UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Flooding the Zone: An Agent-based Exploration

Permalink

https://escholarship.org/uc/item/8z32x661

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Hahn, Ulrike
Assaad, Leon
Schöppl, Klee

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Flooding the Zone: An Agent-based Exploration

Ulrike Hahn (u.hahn@bbk.ac.uk)
Centre for Cognition, Computation, and Modelling, Birkbeck College, University of London, London, WC1E 7HX U.K.

Leon Assaad (L.Assaad@campus.lmu.de)
Munich Center for Mathematical Philosophy, LMU Munich, Ludwigstraße 31, 80539 Munich, Germany

Klee Schöppl (l.u.schoppl@rug.nl)
Department of Theoretical Philosophy, University of Groningen, Groningen, 9712 GL NL

Abstract

Online public discourse faces many threats such as human and bot networks spreading disinformation or harassment campaigns aimed at excluding certain voices. One such threat is the strategy of 'flooding the zone': intentionally pumping into the discourse information that is irrelevant to, or distracting from, an important issue. This technique is employed by both individual and state actors with seeming success. How and why that technique is successful, by contrast, is less well understood. In this paper we use agent-based modelling to help elucidate the disruptive impact of flooding the zone on communication itself. Specifically, we probe the ways in which flooding hampers the spread of relevant information and show consequences of this even for idealised, rational, actors.

Keywords: Argumentation; Agent-based modelling; Communication; Normative reasoning; Polarisation.

Introduction

Over the last two decades, public discourse has increasingly shifted to online social media (Sunstein, 2018). Not only have online social media connected unprecedented numbers of people (Leetaru, Wang, Cao, Padmanabhan, & Shook, 2013), they have arguably altered even local community exchange (Yardi & Boyd, 2010). The lack of gatekeeping facilitates access to our information ecosystem in novel, unprecedented ways. On the one hand, online social media give ordinary people potentially large influence in ways that may be beneficial for the rapid dissemination of crisis information or for grass roots activism (Tufekci & Wilson, 2012; Tufekci, 2017; Mundt, Ross, & Burnett, 2018). On the other hand, there is now evidence to suggest that online social media have harmed democracy (Lorenz-Spreen, Oswald, Lewandowsky, & Hertwig, 2023), potentially fuelling or polarising distrust (Ceron, 2015; Klein & Robison, 2020), and spreading misinformation (Del Vicario et al., 2016). One area of concern stems from the fact that lack of gatekeeping may make the discourse vulnerable to disruption by hostile actors. This includes state actors who may seek to undermine discourse with targeted mis- and disinformation (Rid, 2020). Disruptive techniques, however, are not limited to misinformation. An important tool in the toolkit of disinformation strategies is 'Flooding the Zone' (FtZ). Appropriating what was originally a sports term, FtZ refers to a strategy of sowing confusion and distrust by pumping out large amounts of distracting, conflicting, or irrelevant information (Starr, 2020; Jasser Jasser & Garibay, 2021). Because human resources, such as attention or communication bandwidth, are finite, FtZ can serve to effectively mask relevant information by making it harder and harder to find. As a result, FtZ is employed as a censorship tool by authoritarian governments in order to suppress unfavourable news stories (for examples, see, e.g., Roberts, 2018; Jasser Jasser & Garibay, 2021). It was also recommended as a strategy for neutralising the influence of traditional media by Steve Bannon in a 2018 quote: "The real opposition is the media. And the way to deal with them is to flood the zone with shit" (Starr, 2020).

While the existence of FtZ as a real-world strategy for information disruption is well-established, we know of little work examining its potential effectiveness. In particular, we know of no research that has sought to explore the impacts of scale: how much distracting information is required in order to successfully degrade information exchange? And how might this change as a function of network structure or discourse characteristics?

Acquiring a better understanding of these issues seems important for anyone concerned about the information environment, last but not least, because the *effects* of FtZ might also be achieved inadvertently by scores of otherwise well-intentioned individual actors contributing and amplifying low quality information due to lack of expertise.

A full understanding of the disruptive impact of irrelevant or distracting information in the real world will require combining observational data, simulations, along with a detailed empirical understanding of human information acquisition and consumption. As such, understanding FtZ constitutes a large-scale research project in its own right. Our goals in this paper are thus necessarily rather modest: specifically, we use agent-based modelling to map the functional relationship between volume of irrelevant information and information acquisition in networks of optimal agents. This, we hope, will help provide a frame of reference for understanding the dynamics and power of FtZ.

