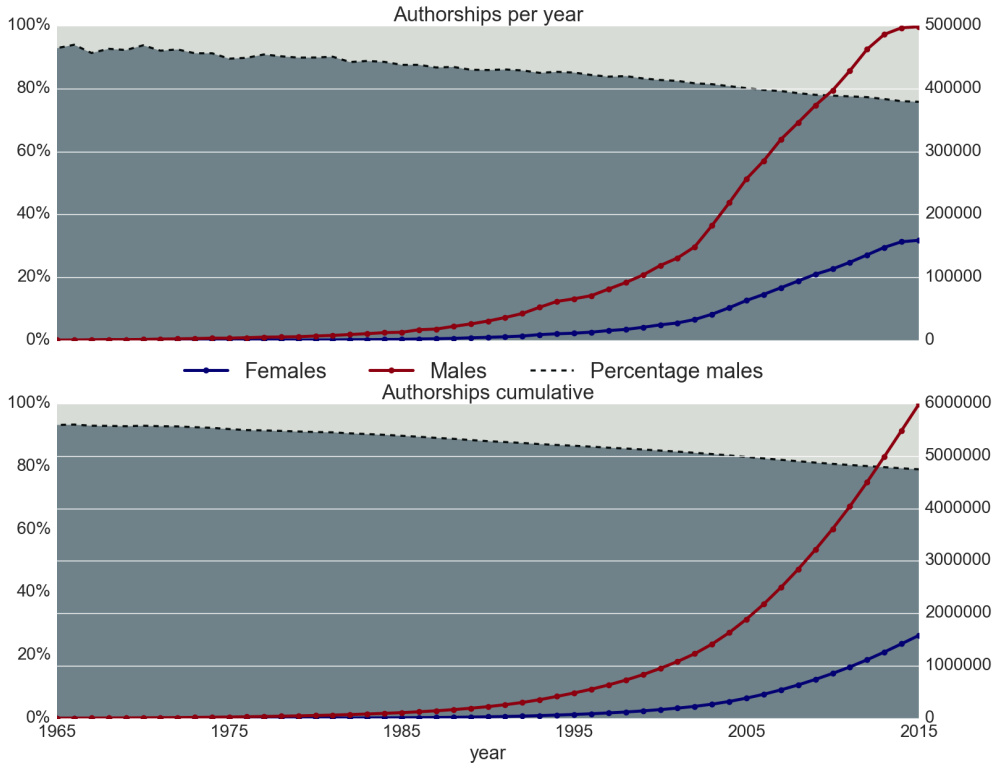

Figure 21: Average authorships over time.

5.3.3 Journal articles and conference papers.

Between the two main types of publications in DBLP, journal articles and conference and workshop papers (in proceedings), no difference is present regarding distribution or trends. Again, by looking at 24, one can see that the patterns followed seem very similar to the trend of general authorship gender distribution in figure 21. As the percentage of female authorships lies around 24%, the average authorship for both types of publications lies below that of male authorships.

For the last part of this analysis, conferences and workshops will be analysed specifically. In figure 25 we can see that conference size does seem to influence authorship gender distribution. In this analysis all conferences and workshops were included which have more than 250 total authorships. These were then distributed over three groups, based on total authorship. Be aware that the average number of authorships is a lot lower for the small group in comparison to the middle group. The same is true for the middle group in comparison to the large group. The lower graph shows that gender distribution is roughly the same for these three groups. In this graph, the distribution of the bigger AI conferences, included in the DBLP bibliography, is also shown. It has about 3.5% more male authorships compared to the average. Finally, the last two

