Polarizing Opinion Dynamics with Confirmation Bias

Tianyi Chen[1], Xu Wang[1], and Charalampos E. Tsourakakis ✉[1,2]

[1] Boston University, Boston, USA
{ctony,xuwang,ctsourak}@bu.edu
[2] ISI Foundation, Turin, Italy

Abstract. Social media and online networks have enabled discussions between users at a planetary scale on controversial topics. However, instead of seeing users converging to a consensus, they tend to partition into groups holding diametric opinions. In this work we propose an opinion dynamics model that starts from a given graph topology, and updates in each iteration both the opinions of the agents, and the *listening* structure of each agent, assuming there is confirmation bias. We analyze our model, both theoretically and empirically, and prove that it generates a listening structure that is likely to be polarized. We show a novel application of our model, specifically how it can be used to find polarized niches across different Twitter layers. Finally, we evaluate and compare our model to other polarization models on various synthetic datasets, showing that it yields equilibria with unique characteristics, including high polarization and low disagreement.

1 Introduction

Nowadays opinions are increasingly shaped by online interactions that take place on social media platforms, including Facebook, Twitter, Reddit, Instagram among others. Users are exposed to a diversity of opinions at a planetary scale, and to authoritative sources of information. However, instead of observing a convergence of opinions on important social topics, we observe an increasing amount of polarization [24], and the widespread of misinformation [4]. We observe that users cluster into groups with diametrically opposite opinions on numerous important social issues such as gun control, and vaccination against COVID-19. The negative implications of this phenomenon can be devastating, leading to a widening political divide, conflict, and radicalization [8]. Furthermore, polarization facilitates the spread of misinformation [7]. One of the roots of evil behind the polarization phenomenon is a human cognitive bias, *confirmation bias*. Specifically, biased assimilation or confirmation bias is the phenomenon according to which individuals process new information in a biased way towards existing beliefs, or expectations. Lord et al. have shown that two people with initially different opinions/conflicting views can examine the same evidence and find reasons to increase the strength of their existing opinions [25]. The groups of users with homogeneous opinions are also known as echo chambers [10].

Understanding how polarization and echo chambers naturally emerge on social media, is a subject of paramount importance, and of interest to a diverse group of researchers, in sociology, economics, and computer science. In recent years, numerous polarizing opinion dynamics models [6, 12, 15, 19, 22] have been proposed. In this work, we propose a simple model of opinion dynamics that extends the model of opinion dynamics used by Abebe et al. [3], that extends the classic work of Friedkin and Johnsen [13,17]. Our model is iterative (i.e., discrete time), and initially starts with a directed network structure G of n agents with confirmation bias that have an initial opinion on a given topic. We also assume that arcs are weighted and normalized; the weight $W_{u\to v}$ corresponds to the social influence strength of u to v, namely how much v "listens" to u. In each iteration, the agents update their opinion by taking into account their initial opinion, and the opinions of their in-neighbors, and also adapt the *listening structure*. Originally, the listening structure is identical to the input network, but the agents due to their confirmation bias tend to strengthen connections towards neighbors that share the same type of opinion as theirs. We focus on unidimensional opinions, that are normalized in the range of $[-1, 1]$; it is known that despite their unidimensionality, they can capture opinions over a multidimensional set of issues due to the political "left-right" spectrum [9,16]. Our two main contributions are the following:

- We propose a Friedkin-Johnsen type model of opinion dynamics that incorporates confirmation bias (FJCB) to modify in each iteration the *listening structure* of each agent. We analyze the dynamics, by finding the equilibrium opinions and listening structure, and by proving the structural properties of the listening structure.

- We perform several experiments, both on synthetic and real datasets, and compare our method to other polarizing opinion dynamics models with respect to different measures. We show that our method yields listening structures with high polarization, and low opinion disagreement, resembling echo chambers.

- We show a novel application of our model in predicting the ideological community participation of Twitter users in the retweet network (i.e., the graph formed by retweets), using information from the follow network.

Notation. We use the terms agent and node interchangeably. For any agent u, we denote its opinion at time t as $x_u(t)$, and its initial opinion as s_u. A directed edge (v, u) means that node u *listens* to node v. The initial listening structure is the weighted graph $G(V, E, w)$ and the weights of the incoming edges (if there exist any) to any node u sum up to 1. The weight w of the edge (v, u) captures the influence of v on u. The in-neighborhood of a node u at time t is denoted as $N_u^-(t)$. While the weights of the edges may change, $N_u^-(t+1) \subseteq N_u^-(t)$, i.e., no edges are added but some may be deleted. Similarly, the listening structure at time t is $\mathcal{L}(t)$, and initially it is equal to G. The weight of an edge (v, u) at time t is denoted as $W_{v\to u}(t)$. To denote the equilibrium, we use $\star$ as a superscript.

