GAP.11 handout for *Rational issue polarisation among agents with perfect memory*

Felix Kopecky

DebateLab, Karlsruhe Institute of Technology

15 September 2022

1 What is polarisation?

Groups of agents can polarise in at least two senses: *affectively* and regarding their *issue positions*. That is an important conceptual finding from sociology (Iyengar et al., 2012; Mason, 2013).

Issue polarisation: Differences of opinions on a disputed issue are strengthened. Rising polarisation in this sense is marked by increasing variance, in-group consistency and out-group issue disagreement.

Affective polarisation: This second type tracks agents' attitude and behaviour toward agents that are perceived as out-group strangers. Characteristics include rising anger or perceiving them as a threat.

Since they track different attributes of a population, it is possible that the types of polarisation develop independently of each other in real-world scenarios. And in fact they seem to do so: Mason (2013, p. 142) observes that Americans are now "increasingly angry at each other, while still agreeing on most issues".

Recent computational models of polarisation dynamics (O'Connor & Weatherall, 2018; Singer et al., 2019) turn out to be models of issue polarisation. In these models, agents polarise due to limited access to their epistemic surroundings, including other agents: agents in O'Connor and Weatherall (2018) can only communicate with other agents if their mutual trust is high enough; agents in Singer et al. (2019) can only hold a couple of beliefs in memory.

These limitations question whether these models actually establish the possibility of polarisation dynamics under condition of *epistemic rationality*. Hence the motivation for the present paper: can issue positions polarise in a computational model even if the agents are not restricted in these ways? The answer is "yes".

2 Agent-based debate models

The model presented in this paper belongs to the family of models based on the theory of dialectical structures. Betz (2013) explores agreement and truth-conduciveness dynamics in a similar model.

2.1 Agents and their beliefs

Agents' belief systems are modelled as mappings from the set of sentences under discussion to truth values. In the base model, these are either True or False, e.g.:

$$\text{Position } A: \ \{p_1 \to \text{True}, p_2 \to \text{False}, \dots, p_n \to \dots\} \quad (1)$$

Here, n is the number of items in the sentence pool. In the base model, $n = 20$.

Distances between agents are measured as differences in their belief systems. In the base model, this is the normalised Hamming distance, which is defined as the number of differently evaluated items relative to the total number of items. As an example, let Position A be $\{p_1 \to \text{True}, p_2 \to \text{False}\}$ and Position B be $\{p_1 \to \text{True}, p_2 \to \text{True}\}$, then $\text{HD}(A, B)/n = 1/2$.

2.2 Argumentation

Agents introduce arguments into the publicly shared debate and react to the introductions of others. An argument is modelled as an implication from a set of premises to a conclusion. All sentences are drawn from the sentence pool:

$$\underbrace{((p_a \wedge ... \wedge p_b) \Rightarrow p_c)}_{\text{Argument 1}} \wedge \underbrace{((p_d \wedge ... \wedge p_e) \Rightarrow p_f)}_{\text{Argument 2}} \wedge ... \quad (2)$$

Agents *respond* to introductions by determining whether their belief system still satisfies, in the logical sense, the Boolean formula that expresses the current debate. In other words, they are solving a SAT problem. In case their belief system is UNSAT relative to the current debate, they are moving to a new belief system by solving a MaxSAT problem: which belief system that satisfies the debate has minimum HD to their previous beliefs?

For example, the belief system $\{p_1 \rightarrow \text{True}, p_2 \rightarrow \text{False}, p_3 \rightarrow \text{True}\}$ is valid at the start of a debate, when no arguments are introduced (as all beliefs are). But after introduction of the argument $(p_1 \wedge \neg p_2) \Rightarrow \neg p_3$, this belief system becomes irrational.

Some normative accounts of belief revision support updating to the closest neighbour as a rational choice. This includes Quine and Ullian (1978, pp. 66–67) (*conservatism* as a virtue), or the *coherence theory* by Gärdenfors (1992, p. 8).

At each turn of a model run, two agents are paired for the introduction of an argument. All agents in the population respond to the introduced argument, but it is devised according to the two selected agents' belief systems. Arguments are introduced according to different *strategies*. CONVERT and ATTACK are two such strategies:

ATTACK: Agent A introduces an argument with premises it accepts and a conclusion that is the negation of one of the sentences that B accepts.

