Computational Social Choice: Spring 2015

Ulle Endriss
Institute for Logic, Language and Computation
University of Amsterdam

Plan for Today

So far we have (implicitly) assumed that agents truthfully report their judgments and have no interest in the outcome of the aggregation.

What if agents instead are strategic? Questions considered:

- What does it mean to prefer one outcome over another?
- When do agents have an incentive to manipulate the outcome?
- What is the complexity of this manipulation problem?
- What other forms of strategic behaviour might we want to study?

F. Dietrich and C. List. Strategy-Proof Judgment Aggregation. *Economics and Philosophy*, 23(3):269–300, 2007.

U. Endriss, U. Grandi, and D. Porello. Complexity of Judgment Aggregation. *Journal of Artificial Intelligence Research (JAIR)*, 45:481–514, 2012.

D. Baumeister, G. Erdélyi, O.J. Erdélyi, and J. Rothe. Computational Aspects of Manipulation and Control in Judgment Aggregation. Proc. ADT-2013.

Example

Suppose we use the premise-based procedure (with premises = literals):

	p	q	$p \vee q$
Agent 1	No	No	No
Agent 2	Yes	No	Yes
Agent 3	No	Yes	Yes

If agent 3 only cares about the conclusion, then she has an incentive to *manipulate* and pretend she accepts p.

Strategic Behaviour

What if agents behave *strategically* when making their judgments?
Meaning: what if they do not just truthfully report their judgments (implicit assumption so far), but want to get a certain outcome?

What does this mean? Need to say what an agent's *preferences* are.

- Preferences could be completely *independent* from true judgment.

  But makes sense to assume that there are some *correlations*.

- Explicit *elicitation* of preferences over all possible outcomes (judgment sets) not feasible: exponentially many judgment sets.

  So should consider ways of *inferring* preferences from judgments.

Preferences

True judgment set of agent $i \in \mathcal{N}$ is J_i. The preferences of i are modelled as a weak order $\succeq_i$ (transitive and complete) on 2^Φ.

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

Thus: closeness-respecting $\overset{\Rightarrow}{\not\Leftarrow}$ top-respecting

Hamming Preferences

Example for a closeness-respecting preference order:

$$J \succeq_i^H J' \quad \textit{iff} \quad H(J, J_i) \leqslant H(J', J_i),$$

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

We say that agent i *has Hamming preferences* in this case.

Strategy-Proofness

Each agent $i \in \mathcal{N}$ has a truthful judgment set J_i and preferences $\succeq_i$.

Agent i is said to *manipulate* if she reports a judgment set $\neq J_i$.

Consider a resolute judgment aggregation rule $F : \mathcal{J}(\Phi)^n \to 2^\Phi$.

Agent i has an *incentive to manipulate* in the (truthful) profile $\boldsymbol{J}$ if $F(\boldsymbol{J}_{-i}, J_i') \succ_i F(\boldsymbol{J})$ for some $J_i' \in \mathcal{J}(\Phi)$.

Call F *strategy-proof* for a given class of preferences if for no truthful profile any agent with such preferences has an incentive to manipulate.

Example: strategy-proofness for all closeness-respecting preferences

Remark: No reasonable rule will be strategy-proof for preferences that are not top-respecting (even if you are the only agent, you should lie).

Strategy-Proof Rules

Strategy-proof rules exist. Here is a precise characterisation:

Theorem 1 (Dietrich and List, 2007) F *is strategy-proof for closeness-respecting preferences iff* F *is independent and monotonic.*

Recall that F is both independent and monotonic *iff* it is the case that $N_\varphi^{\boldsymbol{J}} \subseteq N_\varphi^{\boldsymbol{J'}}$ implies $\varphi \in F(\boldsymbol{J}) \Rightarrow \varphi \in F(\boldsymbol{J'})$.

How to read the theorem exactly? In its strongest possible form:

- If F is independent and monotonic, then it will be strategy-proof for *all* closeness-respecting preferences.
- Take any *concrete* form of closeness-respecting preferences. If F is strategy-proof for them, then F is independent and monotonic.

Discussion: Is this a positive or a negative result?

F. Dietrich and C. List. Strategy-Proof Judgment Aggregation. *Economics and Philosophy*, 23(3):269–300, 2007.

Proof Sketch

Claim: F is *S-P* for *closeness-respecting* preferences $\Leftrightarrow$ F is *I & M*

($\Leftarrow$) *Independence* means we can work formula by formula.
Monotonicity means accepting a truthfully believed formula is always better than rejecting it. ✓

($\Rightarrow$) Suppose F is *not* independent-monotonic. Then there exists a situation with $N_\varphi^{\boldsymbol{J}} \subseteq N_\varphi^{\boldsymbol{J'}}$ and $\varphi \in F(\boldsymbol{J})$ but $\varphi \notin F(\boldsymbol{J'})$.

One agent must be first to cause this change, so w.l.o.g. assume that only agent i switched from $\boldsymbol{J}$ to $\boldsymbol{J'}$ (so: $\varphi \notin J_i$ and $\varphi \in J_i'$).

If φ (and its complement) is the only formula whose collective acceptance changes, then this shows that manipulation is possible: if others vote as in $\boldsymbol{J}$ and agent i has the true judgment set J_i', then she can benefit by lying and voting as in J_i. ✓

Otherwise: similar argument (see paper for details).

