Computational Social Choice: Spring 2019

Sirin Botan

Institute for Logic, Language and Computation
University of Amsterdam

May 8, 2019

Plan for Today

We'll look at strategic behavior in Judgment Aggregation—focus on manipulation of the outcome by agents. We've seen this in voting, but what does it look like in JA...?

- What does it mean for an agent to prefer one outcome over another?
- When do agents have an incentive to manipulate?
- How does manipulation in JA relate to manipulation in voting?

We will also go over some other types of strategic actions.

Premise-Based rule: Example

Suppose the agents only care about the outcome of the conclusion.

	a	b	$c \leftrightarrow (a \wedge b)$	c
Agent 1	Yes	Yes	Yes	Yes
Agent 2	Yes	No	Yes	No
Agent 3	No	Yes	Yes	No
Majority	Yes	Yes	Yes	Yes

Preferences of Agents

In voting, you submit your preferences over outcomes, in JA you submit one outcome only.

- Preferences could be completely independent from the true judgment of the agent...
- ...But we usually assume they are not.
- (We could explicitly elicit the agents' preferences over all possible outcome, but there are exponentially many possible outcomes!)

So we have ways of inferring the preferences from the judgments.

Closeness-respecting Preferences

Let $\succeq_i$ be the preference order of agent i over outcomes.

- $\succeq_i$ is top-respecting iff $J_i \succeq_i J$ for all $J \in 2^\Phi$.
- $\succeq_i$ is closeness-respecting iff $(J_i \cap J') \subseteq (J_i \cap J)$ implies $J \succeq_i J'$ for all $J, J' \in 2^\Phi$.

If $\succeq_i$ is closeness-respecting, then it is top-respecting.

Example:

If $J_i = \{a, b, c\}$, $J = \{a, b, \neg c\}$, $J' = \{a, \neg b, \neg c\}$: $J \succ_i J'$.

What if $J = \{a, b, \neg c\}$, $J' = \{a, \neg b, c\}$?

Hamming Preferences

The most commonly used closeness-respecting preference order is the one induced by the *Hamming distance*. We call these Hamming preferences:

- $J \succeq_i J'$ iff $H(J, J_i) \leqslant H(J', J_i)$,

where $H(J, J_i) = |J \setminus J_i|$ is the *Hamming distance*.

Strategyproofness

Let J_i be agent i's truthful judgment set.

- A manipulation is when she reports a set $J_i' \neq J_i$.
- She has incentive to do so in a profile $\boldsymbol{J}$ if there is some judgment set $J_i' \neq J_i$, such that $F(\boldsymbol{J}_{-i}, J_i') \succ_i F(\boldsymbol{J}_{-i}, J_i)$.
- A rule F is strategyproof for a class of preferences, if no agent with preferences in that class ever has an incentive to manipulate.

Axioms: One Old and One New

Note: $\boldsymbol{J} =_{-i} \boldsymbol{J'}$ means for all agents $j \neq i$, $J_j = J'_j$.

- Independence: for any $\varphi \in \Phi$ and any two profiles $\boldsymbol{J}$ and $\boldsymbol{J'}$, if $\varphi \in J_i \Leftrightarrow \varphi \in J'_i$ for all $i \in N$, then $\varphi \in F(\boldsymbol{J}) \Leftrightarrow \varphi \in F(\boldsymbol{J'})$.
- Monotonicity: Additional support should not "harm".
  - for any $\varphi \in \Phi$ and profiles $\boldsymbol{J}$ and $\boldsymbol{J'}$, $\boldsymbol{J} =_{-i} \boldsymbol{J'}$, and $\varphi \in J'_i \setminus J_i$ for some agent $i \in N$: $\varphi \in F(\boldsymbol{J}) \Rightarrow \varphi \in F(\boldsymbol{J'})$.

A Characterization Result

Theorem (Dietrich and List, 2007) F is strategyproof for *all closeness-respecting preferences* iff F is independent and monotonic.

F. Dietrich & C. List. Strategy-proof Judgment Aggregation. *Economics and Philosophy*, 23(3), 2007.

Independent and Monotonic Rules

Recall quota rules from yesterday:

$$F_q(\boldsymbol{J}) = \{\varphi \mid |N_\varphi^{\boldsymbol{J}}| \geqslant q(\varphi)\}.$$

These are the main class of Independent & Monotonic rules. Known that they cannot not guarantee a consistent and complete outcome.

🌟 Can you think of any other Independent & Monotonic rules?

Proof.

Theorem (Dietrich and List, 2007) F is strategyproof for *all closeness-respecting preferences* iff F is independent and monotonic.

- ⇐ *Independence* means we can look at each formula individually. Monotonicity means it's always better to accept a formula you like. ✓
- ⇒ Suppose F is strategyproof for the class of closeness-respecting preferences. Need to show Monotonicity and Independence.

