Computational Social Choice 2022

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

http://www.illc.uva.nl/~ulle/teaching/comsoc/2022/

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*, 2007.

D. Baumeister, J. Rothe, and A.-K. Selker. Strategic Behavior in Judgment Aggregation. In U. Endriss (ed.), *Trends in Computational Social Choice*, 2017.

Example

Suppose we use the premise-based rule (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 that she actually 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.

Note: Do not confuse this with the issue of embedding preferences.

Preferences

Suppose the true judgment set of agent $i \in N$ is J_i. Let us model the preferences of i as a weak order $\succcurlyeq_i$ (transitive and complete) on 2^Φ.

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

Exercise: *Show that closeness-respecting preferences are top-respecting, but that the opposite need not be the case.*

Hamming Preferences

Example for a concrete choice of preference order:

$J \succcurlyeq_i^H J'$ iff $H(J, J_i) \leqslant H(J', J_i)$,

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

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

Exercise: *Show that Hamming preferences are closeness-respecting.*

Strategyproofness

Each agent $i \in N$ has a true judgment set J_i and true preferences $\succcurlyeq_i$.

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 we have $F(\boldsymbol{J}_{-i}, J_i') \succ_i F(\boldsymbol{J})$ for some lie $J_i' \in \mathcal{J}(\Phi)$.

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

Example: "strategyproofness for all closeness-respecting preferences"

Remark: No reasonable rule will be strategyproof for preferences that are not top-respecting (even if you are the only agent, you should lie). So some restrictions on preferences are unavoidable (and perfectly ok).

Strategyproof Rules

Strategyproof rules exist. Here is a precise characterisation:

Theorem 1 (Dietrich and List, 2007) F *is strategyproof for all 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 $\varphi \in F(\boldsymbol{J}) \Rightarrow \varphi \in F(\boldsymbol{J'})$.

Discussion: *Is this a positive or a negative result?*

F. Dietrich and C. List. Strategy-Proof Judgment Aggregation. *Economics and Philosophy* 2007.

Proof

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

($\Leftarrow$) By *independence*, can analyse what happens formula by formula. By *monotonicity*, always better to accept than reject truly held φ. ✓

($\Rightarrow$) Assume F is *not* independent and monotonic. Need to show F is *not* SP for *at least one* choice of closeness-respecting preferences.

By assumption, $N_\varphi^{\boldsymbol{J}} \subseteq N_\varphi^{\boldsymbol{J'}}$ and $\varphi \in F(\boldsymbol{J})$ but $\varphi \notin F(\boldsymbol{J'})$ for some φ.

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 $\varphi \in J_i'$).

Now consider a scenario where *agent i's true judgment set is J_i'* and where she *only cares about φ* (which is closeness-respecting!).

We have found a scenario where agent i has an incentive to lie. ✓

Complexity of Manipulation

The only independent-monotonic rules we saw are the quota rules (which come with their own set of problems).

So strategyproofness is 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* (strategyproof).
- Should have an *easy outcome determination problem* (otherwise argument about intractability providing protection is fallacious).

Thus: *premise-based rule* (with premises = literals) 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:

$\textsc{Manip}(F)$
Input: Agenda Φ, profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$, agent $i \in N$
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 $\succcurlyeq_i^H$ 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 rule for literals being premises and an agenda closed under propositional variables (so: $\textsc{OutDet}$ is easy).

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

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

Exercise: *Any ideas for how to approach the NP-hardness proof?*

U. Endriss, U. Grandi, and D. Porello. Complexity of Judgment Aggregation. *Journal of Artificial Intelligence Research (JAIR)*, 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.

Let $n = 3$. Define $\psi := q_1 \vee (\varphi \wedge q_2)$. Construct an *agenda* Φ consisting of:

- premises $p_1, \ldots, p_m$, 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). ✓

Group Manipulation

A rule is *group-strategyproof* if no coalition $C \subseteq N$ can ever benefit from manipulating: $F(\boldsymbol{J'}) \succ_i F(\boldsymbol{J})$ for all $i \in C$ for some $\boldsymbol{J} =_{-C} \boldsymbol{J'}$.

Quota rules are *not* group-strategyproof for Hamming preferences:

	p	q	r	$\neg p$	$\neg q$	$\neg r$
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

If agents 1–3 swap the highlighted judgments, they'll all do better.

So group-SP is more rare than SP. Botan *et al.* give a characterisation.

S. Botan, A. Novaro, and U. Endriss. Group Manipulation in Judgment Aggregation. AAMAS-2016.

Bribery and Control

Baumeister *et al.* 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?

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. ADT-2011.

D. Baumeister, G. Erdélyi, O.J. Erdélyi, and J. Rothe. Computational Aspects of Manipulation and Control in Judgment Aggregation. 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?
- *Strategyproofness possible*, but rare (requires independence and monotonicity for closeness-respecting preferences)
- Good news: *manipulation* is computationally *intractable* for the premise-based rule with Hamming preferences

  But: just a worst-case result (no empirical studies to date)
- Briefly: (complexity of) other forms of strategic behaviour

What next? Agenda restrictions that guarantee collective consistency.