2 Related work

Opinion dynamics is the study of how agents interact with one another and reach (or perhaps not) consensus. It has been a topic of intense study by multiple disciplines. We discuss two important models that lie close to our work. DeGroot introduced a continuous opinion dynamics model [13]. The model is based on repeated averaging. Specifically an agent updates her opinion to the weighted average of her neighbors' and her own opinion from the previous time step. Friedkin and Johnsen [18] extended the DeGroot model by including in their model that each individual has certain innate beliefs. Other models allow each agent to have a different degree of "stubbornness", see [3]. The stubbornness of the agents is measured by a vector $\alpha \in [0,1]^V$, where value of α_i close to one means that agent i is more resistant towards keeping their own innate opinion. According to this model (**FJ**), the opinion $x_i(t+1)$ of node i at time $t+1$ is equal to

$$x_i(t+1) = \alpha_i s_i + (1-\alpha_i)\frac{\sum\limits_{j\in N_i} w_{ij}x_j(t)}{deg(i)}. \tag{1}$$

Here, $deg(i) = \sum\limits_{j\in N_i} w_{ij}$ is the weighted degree of node i, and s_i be the innate belief of node i. Recently Auletta et al. [5] extended this model by evolving stubbornness and social relations, and proved such dynamics converges to a consensus with reasonable conditions. The interested reader may refer to the survey by Mossel and Tamuz and references therein for more related work on opinion dynamics [26].

Garimella et al. [20] proposed a pipeline that constructs Twitter network datasets with controversial topics from different domain. The pipeline first builds a conversation graph related to a topic, e.g. Twitter follow graph and retweet graph; Then it splits the graph into two partitions with a graph partitioning algorithm, e.g. METIS [23]; Finally the controversy of this topic is measuresd by how well two partitions are connected.

Polarization and disagreement. We use the notions of polarization and disagreement as introduced by Musco, Musco, and Tsourakakis [27]. While these notions were introduced for the equilibrium point x^* of a convergent opinion dynamics model, the same notions are applicable to any vector of opinions x in a graph $G(V,E)$. Let $\bar{x}$ be the mean-centered vector, i.e., $\bar{x} = x - \frac{x^T\vec{1}}{n}\vec{1}$. The disagreement $d_{uv}(x)$ of edge (u,v) is defined as $d_{uv}(x) = w_{uv}(x_u - x_v)^2$, and the total disagreement $D_G(x)$ and polarization $P(x)$ that intuitively captures how agents' opinions deviate from the average opinion are defined respectively as

$$D_G(x) = \sum_{(u,v)\in E} d_{uv}(x), \quad \text{and} \quad P(x) = \bar{x}^T\bar{x}. \tag{2}$$

Polarization models. In recent years, there has been an increased interest in developing polarization models. Dandekar et al. [12] propose the model Biased

Opinion Formation (**BOF**), and prove that under certain conditions it can explain why extreme polarization occurs. In their model, each agent at time $t+1$ updates its opinion according to the following equation:

$$x_i(t+1) = \frac{W_{ii}x_i(t) + (x_i(t))^{b_i}s_i(t)}{w_{ii} + (x_i(t))^{b_i}s_i(t) + (1-x_i(t))^{b_i}(d_i - s_i(t))}. \quad (3)$$

Here, $s_i(t) = \sum_{j \in N_i} w_{ij}x_j(t)$ is the weighted sum of the opinions of i's neighbors, d_i is i's weighted degree and $b_i \geq 0$ is a bias parameter.

Given d topics, Hazla et al. [21] model an agent's opinion $u \in \mathcal{R}^d$ as a vector that lies on a d dimensional Euclidean sphere. Any global intervention v affects the opinion vector of the agent proportional to $\langle u, v\rangle \cdot v$. They show that opinions polarize if there are one or more influencers sending interventions strategically, heuristically, or randomly. Gaitonde et al. [19] further generalize the study by proving the polarization of opinions exhibits with higher opinion dimension and network interactions. Vicario et al. [14] develop two variants of Bounded Confidence Model(BCM), i.e., a class of models where two agents interact only if they are connected and their opinions are close enough. Specifically, when the distance between opinions from two connected agents is larger than a pre-defined tolerance, with certain probability, their first model rewire such connection while the second model push agents' opinions further. Close to this is the ECHO model proposed by Sasahara et al. [29]. Derived from BCM, it simulates the phenomenon of biased assimilation on online social media. Specifically, each agent expresses its opinions by posting *messages*, and receiving information from neighbors through checking *messages* within a sized *screen*. Furthermore, at the end of each iteration, ECHO also rewire the connection between two agents if the distance between their opinions is larger than the tolerance, thus is able to create echo chambers.

