Computational Social Choice: Spring 2019

Sirin Botan

Institute for Logic, Language and Computation
University of Amsterdam

May 8, 2019
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.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>( c \leftrightarrow (a \land b) )</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent 1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Agent 2</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Agent 3</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Majority</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
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 $J = \{a, b, \neg c\}$, $J' = \{a, \neg b, c\}$?
The most commonly used closeness-respecting preference order is the one induced by the \textit{Hamming distance}. We call these Hamming preferences:

\begin{itemize}
  \item $J \succeq_i J'$ iff $H(J, J_i) \leq H(J', J_i)$,
\end{itemize}

where $H(J, J_i) = |J \setminus J_i|$ is the \textit{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 $J$ if there is some judgment set $J_i' \neq J_i$, such that $F(J_{-i}, J_i') \succ_i F(J_{-i}, J_i)$.
- 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: $J = _i J'$ means for all agents $j \neq i$, $J_j = J'_j$.

- Independence: for any $\varphi \in \Phi$ and any two profiles $J$ and $J'$, if $\varphi \in J_i \iff \varphi \in J'_i$ for all $i \in N$, then $\varphi \in F(J) \iff \varphi \in F(J')$.
- Monotonicity: Additional support should not “harm”.
  - for any $\varphi \in \Phi$ and profiles $J$ and $J'$, $J = _i J'$, and $\varphi \in J'_i \setminus J_i$ for some agent $i \in N$: $\varphi \in F(J) \Rightarrow \varphi \in F(J')$. 
A Characterization Result

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

Recall quota rules from yesterday:

\[ F_q(J) = \{ \varphi \mid |N^J_\varphi| \geq q(\varphi) \}. \]

These are the main class of Independent & Monotonic rules. Known that they cannot 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 \( J \) and \( J' \), \( J =_{-i} J' \), and \( \varphi \in J'_i \setminus J_i \) for some agent \( i \in N \): \( \varphi \in F(J) \Rightarrow \varphi \in F(J') \).

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

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

Since \( \varphi \in F(J) \), \( J_i \) disagrees with \( F(J) \) on \( \varphi \), and thus, since \( F \) is strategyproof, must disagree with \( F(J') \) on \( \varphi \), so \( \varphi \in F(J') \). \( \checkmark \)
Proof cont. ➤

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

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

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

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

Suppose for contradiction that at step $k$, the collective judgment on $\varphi$ 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 \( J = C \upharpoonright J' \), where \( J \) is the “truthful” profile, \( F(J') \succ_i F(J) \) for all \( i \in C \).

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

<table>
<thead>
<tr>
<th></th>
<th>( \varphi_1 )</th>
<th>( \varphi_2 )</th>
<th>( \varphi_3 )</th>
<th>( \neg \varphi_1 )</th>
<th>( \neg \varphi_2 )</th>
<th>( \neg \varphi_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent 1</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Agent 2</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Agent 3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Agent 4</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Agent 5</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Majority</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Connection to Gibbard-Satterthwaite Theorem

Theorem (Gibbard-Satterthwaite) Any resolute SCF for $\geq 3$ 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 $J$ and $J'$ such that $\varphi \in F(J)$ and $\varphi \not\in F(J')$.

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?

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.