An Agent-Based Model of Flooding the Zone

In order to study the impact of FtZ on discourse, we need a modelling framework in which agents exchange individual arguments, pieces of evidence, or reasons in support of a claim, and then aggregate these into an overall belief on the issue at hand. Unfortunately, the bulk of agent-based models of opinion dynamics to date do not provide that level of granularity. Typically, these models involve either the dif-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

fusion of a single numerical opinion via a process of contagion (Centola, 2018) or averaging (Hegselmann & Krause, 2006). A notable exception, here, is the model by Mäs and Flache (2013) which involves agents exchanging 'arguments' in the form of positive or negative valenced vector elements of equal strength (i.e, +1 or -1, with aggregate opinion calculated as the sum across these elements). Another is the recent model by Assaad et al. (2023), which models communication as the exchange of 'arguments' derived from a ground truth world defined by a Bayesian belief network (BN). BNs are graphical representations of multi-variable relationships (Pearl, 1988, 2000; Korb & Nicholson, 2010; Scutari & Denis, 2021) widely used as both theoretical and practical tools. Specifically, BNs summarise the way variables do and, more importantly, do not influence one another in a graphical representation that simplifies Bayesian calculations. BNs thus have a normative (Bayesian) foundation and they connect to extant work on argumentation (such as Bayesian models of argument generation, see Zukerman, McConachy, & Korb, 1998; Zukerman, McConachy, Korb, & Pickett, 1999; Jitnah, Zukerman, McConachy, & George, 2000; Keppens, 2019; Timmer, Meyer, Prakken, Renooij, & Verheij, 2015 or argument quality, see Hahn, 2020). While the use of Bayesian networks as the basis for agent-based information exchange is not new, what is different about Assaad et al.'s NormAN framework (short for *Normative Argument Exchange Across Networks*) is that the framework uses the ground truth 'world' of the underlying BN to stochastically generate a true state of a claim (hypothesis) at issue in the discussion, along with the evidence for it that could be discovered in principle. Agents receive evidence about that world (through inquiry) and may communicate that evidence to others as arguments, as well as receive it in turn. Whatever evidence agents receive, they aggregate optimally via Bayes' rule. To do so, they too, draw on a BN, which in the basic version of the model is a veridical model of 'the world', that is, in essence, a matching (subjective) BN 'in the agent's head'.

This means that the NormAN framework allows one to probe both the diffusion of arguments under different communication rules adopted by agents, as well as to examine the impact on agents' beliefs given a plausible distribution of evidence as generated by the causal structure of the world. Against that baseline, we can assess the disruptive potential of additional irrelevant or poor-quality information.

Simulations

For our initial simulations, we implemented a new version of the model that can be found on OSF.[1] In our extension, rather than importing pre-constructed Bayes' nets from the extant literature as in the base-version of NormAN, we used a simplified BN for the generation of evidence: In each run, the model generates a manually selected number of conditionally independent pieces of evidence (cf. the BN in Assaad & Hahn, 2024). The likelihood ratios of these pieces (or 'tests' of the hypothesis) were randomly assigned before each simulation run: some are therefore more diagnostic than others. In addition, our graphs are extended by a chosen number of non-diagnostic 'flood items'. We also implemented stopping conditions sensitive to whether a piece of evidence is diagnostic, or part of FtZ, so as to allow the exploration of when all (informative) arguments are fully distributed. Finally, we implemented an optional toggle to limit agents' processing capacity to mirror their communicating only one argument per round, as a representation of humans' limited information processing capacities.

More concretely, suppose a simulation in which agents deliberate whether a central hypothesis H is true or false, on the basis of three available pieces of useful evidence (i.e., the observable results of diagnostic tests). NormAN stochastically initializes the hypothesis as true or false in this scenario according to its base rate, and then determines the value of each test by conditionalising on the value of H: if the hypothesis is true (H) and the test has a true positive rate of $P(E|H) = 0.7$, then it has a 70% chance of being positive (i.e., E, instead of $\neg E$). At the start of a simulation, agents may receive these pieces of evidence through inquiry, upon which they 'optimally' update their belief on it ($P(H|E)$), using the underlying probabilities and Bayes' rule. Agents can then communicate these test results to their interlocutors as arguments for or against the hypothesis. Once all pieces of this diagnostic evidence are thus fully distributed across the network of agents, the community of agents will possess all relevant information.