3 Proposed Model

Our proposed model is iterative. At round/time 0, each node u holds its original value s_u. Each round consists of two steps during which agents update (i) their expressed opinion, and the (ii) the strength of their connections. For the former step we use a popular variation of the Friedkin-Johnsen model that incorporates stubbornness parameters $\{\alpha_i\}_{i \in V}$ [3]. Equation (4) describes how each node u updates its value from round $t-1$ to t.

$$x_u(t) = \alpha_u s_u + (1-\alpha_u) \sum_{v \in N_u^-(t-1)} W_{v \to u}(t-1)x_v(t-1) \quad \forall u \in V, t \in \mathbb{N} \quad (4)$$

Initially, $x_u(0) = s_u$ for all $u \in V$. Once the agents update their values, they update the strength of their *incoming* connections; an agent can only control how much influence other nodes exert on her, rather than how much influence she can exert on her neighbors. The next step of the proposed method encodes a

well-known human bias, the *confirmation bias* (aka biased assimilation) that is the psychological tendency to value more evidence, regardless its validity, that reinforces already held beliefs [28]. We do this as follows: an agent i prefers to increase the (relative) strength of connection $j \to i$ according to the values of its endpoints and a positive parameter η that adjusts the changing scale:

$$W_{i\to j}(t) = \max(0, W_{i\to j}(t-1) + \eta x_i(t) x_j(t)) \qquad \forall (i,j) \in A(G(t-1)) \quad (5)$$

By convention, an arc of weight of zero does not exist, and we delete an arc if its weight become nonpositive. Observe that when $\eta = 0$, edge weight does not change at all. And the edges can quickly got enhanced/eliminated as we increase the value of η, thus reveal community information, or even reshape the graph structure. Once the weights are updated according to Equation (5), each node updates the weights of its incoming edges according to the following equation:

$$W_{i\to j}(t) = \begin{cases} \frac{W_{i\to j}(t)}{\sum_{k\in N^-(j)} W_{k\to j}(t)} & \text{if } \sum_{k\in N^-(j)} x_k(t) W_{k\to j}(t) \neq 0 \quad (\underline{\text{Case I}}) \\ W_{i\to j}(t-1) & \text{if } \sum_{k\in N^-(j)} x_k(t) W_{k\to j}(t) = 0 \quad (\underline{\text{Case II}}) \end{cases} \quad (6)$$

In Case I, once we normalize according to Equation (6), the weights of the incoming arcs sum up to one. If the incoming influence is equal to 0, we leave the weights as they were in the previous iteration. This completes one full iteration. The model proceeds to the next iteration, and continues until convergence or until the maximum number of iterations is reached. Under the assumption that an equilibrium point exists, we can find it. This is stated as the next theorem. Understanding the convergence of our dynamics is an interesting open question.

Theorem 1. *Suppose the dynamical system converges to an equilibrium point (x^*, W^*). At equilibrium, the opinion of any node $u \in V(G)$ satisfies:*

$$x_u^\star = \begin{cases} \alpha_u s_u + (1-\alpha_u) \frac{\sum_{k\to u} x_k^{\star 2}}{\sum_{k\to u} x_k^\star} & \text{if } \sum_{k\in N_u^{-\star}} x_k^\star \neq 0, \sum_{k\in N_u^{-\star}} W_{k\to u}^\star x_k^\star \neq 0, \\ \alpha_u s_u & N_u^{-*} = \{v : W_{v\to u}^\star > 0\} = \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Proof. We consider three cases, depending on the listening structure at equilibrium, and the social influence of the in-neighbors of a node.
Case I: Consider a node u at equilibrium whose local listening structure satisfies $\sum_{k\in N_u^{-\star}} x_k^\star \neq 0$ and $\sum_{k\in N_u^{-\star}} W_{k\to u}^\star x_k^\star \neq 0$. By the definition of an equilibrium we obtain that the following equations are satisfied

$$x_u^\star = \alpha_u s_u + (1-\alpha_u) \sum_{k\to u} W_{k\to u}^\star x_k^\star \quad (8)$$

$$W_{v\to u}^* = \frac{W_{v\to u}^* + \eta x_v^* x_u^*}{\sum_{k\to u} \left(W_{k\to u}^* + \eta x_k^* x_u^*\right)} = \frac{W_{v\to u}^* + \eta x_v^* x_u^*}{1 + \eta x_u^* \sum_{k\to u} x_k^*} \quad (9)$$

Simplifying equation (9), yields that the edge weight at equilibrium satisfies

$$W^\star_{v\to u}\eta x^\star_u \sum_{k\to u} x^\star_k = \eta x^\star_u x^\star_v \Rightarrow W^\star_{v\to u} = \frac{x^\star_v}{\sum_{k\to u} x^\star_k}.$$

By substituting this value in Equation (8), we obtain the following expression:

$$x^\star_u = \alpha_u s_u + (1-\alpha_u)\sum_{k\to u} W^\star_{k\to u} x^\star_k = \alpha_u s_u + (1-\alpha_u)\sum_{k\to u} \frac{x^\star_k}{\sum_{k\to u} x^\star_k} x^\star_k \to$$
$$x^\star_u = \alpha_u s_u + (1-\alpha_u)\frac{\sum_{k\to u}(x^\star_k)^2}{\sum_{k\to u} x^\star_k}.$$

Case II: Suppose there exists a node u at equilibrium such that $\sum_{k\in N^-_u} W_{k\to u} x^\star_k = 0$. This can happen in two cases, depending on whether the node listens to some or none of the rest of the nodes.

Case (a): Node u is not listening to any node at equilibrium, i.e., $N^{-\star}_u = \emptyset$. In this case $x^\star_u = \alpha_u s_u$.