Proof cont. ➡

Monotonicity: for any $\varphi \in \Phi$ and profiles $\boldsymbol{J}$ and $\boldsymbol{J'}$, $\boldsymbol{J} =_{-i} \boldsymbol{J'}$, and $\varphi \in J'_i \setminus J_i$ for some agent $i \in N$: $\varphi \in F(\boldsymbol{J}) \Rightarrow \varphi \in F(\boldsymbol{J'})$.

Take $\varphi \in \Phi$, $i \in N$, $\boldsymbol{J} =_{-i} \boldsymbol{J'}$, with $\varphi \notin J_i$ and $\varphi \in J'_i$, and $\varphi \in F(\boldsymbol{J})$.

Define preference relation $\succeq_i$ such that $J \succeq_i J'$ iff J_i agrees with J but not J' on φ, or agrees with both on φ, or agrees with neither on φ. This is a closeness-respecting preference, and thus, F *is strategyproof for agents with such preferences*.

Since $\varphi \in F(\boldsymbol{J})$, J_i disagrees with $F(\boldsymbol{J})$ on φ, and thus, since F is strategyproof, must disagree with $F(\boldsymbol{J'})$ on φ, so $\varphi \in F(\boldsymbol{J'})$. ✓

Proof cont. ➡

Independence: for any $\varphi \in \Phi$ and any two profiles $\boldsymbol{J}$ and $\boldsymbol{J'}$, if $\varphi \in J_i \Leftrightarrow \varphi \in J'_i$ for all $i \in N$, then $\varphi \in F(\boldsymbol{J}) \Leftrightarrow \varphi \in F(\boldsymbol{J'})$.

Take $\varphi \in \Phi$ and two profiles $\boldsymbol{J}, \boldsymbol{J'}$ such that for all $i \in N$: J_i and J'_i agree on φ.

$$(J_1, \ldots, J_n) \to (J'_1, \ldots, J_n) \to \cdots \to (J'_1, \ldots, J'_n).$$

- $J \succeq_i J'$ iff J_i agrees with J but not J' on φ, or agrees with both on φ, or agrees with neither on φ.

Suppose for contradiction that at step k, the collective judgment on φ changes. Then agent k can manipulate the rule (either with J_k as her truthful judgment set or J'_k), which contradicts our assumption of SP. ✓

Group Manipulation

A rule is group-strategyproof if there is no $C \subseteq N$ such that for some $\boldsymbol{J} =_{-C} \boldsymbol{J'}$, where $\boldsymbol{J}$ is the "truthful" profile, $F(\boldsymbol{J'}) \succ_i F(\boldsymbol{J})$ for all $i \in C$.

Quota rules are not strategyproof for groups of manipulators with Hamming preferences.

	φ_1	φ_2	φ_3	$\neg\varphi_1$	$\neg\varphi_2$	$\neg\varphi_3$
Agent 1	No	Yes	Yes	Yes	No	No
Agent 2	Yes	No	Yes	No	Yes	No
Agent 3	Yes	Yes	No	No	No	Yes
Agent 4	No	No	No	Yes	Yes	Yes
Agent 5	No	No	No	Yes	Yes	Yes
Majority	No	No	No	Yes	Yes	Yes

S. Botan, A. Novaro, & U. Endriss. Group Manipulation in Judgment Aggregation. *AAMAS*, 2016.

Connection to Gibbard-Satterthwaite Theorem

Theorem (Gibbard-Satterthwaite) Any resolute SCF for $\geqslant 3$ alternatives that is surjective and strategyproof is a dictatorship.

Theorem (Dietrich & List) For a conjunctive, disjunctive or preference agenda, an aggregation rule F returns a consistent and complete outcome, satisfies responsiveness and strategyproofness for all closeness-respecting preferences if and only if F is a dictatorship.

Responsiveness: for any $\varphi \in \Phi$ there exists two profiles $\boldsymbol{J}$ and $\boldsymbol{J'}$ such that $\varphi \in F(\boldsymbol{J})$ and $\varphi \notin F(\boldsymbol{J'})$.

A. Gibbard. Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4), 1973.
M.A. Satterthwaite. Strategy-proofness and Arrow's Conditions. *Journal of Economic Theory*, 10, 1975
F. Dietrich & C. List. Strategy-proof Judgment Aggregation. *Economics and Philosophy*, 23(3), 2007.

Other Forms of Strategic Behavior

- Bribery: given a budget & costs (of agents), can I bribe some of the agents to get a more preferred outcome?
- Control: Can I get a more preferred outcome by deleting or adding agents?
- Agenda Manipulation: Can I add or remove items from the agenda to get a more preferred outcome?

D, Baumeister, G, Erdélyi, O, Erdélyi & J, Rothe. Bribery and Control in Judgment Aggregation. *COMSOC*, 2012.
F. Dietrich. Judgment Aggregation and Agenda Manipulation. *Games and Economic Behavior*, 95, 2016.

Last Slide

Summary:

- We defined several types of preferences for agents based on their true judgments
- We proved the characterization result by Dietrich & List
- We saw an impossibility result related to the Gibbard-Satterthwaite Theorem
- We noted some examples of other strategic behaviors

Next week: Advanced Axiomatics of Judgment Aggregation & Complexity of JA.