ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Complex Social Systems:
Multi-dimensional opinion dynamics

Project Report

Modelling Multi-dimensional Opinion and Polarization Formation

Michael Andres, Florian Dorner, Gian Luca Gehwolf, Fabian Hafner, David Metzger

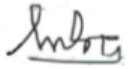
Zurich
December 2020

Abstract

We investigate multi-dimensional opinion and polarization formation in networks of agents, both with and without the presence of social bots. We reproduce and generalize two models featuring opinion formation based on weighted balance theory (WBT), as well as the coevolution of social networks and opinion dynamics, respectively. In computational experiments we explore the effects of various modelling assumptions and the robustness of model behaviour to parameter changes. Both the precise form of the network evolution dynamics and the extent of "evaluative extremeness" determining how much people are influenced by others' opinions play an important role: With high level of evaluative extremeness, we observe strong polarization along a single axis of disagreement but opinions become more diverse if network ties are restricted to others with similar opinions. The effect of social bots is similarly modulated by the network: While bots spreading extreme opinions can increase polarization and bots with more neutral messaging can foster moderation, both are most effective if placed in highly connected nodes. Correspondingly, the effectiveness of bots greatly suffers if network ties are restricted to nodes with similar opinions.

Agreement for free-download

We hereby agree to make our source code for this project freely available for download from the web pages of the SOMS chair. Furthermore, we assure that all source code is written by ourselves and is not violating any copyright restrictions.



Michael Andres

Florian Dorner

Gian Luca Gehwolf

Fabian Hafner

David Metzger

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

MULTIDIMENSIONAL OPINION DYNAMICS

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	First name(s):
Andres	Michael
Dorner	Florian
Gehwolf	Gian Luca
Hafner	Fabian
Metzger	David

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	Signature(s)
Zürich, 4.12.2020	Florian Dorner
	Michael Andres
	David Metzger
	Fabian Hafner
	Gian Luca Gehwolf

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Contents

1 Individual contributions

Florian Dorner proposed the abstract structure of the generalized model, contributed the initial implementation of the abstract generalized model and conducted and wrote about the experiments on bots. Fabian Hafner implemented the reproduction of the Weighted Balance Model, investigated the properties of the Generalized WBT Model and conducted the corresponding experiments. Gian Luca Gehwolf conducted a extensive literature review, wrote the introduction and motivations section, contributed to the implementation section, and commented and gave feedback to the coding of the models. David Metzger conducted the experiments for the coevolution model and contributed to the corresponding sections, as well as some parts of the Generalized Model. He also restructured the code for easier reproducibility. Michael Andres introduced a simplified structure to the implementation of the Generalized Model, reviewed and commented the code, conducted analysis of the network structures in gephi and took the lead in writing the summary.

Everyone proofread the report and contributed a variety of small improvements through their feedback.

2 Introduction and Motivations

The divergence of opinions on certain issues can lead to the formation of several conflicting or contrasting groups. The (re-)emergence of populism in certain countries in Europe and in the US has led to divided societies with strongly opposing opinions on various issues. This political polarization in democratic societies has reached a point, where it poses a threat to political stability [Abramowitz and Saunders, 2008, Hare and Poole, 2014]. In the field of sociology, studies on dynamics of opinion formation have a long history. Additionally, the physics community has also contributed to this field, where studies on social networks and the dynamics of networks have been performed [Helbing, 1994, Laguna et al., 2004, Hegselmann et al., 2002, Holme and Newman, 2006].

Holme and Newman [2006] developed a model that combines opinion dynamics with assortative network formation. They argue that individuals can become like-minded because they are connected via the network, but at the same time they can also form a connection

in network because they are like-minded. Previous models only accounted for one or the other [Castellano et al., 2003, Deffuant et al., 2002, Liggett, 2012, Castellano et al., 2000, McPherson et al., 2001, Sood and Redner, 2005, Sznajd-Weron and Sznajd, 2000]. Holme and Newman's dynamic, non-equilibrium model reaches a consensus state within a finite time and the authors analyze the distribution of community sizes in the converged model. They find a phase transitions in their model such that fundamental changes in the social structure of the community can sometimes be caused by small changes in the parameters of the system. One limitation of Holme and Newman's model is that opinions are modeled as unidimensional. This can be problematic, as opinions about different topics are not always strongly correlated and people care more deeply about certain issues than others [Krosnick, 1990].

Still, opinion dynamic models that try to account for polarization, defined as "division into two conflicting or contrasting groups" (American Heritage Dictionary, 2011), have mostly focused on one-dimensional opinion models ([Lorenz, 2007, Flache et al., 2017, Jager and Amblard, 2005, Salzarulo, 2006]). On the other hand, there exist only a few multidimensional opinion dynamics models [Huet and Deffuant, 2010, Flache and Mäs, 2008, Flache and Macy, 2011, Schweighofer et al., 2020]. All these models can only generate a form of polarization under special conditions, such as high correlation of demographic attributes or complex social structures, with the exception of the model developed by Schweighofer et al. [2020] which does not assume any underlying complex social or logical structures. With their Weighted Balance Theory (WBT) model, they are able to produce hyperpolarization: the emergence of opinion extremeness (opinions far away from the center) and issue constraint (positions on different issues strongly correlate, [Converse, 1964]). In the model, opinion change is induced through two modes: cognitive (issue positions) and affective components (interpersonal attitude), both influencing each other dynamically, e.g. human beings tend to agree more with people they like, and conversely they tend to disagree with people they dislike. However, the model is based on a strong approximation and assumes that each opinion on an issue contributes equally to the interpersonal attitude towards an other individual. As previously described, this contradicts Krosnick's (1990) idea that some people care more deeply about certain issues and less about others. Additionally, Schweighofer et al. [2020] mention that although their model does not require complex network structures to generate polarization, complex network structures can play a role in

polarization dynamics, which previous models have already shown [Deffuant et al., 2013, Manzo and Baldassarri, 2015, Hofstede et al., 2018].

Opinion dynamics can also be influenced by factors other than actors and their networks. Artificial actors, also known as social bots, can have an impact on opinion dynamics. By deliberately trying to inflate the importance of certain topics via social media platforms such as Instagram, Twitter, or Facebook, bots have tried to influence recent presidential election campaigns in the US, France, and to a lesser extent in Germany [Kupferschmidt, 2017, Bohannon, 2017]. There exist a few opinion dynamic models that account for the influence of bots [Ross et al., 2019, Stella et al., 2018, el Hjouji et al., 2018]. These models all show that bots have a significant impact on opinion dynamics. Ross et al. [2019] and el Hjouji et al. [2018] highlight that only a small number of bots is needed to influence opinion dynamics. Stella et al. [2018] provide evidence that social bots target mainly human influencers but generate semantic content depending on the polarized stance of their targets.

In this project, we study what effects network structures have on opinion polarization formation given that the WBT model developed by Schweighofer et al. [2020] does not account for it. To achieve this, we combine the WBT model with the coevolution model developed by Holme and Newman [2006]. This combined model we term "the generalised WBT model". Additionally, we simulate the effects of introducing bots in the generalised WBT model.

The remainder of this paper is structured as follows. Section 3 first introduces the theoretical background of the opinion dynamic models developed by Holme and Newman [2006] and Schweighofer et al. [2020]. This is followed by a theoretical description of the generalised WBT model including the introduction of the bots to this model. Section 4 gives an overview on how we implemented our models in Python. In Section 5, the results of our simulations using the different models are presented and discussed. In the final section, we summarise the results of our simulations, discuss the limitations of the models and give an outlook for possible future work.

3 Description of the Model

3.1 First model: Coevolution of networks and opinions

Our first model is based on Holme and Newman [2006] and models the coevolution of opinions within a network of interacting people and the shape of the network. Each person in the network is modelled as one of the N vertices in a simple undirected graph G with M edges representing connections between persons. Furthermore, each person has an opinion o from a set of opinions O which is stored in the vertex corresponding to the person. The

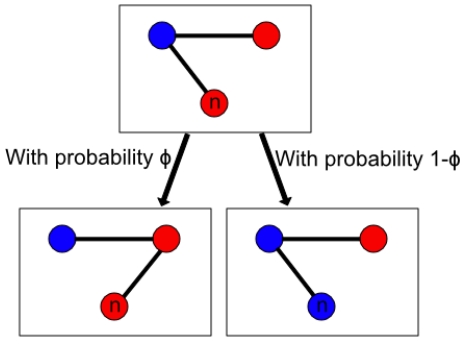


Figure 1: Dynamics of the basic coevolution model

graph G is initialized with edges and opinions sampled uniformly and the dynamics work as depicted in Figure 1: At each time step t, a vertex n is picked at random. If it has degree zero (no connections to other nodes), nothing happens. Else, with probability ϕ, a random edge connected to n, (n, m) is selected and changed to (n, m') where m' is a vertex with the same opinion as n, thus $o(m') = o(n)$. If no such m' exists, nothing happens. In other words, n gives up an old connection to connect to someone with the same opinion. Otherwise, with probability $1 - \phi$, a random vertex l that is directly connected to n by an edge is selected and n's opinion $o(n)$ is set to $o(l)$. This means that n changes their opinion to the opinion of one of their connections. Once all vertices in a connected component of G have the same opinion, the opinions of all vertices n in the component stay fixed as none of the vertices is connected or can connect to a vertex m with a different opinion $o(m) \neq o(n)$. Thus convergence of opinions is achieved once opinions are constant on every connected component. We call the order of a connected component community size and denote it by

s, while the largest community size is denoted by S.

3.2 Second model: Weighted Balance Theory

The second model we studied is based on the Weighted Balance Theory (WBT) developed by Schweighofer et al. [2020]. In this paper, the authors try to explain the simultaneous emergence of the socio-political phenomena of opinion extremeness and issue constraint[1], which they call *hyperpolarization*. In the case of hyperpolarization, two groups of agents with diametrically opposed opinion vectors emerge. This can be compared e.g. to a two-party system splitting the population into 'conservative' and 'liberal' clusters.

The paper models the opinions of agent i as vector

$$\mathbf{o}(i) = \begin{pmatrix} o_1(i) \\ \vdots \\ o_D(i) \end{pmatrix} \in [-1, 1]^D$$

where D is the total number of opinions. Each entry represents the opinion about a certain topic and can range from complete opposition (represented by -1) to complete support (represented by +1).

The WBT model is based on the work on Balance Theory by Heider [1946], which focuses on the simpler cases of a single dimensional opinion vector with categorical value ± 1. Balance Theory proposes a 'triad' of two agents, which describes their opinion about a certain topic $o_d(i)$, $o_d(j)$ and their interpersonal attitude A_{ij}. This triad is is called *balanced*, whenever either both persons support the issue and have a positive attitude to each other, or have a negative attitude of each other and disagree about the policy issue. If positive opinions are encoded as 1 and negative ones as -1, the triad is thus balanced whenever the product of the persons' opinions about the policy is equal to their attitude about each other or equivalently the product of both opinions about the policy and the interpersonal attitude is 1.

The WBT model now expands upon this construct. Instead of the product, WBT uses the

[1] referring to the perceived correlation between opinions on different topics.

signed geometric mean

$$\mathrm{SGM}(x, y) := \mathrm{sign}(xy)\sqrt{(|xy|)}$$

to account for continuous opinions and ensure that issue-induced interpersonal attitudes have similar strength as the opinions about the issue.

In the case of multidimensional opinions about a set of issues D[2], the interpersonal attitude of agent i about person j and vice versa is calculated as a monotonously growing transformation f of the arithmetic mean

$$A_{ij} := f\left(\frac{1}{D}\sum_{d=1}^{D}\mathrm{SGM}(o_d(i), o_d(j))\right).$$

The functional form of f is chosen to preserve the sign to reflect peoples' tendency to agree with people they like and disagree with those they do not like – with the second relation often referred to as the backfire-effect [Wood and Porter, 2019, Bail et al., 2018]. Empirical evidence from survey data for the 2016 US presidential election suggests a roughly sigmoidal form [Schweighofer et al., 2020], such that f is chosen to be

$$f(x) = \mathrm{sign}\,(x)|x|^{1-e}$$

where the free parameter e is called *evaluative extremeness* and determines how strongly small amounts of (dis-)agreement are reflected in peoples' displayed attitude towards each other. It is worth noting that $A_{ii} \neq 1$ is possible.

Opinions are initialized uniformly randomly in $[-1, 1]$ for every agent i and every component $o_d(i)$ of the respective opinion vector $\mathbf{o}(i)$. Then, in each model step, we iterate over all agents in random order and for each agents i another agent j is sampled. Then, the interpersonal attitude A_{ij} is calculated and an agent i incrementally adjusts their opinion about issues $o(i)$ to better match the balance dictated by Weighted Balance Theory by slowly adjusting each of their opinions contained in the vector $\mathbf{o}(i)$ towards the optimally

[2]We refer to the set of opinions as D as well as to that set's cardinality.

balanced state

$$\mathbf{b}^{ij} := \begin{pmatrix} b_1^{ij} \\ \vdots \\ b_D^{ij} \end{pmatrix} := \begin{pmatrix} \text{SGM}(o_1(j), A_{ij}) \\ \vdots \\ \text{SGM}(o_D(j), A_{ij}) \end{pmatrix}$$

.

The unilateral update is performed as

$$\mathbf{o}(i) \leftarrow \mathbf{o}(i) + \alpha \left[\mathbf{b}^{ij} - \mathbf{o}(i)\right] + \xi(0, z)$$

with a fixed update parameter α[3] and noise $\xi(0, z)$ generated by a normal distribution with mean 0 and standard deviation z. The random noise accounts for effects and influences not captured by the model and including it ensures robustness of our results. Agents will try to match the opinions about issues of people they like ($A_{ij} > 0$) while tending to disagree about issues O of people they dislike ($A_{ij} < 0$). The algorithm is graphically shown for a simple case of a two-dimensional opinion vector in Figure 2.