Case (b): Suppose $N^{-\star}_u \neq \emptyset$. Without loss of generality, we can partition the in-neighborhood in three sets $S_{pos}, S_{neutral}, S_{neg}$, depending on whether the nodes have positive, neutral, or negative opinion respectively. Clearly, $S_{pos}, S_{neg} \neq \emptyset$, and

$$\sum_{k\in S_{pos}} W^\star_{k\to u} x^\star_k + \sum_{k\in S_{neg}} W^\star_{k\to u} x^\star_k = 0.$$

Furthermore, node u at equilibrium has to be neutral, i.e., the term $x^\star_u = a_u \cdot s_u$ has to be equal to 0. If not, then $x^\star_u \neq 0$, and this contradicts the equilibrium property of the edge weights from S_{pos}, S_{neg} to u, i.e., $\exists k \in S_{pos}, k' \in S_{neg}$ such that $W_{k\to u}$ will increase, $W_{k\to u}$ will decrease (i.e., if $x^\star_u > 0$) or vice versa if $x^\star_u < 0$.

Case III: If $\sum_{k\in N^{-\star}_u} x^\star_k = 0$ and $\sum_{k\in N^-_u} W_{k\to u} x^\star_k \neq 0$, then with a similar reasoning, we obtain $x^\star_u = 0$. ■

It is worth mentioning that in case II(b), one can also prove using contradiction that there exist no edges between S_{pos}, S_{neg}. In the following, we prove important structural properties of the equilibrium, that show that our model achieves a certain type of polarization, and exhibits an interesting structure depending the setting of the various model parameters.

Lemma 1. *Let $\mathcal{L}_{\text{eq}}$ be the listening structure at equilibrium. Consider a non-neutral node u whose local listening structure $N^{-\star}_u$ satisfies $\sum_{k\in N^{-\star}_u} x^\star_k \neq 0$ and $\sum_{k\in N^{-\star}_u} W^\star_{k\to u} x^\star_k \neq 0$. Then, all the nodes $v \in N^-_{\text{eq}}(u)$ that node u listens to, share the same opinion, i.e., they have the same opinion sign.*

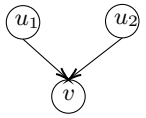


Fig. 1: At equilibrium $W^\star_{u_i \to v} = \frac{x^\star_{u_i}}{Z}, i = 1, 2$ where Z is a normalizing constant. Since the edge weights $W^\star_{u_1 \to v}, W^\star_{u_2 \to v} > 0$ are positive, we observe $sgn(x^\star_{u_1}) = sgn(x^\star_{u_2}) = sgn(Z)$. For details, see Lemma 1.

Proof. The statement trivially holds for any node with in-degree 0 or 1. Consider an arbitrary node node v with at least two in-coming neighbors; let $u_1, u_2 \in N^-(v)$ be two arbitrary such in-neighbors of v in $G^\star$ as shown in Figure 1. At equilibrium the positive edge weights $W^\star_{u_1 \to v}, W^\star_{u_2 \to v}$ satisfy $W^\star_{u_i \to v} = \frac{x^\star_{u_i}}{Z}, i = 1, 2$ where Z is a normalizing constant, as shown in Case I of Theorem 1. Since the edge weights are positive, i.e., $W^\star_{u_1 \to v}, W^\star_{u_2 \to v} > 0$, we obtain $sgn(x^\star_{u_1}) = sgn(x^\star_{u_2}) = sgn(Z)$. Thus, at equilibrium $sgn(x^\star_u) = \sigma \in \{-1, +1\}, \forall u \in N^{-\star}_u$. ■

From now on, if the in-neighbors of a node u have a negative (positive) opinion, we will refer to the in-neighborhood as negative (positive). Can a node u whose in-neighborhood is negative have a positive opinion $x^\star_u$ at equilibrium? The answer to this question is not immediately clear, even when one considers a "stubborn" node u with $\alpha_u = 1$ with a positive initial opinion s_u. While it is clear that node u will never change opinion, it is not (perhaps) clear why u will be connected to nodes of opposite opinion. Should not the edge weight $W_{v \to u}$ between $v \in N^-_u$ and u decrease gradually according to equation (5), and become zero eventually? The answer is no. To see why consider a graph with a single edge $v \to u$, i.e., node u has in-degree 1, and let $x_u > 0 > x_v$. If the edge weight $W_{v \to u}$ is not zeroed-out after the decrease by $\eta x_u x_v$ according to Equation (5), then after the normalization step in equation (6) it remains 1. Notice that in our toy example, the edge weight does not become zero, when the parameter η satisfies $1 + \eta x_u x_v > 0$, or equivalently $\eta < \frac{1}{|x_u||x_v|}$.

The next lemma answers this question more generally. We consider one case for the sign of the in-neighborhood at equilibrium, the other case is symmetric and treated in a similar way. The following lemma is proved in Appendix.