Non-diagnostic 'flood items' work in the same way, except that they do not help the agents reason about the hypothesis, as they are not diagnostic either way. In our example, suppose the agents' discourse is diluted by ten arguments resulting from FtZ, then this may hinder their attempts at fully distributing the useful evidence.

Results

In our basic simulations, we focused on time to convergence: that is, we examined how many time steps it takes before all agents in a network have received all pieces of available evidence. Needless to say, this time to convergence will depend not only on the volume of evidence available and the number of agents, but also on agents' chosen communication rule. For our baseline simulations, we let agents choose what argument to communicate from their available store *at random*. While this memory-less rule is inefficient as agents can repeatedly communicate the same piece of information to the same recipients, it is unbiased. This means that unlike other communication rules that have been examined within the NormAN framework that are sensitive to argument strength (e.g., Assaad et al., 2023; Assaad & Hahn, 2024) or argument polarity (i.e., for or against the hypothesis Schöppl & Hahn, 2024), random sharing leads to eventual convergence of beliefs across agents.

On each run, we initialised the model such that each agent was randomly assigned one piece of evidence from the total

[1] https://osf.io/ux34v/?view_only=309169c

available set, and, at each subsequent step selected a single piece of evidence to communicate to their neighbours (as one might do, for example, on Facebook or Twitter) with probability 1. As can be seen from Fig. 1, for a fixed network topology (i.e., without dynamic rewiring of connections), time to convergence, that is, the number of time steps in the model until all agents have the same set of evidence, is a linear function of the amount of evidence available.

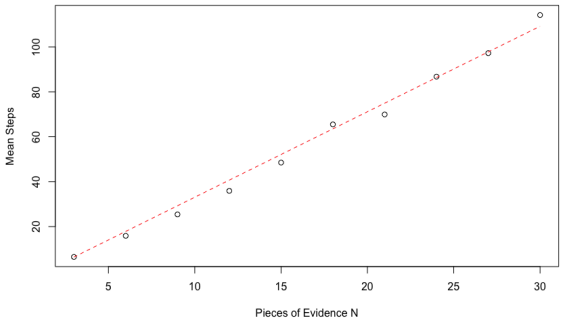


Figure 1: The x-axis shows the number of in principle available pieces of evidence N, the y-axis the mean number of steps to complete knowledge. Each data point represents the average over 30 runs, for a total of 300 runs. This particular simulation involves a perfect lattice of 10 agents, with $k = 2$ (based on Watts and Strogatz (1998), where agents have $k \cdot 2$ neighbours) the red line plots the relationship $\frac{ax}{30} - 5$, where a is the convergence time for 3 items.

At the same time, we can ask how this compares to systematically reducing the probability of communication between agents. Specifically, we can track how the time to complete knowledge changes as we gradually lower the probability that agents will communicate a piece of evidence on a given time step from probability 1 to 0. Figure 2 reveals the functional relationship to be a power law.

Against this backdrop, we can now examine the functional form of increasing the number of irrelevant distractors on the time to convergence with respect to a fixed number of 'real' evidence items, that is, arguments with genuine diagnostic value vis à vis the hypothesis at hand. The results of this are shown in Figure 3, for different numbers of distractors in addition to 10 relevant items. The mean step number now corresponds to the time to complete knowledge for the 10 evidence items themselves (regardless of whether the distractors reach every agent or not). We chose the number of flood items $N = [0, 1, 3, 4, 7, 10, 15, 23, 40, 90]$ in order to correspond roughly to the proportions of real evidence items within the total set (i.e., real evidence + flood distractors) of $p = [.1, .2., .3, .4, .5, .6, .7, .8, .9, 1]$ so as to match the probability of communication plots of Fig 2. The parameters otherwise mirrored those of Figures 1 and 2.

Crucially, the three different panels of Figure 3 plot *the same data* in three different ways, showing the two relationships demonstrated in Figs. 1 and 2 to be one and the same. The middle panel shows the same linear relationship seen in Fig. 1. The middle panel simply reorders these data points in steps that correspond to the evidence proportions, and, in so doing, replicates the basic pattern created by reducing communication probability seen in Fig. 2. For completeness, the left-hand panel plots the results of each individual run (50 per level of N for a total of 500 runs), as opposed to the means, illustrating how the variance increases with the number of communicated items.