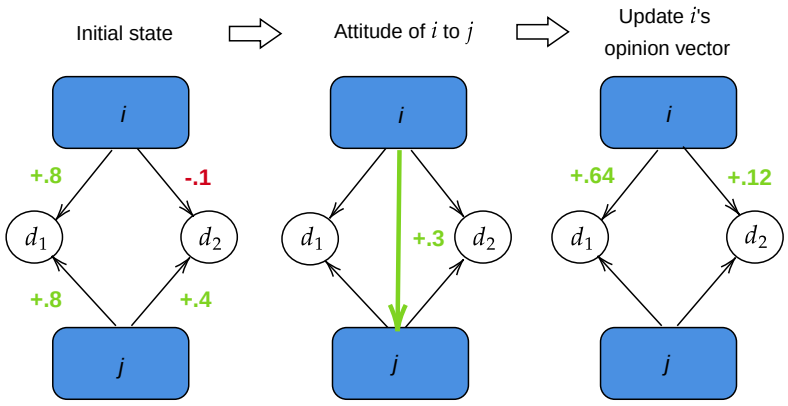


Figure 2: Schematic of opinion exchange under the WBT model with $\alpha = 0.5$. Adapted from: Schweighofer et al. [2020]

The model is considered to have converged once the difference of the absolute entries of

[3] Schweighofer et al. [2020] state that it is reasonable to assume that opinions have a certain degree of inertia and do not change completely upon a single interaction with another person. This fact is represented by α.

consecutive opinion matrices

$$O := \begin{pmatrix} \mathbf{o}(1)^{\mathrm{T}} \\ \vdots \\ \mathbf{o}(N)^{\mathrm{T}} \end{pmatrix} \in [-1,1]^{D \times N}$$

does not change significantly, or more formally:

$$|O_t - O_{t-1}| := \sum_{i=1}^{N} \sum_{d=1}^{D} |o_d(i)| \leq \chi(N, D, z)$$

where $\chi(N, D, z)$ is a threshold dependent on the number of agents N, number of opinions D and standard deviation of the noise-level z. To control for robustness we require that at least 5 consecutive time steps meet this criterion.

3.3 Generalized Model

We propose a set of extensions to Holme and Newman [2006]'s model: First, we replace the discrete opinions o by a vector of continuous opinions $\mathbf{o} \in \mathbb{R}^k$ like in the WBT model. Then, the condition on the active node n's opinion $o(n)$ and another node m's opinion $o(m)$ for m being included in the sampling of new connections can be generalized from the equality operator to a general map

$$\text{connect} : \mathbb{R}^k \times \mathbb{R}^k \to \{0, 1\}.$$

Similarly, the update of $o(n)$ based on a connected node m's opinion $o(m)$ can be generalized from the projection of the second dimension

$$\text{update}(o(n), o(m)) = o(m)$$

to a general (stochastic) map

$$\text{update} : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k.$$

Similarly, the initialization of opinions, the way 'active' nodes are sampled and the convergence criterion can be modified.

3.3.1 Generalized WBT model

In particular, we can obtain the WBT model by setting $M = (N^2 - N)/2$ such that G is fully connected, $\phi = 0$, connect arbitrary and

$$\text{update}(\mathbf{o}(i), \mathbf{o}(j)) = \mathbf{o}(i) + \alpha \left[\mathbf{b}^{ij} - \mathbf{o}(i)\right] + \xi(0, z).$$

As the WBT model updates all its nodes in (a random) sequence rather than sampling nodes independently at each time step, we have to draw active nodes without replacement rather than independently at each step (starting over once no more nodes are left to be drawn) to match the WBT model's fine dynamics. Note that while Schweighofer et al. [2020] refer to a whole pass through all nodes as a step, we choose to refer to a single node update as a step in the following for the sake of generality.

This framing of the WBT model immediately suggests extensions to arbitrary graphs by keeping $\phi = 0$ but reducing the amount of edges M or initializing the graph G with a specific structure. Similarly, by introducing a connection criterion and increasing ϕ, the WBT model can be extended to dynamically changing graphs. The connection criterion could be based on a positive mutual opinion A_{ij}, matching signs for all opinion dimensions $o_d(i)$ and $o_d(j)$, a positive scalar product $\mathbf{o}(i) \cdot \mathbf{o}(j)$, or based on a threshold on the difference in opinions $|\mathbf{o}(i) - \mathbf{o}(j)| < \varepsilon(N, D, z)$.

3.4 Bots

To investigate the effect of bots or other malicious actors that actively try to spread their opinion on the generalized Weighted Balance model, we consider three different bot models: In all three, bots work like normal nodes but we modify the update function, such that bots never change their opinions. In the first model, all bots have the same opinion at the extreme end of the spectrum meaning that $o_d(i) = 1$ for all opinion dimensions d. The

second model is like the first, but this time, bots act on both sides of the opinion spectrum such that half of the bots have $o_d(i) = -1$ for all opinion dimensions d, while the rest is on the other extreme, as before. In the third model, bots have neutral opinions on all topics such that $o_d(i) = 0$ for all opinion dimensions d. Next, bots can either be placed in random nodes, or more selectively in the most (or least) connected nodes according to the degree. Furthermore, we consider both cases in which bots connect to other nodes in the same way all nodes do and cases, in which the bots' connect function is modified such that they sample their new connections from the full set of nodes rather than being restricted by the same connect function as other nodes.

4 Implementation

All the opinion dynamics models in this work are implemented as agent-based models in `Python`. We implemented the generalized model described in subsection 3.3 in `model.py` in the class `coevolution_model_general`. As all of the other models we considered are special cases of the model and were easily implemented as subclasses.
In the generalized model, the network graph is first initialized as a random graph with a predefined number of *vertices* and *edges* and a discrete or continuous opinion value attached to each node. The opinions of the *vertices* and the *edges* are updated in the `step` function using the specific `update` and `connect` functions as specified in the respective subclasses. In a simulation run, the `step` function is repeated as long as the `converge` function, which is defined differently in each subclass, returns a false value indicating that the model has not converged. For some of our graphs, we also introduced a strict step limit to save time.

4.1 Coevolution of networks and opinion model

The coevolution model is implemented as a subclass of `coevolution_model_general` called `holme` in `model.py`. In this subclass, we set the `update` and `connect` functions to

$$\text{update}(o(n), o(m)) = o(m)$$

and

$$\text{connect} = [o(n) == o(m)]$$

to match the behaviour of the coevolution model as described in section 3.1.

4.2 Weighted Balance Theory Model

While the Weighted Balance Theory model can easily be implemented as a subclass of `coevolution_model_general` as described in 4.3 and we did this in `weighted balance`, we also provide a second simpler implementation in `WBT_model.py` and verified that both versions produce the same behaviour.

In both cases, we chose

$$\varepsilon(N, D, z) = NDz$$

as convergence criterion, as described by Schweighofer et al. [2020].

The simpler implementation can be run via the function run_model(N, D, e, z, α) which, per default, computes time steps until the model converges. Alternatively, an opinion matrix O and attitude matrix A can be initiated and then each timestep can be calculated individually with the update function update_model(A, O, e, z, α) function. Additionally, the hyperpolarization parameter can be calculated with the function H(O).

4.3 Generalized WBT Model

Similar to the coevolution model, we constructed a subclass `weighted_balance_general` in `model.py` for the generalized WBT model, which uses

$$\text{update}(\mathbf{o}(i), \mathbf{o}(j)) = \mathbf{o}(i) + \alpha \left[\mathbf{b}^{ij} - \mathbf{o}(i)\right] + \xi(0, z)$$

and different `connect` functions.

4.4 Bots

We built a subclass `weighted_balance_general` of `coevolution_model_general` to introduce bots. For this, we added a dummy dimension indicating whether a node is a bot to the opinion dimensions. Then, we modified the update function used in 4.3 to keep the dummy dimension constant and ignore it for the calculation of the update, unless it is equal to 1, in which case the updating node is a bot and the update function just returns the node's initial opinion (including the dummy dimension). Similarly, we modified the connect function to ignore the dummy variable.

However, the class also allows for changing the connect function and we experimented with setting connect to `TRUE` for all other nodes when the active node is a bot, such that bots sample their connections from all nodes rather than just ones with similar opinions. Furthermore, the class allows for various ways of setting bots' opinions and placing them in specific nodes, which are further explained in the code.

While we only implemented bots in a dynamic graph setting akin to the generalized WBT model, the basic WBT model can be recovered by using a fully connected graph and $\phi = 0$. Similarly, setting $\phi = 0$ allows for a WBT model with bots on static graphs.

Lastly, we adapted the convergence criterion to $\varepsilon(N, D, z) = (1 - \phi)(N - N_{bots})Dz$ to account for the smaller expected change in the opinion matrix because of bots not changing their opinions and edge changes happening instead of opinion updates with probability ϕ.

5 Simulation Results and Discussion

5.1 First model: Coevolution of networks and opinions

In their analysis, Holme and Newman [2006] focus mainly on the community size s given by the order of the connected components with internal consensus that form at convergence. They investigate how the community sizes are distributed and how they vary with the parameter Φ, i.e. relative frequency of opinion change vs. edge change. We attempted to reproduce their findings. In Figure 3 we see an exemplary run in a small graph with $\phi = 0.5$,

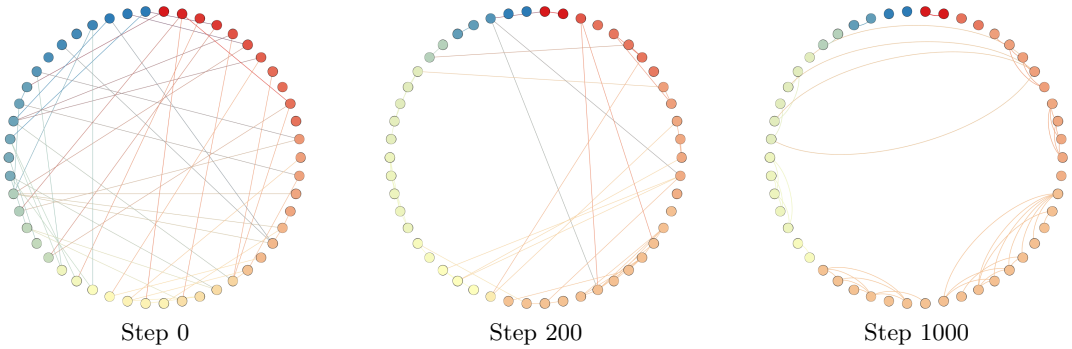


Figure 3: Various Steps of the Holme Model with 50 nodes and 50 edges and with $\phi = 0.5$. Each node has one opinion which is represented by the color and placement in the circle, where blue and red are the relative extremes. Graphs are drawn with Gephi using the Circular Layout.

where multiple components emerge. Because of our limited computational resources, we had to reduce the size of the network compared to the original paper. Our results for the distribution of the community sizes in a graph with 640 vertices were less smooth but qualitatively compatible with the results from the paper, which we can see in Figure 4: for very small Φ, i.e. for high edge change and low opinion change, we find configurations with one giant component and many small components, whose numbers decay exponentially with the size. For Φ around 0.46 we find more distributed sizes. They fit a power law for sizes between 8 and 80. In contrast to Holme and Newman [2006] we frequently find giant components in the converged model. For Φ close to 1, where the opinions remain as initialized, the communities are formed by nodes with the same initial opinion. The resulting distribution is close to a multinomial distribution with a mean equal to $\gamma = 10$, the average number of nodes per opinion.

To take a closer look at the phase transition found by Holme and Newman [2006], we ran the model for many different values of Φ, which can be seen in Figure 5. Around $\Phi = 0.45$ the model transitions from a consensus state where most individuals hold the same opinion to one where there is a diverse range of opinions in the population.

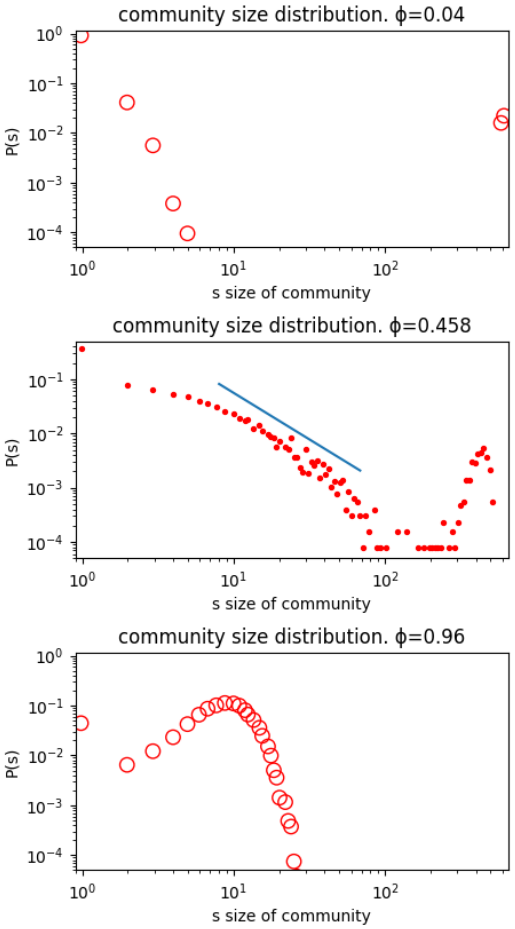


Figure 4: Histograms of community sizes in the consensus state for different values of Φ. Distribution of the community sizes over 400 iterations with $N = 640$ vertices. $\gamma = 10$, $k = 4$

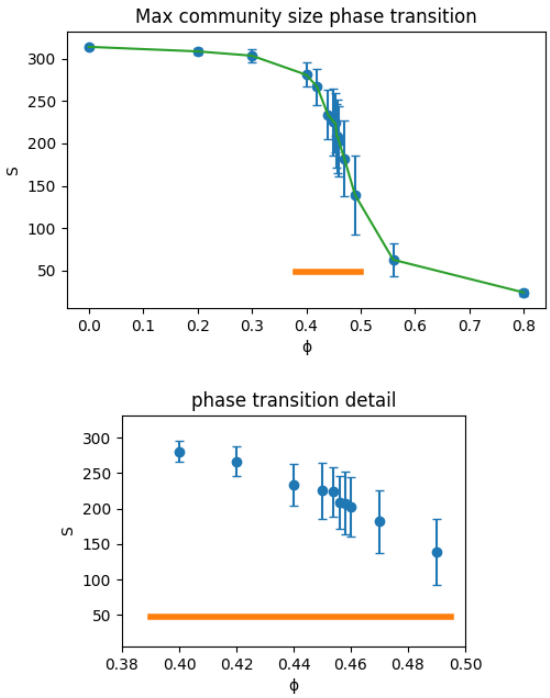


Figure 5: Size of the largest community for different Φ, average and standard error over 80 iterations with $N = 320$ vertices. $\gamma = 10$, $k = 4$

5.2 Second model: Weighted Balance Theory

To evaluate the extent of hyperpolarization in their model, Schweighofer et al. [2020] suggest the hyperpolarization measure

$$H(O) := \frac{4}{N^2} \sum_{1=i<j}^{N} \left(\frac{\delta(o(i), o(j))}{\delta_{\max}} \right)^2,$$

where δ denotes a metric (we use the euclidean metric) and δ_{max} is the maximal value of this metric for two points in the hypercube $[-1,1]^D$. In the case of maximal hyperpolarization[4], we have $H(O)$=1: Without loss of generality, we can assume that the first $\frac{N}{2}$ nodes have opinion 1 across all dimensions, while the rest of the nodes have opinion -1. Then, all terms in the sum are either 0 if i and j are both larger or both smaller than $\frac{N}{2}$, or 1 if they are on different sides of $\frac{N}{2}$. Because we only sum over terms with $i < j$, all terms with $i > \frac{N}{2}$ and all terms with $j < \frac{N}{2} + 1$ fade and the sum reduces to

$$\sum_{1=i<j}^{N} \left(\frac{\delta(o(i), o(j))}{\delta_{\max}} \right)^2 = \sum_{i=1}^{\frac{N}{2}} \sum_{j=\frac{N}{2}}^{N} 1 = \left(\frac{N}{2} \right)^2 = \frac{N^2}{4},$$

such that $H(O) = 1$ as claimed. On the other hand, $H(O) = 0$ if and only if all nodes have the same opinion.