Lemma 2. *Consider a node u with a negative in-neighborhood N^{-*}_u. Then, the opinion of node u at equilibrium can be positive when conditions (i)-(v) hold:*

$$\begin{array}{llll} & (i)\ \ s_u > 0 & \text{and} & (ii)\ \frac{s_u}{\sum_{v \in N^{-*}_u} W^\star_{v \to u} |x^\star_v|} > \frac{1-\alpha_u}{\alpha_u} \\ \text{and} & (iii)\ W^\star_{v \to u} = \frac{1}{|N^{-*}_u|} \quad \forall v \in N^{-*}_u & & \\ \text{and} & (iv)\ x^\star_v = c & \text{and} & (v)\ \eta < \frac{1}{|N^{-*}_u||x^\star_u||x^\star_v|} \end{array}$$

To summarize our results, the typical case of a node u at equilibrium is to be listening to nodes with the same opinion, unless a set of complicated conditions

hold. Furthermore, neutral nodes (as shown in Case II b) can be listening to both negative, and positive opinions. Our experimental results show that almost always on real data, or on data with random opinions, we obtain polarized echo-chambers, unless we create artifacts as described, e.g., in Lemma 2.

4 Experiments

4.1 Experimental setup

Real-world Datasets. We use five publicly available Twitter datasets [11, 20] to evaluate our model. The datasets are summarized in Table 1. Each dataset focuses on a single controversial topic. A *follow* graph and a *retweet* graph are collected based on hashtags related to the topic, and they are converted into undirected graphs, for details see [11, 20]. Both graphs are partitioned into two communities using METIS [23], that can thought of as echo chambers with users with diametric opinions (i.e., positive vs negative). For any Twitter user u, given its neighborhood N_u and the community C_u it belongs to, we assign u's *ideological community participation* (or polarity in short) as $sgn(C_u)\frac{|N_u \cap C_u|}{|N_u|}$. We say that a user u is *persistent* if its polarity in the follow layer is equal to its polarity in the retweet layer.

Topic	# follows	# retweets	# common nodes	# *persistent* nodes
russia march [20]	16 471	2 951	482	302
debate [20]	344 088	44 174	5 015	580
beefban [20]	6 026	1 978	284	120
baltimore [20]	28 291	4 505	356	61
vaxnovax [11]	1 806 164	68 543	17 650	2 753

Table 1: Description of Twitter follow and retweet graphs induced by topic hashtags.

Competitors. We compare our model to three polarizing opinion dynamics models, the FJ model [18], the ECHO model [29] [3], and the BOF model [12]. For ECHO, the numerous parameters of the model (see [29]) are set to the default values, unless specified otherwise.

Machine. All experiments run on a laptop with 3.10GHz Intel Core i5-7267U CPU and 8GB of main memory. The code is written in Python 3 and will become publicly available upon publication.

4.2 Synthetic Experiments

Evaluation. In order to understand the emergence of polarized communities we use the stochastic block-model [2]. We construct two equal-sized communities

[3] `https://github.com/soramame0518/echo_chamber_model`

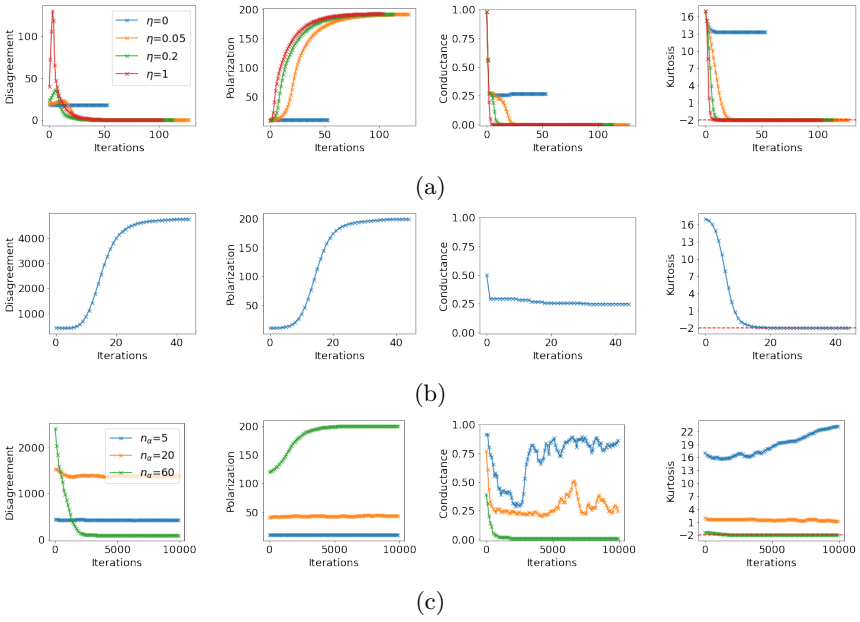


Fig. 2: Disagreement, polarization, conductance, and kurtosis per iteration for models on stochastic block model graphs with $p = 0.2$ and $q = 0.05$ (a): FJCB with different η, note with $\eta = 0$ it becomes FJ, (b): BOF, (c): ECHO with different number of initial stubborn nodes n_α.

with 100 nodes per each. The probability of an edge between two nodes from the same (different) community (communities) is 0.2(0.05). Metrics including polarization, disagreement (see Eq. (2)), conductance of the nodes with positive opinions, and kurtosis of opinions are reported to show how echo chambers are created by our model. Specifically, given a node subset S, we can convert the graph to undirected and calculate its conductance. Kurtosis is a unitless measure of a distribution's shape, and becomes a smaller value when the distribution has a lower tendency for producing extreme values. Formally, given any distribution D with mean μ, variance σ and random variable $X \sim D$, its kurtosis is defined as $Kurt = E[(\frac{X-\mu}{\sigma})^4]$. A completely bimodal distribution is reflected by a kurtosis value of -2.