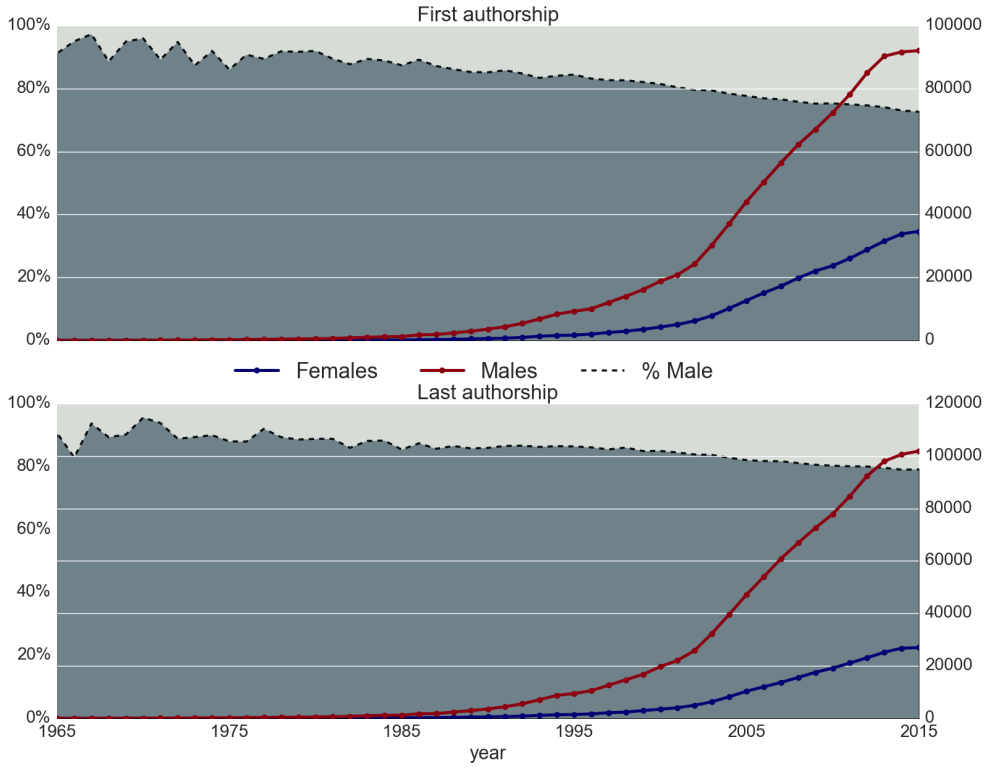

Figure 22: Dominant authorship position per year.

bars in the lower graph are the two extremes out of all conferences. One has nearly 100% male authorship, while the other just has above 30% male authorship. These two conferences are appropriately, but sadly, titled: 'Fachtagung Prozessrechner' and 'Nursing Informatics'.

6 Conclusion

Analyzing the gender disparities within the DBLP bibliography required the inference of gender authors. The gender classification that was performed, was done by name and profile pictures. Individual classification performed best when gender was classified by name. A reason for this is that not every author has a suitable picture available. Genderize.io performed gender classification better than the other two classifiers, obtaining an F_2-score of 0.864 on the DBLP validation set and 0.892 on the Wikipedia validation set. The classification of the authors in the actual DBLP data set was performed by a combined classifier. Triple classification, with Genderize.io being the main classifier, performed better than the other combinations. The final F_2-score achieved by this combined classifier, resulting from the DBLP validation, is 0.873.

