Opinion Dynamics with Backfire Effect and Biased Assimilation

Xi Chen[1*], **Panayiotis Tsaparas**[2], **Jefrey Lijffijt**[1], **Tijl De Bie**[1]
[1]Dept. of Electronics and Information Systems, IDLab, Ghent University
[2]Department of Computer Science and Engineering, University of Ioannina
[1] {firstname.lastname}@ugent.be, [2] tsap@cs.uoi.gr

Abstract

The democratization of AI tools for content generation, combined with unrestricted access to mass media for all (e.g. through microblogging and social media), makes it increasingly hard for people to distinguish fact from fiction. This raises the question of how individual opinions evolve in such a networked environment without grounding in a known reality. The dominant approach to studying this problem uses simple models from the social sciences on how individuals change their opinions when exposed to their social neighborhood, and applies them on large social networks.

We propose a novel model that incorporates two known social phenomena: (i) *Biased Assimilation*: the tendency of individuals to adopt other opinions if they are similar to their own; (ii) *Backfire Effect*: the fact that an opposite opinion may further entrench someone in their stance, making their opinion more extreme instead of moderating it. To the best of our knowledge this is the first model that captures the Backfire Effect. A thorough theoretical and empirical analysis of the proposed model reveals intuitive conditions for polarization and consensus to exist, as well as the properties of the resulting opinions.

1 Introduction

Recent years have seen an increasing amount of attention from the computational social sciences in the study of opinion formation and polarization over social networks, with applications ranging from politics to brand perception [Conover *et al.*, 2011; Gionis *et al.*, 2013; Akoglu, 2014]. Much of this research leverages pre-existing opinion formation models that have been studied for decades [Jackson, 2008; Castellano *et al.*, 2009]. These models formalize the fact that people form their opinions through interactions with others. One of the best-known models is DeGroot's model [DeGroot, 1974], which considers an individual's opinion as dynamic, assuming that it is updated as the weighted average of the individual's current opinion and those of her social neighbors. The weights represent the strength of the social connections.

DeGroot's model is elegant and intuitive and it guarantees that the opinions converge towards a consensus [DeGroot, 1974; Jackson, 2008]. Yet, the opinions cannot polarize, contradicting empirical observations [Baron *et al.*, 1996; Gilbert *et al.*, 2009]. Variants of DeGroot's model have been proposed that incorporate *biased assimilation* [Krause, 2000; Dandekar *et al.*, 2013], which is also known as *confirmation bias* or *myside bias* and refers to the phenomenon where information that corroborates someone's beliefs affects those beliefs more strongly than information that contradicts it [Lord *et al.*, 1979]. Incorporating biased assimilation has been shown to potentially lead to polarization [Dandekar *et al.*, 2013] or opinion clustering [Krause, 2000].

An extreme manifestation of confirmation bias is a behavior known in social psychology as the *Backfire Effect* [Nyhan and Reifler, 2010; Allahverdyan and Galstyan, 2014]. It refers to the fact that, when an individual is faced with information that contradicts their opinion, they will not only tend to discredit it, but they will also become more entrenched and thus extreme in their opinion. The backfire effect may help explain the emergence of polarization. Yet, it has so far been overlooked by existing opinion formation models.

Motivated by these observations, we propose the BEBA model, a novel opinion formation model that simultaneously models the Backfire Effect and Biased Assimilation. BEBA depends on a single—intuitive, node-dependent—parameter β_i, which we call the *entrenchment* of node i. It captures both the tendency of node i to become more entrenched by opposing opinions and the bias towards assimilating opinions favorable to its own. Our main contributions are:

- We propose the BEBA model of opinion formation, which accounts for both the Backfire Effect and Biased Assimilation (Section 3). To the best of our knowledge BEBA is the first model that incorporates the Backfire Effect.
- We theoretically analyze the BEBA model in Section 4, studying conditions for reaching consensus or polarization.
- In Section 5 we empirically evaluate, on real and synthetic data, the effect of both network topology and initial opinions on polarization / convergence.