Initially, we randomly pick n_α *stubborn nodes* from each community and set their opinions to +1 and -1 respectively, and their stubbornness parameter is set to be 1. For FJCB and BOF $n_\alpha = 5$, and for ECHO n_α is ranged between 5, 20 and 60. The initial opinions of the rest of the nodes are set to 0, and their stubbornness parameter to 0.001. Figures 2(a), (b) and (c) plot the four metrics versus the iteration for FJCB, BOF, and ECHO respectively. Note when we set $\eta = 0$, FJCB becomes FJ. Recall the listening structures are directed, but in the case of conductance we consider it to be undirected. For ECHO, we set its confidence distance $\epsilon = 1.01$ in order for the model to work, i.e., opinions can start propagating in the beginning. Some observations follow:

• All methods except FJ can return equally polarized opinions at equilibrium, reaching a value of 200. ECHO requires a large number of stubborn nodes for this to happen (i.e., $n_\alpha = 60$), in contrast to FJCB and BOF. The kurtosis becomes -2, reflecting a completely bimodal distribution.

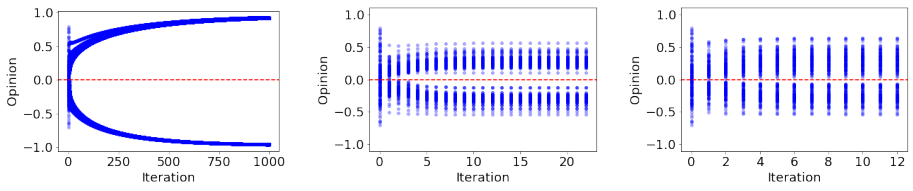
• FJCB converges after at most 120 iterations, for all η values. With non-zero η values, it disconnects the graph into two components whose nodes have different signs of opinions. This happens, within the first 20 iterations, as the conductance drops to 0, even for small values of η. In all our experiments FJCB reaches an equilibrium. FJ, i.e. FJCB with $\eta = 0$, shows limited polarization ability as the changes of its metrics are insignificant compared to other methods.

• FJCB shows the emergence of two polarized niches, with high polarization and low disagreement. BOF (by design) does not alter the listening structure of the network, and therefore disagreement remains high. Our model clearly contrasts the initial graph topology, to the final listening structure that is polarized, and is consistent with the creation of echo chambers [1].

• Interestingly, in Figure 2(a) we observe a rise of disagreement before the final drop. This happens because initially the majority of the nodes are neutral. As their opinions start changing, an increasing number of neighbors produce non-zero disagreement terms, which results in the increase of disagreement. As the method starts converging to echo-chambers, disagreement starts decreasing.

•ECHO does not perform well with respect to creating echo-chambers when there is a small number of stubborn nodes, resulting in an unpredictable performance, as indicated by the fluctuation in the conductance.

Study the effect of η. We also study the effect of parameter η. Figure 2(a) shows the changes we observe with different η values. Empirically we find larger η leads to faster convergence as FJCB can break connections between agents with opposite opinions with fewer iterations. When η becomes 0, the model becomes to FJ with static network structure and loses the ability to simulate extreme polarization.



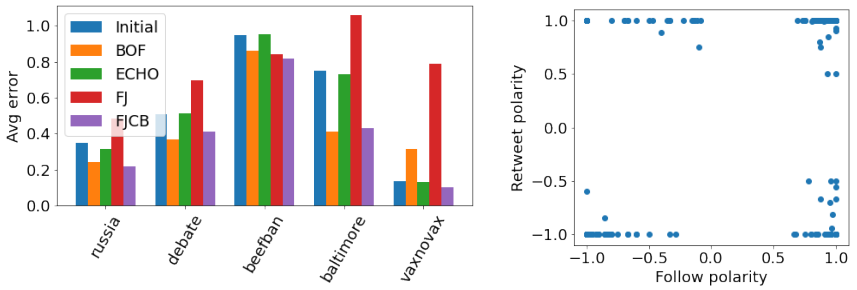
(a) (b) (c)

Fig. 3: Scatter plot of agent opinions over time according to the FJCB model with different choice of α values, and initial opinions sampled from $\mathcal{N}(0, 0.3)$. For any non-stubborn agent u, its stubbornness parameter is set to (a) $\alpha_u = 0.001$. (b) α_u sampled from the half-normal distribution, and then normalized to be in the range of $(0, 1)$. (c) α_u sampled uniformly at random from $(0, 1)$.

Effect of $\boldsymbol{\alpha}$ values. What is the effect of the stubbornness parameters $\{\alpha_u\}_{u \in V}$ on the final opinions? Our results are shown in Figure 3.