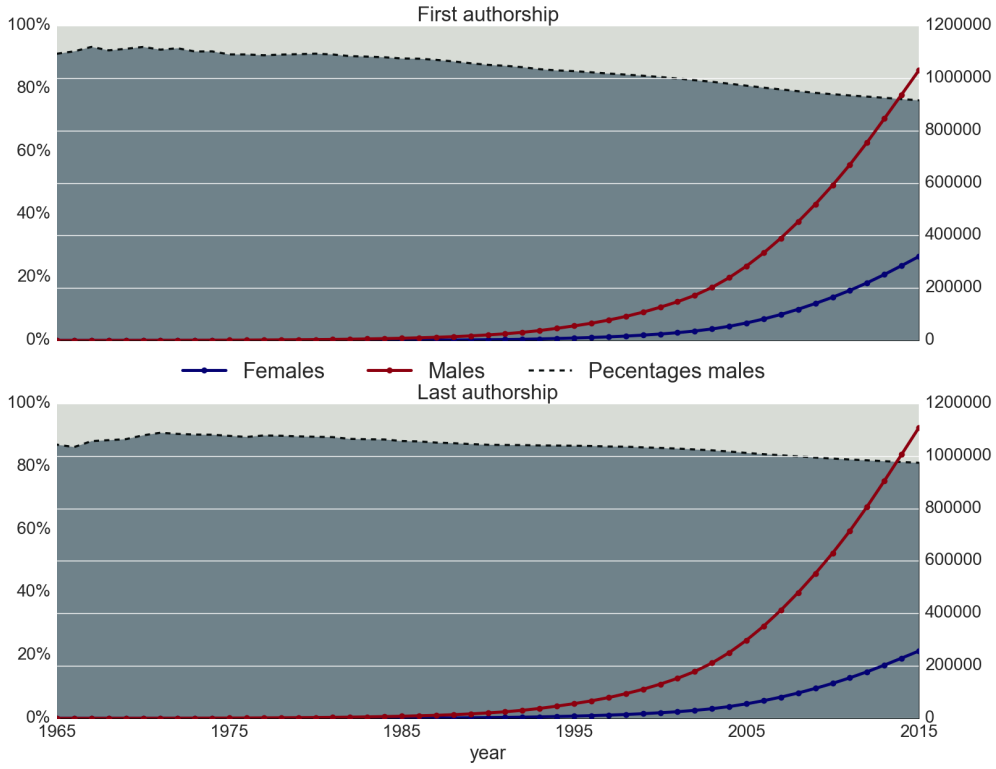

Figure 23: Dominant authorship position cumulative.

These accuracy results are on par or above that of similar contemporary research involving gender classification.

Due to images being unavailable or unsuitable in the actual DBLP data set, a combination of the two classifiers that classify gender by first name was used. The main classifier of this combination is Genderize.io, while SexMachine performed secondary classification. The resulting $\boldsymbol{F_2}$-score is 0.873. Applying the final classification on the DBLP authors resulted in inferring gender for 75.8% of the authors. The percentage of females and males in the set of classified authors is, respectively, 26.1% and 73.9%.

The analysis of gender disparities within the DBLP bibliography was done for the set of authors and publication authorships for which gender was inferred. In terms of general authorships, women are less productive than male scientists, having on average 1.57 less authorships. Trends in the change in gender distribution regarding authorships, show that absolute females authorship is increasing. However, as no temporal data was used on the total amount of female and male scientists, we cannot say if average female authorship is also increasing over time.

The results of gender distribution, regarding first and last authorship, are similar to those of gen-

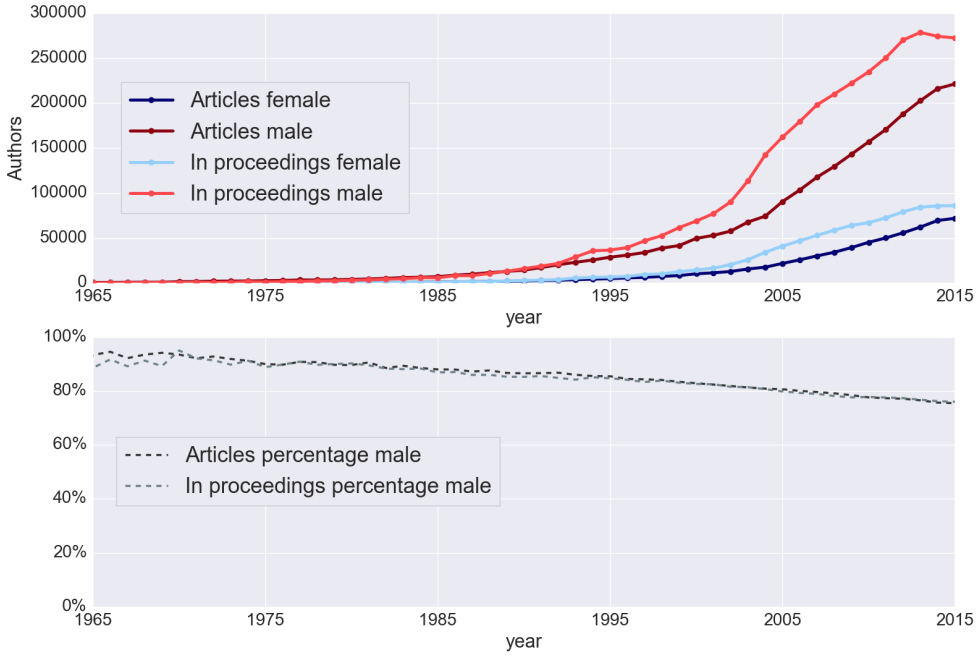

Figure 24: Journal articles and workshop papers.

eral authorships. First authorship position is relatively more often claimed by female scientists than male scientists. This is the only aspect within the analysis where a disparity favours females instead of males. Last authorship is, once again, more often claimed by males.