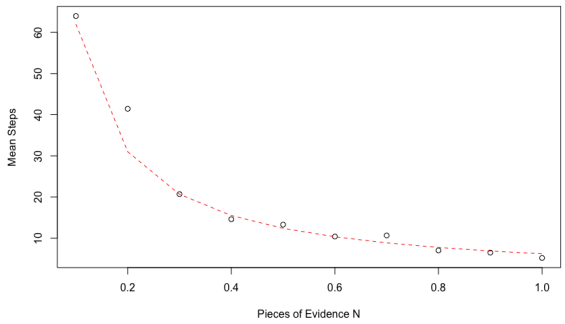


Figure 2: The x-axis shows the probability of communication at each time step, the y-axis mean number of steps to convergence. Each data point represents the average over 30 runs, for a total of 300 runs. Network size and topology as in Fig. 1. The red line plots the relationship $a + \frac{1}{x}$, where a is the convergence time for the 3 items at chattiness $p = 1$ (i.e., agents share one piece of evidence each round).

Consequences for agent beliefs

That helps understand the impact of flooding, but one might ask, nevertheless, why this matters. To appreciate its relevance it helps to understand that slowing the spread of information may lead to a less informed population (as is the very point of FtZ as a method of censorship) and that incomplete information may give rise to polarisation—even among rational agents (see e.g., Mäs & Flache, 2013; Kopecky, 2022, 2024; Assaad et al., 2023; Assaad & Hahn, 2024; Schöppl & Hahn, 2024).

In the real world, deliberation time may itself be fixed: imagine, for example, an election or a referendum. And even where no explicit deadline exists, topics of current interest are overtaken by new events. To illustrate the effects of this, we reran the above simulations, but stopped them at 50 steps, and probed, instead, the accuracy of agents' beliefs and their polarisation. We measured accuracy as the squared deviation from the underlying true state of the world on that model run (the higher the value, the less accurate the population) and polarisation with the variance of beliefs (as in e.g., Olsson, 2013; Schöppl & Hahn, 2024). The latter is an imperfect[2]

[2]For an extended discussion see Hahn, Merdes, & von Sydow, 2024, Appendix.

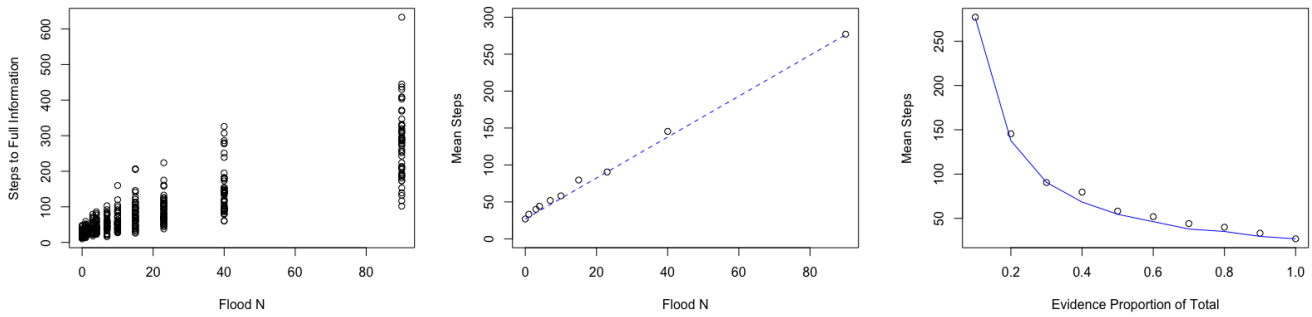
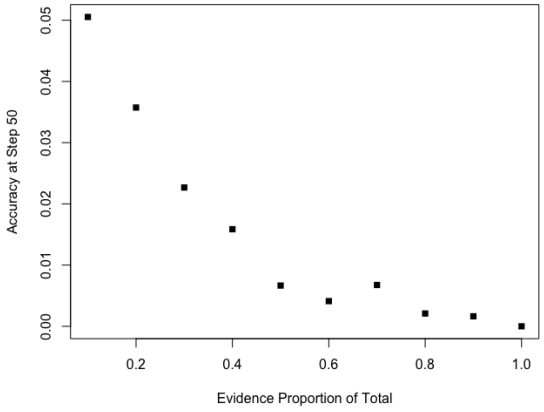
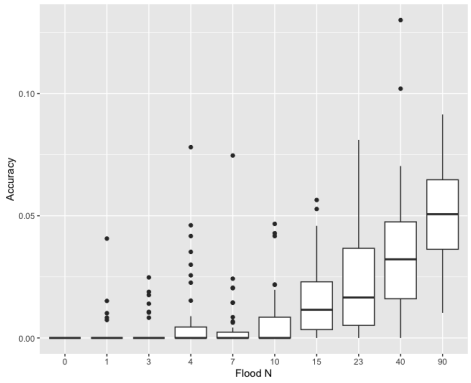