We use again the stochastic block model with $n_\alpha = 1$. Additionally, we sample for every non-stubborn node an initial opinion from the normal distribution $N(\mu = 0, \sigma^2 = 0.3^2)$, and adjust the signs according to the block structure, i.e., nodes within the same block have the same opinion, but opposite across blocks. We then assign stubbornness parameters to the non-stubborn nodes in different ways, and plot the polarity vs. the iteration. For Figure 3(a) we set all stubbornness parameters to be 0.001. We see that at equilibrium the non-stubborn nodes have converged to the sign of the stubborn node from their own block. For Figure 3(b), the stubbornness parameters are sampled from the half-normal distribution, and then normalized to be in the range of $(0, 1)$. We observe that due to the presence of fairly stubborn nodes, they do not converge to ± 1 as in Figure 3(a), but still the resulting distribution is bimodal, and polarized. Finally, for Figure 3(c) we sample the stubbornness parameters from the uniform distribution in $(0, 1)$. We observe the same behavior as in (b), with less polarization due to the presence of more stubborn nodes.

4.3 Predicting Echo Chamber Participation



(a) (b)

Fig. 4: (a) Average error of opinion predicted by methods on five Twitter networks. (b) Scatterplot of polarities of the retweet layers vs. the follow layer for the Beefban dataset.

Polarized niches across Twitter layers. In this application, we are interested in understanding how the communities in the *follow* layer of Twitter can predict the structure of echo-chambers in the *retweet* layer. As mentioned in Section 4.1, for any node u we define its polarity in a given Twitter layer as $sgn(C_u)\frac{|\{N_u \cap C_u\}|}{|N_u|}$. Notice that the sign function preserves the polarity of the node with respect to its community C_u, whereas the second term $\frac{|\{N_u \cap C_u\}|}{|N_u|}$ measures the cohesiveness of the node within its community; in the extreme case where $N_u \subseteq C_u$ the cohesiveness is equal to 1. We show here that we can accurately predict the

polarities in the *retweet* layer for the non-persistent users, if we have the *follow* graph and all persistent users. As a baseline, we use the method *Initial* that naively predicts the polarity of a node in the *retweet* layer to be equal to the polarity in the *follow* layer.

On all Twitter datasets, we apply FJCB, FJ, BOF, ECHO. We run these models as follows: we consider the polarities in the *follow* layer as the initial opinions of the agents, and we use the opinions at equilibrium as our predictor for the non-persistent nodes. The follow graph is used as the initial *listening structure*. In both FJCB, FJ and BOF, the stubbornness parameters for persistent users are set to 1, and 0.001 for the rest. We also fix the opinions of persistent users in ECHO throughout the iterations. In Figure 4, we report the average error in the ℓ_1 norm, i.e., the average difference between the predicted opinions and the *retweet* polarities of all users. We observe FJCB has comparable performance to BOF, and both models outperform the *Initial* baseline, but not always. The performance of ECHO is close to *Initial*. FJ has the worst performance over four dataset as it tends to give highly biased predictions, see Figure 5 in Appendix for an example. FJCB is capable of predicting the *retweet* polarities on vaxnovax dataset with less than 0.1 average error. By further investigation, we find the retweet network of this topic is composed of two almost disconnected communities, making its polarities highly concentrated to 1 and -1. All methods cannot accurately predict the *retweet* polarities on Beefban dataset, as the partitions of its follow and retweet graphs are pretty much orthogonal. In Figure 4(b), we see blocks of users on all four corners, since a large number of users belong to different communities on two layers.

Interpretation Given a topic, why models of opinion dynamics can predict the structure of the *retweet* layer with the *follow* layer? In general, *follow* is a long-term connection that can exist due to various reasons, including friendship or interest of an agent in *any* posted content by the other agent. Such a relationship describes a user's community belonging and stance comprehensively from a high level. On the other hand, retweets without quote usually indicate endorsement [20]. In our Twitter datasets, the retweet connections can be considered as users' agreements with respect to a specific topic. Therefore, it is reasonable to regard users' *follow* polarities as their initial opinions, and the *retweet* polarities as their final opinions after all the propagation of information.

5 Conclusion

In this work, we propose FJCB, a Friedkin-Johnsen (FJ) opinion dynamics model that in addition to the classic FJ model incorporates confirmation bias. The model iteratively updates both the opinions of the agents and the listening structure, i.e., to whom each agent listens. We analyze the dynamics, and show that at equilibrium the listening structure is in principle polarized, but it also exhibits interesting structure due to the existence of neutral nodes. We evaluate our model both on synthetic and real data, showing the effect of the various parameters, and its applicability to predicting the echo chamber community participation. An interesting open direction is to prove the convergence of our dynamics, and explore more properties of the equilibrium.