Lastly, the type of publication does not seem to influence the gender distribution of authorships. No differences could be found between journal articles and conference and workshop papers. The gender distribution of both is equal, meaning they follow the pattern of general authorship gender distribution. Finally, conferences were analysed. Within these results, once again no large differences could be found compared to general authorship. Conference size does not affect gender distribution. The subject of the conference, at least for artificial intelligence, does affect gender distribution. Relative male authorships of AI conferences, consisting of AAAI and IJCAI, lie 3.5% above the average conference authorship gender distribution.

Concluding this section, gender inference with minimal data can obtain in good accuracy results. Being able to classify gender correctly in many cases, gender disparities within the DBLP bibliography could be analysed. In most aspects, women are relatively underrepresented compared to men. Why women are relatively less productive on average cannot be said from the perspective of this research. It is up to other disciplines to perform this analysis. Hopefully, when academia has provided us with enough insight on gender disparities, solutions will be found to make all scientific disciplines, and maybe even all of society, more egalitarian.

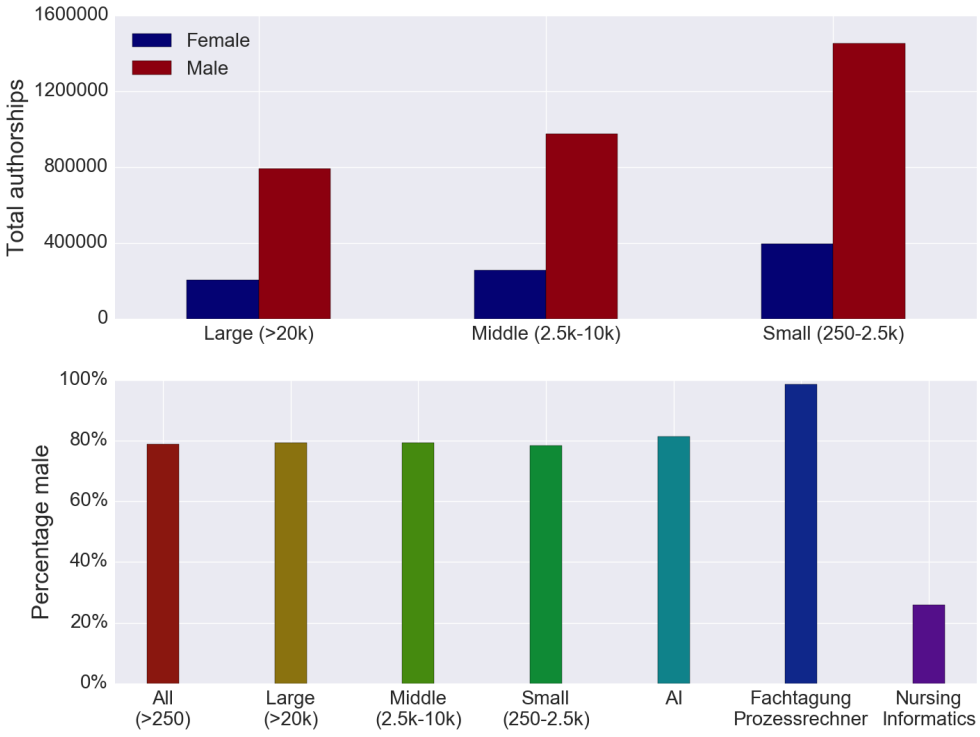

Figure 25: Conferences.

6.1 Discussion

The performance of gender classification based on first names is high on average. When a data set includes many Western names it performs very well, but its performance degrades when many non-Western names are included. Increasing this performance could be done by collecting data as country of origin. Another way to increase the performance is combining it with gender classification based on images. Unfortunately, the DBLP data set does not directly include pictures of the authors. URLs of homepages were scarcely available. For this set of authors, for which an URL was provided, crawling the homepages for profile pictures was not a reliable method. For most of these cases, no face was recognized in any of the images. The performance, measured in speed, was also very low. A more reliable method would be querying the authors name on Google images, and using the first 3 or 5 pictures of these results for gender classification through images.

The analysis of gender disparities was performed based solely on gender distribution of authorships. Additional information that could be included is the authors h-index and citation data. This data would have to be scraped and Google proved to have efficient methods to prevent this.