First, we tried to reproduce the emergence of hyperpolarization under the conditions set in the original paper, which can be seen in Figure 6. In this three-dimensional cube every dimension responds to an opinion d_i. This low-dimensional model already shows interesting behavior: After random sampling, the agents' opinions are initially drawn to the center $(0, 0, 0)$. However shortly after, two distinct groups of like minded agents begin to form, repulse each other and eventually end up in opposing corners of the cube.

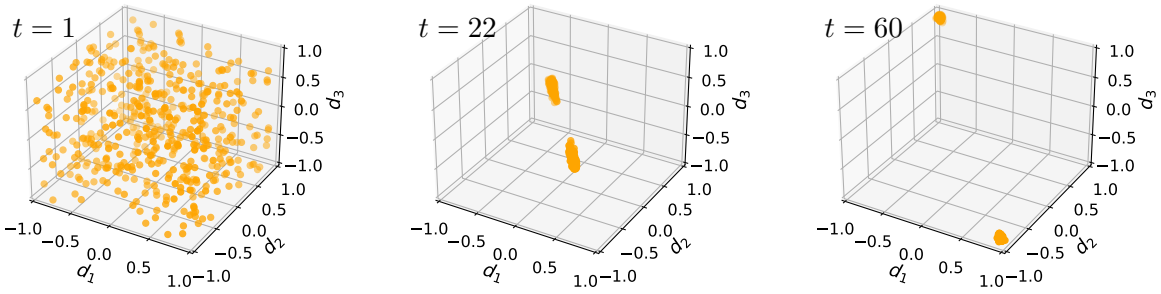
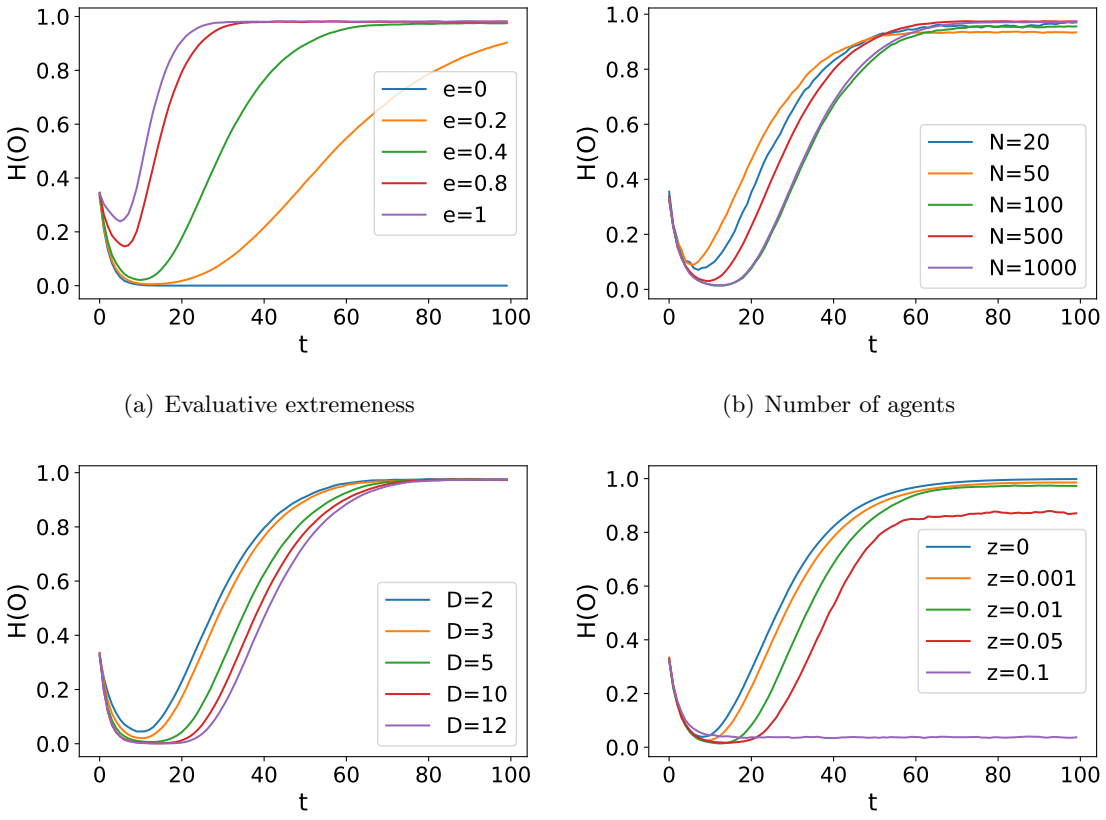


Figure 6: Weighted Balance Model with randomly sampled initial opinions and parameters $N = 500$, $D = 3$, $e = 0.3$, $\alpha = 0.4$, $z = 0.01$. In the final time-step depicted the models shows a hyperpolarized state with $H \simeq 0.85$. The corners towards which the model converges are random, thus depending on the random starting conditions, the model always ends up converging into diametrically opposed stable states.

[4] which can only happen with an even number of nodes

Next, in order to better understand the model behavior under parameter changes, we varied all parameters and plotted the resulting changes in hyperpolarization $H(O)$ in Figure 7.



(a) Evaluative extremeness

(b) Number of agents

(c) Opinion dimensions

(d) Noise level

Figure 7: Convergence behaviour to a hyperpolarized state for various variations of the parameter (a) e, values of $N = 500$, $D = 3$, $z = 0.01$ and $\alpha = 0.4$ kept constant; (b) N, values of $D = 3$, $e = 0.4$, $z = 0.01$ and $\alpha = 0.4$ kept constant; (c) D, values of $N = 500$, $e = 0.4$, $z = 0.01$ and $\alpha = 0.4$ kept constant; (d) z, values of $N = 500$, $D = 3$, $e = 0.4$ and $\alpha = 0.4$ kept constant. From these Figures one can see that the evaluative extremeness e is the main driver of hyperpolarization. N and D have major influence on the convergence speed, while z only produces complete hyperpolarization for $z \lesssim 0.05$.

We observe that the qualitative behaviour does not fundamentally change for most parameter perturbations. Important exceptions are high values of the noise level z and very low evaluative extremeness e, which was also highlighted by [Schweighofer et al., 2020] and can be seen in Figure 8. In the Figure we still see the polarization into two clusters, however

due to the evaluative extremeness being 0, they do not drift towards the corners but rather stay close to the center.

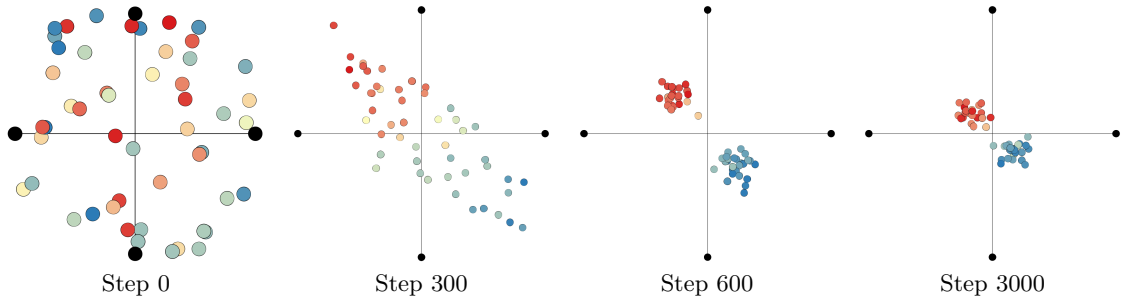


Figure 8: Various Steps of the Weighted Balance Model with 50 nodes and in a complete graph.Further f(x)=x (i.e. e=0). The x and y axis represent one opinion each, while coloring represents the third opinion. Black nodes are opinion extremes with values of +/-1 for the first image. Second, third and fourth image have black coordinates of value +/-0.2. Edges are not drawn.

These observations lead to the conclusion that the configuration of the basic WBT model at convergence is constrained to a set of the size of half the number of vertices in a hypercube, i.e. $2^D/2$. These final states are always composed of opposite clusters of agents with opposing opinion vectors. The model does not allow for 'moderate' nor more uniformly distributed states and is thus quite limited in its behavior.

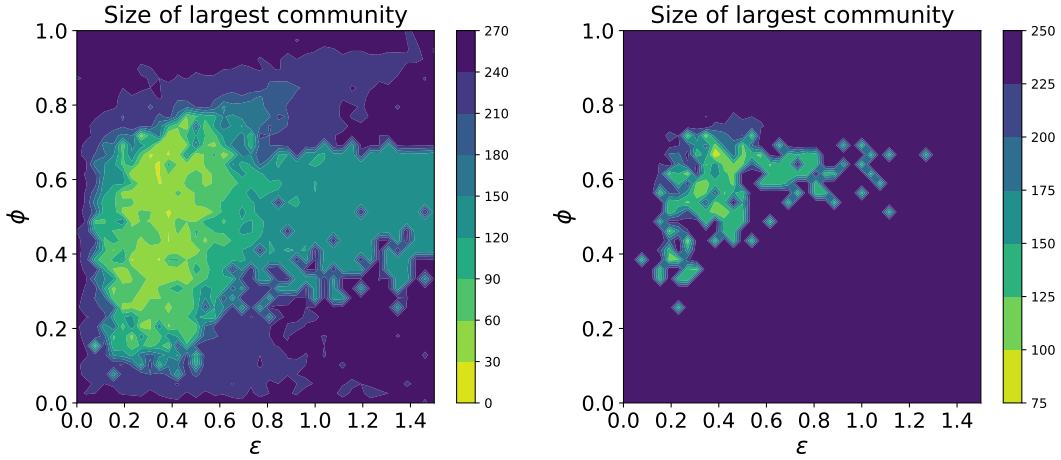
5.3 Generalized Weighted Balance Model

For the general model, we embedded the WBT model into a network similar to the coevolution model but with multiple opinion dimensions. The defining feature characterizing the behavior in this model lies in the form of the connect-function. While we tested several functional forms, e.g. one based on the angle or matching signs of the opinions, we ultimately decided to use a criterion based on the euclidean distance between two agents' opinion vectors:

$$\text{connect}(\mathbf{o}(i), \mathbf{o}(j)) = \begin{cases} 1 & \text{if } |\mathbf{o}(i) - \mathbf{o}(j)| < \varepsilon(N, D, z) \\ 0 & \text{else} \end{cases}$$

This approach relates to the sociological observation that people tend to form connections with those that do not stray too far from their own opinions (Bahns et al. [2017]) and often form echo chambers that further fortify their stances.

In order to reduce complexity, we decided to focus on the role of the probability to form a new connection ϕ and the distance parameter ε which indicates a node's "sphere of influence", to nodes in which it can connect. We find distinctly different behavior depending on the choice of these parameters as can be seen in the contour plot of the maximum (non-overlapping) component size S (figure 9).



(a) $M = 500$ (b) $M = 1000$ (color bar with different scale)

Figure 9: Contour plot for WBT general $S(\varepsilon, \phi)$. $N = 250$, $D = 5$, $e = 0.4$, $z = 0.01$, $\alpha = 0.3$. With $S \simeq N/2$ the system exhibits splitting into two similarly sized communities (cf. Figure 10), which is not apparent from this contour plot. With increasing number of edges, the network exhibits less phases leading to mostly uniform alignment except for a small region.

Next, we looked at the network evolution of particular set of parameters which can be seen in Figures 10 (splitting into two main clusters) and Figure 11 (fragmentation into several clusters). There, the graphs are qualitatively depicted with their vertices and edges, however the opinions are not apparent in this visualization. For this, we reduced the model to two and three opinion dimensions, which will be discussed later (cf. Figure 12).