References

1. Concept of echo chamber. `https://en.wikipedia.org/wiki/Echo_chamber_(media)`.
2. E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2015.
3. R. Abebe, J. Kleinberg, D. Parkes, and C. E. Tsourakakis. Opinion dynamics with varying susceptibility to persuasion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1089–1098, 2018.
4. H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.
5. V. Auletta, A. Fanelli, and D. Ferraioli. Consensus in opinion formation processes in fully evolving environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6022–6029, Jul. 2019.
6. F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4):048301, 2020.
7. A. Bessi, F. Petroni, M. Del Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi. Viral misinformation: The role of homophily and polarization. In *Proceedings of the 24th international conference on World Wide Web*, pages 355–356, 2015.
8. L. Boxell, M. Gentzkow, and J. M. Shapiro. Cross-country trends in affective polarization. Technical report, National Bureau of Economic Research, 2020.
9. J. Brandts, A. E. Giritligil, and R. A. Weber. An experimental study of persuasion bias and social influence in networks. *European Economic Review*, 80:214–229, 2015.
10. M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.
11. A. Cossard, G. De Francisci Morales, K. Kalimeri, Y. Mejova, D. Paolotti, and M. Starnini. Falling into the echo chamber: The italian vaccination debate on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):130–140, May 2020.
12. P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.
13. M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.
14. M. Del Vicario, A. Scala, G. Caldarelli, H. Stanley, and W. Quattrociocchi. Modeling confirmation bias and polarization. *Scientific Reports*, 7, 06 2016.
15. M. Del Vicario, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. Modeling confirmation bias and polarization. *Scientific reports*, 7(1):1–9, 2017.
16. P. M. DeMarzo, D. Vayanos, and J. Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly journal of economics*, 118(3):909–968, 2003.
17. N. E. Friedkin and E. C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.
18. N. E. Friedkin and E. C. Johnsen. Social positions in influence networks. *Social Networks*, 19(3):209–222, 1997.

19. J. Gaitonde, J. Kleinberg, and É. Tardos. Polarization in geometric opinion dynamics. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 499–519, 2021.
20. K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1), jan 2018.
21. J. Hazla, Y. Jin, E. Mossel, and G. Ramnarayan. A geometric model of opinion polarization. *CoRR*, abs/1910.05274, 2019.
22. J. Hk, Y. Jin, E. Mossel, G. Ramnarayan, et al. A geometric model of opinion polarization. Technical report, 2021.
23. G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *Siam Journal on Scientific Computing*, 20, 01 1999.
24. E. Klein. *Why we're polarized.* Simon and Schuster, 2020.
25. C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.
26. E. Mossel, O. Tamuz, et al. Opinion exchange dynamics. *Probability Surveys*, 14:155–204, 2017.
27. C. Musco, C. Musco, and C. E. Tsourakakis. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 World Wide Web Conference*, pages 369–378, 2018.
28. R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
29. K. Sasahara, W. Chen, H. Peng, G. Ciampaglia, A. Flammini, and F. Menczer. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4:1–22, 05 2021.

A Appendix

A.1 Proof of Lemma 2

Proof. Consider the opinion $x_u^\star$ of node u at equilibrium with a negative in-neighborhood, it has to satisfy equation (4), i.e.,

$$x_u^\star = \alpha_u s_u + (1-\alpha_u) \sum_{v \to u} W_{v\to u}^\star x_v^\star.$$

When $x_u^\star > 0$, it is necessary that $s_u > 0$, otherwise $x_u^\star < 0$ since the second term in the summation is negative. By rearranging the inequality $\alpha_u s_u + (1-\alpha_u)\sum_{v\to u} W_{v\to u}^\star x_v^\star > 0$ we obtain $\frac{s_u}{\sum_{v\to u} W_{v\to u}|x_v^\star|} > \frac{1-\alpha_u}{\alpha_u}$. Furthermore, $W_{v\to u} = \frac{1}{|N_u^{-*}|}$ for all in-neighbors $v \in N_u^{-*}$. To see why, for the sake of contradiction, assume without loss of generality[4] that there exists an arc $v \to u$ such that $W_{v\to u}^\star < \frac{1}{|N_u^{-*}|}$. Observe that each arc weight is updated in every iteration according to Equations (5) and (6). It is straight-forward to check that in that case $W_{v\to u}^\star$ will decrease in an iteration, contradicting its equilibrium property.Furthermore, in order for all the incoming arcs to u have the same weight,

[4] The sum of the arcs is 1, so if they are not all equal there exists an arc less than the average $\frac{1}{|N_u^{-*}|}$.

the update term $\eta x_v^\star x_u^\star$ must be equal for all $v \in N_u^{-*}$, and it must not zero-out the weight. These two facts imply that $x_v^\star = x$ for some value x for all $v \in N_u^{-*}$, and $\frac{1}{|N_u^{-*}|} - \eta x_v^\star x_u > 0$ which implies the last condition. ■

A.2 Example of section 4.3

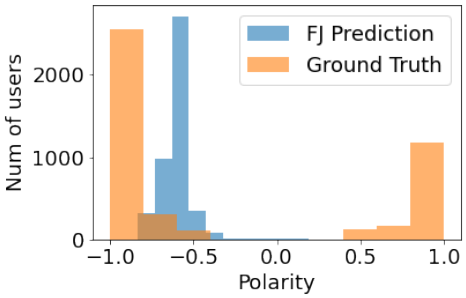


Fig. 5: Histogram of user retweet polarity ground truth and the prediction by FJ model on debate dataset.