Still, these factors could be distributed differently compared to authorships and thus able to provide valuable insights regarding gender disparities. The difference between average authorships for females and males in all aspects could only be compared based on cumulative authorships. It could not be done per year, since no temporal data was gathered on the gender distribution of absolute authors per year. This would have been possible if during the parsing of the data the earliest publication date of the author had been registered for him or her. Unfortunately this aspect was thought of too late, and time did not permit re-parsing the data set.

Finally, the methods used in this research could be applied to all data sets where first names, and optionally pictures, are available. Many scientific bibliographies could be analysed by this, but also other data sets. Think of online petitions, does the subject of a petition affect the gender distribution of all persons that sign the petition?

References

Cassidy R Sugimoto, Vincent Lariviere, CQ Ni, Yves Gingras, Blaise Cronin, et al. Global gender disparities in science. *Nature*, 504(7479):211–213, 2013.

Catherine Ashcraft and Sarah Blithe. Women in it: The facts. *National Center for Women in Information*, 2009.

Cedric Herring. Does diversity pay?: Race, gender, and the business case for diversity. *American Sociological Review*, 74(2):208–224, 2009.

Lydna Gratton, Elisabeth Kelan, Andreas Voigt, Lamia Walker, and Hans-Joachim Wolfram. Innovative potential: men and women in teams. *The Lehman Brothers Centre for Women in Business. London Business School*, 2007.

J McGrath Cohoon, Sergey Nigai, and Joseph Jofish Kaye. Gender and computing conference papers. *Communications of the ACM*, 54(8):72–80, 2011.

Michael Ley. Dblp computer science bibliography, 2005.

Moshe Y Vardi. What can be done about gender diversity in computing? a lot!, 2015.

Douglas Baumann, Susanne Hambrusch, and Jennifer Neville. Gender demographics trends and changes in u.s. cs departments. *Commun. ACM*, 54(11):38–42, November 2011. ISSN 0001-0782. doi: 10.1145/2018396.2018410. URL http://doi.acm.org/10.1145/2018396.2018410.

Bin Lin and Alexander Serebrenik. Recognizing gender of stack overflow users. In *Proceedings*

of the 13th International Workshop on Mining Software Repositories, pages 425–429. ACM, 2016.

Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. *arXiv preprint arXiv:1603.04322*, 2016.

C Blevins and L Mullen. Jane, john... leslie? a historical method for algorithmic gender prediction. *Digital Humanities Quarterly*, 9(3), 2015.

Ergin Elmacioglu and Dongwon Lee. On six degrees of separation in dblp-db and more. *ACM SIGMOD Record*, 34(2):33–40, 2005.

Claudio Schifanella, Luigi Di Caro, Mario Cataldi, and Marie-Aude Aufaure. D-index: a web environment for analyzing dependences among scientific collaborators. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1520–1523. ACM, 2012.

Michael Ley. Dblp xml requests, 2009.

Appendices

A Validation results

All accuracy measures for both validation sets are listed in this section. The row labels stand for the corresponding classifiers, S being SexMachine, G being Genderize.io and F being the image classifier, F.A.C.E. The order of classifiers indicates which classification was used first, e.g. **G S** means Genderize's classification was first used and all remaining unknown cases were assigned SexMachine's classification. Figures 26 and 27 and tables 9, 10 and 11 are results for the DBLP validation set. The results for the Wikipedia validation set can be found in figures 28 and 29 and tables 12, 13 and 14.