We compared the general model with the WBT model and found that with the parameters chosen, the opinions converge again to the corners of the opinion space, but conversely to

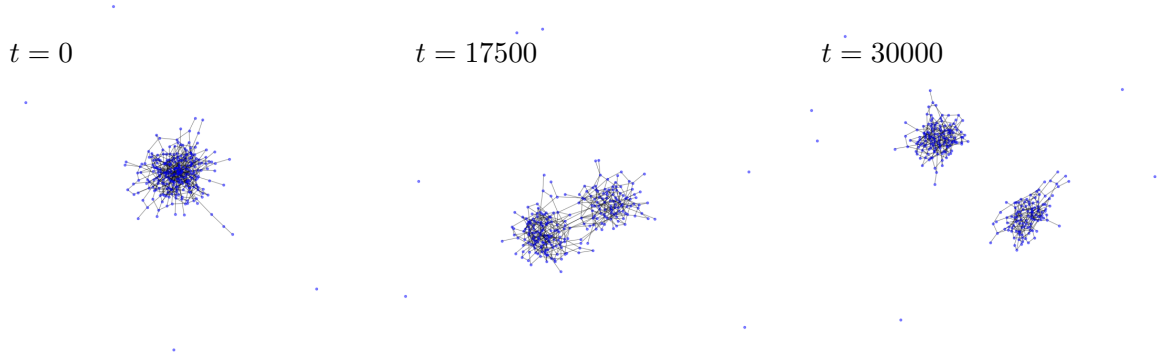


Figure 10: Exemplary convergent model run for $N = 250$, $M = 500$, $|D| = 5$, $\varepsilon = 1$, $\phi = 0.5$, $z = 0.01$, $\alpha = 0.4$, $S = 129$. The model displays gradual splitting into two larger hyperpolarized communities as well as some singular nodes. This outcome corresponds to the outcome of the simpler WBT model. Note that the timesteps correspond to a single update for a node contrary to an update on the all nodes as previously defined.

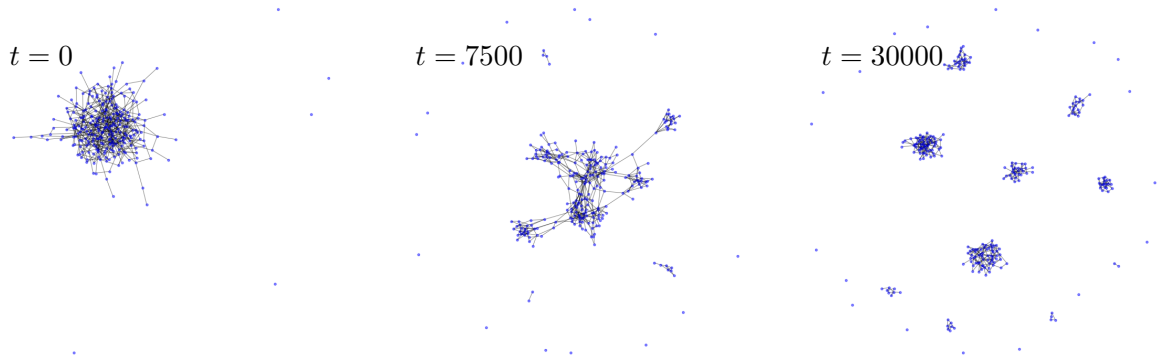


Figure 11: Exemplary convergent model run with $N = 250$, $M = 500$, $|D| = 5$, $\varepsilon = 0.6$, $\phi = 0.5$, $z = 0.01$, $\alpha = 0.4$ and $S = 62$. The model displays gradual splitting into 6 larger communities as well as some very small and singular ones. Note that the timesteps correspond to a single update for a node contrary to an update on the all nodes as previously defined.

WBT, they occupy many or all corners, as can be seen in Figure 12. We therefore have a high polarisation for the individual opinions but less hyperpolarization as defined in the metric $H(O)$ within the model proposed by Schweighofer et al; individuals have extreme opinions, but the different opinions are distributed more dispersedly. There also remain a few unconnected nodes that are too far apart from other nodes to reconnect and therefore do not change their opinion.

Moreover, looking at the network structure evolution in Figure 13 we see that the edges

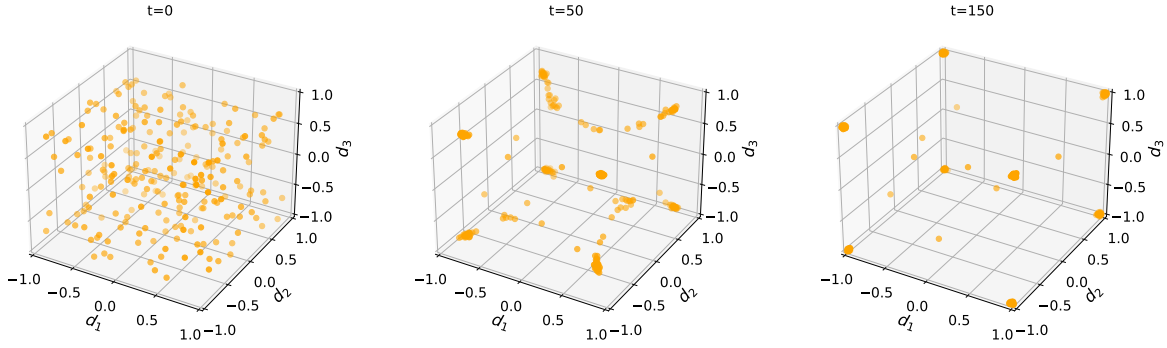


Figure 12: Generalized Model with randomly initialized dynamic graph, $N = 250$, $N_{edges} = 500$, $D = 3$, $e = 0.3$, $\alpha = 0.3$, $\phi = 0.45$, $\varepsilon(N, D, z) = 0.3$. The opinions converge in the corners. We can see opinion fragmentation but not hyperpolarization.

between communities vanish when they move apart in the opinion space. Our extension to the WBT model shows that when taking into account the evolving network structure within a population, opinions do not necessarily converge to a previously observed hyperpolarized state [5].

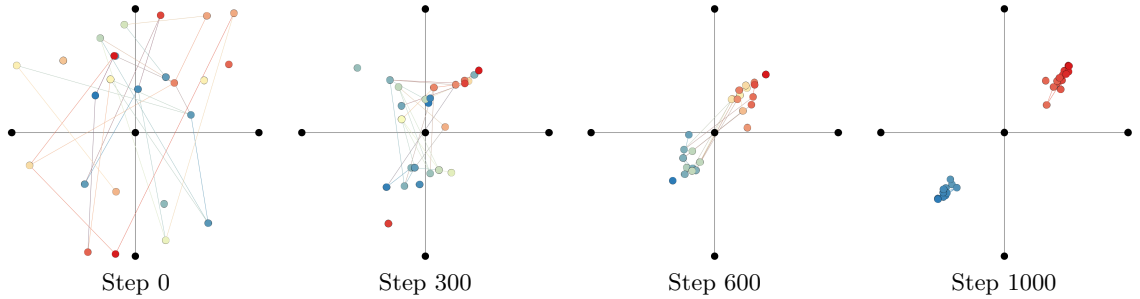


Figure 13: Various Steps of the General Model with 25 nodes and 25 edges, maximal connection distance is 1. The x and y axis represent one opinion each, while coloring represents the third opinion. Black nodes are coordinate extremes with values of +/-1 and the center.

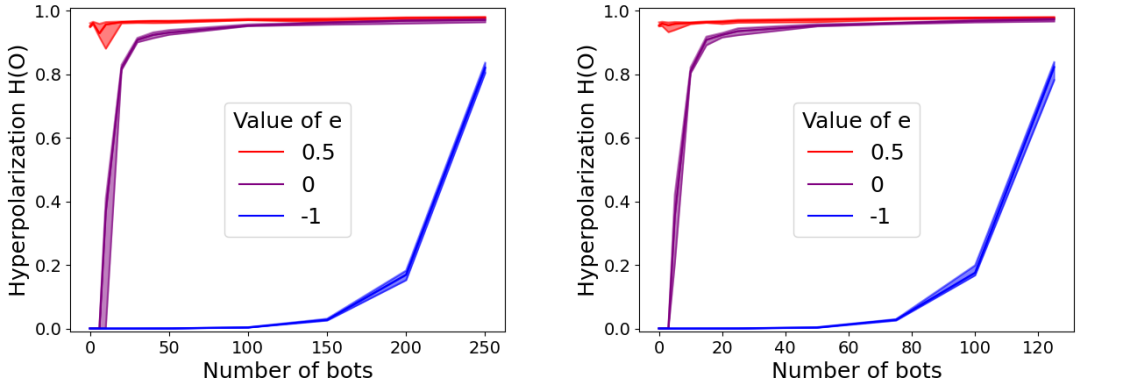
In contrast to the previous WBT model, our generalized model shows a myriad of varying behaviours. The model is far less constrained, however, the number of parameters makes systematic description difficult. Within this section we showed that both classical hyperpolarization as well as fragmentation into similarly sized communities are possible.

[5] However, fine-tuning the parameters (cf. Figure 10) can still lead to analogous results.

5.4 Bots

5.4.1 Fully connected graph

Adding bots that promote extreme opinions to the basic WBT model (fully connected and static graph) can lead to hyperpolarization even without evaluative extremeness, which is illustrated in Figure 14. While near-maximal hyperpolarization happens even without bots for positive values of the evaluative extremeness parameter e, hyperpolarization is low without bots but increases rapidly as bots are added for $e = 0$. Hyperpolarization



(a) Bots with $o = 1$ for all dimensions.

(b) Half of bots with $o = 1$ and the half with $o = -1$ for all dimensions.

Figure 14: Hyperpolarization at convergence/after 100k single updates for Weighted Balance on a fully connected graph, $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$, $f(x) = \text{sign}(x) \cdot |x|^{1-e}$. Area between 10%- and 90%-empirical quantile over 10 runs shaded. Bots excluded from the calculation of H.

also increases with the number of bots for negative evaluative extremeness e, but maximal hyperpolarization is not reached, even when half of the population consists of bots. Notably, there is no difference between having an equal number of bots on both extremes or having twice the amount of bots on a single side. This is because

$$\text{SGM}(\mathbf{o}(i), -\mathbf{o}(j)) = -\text{SGM}(\mathbf{o}(i), \mathbf{o}(j)) = \text{SGM}(-\mathbf{o}(i), \mathbf{o}(j)),$$

which implies

$$A_{ij} = f\left(\frac{1}{D}\sum_{d=1}^{D}\mathrm{SGM}(o_d(i), o(j)))\right) = -f\left(\frac{1}{D}\sum_{d=1}^{D}\mathrm{SGM}(o_d(i), -o(j)))\right) = A'_{ij},$$

for uneven f with $f(-x) = -f(x)$, where A'_{ij} is the value of A_{ij} if the opinion of node j was flipped. But in the absence of noise, the Weighted Balance update just moves $\mathbf{o}(i)$ in the direction of

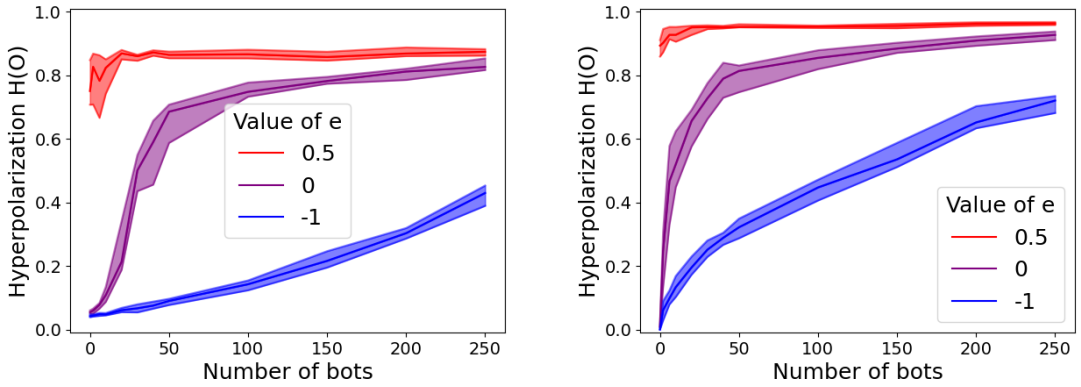
$$\mathrm{SGM}(A_{ij}, \mathbf{o}(j)) = \mathrm{SGM}(-A_{ij}, -\mathbf{o}(j)) = \mathrm{SGM}(A'_{ij}, -\mathbf{o}(j)),$$

implying that the update of node i's opinion is never affected by flipping the counterparts' opinion as long as f is uneven as we use the functional form $f(x) = \mathrm{sign}(x) \cdot |x|^{1-e}$. This symmetry might also explain why the mean opinion is essentially unaffected by the bots as can be seen in Figure A.30. However, it is worth noting that the effect of bots on hyperpolarization diminishes, while the effect on the mean grows if non-bot opinions are initialized with nonzero expectation rather than uniformly from $[-1, 1]^d$, as can be seen in Figure A.23.

5.4.2 Static graphs

Moving from fully connected to more sparse static graphs has three effects that can be seen in Figure 15: First, between-run variability for the same parameters is increased as the graph is initialized randomly. Second, hyperpolarization is smaller for uniformly randomly sampled edges compared to Barabási–Albert graphs, likely because of isolated nodes that remain unaffected by their surroundings. Third, the relative impact of the number of bots on hyperpolarization is lower than in the fully connected case, except for small numbers of bots and negative evaluative extremeness e. Next, Figure 16 shows the effect of the bots' centrality on hyperpolarization: perhaps unsurprisingly, placing bots in the most connected nodes of an exponential Barabási–Albert graph magnifies their effect, while placing them in the least central nodes makes them a lot less effective.

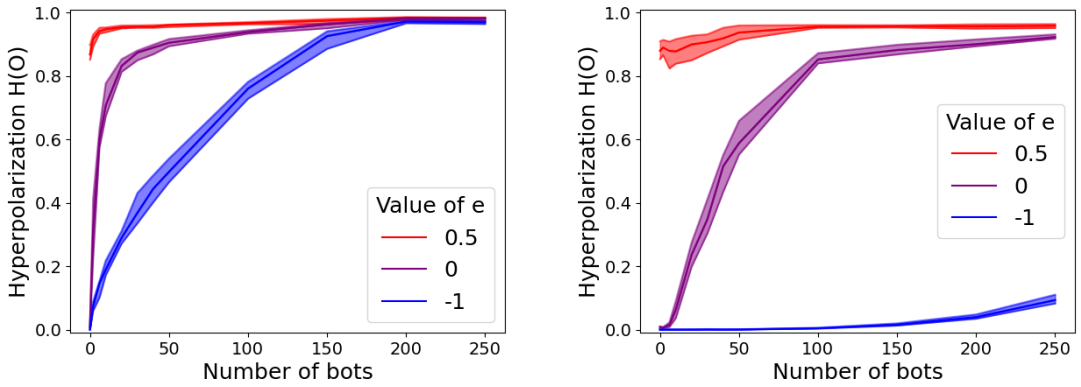
Adding bots that spread neutral opinions reduces hyperpolarization, as can be seen for $e = 0.5$ in Figure 17. However, a large number of such bots can be necessary to combat



(a) Edges sample uniformly

(b) Edges sampled accoruding to Barabási–Albert model

Figure 15: Hyperpolarization at convergence/after 100k single updates for Weighted Balance, static graph with 499 edges, $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$, $f(x) = \text{sign}(x) \cdot |x|^{1-e}$ and varying amount of bots with $o = 1$ for all dimensions. Area between 10%- and 90%- empirical quantile over 10 runs shaded.