[*]Contact Author

2 Related Work

Opinion formation has been studied in diverse research fields, from psychology and social sciences to economics and physics [Jackson, 2008; Castellano *et al.*, 2009]. The former mostly use empirical methods to understand the factors that affect opinion formation, while the latter mostly aim to understand emergent behavior implied by these theories.

Two observations from psychology and social sciences relating to our work are the biased assimilation and backfire effect [Corner *et al.*, 2012; Lord and Taylor, 2009], which state that individuals are more inclined to accept opinions closer to their own [Lord *et al.*, 1979], and that, when exposed to the opposite opinion, individuals entrench themselves in their own opinion [Nyhan and Reifler, 2010; Chong and Druckman, 2007; Herr, 1986], respectively.

We study the common setting where opinions are formalized as real values, formed through social interactions (see [Jackson, 2008] and [Castellano *et al.*, 2009] for surveys). The most popular models include the Voter model [Clifford and Sudbury, 1973; Holley and Liggett, 1975], DeGroot's model [DeGroot, 1974], and the Friedkin-Johnsen model [Friedkin and Johnsen, 1990]. Yet, none of these account for the biased assimilation or backfire effect.

There is work on modeling the fact that users are more influenced by opinions closer to their own. The bounded confidence models [Deffuant *et al.*, 2000; Deffuant *et al.*, 2002; Hegselmann and Krause, 2002] assume that a user is influenced only by opinions that are within ϵ of its own. The work of Kempe et al., [Kempe *et al.*, 2016] assumes that there are different types of opinions and users are influenced by opinions of similar types. Das et al., [Das *et al.*, 2014] consider a biased version of the voter model that biases individuals to adopt similar opinions. The work most closely related to ours is that of Dandekar et al., [Dandekar *et al.*, 2013] who propose a variant of DeGroot's model to capture the biased assimilation effect. In their model, the importance that a node attaches to the opinion of a neighbor depends on their agreement. However, it does not model the backfire effect.

3 Model definition

In this section, we first describe existing models on which our work builds and then introduce our nonlinear opinion formation BEBA model, which is generalized from DeGroot's model, and accounts for both backfire effect and biased assimilation. Finally, we provide a comparison between our BEBA and the related biased opinion formation model on a simple example, to highlight their qualitative differences.

3.1 Preliminaries and background

Notation. Let $G = (V, E)$ denote a connected undirected network, with $V = \{1, ..., n\}$ the set of nodes, and $E \in V \times V$ the set of $m = |E|$ edges, where $(i, j) \in E$ iff $(j, i) \in E$. When the network is weighted, $w_{ij} = w_{ji}$ represents the weight of edge (i, j). We use $N(i)$ to denote the set of neighbors of node i: $N(i) \triangleq \{j \in V | (i, j) \in E\}$.

In the considered models, opinions are real numbers within a fixed interval $[0, 1]$ or $[-1, 1]$, depending on the model. To discriminate between the two, we use x to denote the opinions within $[0, 1]$, and y to denote the opinions that belong to $[-1, 1]$. All models we consider in this work can be defined as dynamical systems, where opinions are updated iteratively. We use $x_i(t)$ (resp. $y_i(t)$) to denote the opinion of node i at iteration (time) $t = 0, 1, 2, \ldots$. We further use $\mathbf{x}(t)$ and $\mathbf{y}(t)$ to denote the opinion vectors for the network at time t. With x_i (resp. y_i) we denote the opinion of node i after convergences for $t \to \infty$ (if that limit exists), and $\mathbf{x}$ (resp. $\mathbf{y}$) to denote the corresponding vectors.

DeGroot's Model. This model [DeGroot, 1974] is an averaging opinion formation model, where the individual's opinion is determined by the average of her own opinion and that of her neighbors. More specifically, it is updated as follows:

$$x_i(t+1) = \frac{w_{ii} x_i(t) + \sum_{j \in N(i)} w_{ij} x_j(t)}{w_{ii} + \sum_{j \in N(i)} w_{ij}} \quad (1)$$

where w_{ii} represents the extent to which the node values its own opinion, and w_{ij} is the strength of the connection/friendship between node i and j. Iterative opinion updates will converge to a stationary state, where every node has the same opinion $x_i = x^*$ [Jackson, 2008]. Therefore, the model always reaches consensus, and never polarizes.