Discussion

So independent and monotonic rule are strategy-proof. But:

- The *only* independent-monotonic rules we saw are the *quota rules*, and they are not consistent (unless the quota is large)
- *None* of the (reasonable) rules we saw that *guarantee consistency* (e.g., max-sum, max-number) are independent.
- The *impossibility* direction of the agenda characterisation result discussed in depth showed that, if on top of independence and monotonicity we want neutrality and if agendas are sufficiently rich (violation of the median property), then the only rules left are the *dictatorships* (which indeed are strategy-proof).

Dietrich and List explore this point and prove a similar result (but w/o using neutrality and for a different agenda property) that is similar to the famous *Gibbard-Satterthwaite Theorem* in voting.

Complexity of Manipulation

So strategy-proofness is very rare in practice. Manipulation is possible.

Idea: But maybe *manipulation* is computationally *intractable*?

For what aggregation rules would that be an interesting result?

- Should *not* be both *independent* and *monotonic* (strategy-proof).
- Should have an *easy winner determination problem* (otherwise argument about intractability providing protection is fallacious).

Thus: *premise-based procedure* is good rule to try

The Manipulation Problem for Hamming Preferences

For a given resolute rule, the manipulation problem asks whether a given agent can do better by not voting truthfully:

$\textsc{Manip}(F)$
Instance: Agenda Φ, profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$, agent $i \in \mathcal{N}$
Question: Is there a $J_i' \in \mathcal{J}(\Phi)$ such that $F(\boldsymbol{J}_{-i}, J_i') \succ_i^H F(\boldsymbol{J})$?

Recall that $\succeq_i^H$ is the preference order on judgment sets induced by agent i's true judgment set and the Hamming distance.

Complexity Result

Consider the premise-based procedure for literals being premises and an agenda closed under propositional variables (so: WinDet is easy).

Theorem 2 (Endriss et al., 2012) $\textsc{Manip}(F_{pre})$ *is NP-complete.*

Proof: *NP-membership* follows from the fact we can verify the correctness of a certificate J_i' in polynomial time.

NP-hardness: next slide

U. Endriss, U. Grandi, and D. Porello. Complexity of Judgment Aggregation. *Journal of Artificial Intelligence Research (JAIR)*, 45:481–514, 2012.

Proof

We prove NP-hardness by reduction from SAT *for formula* φ. Let $p_1, \ldots, p_m$ be propositional variables in φ and let q_1, q_2 be two fresh variables.

Define $\psi := q_1 \vee (\varphi \wedge q_2)$. Construct *agenda* Φ consisting of:

- $p_1, \ldots, p_m$, q_1, q_2
- $m + 2$ syntactic variants of ψ, such as $(\psi \wedge \top)$, $(\psi \wedge \top \wedge \top)$, ...
- the complements of all the above

Consider profile $\boldsymbol{J}$ (with rightmost column having "weight" $m + 2$):

	p_1	p_2	$\cdots$	p_m	q_1	q_2	$q_1 \vee (\varphi \wedge q_2)$
J_1	1	1	$\cdots$	1	0	0	don't care
J_2	0	0	$\cdots$	0	0	1	don't care
J_3	1	1	$\cdots$	1	1	0	**1**
$F_{\text{pre}}(\boldsymbol{J})$	1	1	$\cdots$	1	0	0	**0**

Hamming distance between J_3 and $F_{\text{pre}}(\boldsymbol{J})$ is $m + 3$.

Agent 3 can achieve Hamming distance $\leqslant m + 2$ *iff* φ is satisfiable (by reporting satisfying model for φ on p's and 1 for q_2). ✓.

Bribery and Control

Baumeister et al. (2011, 2012, 2013) also study several other forms of strategic behaviour in judgment aggregation (by an outsider):

- *Bribery:* Given a budget and known prices for the judges, can I bribe some of them so as to get a desired outcome?
- *Control by deleting/adding judges:* Can I obtain a desired outcome by deleting/adding at most k judges?
- *Control by bundling judges:* Can I obtain a desired outcome by choosing which subgroup votes on which formulas?

D. Baumeister, G. Erdélyi, and J. Rothe. How Hard Is it to Bribe the Judges? A Study of the Complexity of Bribery in Judgment Aggregation. Proc. ADT-2011

D. Baumeister, G. Erdélyi, O.J. Erdélyi, and J. Rothe. Control in Judgment Aggregation. Proc. STAIRS-2012.

D. Baumeister, G. Erdélyi, O.J. Erdélyi, and J. Rothe. Computational Aspects of Manipulation and Control in Judgment Aggregation. Proc. ADT-2013.

Summary

This has been an introduction to strategic behaviour in JA:

- *Preferences:* top- or closeness-respecting, Hamming preferences

  Open research question: how to best model preferences in JA?

- *Strategy-proofness possible*, but rare (requires independence and monotonicity for closeness-respecting preferences)

- Good news: *manipulation* is computationally *intractable* for premise-based rule with Hamming preferences

  But: just a worst-case result (no experimental studies to date)

- Briefly: (complexity of) other forms of strategic behaviour

What next?

We will briefly discuss one final topic in JA, namely *truth-tracking*, and then summarise what we have covered in this field.