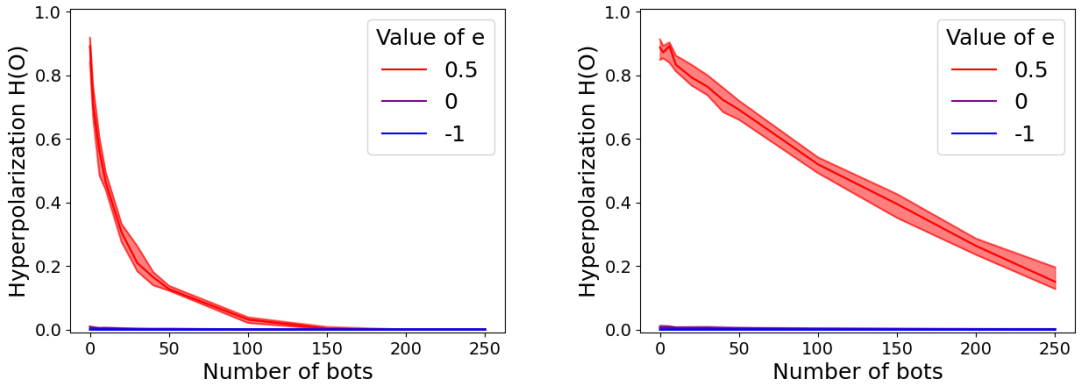
(a) Bots placed in the most connected nodes.

(b) Bots placed in in the least connected nodes.

Figure 16: Hyperpolarization at convergence/after 100k single updates for Weighted Balance with randomly intialized Barabási–Albert graph with 499 edges, $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$, $f(x) = \text{sign}(x) \cdot |x|^{1-e}$ and varying amount of bots with $o = 1$ for all dimensions. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Bots excluded from the calculation of H.

evaluative extremeness, especially if the bots are placed in noncentral nodes.

As can be seen in 18, breaking the symmetry in f can cause a range of interesting behaviour. When only attractive or repulsive forces are considered (f equal to the minimum or maximum of x and 0), the average opinion for every dimension can be influenced by the

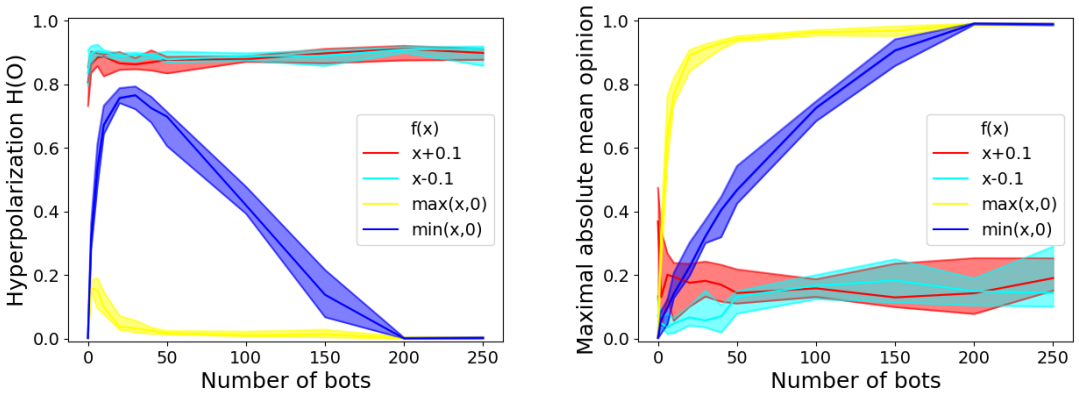


(a) Neutral bots with $o = 0$ for all dimensions placed in the most connected nodes.

(b) Neutral bots with $o = 0$ for all dimensions placed in the least connected nodes.

Figure 17: Hyperpolarization at convergence/after 100k single updates for Weighted Balance on static Barabási–Albert graph with 499 edges, $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$, $f(x) = \text{sign}(x) \cdot |x|^{1-e}$. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Bots excluded from the calculation of H.

bots, unlike in the previous cases. With attractive forces, it converges to the bots' opinion rather quickly, while repulsive forces cause slower convergence to the opposite opinion. Interestingly, the latter coincides with non-monotonous hyperpolarization if bots are placed in the most connected nodes: Without any bots, the repulsive forces mostly balance out such that the unpolarized initial configuration is mostly stable. Then, as bots are added, they cause nodes to take on the opposite extreme opinion, which in turn leads to other nodes being pushed towards the bots' opinion. As even more bots are added, this counteraction quickly weakens, as fewer and fewer connections between non-bot nodes remain. Surprisingly, adding or subtracting a small constant to $f(x) = x$, representing a consistently biased opinion about other people increases hyperpolarization without bots regardless of the biases direction and has a rather small effect on the overall mean, independent of the direction in which the bias points.

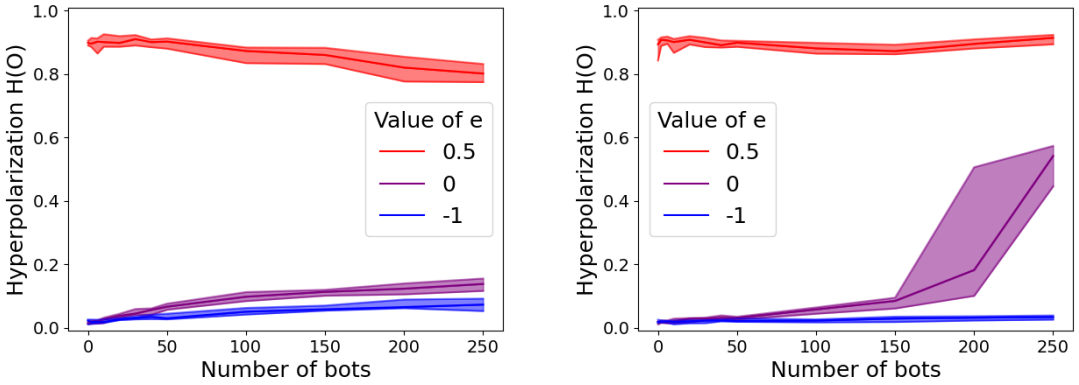


(a) Hyperpolarization.

(b) Maximum (over dimensions) absolute mean opinion.

Figure 18: Weighted Balance model on randomly initialized static Barabási–Albert graph with 499 edges at convergence/after 100k single updates with $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$ and varying amount of bots with $o = 1$ for all dimensions. Bots placed in the most connected nodes. Area between 10%- and 90%-empirical quantile over 10 runs shaded. Bots excluded from the calculation of H and mean opinion.

5.4.3 Dynamic graphs



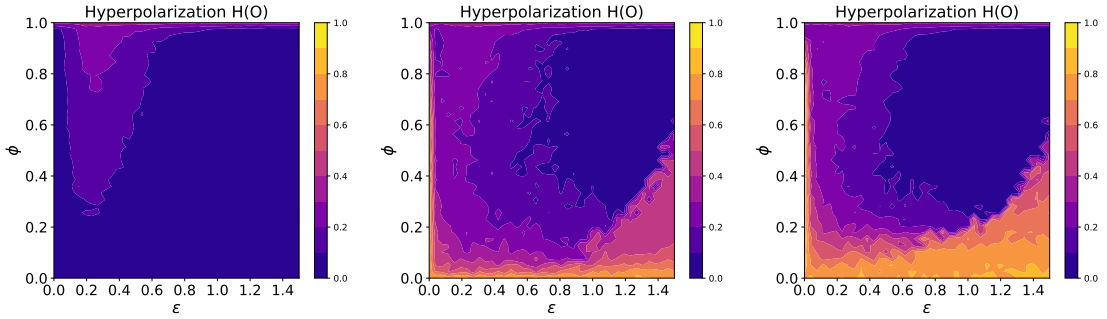
(a) Bots use same connect function as other nodes

(b) Bots can connect to any node, not just close ones

Figure 19: Hyperpolarization in dynamic graph Weighted Balance model on randomly initialized Barabási–Albert graph with 499 edges at convergence/after 100k single updates with $\phi = 0.5$, distance-based connect function with $\varepsilon = 1$ as well as $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$. Varying amount of bots with $o = 1$ for all dimensions. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Bots excluded from the calculation of H.

Adding bots to the Generalized WBT model for values of ε and ϕ that led to moderate

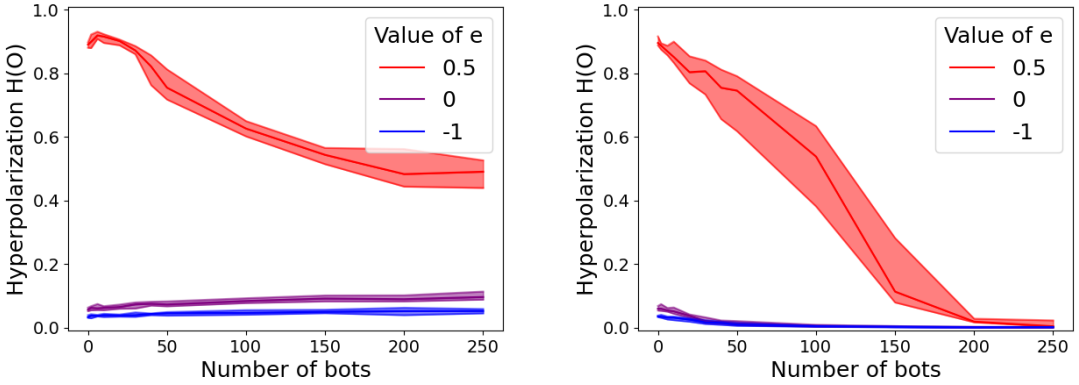
hyperpolarization in Figure 9 produces some interesting results. As can be seen in Figure 19, the influence of bots on hyperpolarization is greatly diminished in dynamic graphs as the bots start with extreme opinions and thus quickly lose most of their connections to non-extreme nodes, even if the bots started out highly connected. Surprisingly, the bots influence is similarly small for moderate numbers of bots, when the connect function is modified such that bots sample their new connections from any node, not just close ones. Interestingly, the same effect can be seen when "competing" bots are placed in opposing corners of opinion space. However, the increase in hyperpolarization with large amount of bots is even more rapid but also more uncertain in that case, as can be seen in Figure A.28. This might be because multiple updates are needed to bring a node close enough to the



(a) No bots (b) 50 bots constrained by ε (c) 50 bots unconstrained by ε

Figure 20: Hyperpolarization for different levels of ε and ϕ in dynamic graph Weighted Balance model on Barabási–Albert graph with 499 edges at convergence/after 100k single updates with $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$. Bots with $o = 1$ for all dimensions placed in random nodes and excluded from calculation of H. Bots have the strongest effect on hyperpolarization for small values of *phi* and large values of ε. Similarly, the effect of allowing bots to connect to more distant nodes is most pronounced for these values.

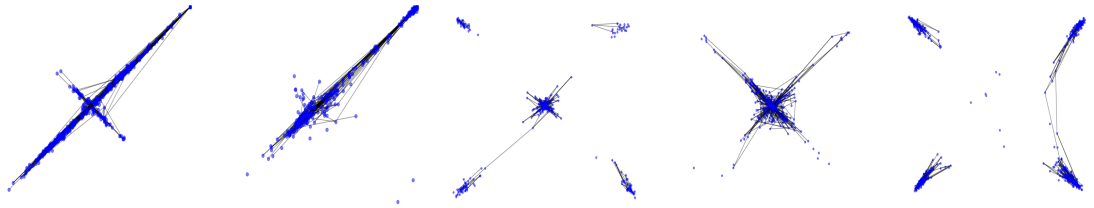
bots, such that it stays connected to them rather quickly replacing connections to bots by sampling its new connections exclusively from non-bots. To investigate this hypothesis, we looked at the hyperpolarization for different values of ε and ϕ for 50 (Figure 20) and 200 (Figure A.29) bots sharing the same extreme opinion . There, allowing bots to connect to arbitrary nodes has the biggest effect for small values of ϕ and large values of ε. This fits our hypothesis as small values of ϕ mean that a node is less likely to update its connections after being connected to a bot. Similarly, for large ε, fewer updates are needed to get a node from outside the bots' sphere of influence to within the sphere, where it is less likely to disconnect from the bots.



(a) Bots connect in the same way as other nodes (b) Bots can connect to any node, not just close ones.

Figure 21: Hyperpolarization in dynamic graph Weighted Balance model on randomly initialized Barabási–Albert graph with 499 edges at convergence/after 100k single updates with $\phi = 0.5$, distance-based connect function with $\varepsilon = 0.6$, as well as $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Varying amount of neutral bots with $o = 0$ for all dimensions.Bots excluded from the calculation of H.

The effect of neutral bots in the dynamic graph is more pronounced, but again small compared to static graphs (Figure 21). Unlike with extreme bots, neutral bots with the ability to connect to arbitrary bots are able to reduce hyperpolarization to a similar extent as in a static graph with bots placed in the least connected nodes.