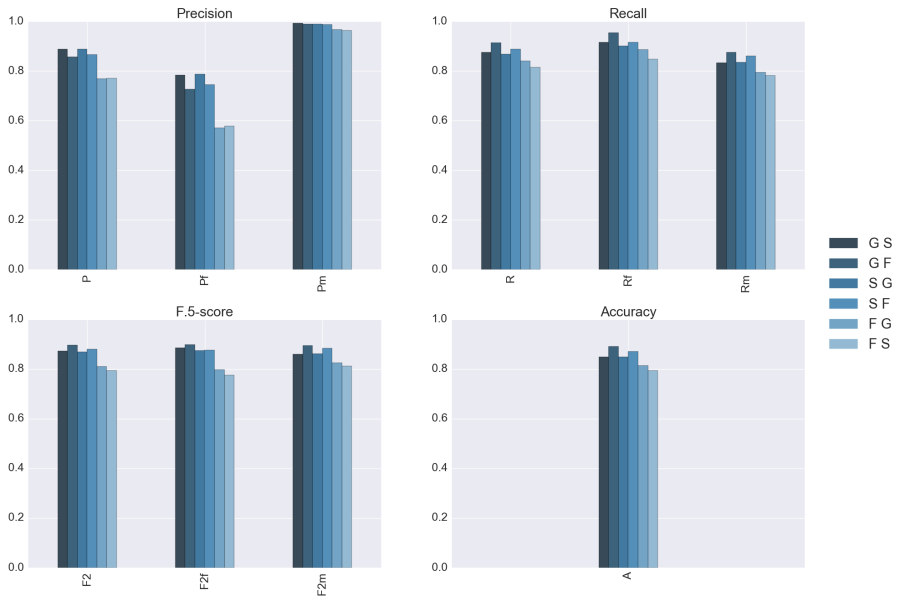

Figure 26: DBLP validation - double classification.

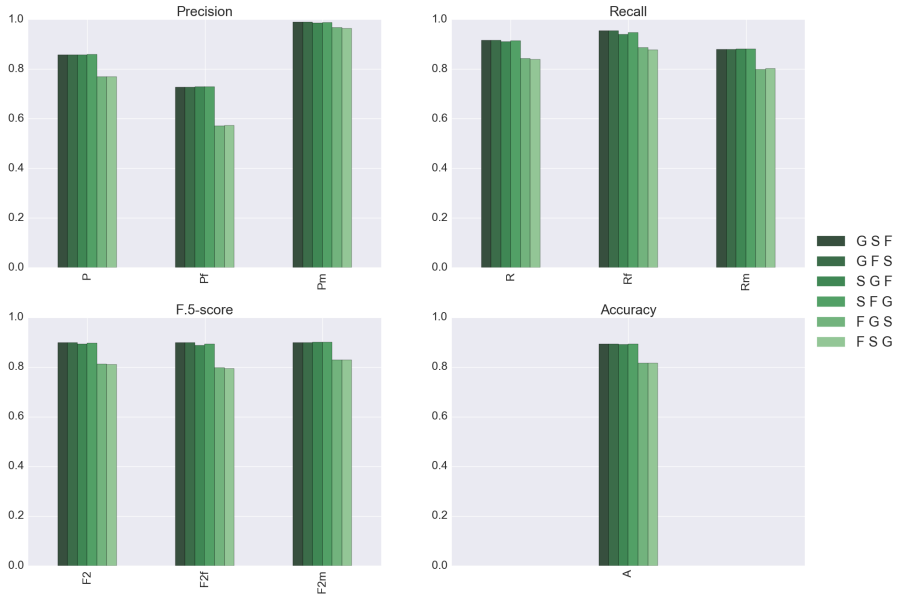

Figure 27: DBLP validation - triple classification.

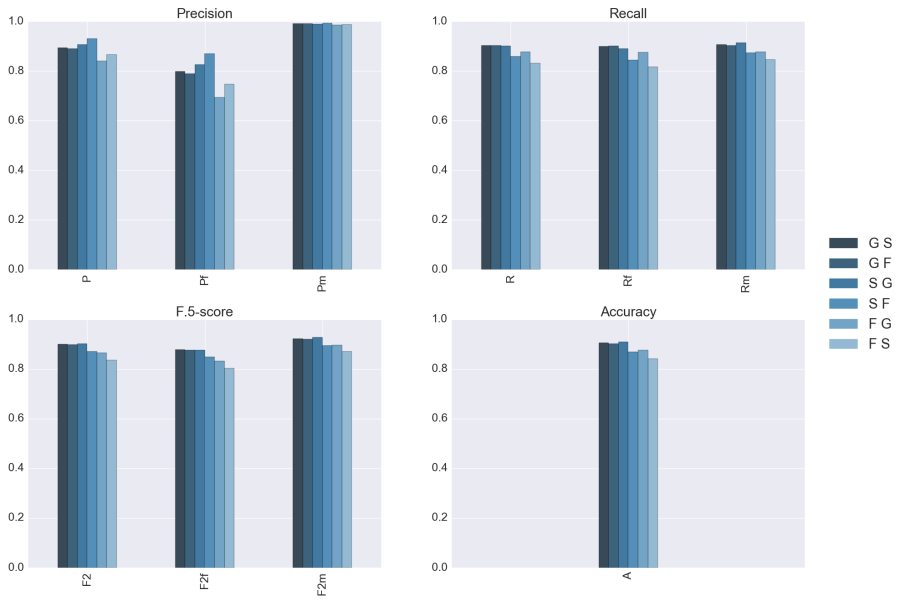

Figure 28: Wikipedia validation - double classification.