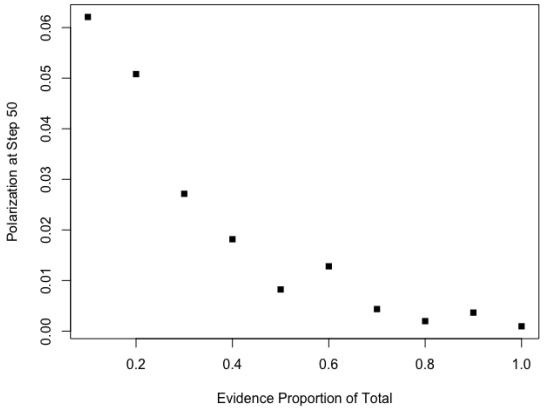
Figure 3: All panels display the same underlying data: the middle panel shows N of flood items (x-axis) against the mean number of steps to complete knowledge by all agents (y-axis). The right panel shows these data relative to an x-axis ordering in terms of evidence proportion. The left panel plots all 500 runs. Underlying network: 10 agents, in a perfect lattice with $k = 2$.



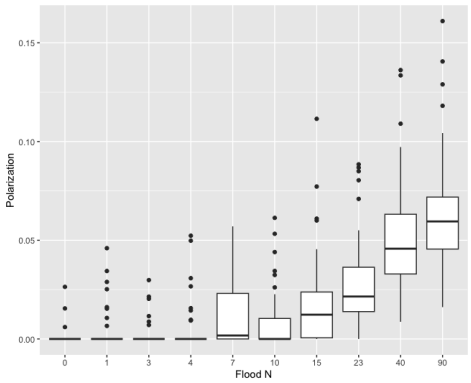
(a) Mean accuracy (y) by evidence proportion (x).



(b) Accuracy (x) by number of flood items (y).



(c) Mean polarisation (y) by evidence proportion (x).



(d) Polarisation (x) by number of flood items (y).

Figure 4: Accuracy and polarisation at time step 50.

measure of polarisation, but nevertheless provides an intuitive and easy-to-compute indication.

The results of these further simulations are shown in the four panels of Fig. 4. The top two panels show the results for accuracy. Panel 4a shows the means (calculated, once again, from 50 runs per condition for a total of 500 runs, all parameters as in Fig. 3), ordered by evidence proportion to match the right hand panel of Fig. 3. Figure 4b shows the underlying data indexed by the number of flood items (to match Fig. 3 left and middle panels, note, though, the categorical, as opposed to numerical ordering of the x-axis). Figures 4c and 4d show the same for polarisation (measured in the same simulation). Observably, the signature of the effects previously seen for steps to complete knowledge show again on these measures (n.b., perfect correspondence is impossible as accuracy and polarisation, in contrast to convergence time, are bounded both from below and above).

Generality of results

At this point, readers may legitimately wonder how specific our results are to the chosen simulations. We address this question in the remainder.

Other networks

First, how dependent are our results on the specific network structures we have modelled? Do they generalize to other networks that differ in type, density, or size? Figure 5 shows the results of changing network type, specifically different network types obtained by changing the rewiring probability of the perfect lattice ($p = 0$) to $p = .2$ for a small world network (Watts & Strogatz, 1998) and $p = 1$ for a random network (Erdős & Rényi, 1961). Displayed are the results of 500 runs each for the perfect lattice used thus far (left panel, labelled 0), for a small world network (middle panel, labelled .2), and a random network (right panel, labelled 1), with all other parameters kept as in the previous simulations. As can be seen, the network structure has some impact but does not fundamentally alter the functional form.