(a) (b) (c) (d) (e)

Figure 22: Weighted Balance model on Barabási–Albert graph with 499 edges after 25 steps per node with $N = 500$, $D = 2$, $\alpha = 0.4$, $z = 0.01$. a) Static graph, $e = 0$, 50 bots ($o = 1$) in least connected nodes. b) Dynamic graph ($\varepsilon = 1, \phi = 0.5$), $e = 0$, 200 bots ($o = 1$) in most connected nodes. c) Dynamic graph ($\varepsilon = 0.6, \phi = 0.5$), $e = 0.5$, 200 neutral bots ($o = 0$) in random nodes. d) Dynamic graph ($\varepsilon = 0.6, \phi = 0.5$), $e = 0.5$, 200 neutral bots ($o = 0$) in random nodes. Bots unaffected by ε. e) Dynamic graph ($\varepsilon = 0.6, \phi = 0.5$), $e = 0.5$, no bots.

6 Summary and Outlook

We implemented, generalized and extended the opinion formation models by Holme and Newman [2006] and Schweighofer et al. [2020]. During the reproduction of the coevolution model we saw very similar effects regarding the community sizes, as well as their phase transition, compared to the original paper. Only minor deviations occurred, which we ascribe to our limitations in computational power resulting in simulations with smaller networks but qualitative findings were not influenced by this. The reproduction of the Weighted Balance Theory model yielded the behaviour reported in the original paper, validating our implementation of the models.

The generalization and unification of the models let us combine the strengths of dynamic networks with attitude-based influence and thus allows us to account for previously neglected effects. The investigation of the effects of various parameters confirmed our theoretical suspicion that the evaluative extremeness parameter e and more generally the choice of the attitude transformation *f function* have a large impact on qualitative model behaviour and often determine the extent of hyperpolarization. Notably, hyperpolarization can also be affected by high levels of noise and is relevantly influenced by the choice of ϕ and ε which specify the network behaviour in the dynamic model.

With the inclusion of evolving networks, we were able to observe fragmentation into independent extreme opinions instead of hyperpolarization along a single axis when nodes prefer to connect to nodes that are very close to them in terms of opinion. This can be seen as an interesting starting point for the examination of effects of various connectivity configurations and shows that social media might not necessarily lead to the high hyperpolarization currently observed in the US: By leveraging the increased control over whom they connect to which is provided by social media, people can avoid societal polarization. However, it is not clear whether the resulting echo chambers and fragmentation are necessarily preferable from a societal point of view, especially if opinions still consistently increase in extremeness, as observed in the generalized WBT model with dynamic graphs.

The impact of bots with fixed opinions on polarization varies but strongly depends on the evaluative extremeness and the connections bots start with and are able to maintain. In our simulations placing bots in the most connected nodes and letting them connect to a wider

range of other nodes increased the effectiveness of bots, both polarizing and mediating ones. This is interesting as Stella et al. [2018] came to the conclusion that social bots (in this case using tweets in twitter) mainly target human influencers on both sides of the opinion spectrum within a network, which can be interpreted as an attempt to leverage the influencers' connections to effectively influence many others' opinions. This is consistent with strategic behaviour that would be implied by extrapolating from our model.

While our model in its current form is already able to provide a wide range of behaviour with different parameter choices (see figure 22), there is an abundance of possible extensions that could be performed. Firstly, the models do not consider factors like home city and hobbies which are important in the formation of real-life social networks, as well as the effect described by Krosnick [1990], concerning interpersonal differences in the relative importance of certain topics and the extent to which they influence a person's attitude towards others. Also, historical data on political opinions and social network structures could be used to better understand which model parameters best describe real world opinion dynamics in different contexts. This might be especially important for the evaluative extremeness parameter, which has a huge impact on polarization produced by our model and has so far only been empirically investigated in the context of the highly polarized 2016 US presidential evaluation. More generally, an empirical investigation of the update dynamics specified by the Weighted Balance Theory model could be useful, as the model has some unintuitive properties: Apart from consistently producing extreme opinions for positive evaluative extremeness, even for isolated clusters of nodes with the same opinion, the model implies that people moderate their opinion when meeting somebody they feel neutral about. Next, the initialization of the model can affect its limiting behaviour and therefore, current trends might be better described with models initialized with pre-set biased opinions other than randomly distributed ones with mean zero. Finally, expanding on our experiments with social bots, it could be interesting to see how their connecting strategy can be improved to maximize their effect on polarization in either direction. This could be done using simple heuristics, or more complex strategies that react to an evolving network, learnt by reinforcement learning. Similarly, further work could identify strategies society and individual citizens can employ to counter the influence of bots that do not come at the price of increased fragmentation.

References

Alan I. Abramowitz and Kyle L. Saunders. Is polarization a myth? *The Journal of Politics*, 70(2):542–555, 2008.

Angela J Bahns, Christian S Crandall, Omri Gillath, and Kristopher J Preacher. Similarity in relationships as niche construction: Choice, stability, and influence within dyads in a free choice environment. *Journal of Personality and Social Psychology*, 112(2):329, 2017.

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

John Bohannon. The pulse of the people. *Science*, 355:470–472, 2017.

Claudio Castellano, Matteo Marsili, and Alessandro Vespignani. Nonequilibrium phase transition in a model for social influence. *Physical Review Letters*, 85(16):3536, 2000.

Claudio Castellano, Daniele Vilone, and Alessandro Vespignani. Incomplete ordering of the voter model on small-world networks. *EPL (Europhysics Letters)*, 63(1):153, 2003.

Philip E. Converse. The nature of belief systems in mass publics. In *Ideology and Discontent*. Free Press, 1964.

Guillaume Deffuant, Frédéric Amblard, Gérard Weisbuch, and Thierry Faure. How can extremism prevail? a study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, 5(4), 2002.

Guillaume Deffuant, Timoteo Carletti, and Sylvie Huet. The leviathan model: Absolute dominance, generalised distrust, small worlds and other patterns emerging from combining vanity with opinion propagation. *J. Artif. Soc. Soc. Simul.*, 16(1), 2013. doi: 10.18564/jasss.2070.

Zakaria el Hjouji, D Scott Hunter, Nicolas Guenon des Mesnards, and Tauhid Zaman. The impact of bots on opinions in social networks. *arXiv preprint arXiv:1810.12398*, 2018.

Andreas Flache and Michael W. Macy. Small worlds and cultural polarization. *The Journal of Mathematical Sociology*, 35(1-3):146–176, 2011. doi: 10.1080/0022250X.2010.532261.

Andreas Flache and Michael Mäs. Why do faultlines matter? a computational model of how strong demographic faultlines undermine team cohesion. *Simulation Modelling Practice and Theory*, 16(2):175 – 191, 2008.

Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4):2, 2017. ISSN 1460-7425. doi: 10.18564/jasss.3521. URL http://jasss.soc.surrey.ac.uk/20/4/2.html.

Christopher Hare and Keith T. Poole. The polarization of contemporary american politics. *Polity*, 46(3):411–429, 2014.

Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5 (3), 2002.

Fritz Heider. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1): 107–112, 1946. doi: 10.1080/00223980.1946.9917275. URL https://doi.org/10.1080/00223980.1946.9917275. PMID: 21010780.

Dirk Helbing. A mathematical model for the behavior of individuals in a social field. *Journal of Mathematical Sociology*, 19(3):189–219, 1994.

Gert Jan Hofstede, Jillian Student, and Mark R. Kramer. The status–power arena: a comprehensive agent-based model of social status dynamics and gender in groups of children. *AI & Society*, 1, 2018. doi: 10.1007/s00146-017-0793-5.

Petter Holme and Mark EJ Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):056108, 2006.

Sylvie Huet and Guillaume Deffuant. Openness leads to opinion stability and narrowness to volatility. *Advances in Complex Systems*, 13(03):405–423, 2010. doi: 10.1142/S0219525910002633.

Wander Jager and Frédéric Amblard. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*, 10:295–303, 2005. doi: 10.1007/s10588-005-6282-2.

Jon A. Krosnick. Government policy and citizen passion: A study of issue publics in contemporary america. *Political Behavior*, 12(1):59–92, 1990.

Kai Kupferschmidt. Social media 'bots' tried to influence the US election. Germany may be next. *Science*, 13, 2017.

MF Laguna, Guillermo Abramson, and Damián H Zanette. Minorities in a model for opinion formation. *Complexity*, 9(4):31–36, 2004.

Thomas Milton Liggett. *Interacting particle systems*, volume 276. Springer Science & Business Media, 2012.

Jan Lorenz. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, 18(12):1819–1838, 2007. doi: 10.1142/S0129183107011789.

Gianluca Manzo and Delia Baldassarri. Heuristics, interactions, and status hierarchies: An agent-based model of deference exchange. *Sociological Methods & Research*, 44(2): 329–387, 2015. doi: 10.1177/0049124114544225.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

Björn Ross, Laura Pilz, Benjamin Cabrera, Florian Brachten, German Neubaum, and Stefan Stieglitz. Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4):394–412, 2019.

Laurent Salzarulo. A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies and Social Simulation*, 9(1):13, 2006.

Simon Schweighofer, Frank Schweitzer, and David Garcia. A weighted balance model of opinion hyperpolarization. *Journal of Artificial Societies and Social Simulation*, 23(3): 5, 2020. ISSN 1460-7425. doi: 10.18564/jasss.4306. URL http://jasss.soc.surrey.ac.uk/23/3/5.html.

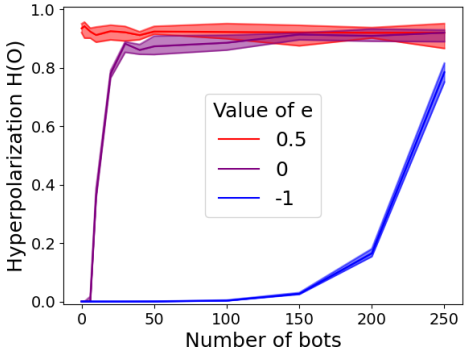
Vishal Sood and Sidney Redner. Voter model on heterogeneous graphs. *Physical review letters*, 94(17):178701, 2005.

Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440, 2018.

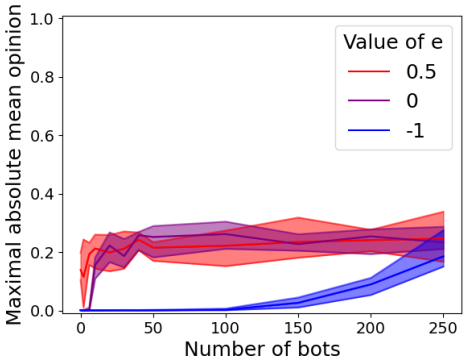
Katarzyna Sznajd-Weron and Jozef Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000.

Thomas Wood and Ethan Porter. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1):135–163, 2019.

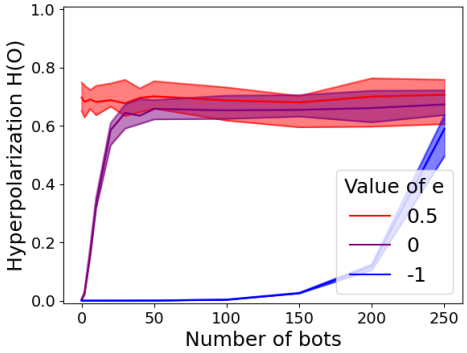
A Appendix



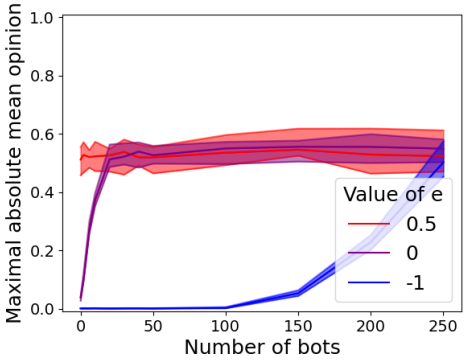
(a) Hyperpolarization: minimal initial opinion at -0.75



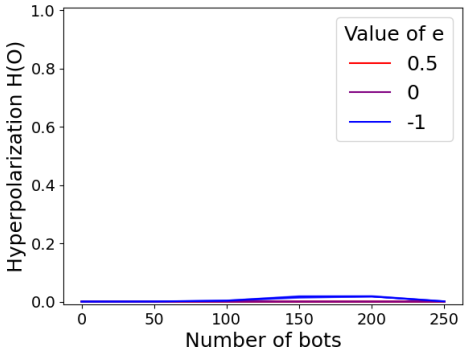
(b) Maximum (over dimensions) absolute mean opinion (graph as in a)



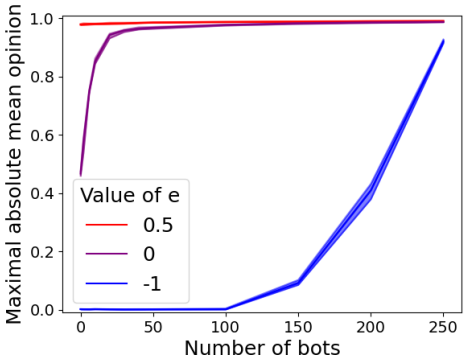
(c) Hyperpolarization: minimal initial opinion at -0.5



(d) Maximum (over dimensions) absolute mean opinion (graph as in c)

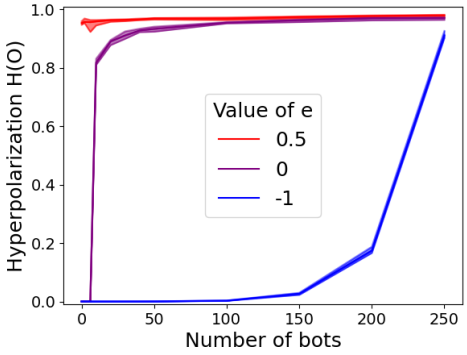


(e) Hyperpolarization: minimal initial opinion at 0

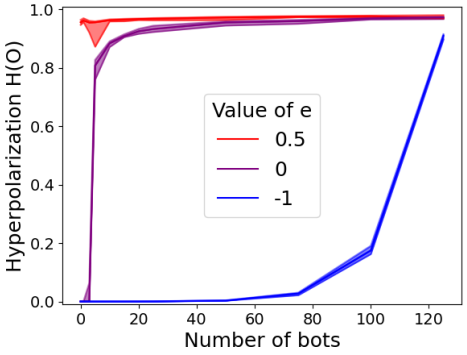


(f) Maximum (over dimensions) absolute mean opinion (graph as in e)

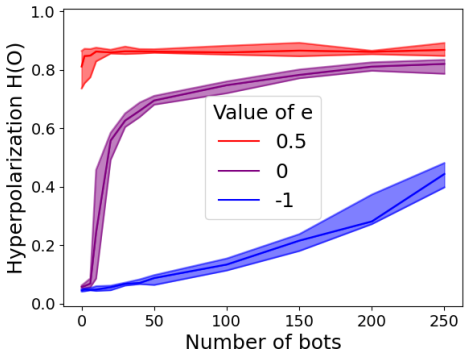
Figure A.23: Weighted balance model on fully connected graph at convergence/after 100k single updates for $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$, $f(x) = \text{sign}(x) \cdot |x|^{1-e}$ and varying amount of bots with $o = 1$ for all dimensions. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Bots excluded from the calculation of H and mean opinion.