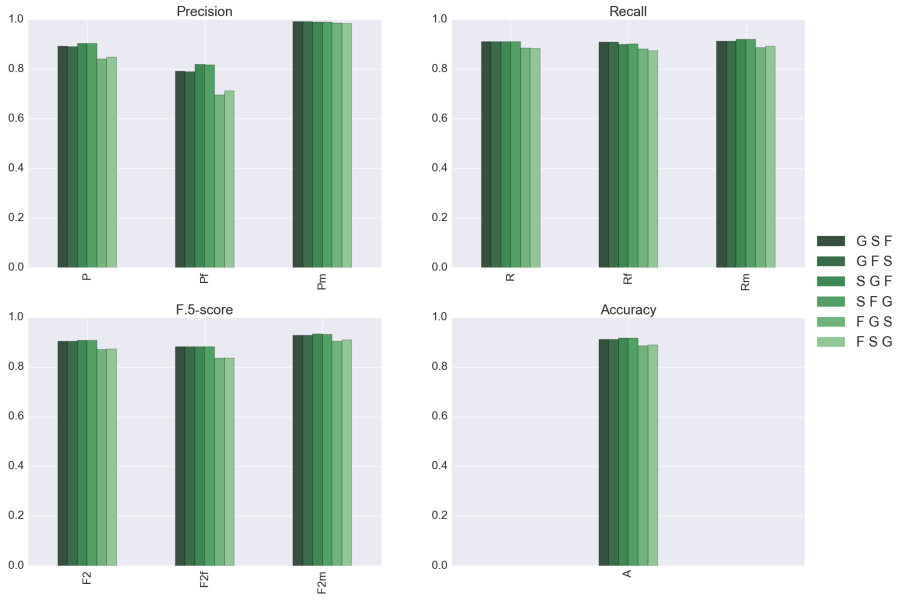

Figure 29: Wikipedia validation - triple classification.

	P_f	P_m	R_f	R_m
G	0.784	0.993	0.916	0.811
S	0.908	0.989	0.824	0.734
F	0.483	0.954	0.557	0.583
G S	0.784	0.993	0.916	0.833
G F	0.727	0.989	0.954	0.875
S G	0.787	0.988	0.901	0.835
S F	0.745	0.986	0.916	0.861
F G	0.571	0.966	0.885	0.795
F S	0.578	0.963	0.847	0.781
G S F	0.727	0.989	0.954	0.879
G F S	0.727	0.989	0.954	0.879
S G F	0.728	0.984	0.939	0.881
S F G	0.729	0.987	0.947	0.881
F G S	0.571	0.966	0.885	0.799
F S G	0.572	0.964	0.878	0.801

Table 9: DBLP validation - Precision and Recall rates.

	$F_{.5f}$	$F_{.5m}$	F_{1f}	F_{1m}	F_{2f}	F_{2m}
G	0.808	0.950	0.845	0.893	0.886	0.842
S	0.890	0.925	0.864	0.842	0.840	0.774
F	0.497	0.846	0.518	0.723	0.541	0.632
G S	0.808	0.956	0.845	0.906	0.886	0.861
G F	0.763	0.964	0.825	0.928	0.898	0.895
S G	0.807	0.953	0.840	0.905	0.875	0.862
S F	0.774	0.958	0.822	0.919	0.876	0.883
F G	0.615	0.926	0.695	0.872	0.798	0.824
F S	0.617	0.920	0.687	0.863	0.775	0.812
G S F	0.763	0.965	0.825	0.931	0.898	0.899
G F S	0.763	0.965	0.825	0.931	0.898	0.899
S G F	0.762	0.962	0.820	0.930	0.887	0.900
S F G	0.764	0.963	0.824	0.931	0.893	0.900
F G S	0.615	0.928	0.695	0.875	0.798	0.828
F S G	0.615	0.926	0.693	0.875	0.793	0.829

Table 10: DBLP validation - F_β-scores.