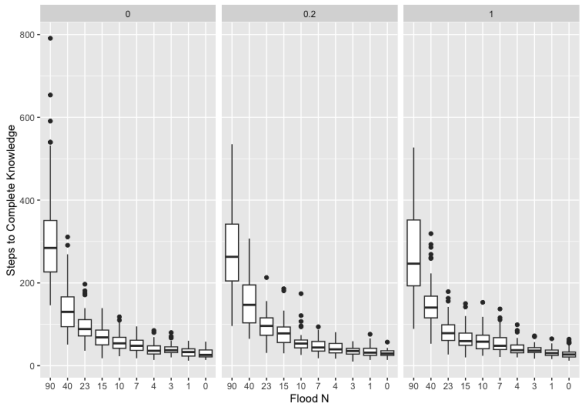


Figure 5: Effects of network structure, comparing the regular lattice used thus far ($p = 0$) to a small world network ($p = .2$) and a random network ($p = 1$). In each sub-panel, the x-axis shows the number of flood items, and the y-axis time to complete knowledge. Simulation parameters otherwise as before, N agents = 10.

Figure 6 shows the results of a further simulation that varies network density, that is, increases the number of connections each agent entertains across three levels by varying the parameter governing the number of connections from $k = 1$ to $k = 2$ (as before) to $k = 4$ (with all remaining parameters kept the same). As expected, greater density, that is, more communication channels, speeds up convergence, but again leaves the basic form unaltered.

Finally, Fig. 7 varies the network size (holding constant all other parameters) from 10 to 50 agents in steps of 10 (again for the regular lattice). And Fig. 8 plots all 2500 data points of that simulation split by panel to allow scrutiny of the effects of network size on the variance.

Four observations stand out from these final two figures: first, as expected, increasing the network size increases time to convergence. In increasing the number of agents, one can either hold constant the absolute number of copies of the evidence initially present in the population, or its proportional prevalence, but not both. In our simulations, we opted for the first variant: each piece of evidence was randomly distributed to the agents exactly once (whereby some agents could receive more than one piece or none at all). In other words, the number of initial seeds for a piece of information does not scale with network size. This matches the most likely real-world circumstance, and, necessarily leads to slowing (all other things equal). Second, network size—unlike the number of distractors—seems to have little impact on the variance. Third, increasing network size, once again, does not change the general functional form. Fourth, and most importantly, the increase in network size interacts with the amount of evidence. Specifically, making the network larger amplifies the effect of increasing distractors. This makes FtZ much more potent.

More realistic communication rules

It should be clear from the preceding that differences in network topology will moderate the effects of FtZ for better or worse, but the general profile would seem to remain robust. Only by increasing communication density between agents beyond the number of items can the degradations observed in Figures 1 to 4 be avoided altogether. This is, in principle, possible in multiple ways. For example, further increasing connectivity and/or increasing the number of messages that can be sent in one step; in other words, outrunning communication limitations. This seems unrealistic in practice.[3]

[3] A factor that is easily accounted for in our extension by toggling on the model's sensitivity to the cost (time, mental resources, etc.) that updating on an argument demands of a recipient ('*limit-listening*'). By default, as long as agents in NormAN receive valuable arguments at any given point in time, no matter whether this information arrives heavily diluted due to FtZ, they will efficiently process this information and be swayed by it. As such, with this toggle turned off, the diluted argument quality will not produce ineffi-

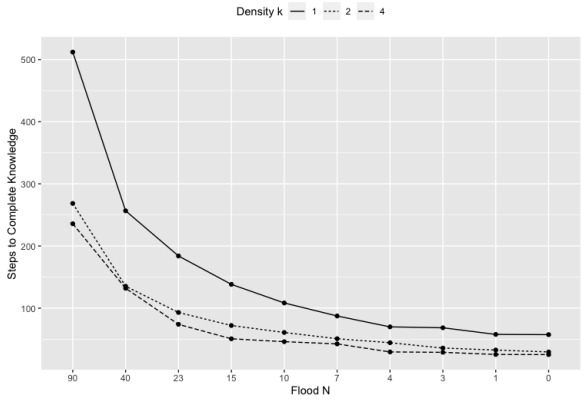


Figure 6: Effects of network density, from k = 1 (top line) to $k = 4$ (bottom line). x-axis number of flood items, y-axis mean time to complete knowledge. Simulation parameters otherwise as before: regular lattice, 10 agents. Each data point represents the mean of 50 runs.