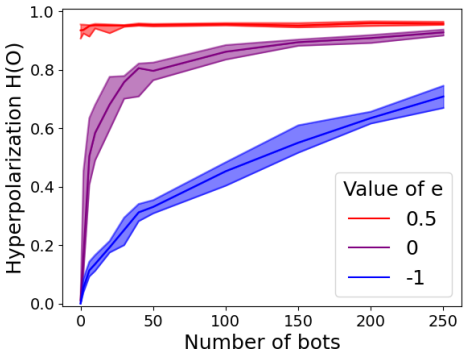
(a) Bots with $o = 1$ for all dimensions. Fully connected graph.



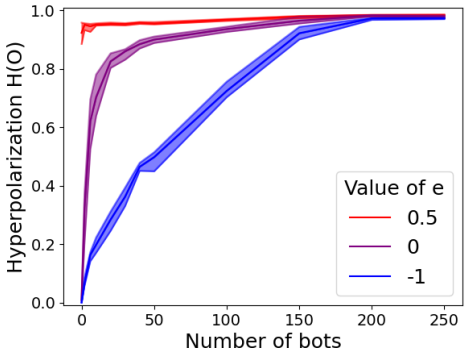
(b) Half of bots with $o = 1$ and other half with $o = -1$. Fully connected graph. Number refers to bots per side.



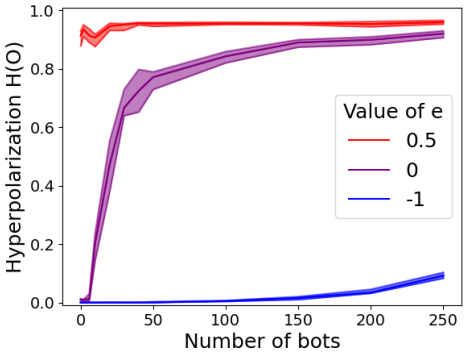
(c) Bots with $o = 1$ placed randomly, 499 uniformly sampled edges



(d) Bots with $o = 1$ placed randomly, Barabási–Albert model with 499 edges

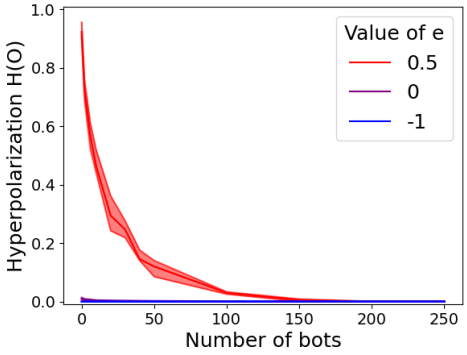


(e) Bots with $o = 1$ placed in most connected nodes, Barabási–Albert model with 499 edges.

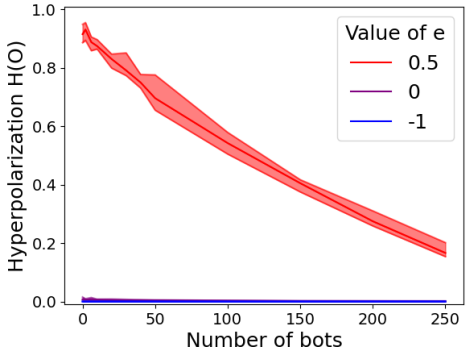


(f) Bots with $o = 1$ placed in least connected nodes, Barabási–Albert model with 499 edges.

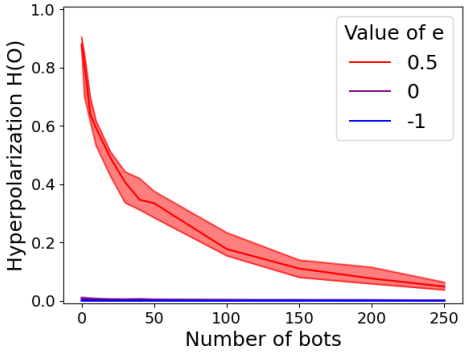
Figure A.24: Hyperpolarization at convergence/after 250k single updates for static graph Weighted Balance model, $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$, $f(x) = \text{sign}(x) \cdot |x|^{1-e}$ and varying amount of bots with $o = 1$ for all dimensions. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Bots excluded from the calculation of H.



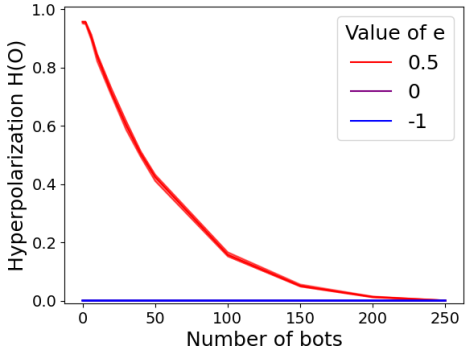
(a) Step limit of 250k, bots placed in most connected nodes



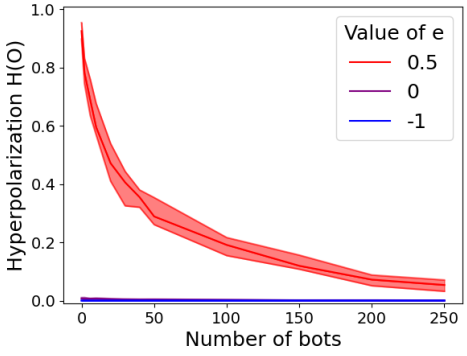
(b) Step limit of 250k, Bots placed in least connected nodes



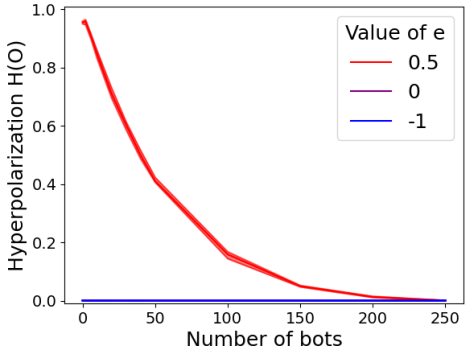
(c) Bots placed randomly



(d) Fully connected graph.

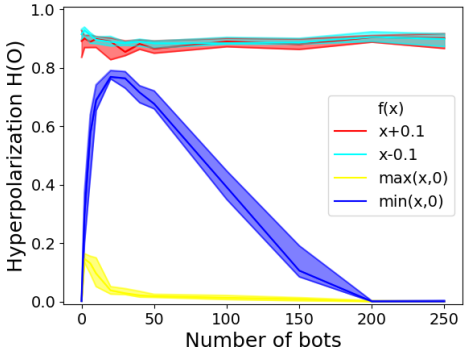


(e) Step limit of 250k, bots placed randomly

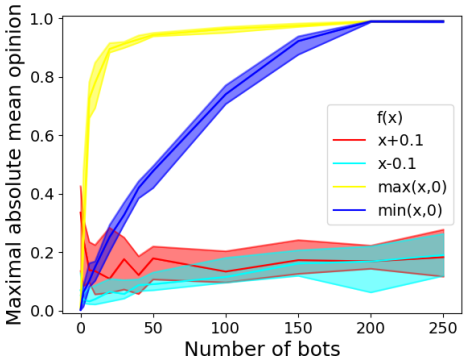


(f) Step limit of 250k, fully connected graph.

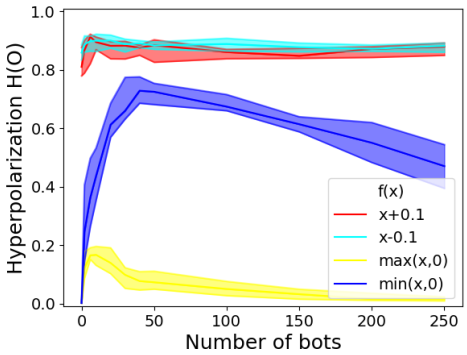
Figure A.25: Hyperpolarization with neutral bots with $o = 0$ for all dimensions in static graph Weighted Balance with 499 edges, $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$, $f(x) = \text{sign}(x) \cdot |x|^{1-e}$. Shown results at convergence/100k single steps and for Barabási–Albert graph unless stated otherwise. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Bots excluded from the calculation of H..



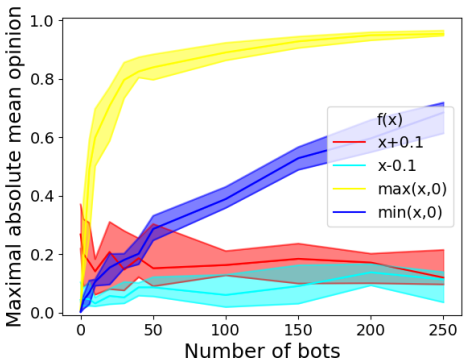
(a) Hyperpolarization: Step limit of 250k



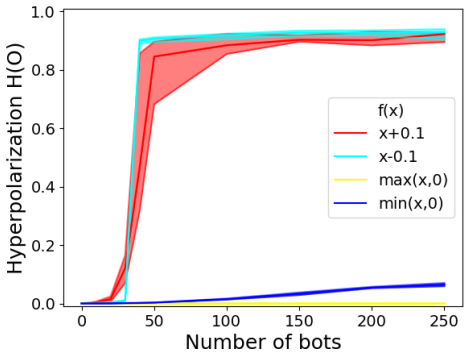
(b) Maximum (over dimensions) absolute mean opinion: Step limit of 250k



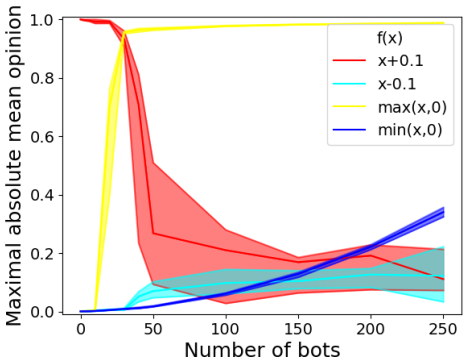
(c) Hyperpolarization: Bots placed randomly



(d) Maximum absolute mean opinion: Bots placed randomly

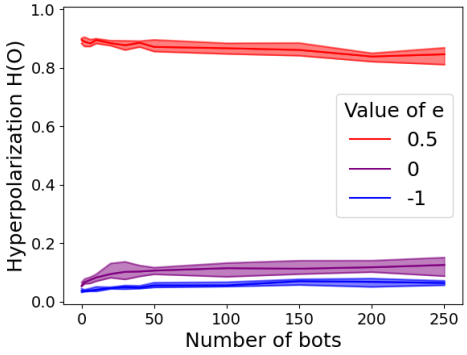


(e) Hyperpolarization: fully connected graph

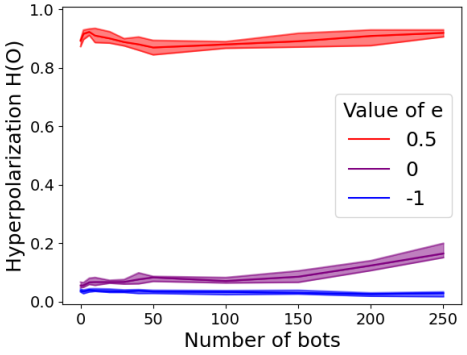


(f) Maximum absolute mean opinion: fully connected graph.

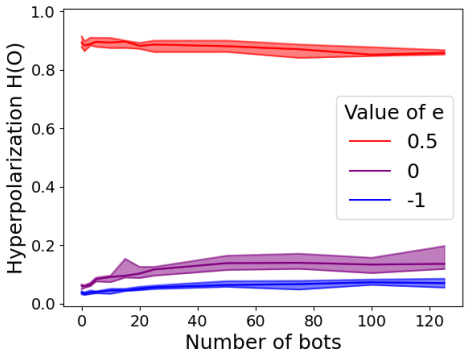
Figure A.26: Static graph Weighted Balance model with 499 edges with $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$ and varying amount of bots with $o = 1$ for all dimensions. Bots placed in the most connected nodes, model at convergence/after 100k single updates and on randomly initialized Barabási–Albert graph unless stated otherwise. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Bots excluded from the calculation of H and mean opinion.



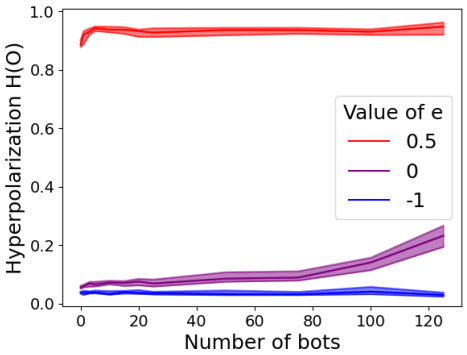
(a) $\varepsilon = 0.6$. Bots with constant opinion 1 and constrained by ε.



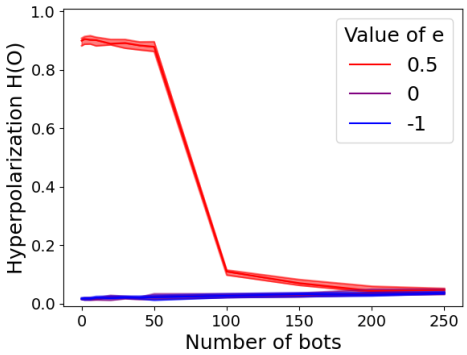
(b) $\varepsilon = 0.6$. Bots with constant opinion 1 and not constrained by ε.



