# Judgment Aggregation June Project: Strategic Behaviour

Zoi Terzopoulou

Institute for Logic, Language, and Computation
University of Amsterdam

7/6/2018

(based on the slides of Ulle Endriss)

# Goals

So far we have (implicitly) assumed that agents just report their judgments *truthfully*. What if agents instead are *strategic*? This brings out game-theoretical considerations...

- What does it mean to *prefer* one outcome over another?
- When do agents have an *incentive to manipulate*?
- What forms of strategic behaviour we might want to study?

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

# Example for the Premise-based Procedure

- You pass the project ($p$) iff you pass both the presentation ($r$) and the paper ($s$): $p \leftrightarrow r \wedge s$.
- Ulle wants you to fail...

| | $r$ | $s$ | $p$ |
|---|---|---|---|
| Sirin: | **Yes** | **Yes** | Yes |
| Ulle: | **Yes** | No | No |
| Zoi: | No | **Yes** | No |
| Committee: | **Yes** | **Yes** | **Yes!** |

| | $r$ | $s$ | $p$ |
|---|---|---|---|
| Sirin: | Yes | Yes | Yes |
| Ulle: | No | No | No |
| Zoi: | **No** | Yes | No |
| Committee: | **No** | Yes | No! |

Ulle lies, and obtains a preferable outcome.

# About Preferences

What does it mean that an agent *prefers* an outcome over another? We 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. (recall example ⭐)
- Explicit elicitation of preferences over all possible outcomes (judgment sets) is hard: exponentially many judgment sets.

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

# Specific Preferences

The true judgment set of agent $i$ is $J_i$, and her preferences of $i$ are modelled as *weak orders* $\succsim_i$ (transitive and complete) on $2^\Phi$.

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

So, closeness-respecting are top-respecting, but not necessarily the other way around. ⭐

A commonly used closeness-respecting preference order is the *Hamming-distance* preference order:

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

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

# Strategy-proofness

Agent $i$ has a truthful judgment set $J_i$ and preferences $\succsim_i$.

She *manipulates* if she reports a judgment set $J_i^* \neq J_i$.

She has an *incentive to manipulate* in the profile $\boldsymbol{J}$ if there exists some judgment set $J_i^* \neq J_i$ such that $F(\boldsymbol{J}_{-i}, J_i^*) \succ_i F(\boldsymbol{J}_{-i}, J_i)$.

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

Note that 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, and we have a precise characterisation of them:

**Theorem (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 that $\varphi \in F(\boldsymbol{J}) \Rightarrow \varphi \in F(\boldsymbol{J}')$.

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

($\Leftarrow$) *Independence* means we can work formula by formula. *Monotonicity* means accepting a truthfully believed formula is always better than rejecting it. ⭐
So these properties together imply strategy-proofness. ✓

($\Rightarrow$) Suppose that $F$ is not independent + monotonic: there exists a situation $N_\varphi^{\boldsymbol{J}} \subseteq N_\varphi^{\boldsymbol{J}'}$ where $\varphi \in F(\boldsymbol{J})$ 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 $\boldsymbol{J}'$: $\varphi \notin J_i$ and $\varphi \in J_i'$.

If $\varphi$ is the only formula of which the 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 (similarly... see paper)

Independent and monotonic rules are strategy-proof. But:

- The *only* independent-monotonic rules we saw are the *quota rules*, and they are only consistent for large quotas.
- *None* of the (reasonable) rules we saw that guarantee consistency (e.g., Kemeny) are independent.
- The impossibility direction of the agenda characterisation result seen yesterday 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)...
- So, any ideas what we can do next? 

# Refined Strategic Behaviour

Remember that so far we have (implicitly) assumed that:

- All the agents *know* what the *actual profile* of judgments is.
- The agent that manipulates thinks that *everyone else remains truthful*.
- Only *one agent* may manipulate *at a time*.

Under these assumptions, strategy-proofness is hard to achieve. But what if we relax them?

Z. Terzopoulou. Manipulating the Manipulators: Richer Models of Strategic Behavior in Judgment Aggregation. MSc Thesis, University of Amsterdam, 2017.

# More Manipulation-related Topics

Of course, the assumptions one could make are endless...
Other people have also studied:

- Manipulation by a *group* instead of an *individual*.
- *Complexity* of manipulation in Judgment Aggregation.

Do you have more ideas? ⭐

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

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

# Other Forms of Strategic Behaviour

These are manipulation cases by an "outsider":

- *Bribery*: Given a budget and known prices for the agents, can I bribe some of them so as to get a preferable outcome?
- *Control by deleting/adding agents*: Can I obtain a preferable outcome by deleting/adding agents in the group?
- *Control by bundling judges*: Can I get a preferable 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.

# Summary of Part A

We formally introduced strategic behaviour in JA:

- *Preferences*: top-,closeness-respecting, Hamming distance.
  Open research question: how best model preferences in JA?
- *Strategy-proofness* possible, but *rare* (requires independence and monotonicity for closeness-respecting preferences).
- Briefly, *refinements* and other *forms* of strategic behaviour.

Next, we move to truth tracking!