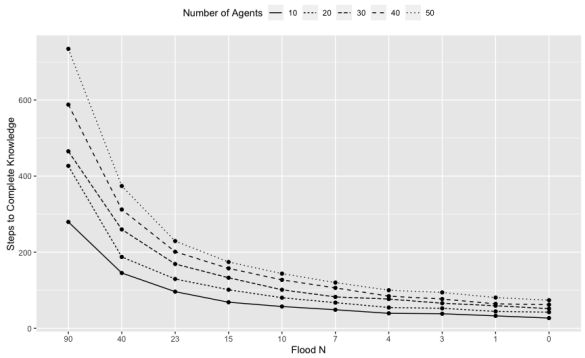


Figure 7: Effects of network size, from 10 (bottom line) to 50 (top line) in steps of 10. x-axis number of flood items, y-axis mean time to complete knowledge. Simulation parameters otherwise as before: regular lattice, k =2. Each data point represents the mean of 50 runs.

The other is to make better use of that limited bandwidth by having agents be more selective about what they send. This will, again, attenuate the effects but in any realistic context is unlikely to eliminate them entirely, because agents can only share evidence they have actually come to possess. And the hostile agents seeding the flood are not playing by the same rules.

We have left open what exactly our distractors might consist of: attractive information about another topic or extremely weak evidence on the issue at hand, because both will function in the same way w.r.t. their negative impact on the spread of genuinely diagnostic evidence. Where distractors consist of unrelated, but attractive information, effective

ciencies on the receiving side of communication, only on the sending side. In this respect, our already grim results on the effects of FtZ still err on the side of optimism.

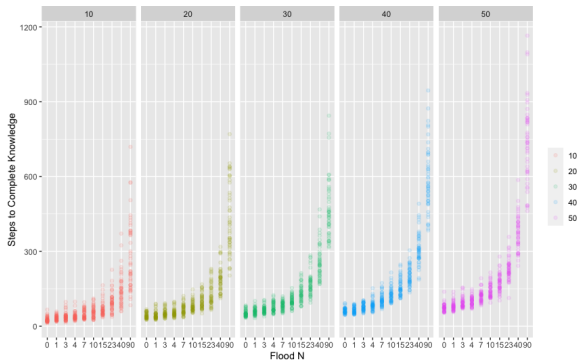


Figure 8: Simulation results shown in Fig. 7 split by network size, from 10 (left most) to 50 (right most) in steps of 10. x-axis number of flood items, y-axis time to complete knowledge.

filtering is made hard by the choice of those filters. Where distractors consist largely of irrelevant on-topic evidence, filtering faces a different challenge. Agents cannot just share 'the best evidence', they can only share *their* best evidence. The overall evidence distribution is unknown and necessarily revealed only gradually via communication (that is the point of communication). As a result, selectively choosing one's evidence *itself* may inadvertently reduce accuracy and create polarisation (Assaad et al., 2023; Assaad & Hahn, 2024; Schöppl & Hahn, 2024), even where the goal is to communicate helpfully by selecting one's best evidence. There are, consequently, no magic bullets for communication across a network.

Conclusion

Our simulations illustrate how flooding the zone can effectively clog communication channels, even for communities of rational reasoners. The process slows information exchange with attendant consequences for accuracy and polarisation. This impact of feeding distractors into the discourse can be re-conceptualised as restricting the probability of transmission across the individual communication links, which means that it will be attenuated by measures to make transmission of relevant information more probable (such as filtering or increasing the density of connection). However, its effects will be greatly exacerbated by network size, making it a potent weapon in practice.

Acknowledgements

U.H. was supported by the European Union, via Horizon 2020 Project PERYCLES. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Horizon 2020 Project PERYCLES. Neither the European Union nor the granting authority can be held responsible for them.

L.A. gratefully acknowledges funding provided by the Konrad- Adenauer-Stiftung (PhD scholarship) and the sup-

port of the Chair of Philosophy of Science at the Munich Center for Mathematical Philosophy (LMU).

K.S. was supported by the research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP). They are grateful for funding from the Netherlands Organization for Scientific Research (NWO) and the Dutch Ministry of Education, Culture and Science (OCW) in the context of its 2017 Gravitation Program (grant number 024.003.025).