CONVERT: Agent A introduces an argument with premises that B accepts to a conclusion that A accepts.

2.3 Measuring polarisation

Debates in the model progress by introductions and responses to introduced arguments. This process continues until the Boolean formula representing the debate is only satisfied by one belief system. At this point, the inferential density D equals 1. Before this point, the density yields a measure of progress: how many belief systems are still available to the agents given the current debate?

After every argument introduction and response, the belief systems of all agents are stored and polarisation measures are applied. Tracking polarisation dynamics consists in plotting these measures against simulation progress in terms of inferential density.

Some polarisation measures work on a population as a whole and do not require prior clustering of agents (Bramson et al., 2017, §§2.1–2.4). But group-based measures often paint a clearer picture of a population's polarisation level. To measure them, antecedent clustering is necessary, for which I used Leiden (Traag et al., 2019) and affinity propagation (Frey & Dueck, 2007), two state-of-the-art, deterministic algorithms agnostic of group size.

These algorithms will structure the population of agents into groups in a list-of-lists format:

$$\Big[\ \underbrace{[a_1, a_4, a_5]}_{\text{Cluster 1}},\quad \underbrace{[a_2, a_3]}_{\text{Cluster 2}},\quad \underbrace{[a_6, a_7]}_{\text{Cluster 3}}\ \Big] \quad (3)$$

Group-based measures, like *group divergence* (Definition 1) are then applied on this structuring.

Definition 1. Group divergence, based on Bramson et al. (2017, §2.7). Let A_τ be the population of agents at debate stage τ. Let δ be the normalised Hamming distance. For a position x_i, $G(x_i)$ is the set of positions of the same group (neighbours), while $G^*(x_i)$ are the out-group positions (strangers) determined by a community structuring algorithm. Note that $|\cdot|$ denotes either the cardinality of a set or the absolute value of a distance, depending on its argument.

$$\text{div}(\tau) := \frac{1}{|A_\tau|} \sum_{i}^{|A_\tau|} \left| \frac{\sum_{j \in G(x_i)} \delta(x_i, x_j)}{|G(x_i)|} - \frac{\sum_{k \in G^*(x_i)} \delta(x_i, x_k)}{|G^*(x_i)|} \right|$$

Note: The egocentric "me" in the measure runs on index i. Its neighbours run on index j, and its strangers on k.

3 Simulation results

Figure 1 shows the evaluation of individual model runs for two argumentation strategies. There, 50 agents de-

bated 20 atomic propositions and were all equipped with the same argumentation strategy.

Tri-polarisation is observable for the ATTACK strategy but not for CONVERT. Evaluating more simulation runs reveals that the CONVERT strategy can lead agents to become more polarised than in the pictured scenario, but the opposite is not true: ATTACK almost never invokes depolarising dynamics.

Figure 2 shows the mean polarisation dynamics from a simulation experiment of 1,000 model runs for each strategy. When agents introduce arguments following the ATTACK strategy, they end up in polarised states significantly more often than when using CONVERT.

4 Conclusion

Simulations run on this model support the case for the possibility of polarisation under condition of epistemic rationality. But they do so without relying on limiting agents' epistemic abilities.

Argumentation strategies had a substantial effect on the obtained levels of issue polarisation. This is evidence for the profound effect that argumentation can have on multi-agent epistemic processes: the way in which agents argue among each other influences their polarisation dynamic. This can motivate increased attention in social epistemology toward argumentation, particularly when it deals with scientific processes.

Importantly, this model is silent on the possibility of *affective* polarisation under epistemic rationality.

References

Betz, G. (2013). *Debate dynamics: How controversy improves our beliefs*. Springer.

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, *84*(1), 115–159.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976.

Gärdenfors, P. (1992). Belief revision: An introduction. In P. Gärdenfors (Ed.), *Belief revision*. Cambridge University Press.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431.

Kopecky, F. (2022). Arguments as drivers of issue polarisation in debates among artificial agents. *Journal of Artificial Societies and Social Simulation*, *25*(1).

Mason, L. (2013). The rise of uncivil agreement: Issue versus behavioral polarization in the American electorate. *American Behavioral Scientist*, *57*(1), 140–159.