	A	$F_{.5}$	F_1	F_2	P	R
G	0.832	0.879	0.869	0.864	0.889	0.864
S	0.752	0.907	0.853	0.807	0.948	0.779
F	0.580	0.671	0.621	0.586	0.719	0.570
G S	0.850	0.882	0.876	0.873	0.889	0.875
G F	0.890	0.863	0.877	0.897	0.858	0.914
S G	0.848	0.880	0.873	0.869	0.887	0.868
S F	0.871	0.866	0.871	0.880	0.866	0.888
F G	0.813	0.771	0.784	0.811	0.769	0.840
F S	0.795	0.769	0.775	0.794	0.771	0.814
G S F	0.893	0.864	0.878	0.898	0.858	0.916
G F S	0.893	0.864	0.878	0.898	0.858	0.916
S G F	0.892	0.862	0.875	0.894	0.856	0.910
S F G	0.893	0.864	0.877	0.897	0.858	0.914
F G S	0.817	0.771	0.785	0.813	0.769	0.842
F S G	0.817	0.771	0.784	0.811	0.768	0.840

Table 11: DBLP validation - Combined gender mean accuracy measures.

	P_f	P_m	R_f	R_m
G	0.798	0.991	0.892	0.895
S	0.891	0.993	0.821	0.859
F	0.444	0.951	0.153	0.127
G S	0.798	0.991	0.900	0.906
G F	0.790	0.991	0.902	0.903
S G	0.825	0.989	0.891	0.913
S F	0.869	0.993	0.844	0.873
F G	0.694	0.985	0.875	0.878
F S	0.746	0.987	0.818	0.847
G S F	0.791	0.991	0.908	0.912
G F S	0.790	0.991	0.908	0.912
S G F	0.818	0.989	0.899	0.919
S F G	0.816	0.989	0.900	0.919
F G S	0.695	0.985	0.882	0.886
F S G	0.712	0.984	0.874	0.892

Table 12: Wikipedia validation - Precision and Recall rates.

	$F_{.5f}$	$F_{.5m}$	F_{1f}	F_{1m}	F_{2f}	F_{2m}
G	0.815	0.970	0.842	0.941	0.871	0.913
S	0.876	0.963	0.855	0.921	0.834	0.883
F	0.322	0.415	0.228	0.225	0.177	0.154
G S	0.816	0.973	0.846	0.947	0.877	0.922
G F	0.810	0.972	0.842	0.945	0.877	0.919
S G	0.838	0.973	0.857	0.950	0.877	0.927
S F	0.864	0.966	0.856	0.929	0.849	0.895
F G	0.724	0.962	0.774	0.928	0.832	0.897
F S	0.760	0.955	0.780	0.911	0.802	0.871
G S F	0.812	0.974	0.846	0.950	0.882	0.927
G F S	0.811	0.974	0.845	0.950	0.882	0.927
S G F	0.833	0.974	0.857	0.953	0.882	0.933
S F G	0.831	0.974	0.856	0.953	0.882	0.932
F G S	0.726	0.964	0.777	0.933	0.837	0.905
F S G	0.739	0.964	0.785	0.936	0.836	0.909

Table 13: Wikipedia validation - F_β-scores.

	A	$F_{.5}$	F_1	F_2	P	R
G	0.895	0.893	0.891	0.892	0.894	0.894
S	0.853	0.920	0.888	0.859	0.942	0.840
F	0.131	0.368	0.226	0.165	0.698	0.140
G S	0.905	0.895	0.896	0.900	0.895	0.903
G F	0.903	0.891	0.893	0.898	0.890	0.902
S G	0.910	0.905	0.903	0.902	0.907	0.902
S F	0.869	0.915	0.893	0.872	0.931	0.858
F G	0.877	0.843	0.851	0.865	0.840	0.876
F S	0.842	0.857	0.846	0.837	0.867	0.832
G S F	0.912	0.893	0.898	0.905	0.891	0.910
G F S	0.911	0.893	0.897	0.904	0.890	0.910
S G F	0.916	0.904	0.905	0.907	0.904	0.909
S F G	0.916	0.903	0.904	0.907	0.903	0.910
F G S	0.886	0.845	0.855	0.871	0.840	0.884
F S G	0.889	0.852	0.860	0.873	0.848	0.883

Table 14: Wikipedia validation - Combined gender mean accuracy measures.