References

Assaad, L., Fuchs, R., Jalalimanesh, A., Phillips, K., Schöppl, K., & Hahn, U. (2023). A bayesian agent-based framework for argument exchange across networks. *arXiv preprint arXiv:2311.09254*.

Assaad, L., & Hahn, U. (2024). Rational polarization: Sharing only one's best evidence can lead to group polarization. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Centola, D. (2018). *How behavior spreads: The science of complex contagions*. Princeton University Press.

Ceron, A. (2015). Internet, news, and political trust: The difference between social media and online media outlets. *Journal of computer-mediated communication*, *20*(5), 487–503.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the national academy of Sciences*, *113*(3), 554–559.

Erdős, P., & Rényi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, *12*(1), 261–267.

Hahn, U. (2020). Argument quality in real world argumentation. *Trends in cognitive sciences*, *24*(5), 363–374.

Hahn, U., Merdes, C., & von Sydow, M. (2024). Knowledge through social networks: Accuracy, error, and polarisation. *Plos one*, *19*(1), e0294815.

Hegselmann, R., & Krause, U. (2006). Truth and cognitive division of labor: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, *9*(3), 10.

Jasser Jasser, D. E., & Garibay, I. (2021). Flooding the zone: a censorship and disinformation strategy that needs attention.

Jitnah, N., Zukerman, I., McConachy, R., & George, S. (2000). Towards the generation of rebuttals in a bayesian argumentation system. In *Inlg'2000 proceedings of the first international conference on natural language generation* (pp. 39–46).

Keppens, J. (2019). Explainable bayesian network query results via natural language generation systems. In *Proceedings of the seventeenth international conference on artificial intelligence and law* (pp. 42–51).

Klein, E., & Robison, J. (2020). Like, post, and distrust? how social media use affects trust in government. *Political Communication*, *37*(1), 46–64.

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

Kopecky, F. (2024). Argumentation-induced rational issue polarisation. *Philosophical Studies*, *181*(1), 83–107.

Korb, K. B., & Nicholson, A. E. (2010). *Bayesian artificial intelligence*. CRC press.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global twitter heartbeat: The geography of twitter. *First Monday*.

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour*, *7*(1), 74–101.

Mäs, M., & Flache, A. (2013). Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PloS one*, *8*(11), e74516.

Mundt, M., Ross, K., & Burnett, C. M. (2018). Scaling social movements through social media: The case of black lives matter. *Social media+ society*, *4*(4), 2056305118807911.

Olsson, E. J. (2013). A bayesian simulation model of group deliberation and polarization. In *Bayesian argumentation* (pp. 113–133). Springer.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kauffman.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). Springer.

Rid, T. (2020). *Active measures: The secret history of disinformation and political warfare*. Farrar, Straus and Giroux.

Roberts, M. (2018). *Censored: distraction and diversion inside china's great firewall*. Princeton University Press.

Schöppl, K., & Hahn, U. (2024). Exploring effects of self-censoring through agent-based simulation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Scutari, M., & Denis, J.-B. (2021). *Bayesian networks: with examples in r*. Chapman and Hall/CRC.

Starr, P. (2020). The flooded zone: How we became more vulnerable to disinformation in the digital era. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age* (p. 67–92). Cambridge University Press.

Sunstein, C. (2018). *# republic: Divided democracy in the age of social media*. Princeton university press.

Timmer, S. T., Meyer, J.-J. C., Prakken, H., Renooij, S., & Verheij, B. (2015). Explaining bayesian networks using argumentation. In *Symbolic and quantitative approaches to reasoning with uncertainty: 13th european conference, ecsqaru 2015, compiègne, france, july 15-17, 2015. proceedings 13* (pp. 83–92).

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Tufekci, Z., & Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of communication*, *62*(2), 363–379.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440–442.

Yardi, S., & Boyd, D. (2010). Tweeting from the town square: Measuring geographic local networks. In *Proceedings of the international aaai conference on web and social media* (Vol. 4, pp. 194–201).

Zukerman, I., McConachy, R., & Korb, K. B. (1998). Bayesian reasoning in an abductive mechanism for argument generation and analysis. In *AAAI/IAAI* (pp. 833–838).

Zukerman, I., McConachy, R., Korb, K. B., & Pickett, D. (1999). Exploratory interaction with a Bayesian argumentation system. In *IJCAI* (pp. 1294–1299).