O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science*, *8*, 855–875.

Quine, W. v. O., & Ullian, J. S. (1978). *The web of belief* (2nd ed.). McGraw-Hill.

Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., Ranginani, A., & Berger, W. J. (2019). Rational social and political polarization. *Philosophical Studies*, *176*(9).

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*, 5233.

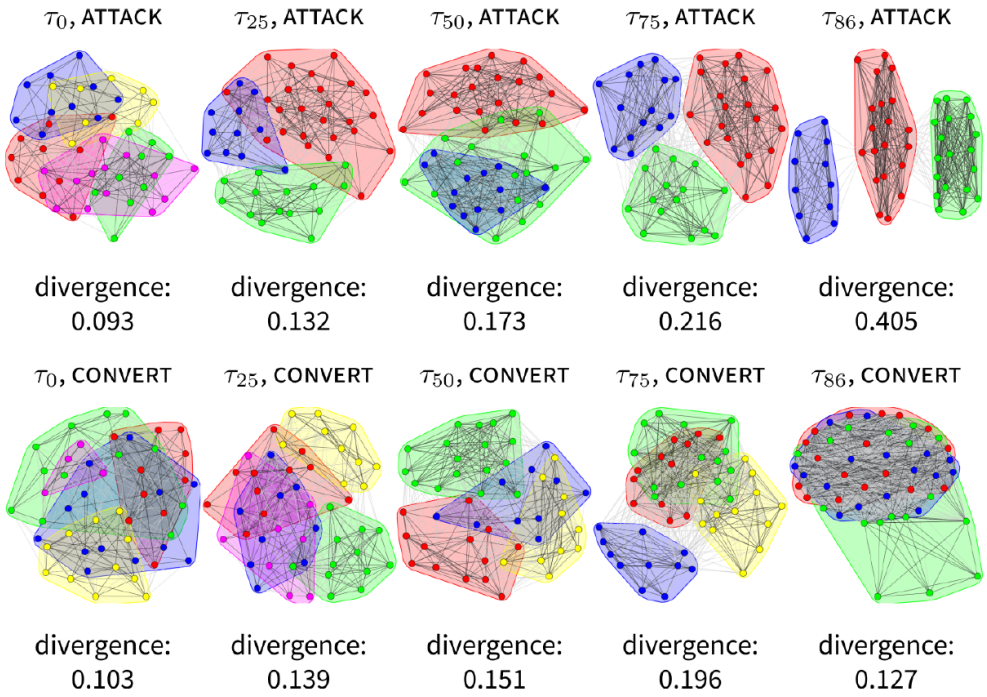

Figure 1: Clusterings obtained on five debate stages from two model runs via the Leiden algorithm. Each node represents an agent, colours signify different clusters (from Kopecky, 2022).

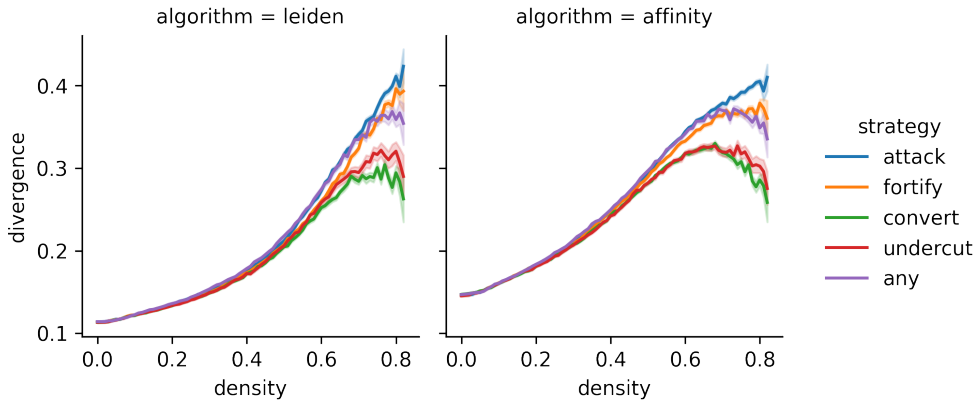

Figure 2: Dynamics of mean group divergence plotted against inferential density for two clustering algorithms and five argumentation strategies (from Kopecky, 2022).