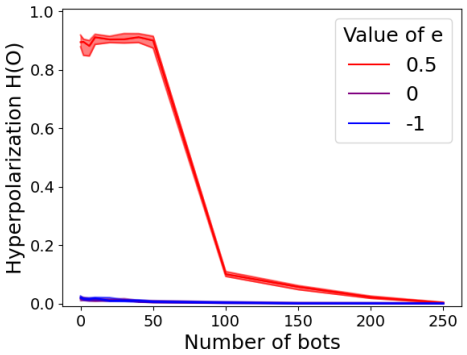
(c) $\varepsilon = 0.6$. Half of bots with constant opinion 1 and -1 each. Bots constrained by ε.



(d) $\varepsilon = 0.6$. Half of bots with constant opinion 1 and -1 each. Bots not constrained by ε.

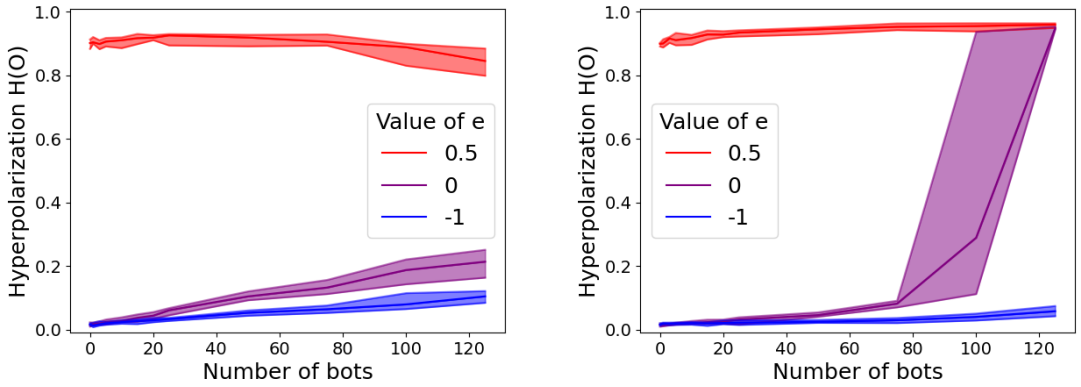


(e) $\varepsilon = 1$. Neutral bots with constant opinion 0. Bots constrained by ε.



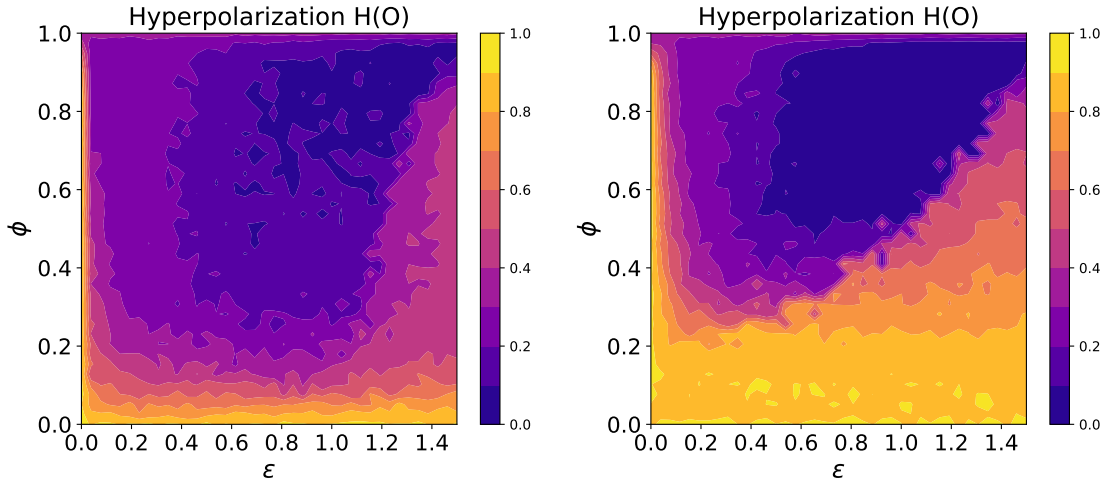
(f) $\varepsilon = 1$. Neutral bots with constant opinion 0. Bots not constrained by ε.

Figure A.27: Dynamic graph weighted Balance model on randomly initialized Barabási–Albert graph at convergence/after 100k single updates with $\phi = 0.5$ and distance-based connect function, as well as 499 edges with $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$. Bots placed randomly. Area between 10%- and 90%-empirical quantile over 10 runs shaded. Bots excluded from the calculation of H.



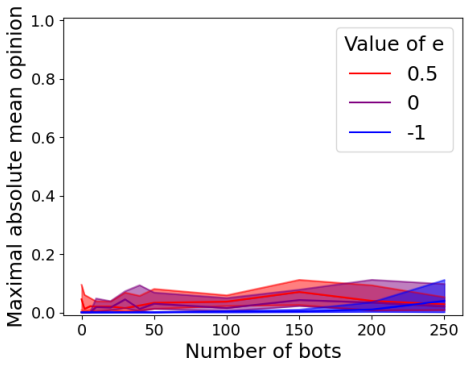
(a) Bots constrained by ε (b) Bots unconstrained by ε

Figure A.28: Weighted Balance model on randomly initialized Barabási–Albert graph at convergence/after 100k single updates with $\phi = 0.5$ and distance-based connect function with $\varepsilon = 1$, as well as 499 edges with $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$. Equal amount of bots with opinion 1 along all dimensions and bots with opinion -1 along all dimensions placed randomly in the graph. Number of bots refers to the number of bots per side.. Area between 10%- and 90%- empirical quantile over 10 runs shaded. Bots excluded from the calculation of H.

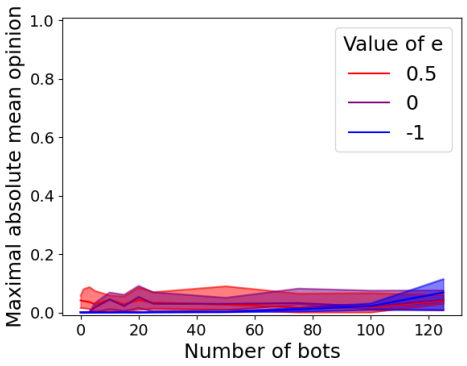


(a) 200 bots constrained by ε (b) 200 bots unconstrained by ε

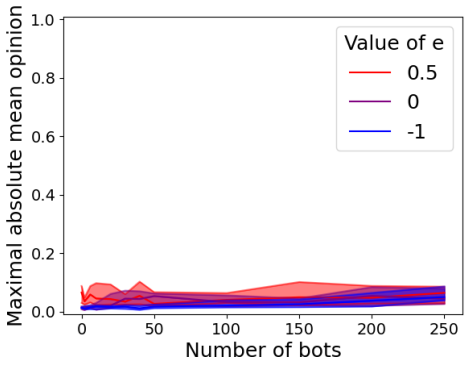
Figure A.29: Hyperpolarization for different levels of ε and ϕ in Barabási–Albert graph with 499 edges at convergence/after 100k single updates with $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$. Bots with $o = 1$ for all dimensions placed in random nodes and excluded from calculation of H.



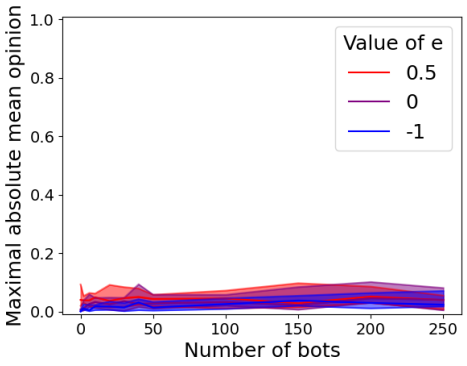
(a) Fully connected graph, bot opinion constant 1. Bots placed randomly.



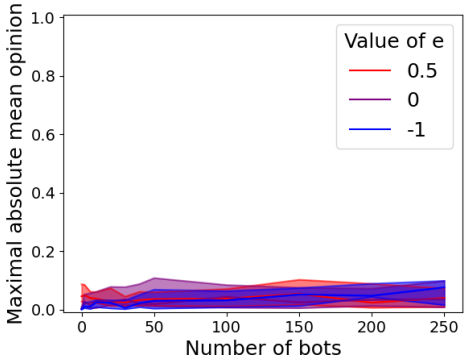
(b) Fully connected graph, half of bots with opinion 1, half with opinion -1. Number of bots refers to number per side. Bots placed randomly.



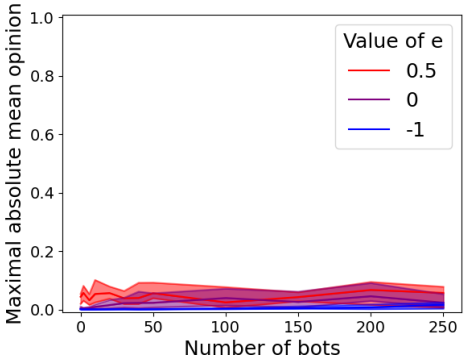
(c) Edges sampled uniformly random, bot opinion constant 1. Bots placed randomly.



(d) Barabási–Albert graph, bot opinion constant 1. Bots placed randomly.

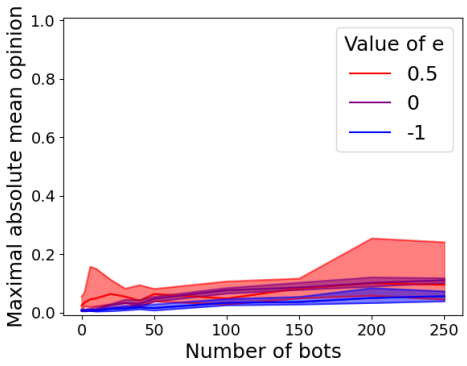


(e) Barabási–Albert graph, bot opinion constant 1. Bots placed in most connected nodes.

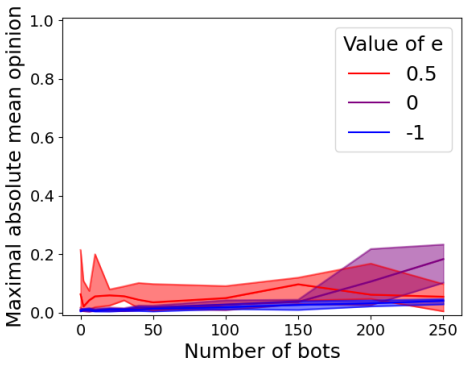


(f) Barabási–Albert graph, bot opinion constant 1. Bots placed in least connected nodes.

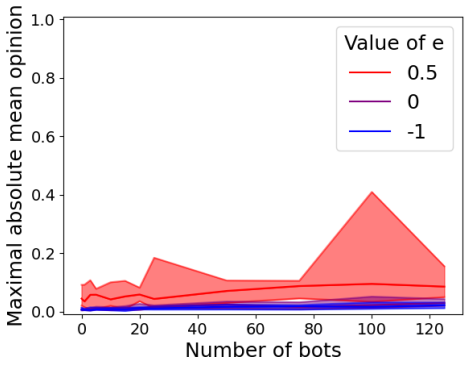
Figure A.30: Maximum (over dimensions) absolute mean opinion (excluding bots) at convergence/after 100k single updates for Weighted Balance model with 499 edges, $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$, $f(x) = \text{sign}(x) \cdot |x|^{1-e}$. Area between 10%- and 90%- empirical quantile over 10 runs shaded.



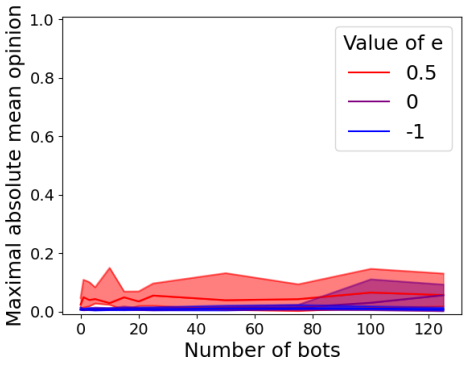
(a) $\varepsilon = 1$. Bots with $o = 1$, affected by ε.



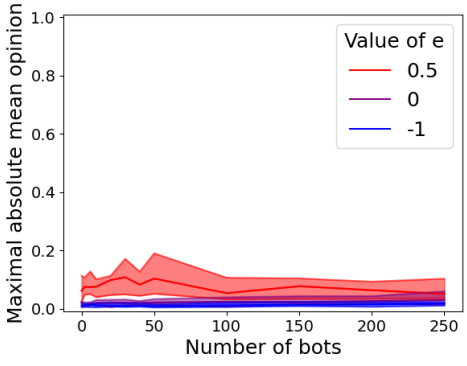
(b) $\varepsilon = 1$. Bots with $o = 1$, unaffected by ε.



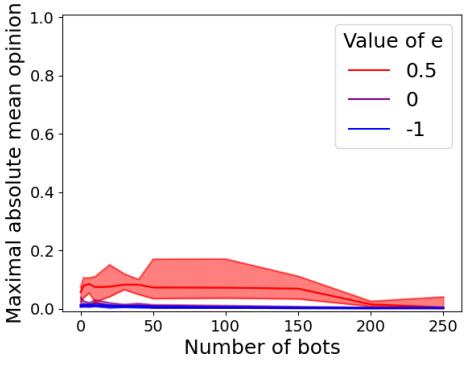
(c) $\varepsilon = 1$. Half of bots with $o = 1$, half with $o = -1$. Number of bots refers to number of bots per side. Bots affected by ε.



(d) $\varepsilon = 1$. Half of bots with $o = 1$, half with $o = -1$. Number of bots refers to number of bots per side. Bots unaffected by ε,



(e) $\varepsilon = 0.6$. Neutral bots with $o = 0$, affected by ε.



(f) $\varepsilon = 0.6$. Neutral bots with $o = 0$, unaffected by ε.

Figure A.31: Maximum (over dimensions) absolute mean opinion (excluding bots) weighted Balance model on randomly initialized Barabási–Albert graph with 499 edges at convergence/after 100k single updates with $\phi = 0.5$, distance-based connect function as well as $N = 500$, $D = 3$, $\alpha = 0.4$, $z = 0.01$. Bots placed in graph randomly. Area between 10%- and 90%- empirical quantile over 10 runs shaded.