Biased Opinion Formation. The BOF model [Dandekar *et al.*, 2013] generalizes DeGroot's to incorporate *biased assimilation*. Given a weighted undirected graph $G = (V, E, w)$, every node $i \in V$ is assigned a bias parameter $b_i \geq 0$. Higher values of b_i means that node i is more biased. The opinion value $x_i(t) \in [0, 1]$ is interpreted as the degree of support for opinion position 1 (i.e., the highest possible opinion value), while $1 - x_i(t)$ is the support for 0. It is defined as

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

where $s_i(t) \triangleq \sum_{j \in N(i)} w_{ij} x_j(t)$ is the weighted sum of i's neighbouring opinions, and $d_i \triangleq \sum_{j \in N(i)} w_{ij}$ is the weighted degree of node i. During the updating process, node i weighs confirming and disconfirming evidence in a biased way: weighing the neighboring support for opinion 1 by $(x_i(t))^{b_i}$, and that for opinion 0 by $(1 - x_i(t))^{b_i}$.

3.2 The BEBA model

We now define the BEBA model, which is a generalization of DeGroot's model that incorporates both biased assimilation and backfire effect. To capture these phenomena, we adapt DeGroot's model by dynamically setting the weights on the edges. Let $\mathbf{y}(t)$ denote the vector of opinions at time t, with $y_i(t) \in [-1, 1]$. Then, rather than using fixed weights as in DeGroot's model, we propose to let the weights be determined by the opinions as well. Specifically, for an edge $(i, j) \in E$ we define the edge weight $w_{ij}(t)$ at time t as

$$w_{ij}(t) = \beta_i y_i(t) y_j(t) + 1.$$

The product $y_i(t) y_j(t)$ captures the degree of (dis)agreement between the opinions of node pair (i, j). The parameter $\beta_i > 0$ models the influence for i that the (dis)agreement with node

j will have on the weight $w_{ij}(t)$: the larger, the stronger the biased assimilation and backfire effects. We will refer to β_i as the *entrenchment parameter* of node i.

Given the weight $w_{ij}(t)$, the opinions in the BEBA model are updated as in DeGroot's model:

$$y_i(t+1) = \frac{w_{ii} y_i(t) + \sum_{j \in N(i)} w_{ij}(t) y_j(t)}{w_{ii} + \sum_{j \in N(i)} w_{ij}(t)} \tag{2}$$

Note that when $\beta_i = 0$, BEBA's update rule is identical to that of DeGroot's (Eq. (1)) for unweighted networks. When $\beta_i \neq 0$, we discriminate two cases depending on $w_{ij}(t)$:

1. $w_{ij}(t) < 0$: This case models the backfire effect where $\beta_i y_i(t) y_j(t) < -1$. Since $\beta_i > 0$, $y_i(t) y_j(t) < 0$, that is, nodes i and j hold opposing views. Multiplying $y_j(t)$ with this negative weight $w_{ij}(t)$ in the summation in the numerator leads to a contribution of the same sign as $y_i(t)$, while adding the negative weight to the denominator reduces it, inflating the resulting quotient. The combination of these two effects models the backfire effect.
2. $w_{ij}(t) > 0$: This case models biased assimilation, including two subcases:
   (a) $0 < \beta_i y_i(t) y_j(t)$: Thus, node i and j have both positive or both negative opinions, resulting in an increased weight $w_{ij}(t)$. In this case node i assimilates the opinion of neighbor j more strongly if the extent of their agreement is stronger.
   (b) $-1 < \beta_i y_i(t) y_j(t) < 0$: Here nodes i and j hold opposing but not too different opinions. In this case, node i critically evaluates the conflicting opinion of node j, but still assimilates it to a reduced extent.

Note that the denominator in Eq. (2) can become 0 resulting in a diverging opinion, or negative causing an unnatural opinion reversal. We consider this situation to be beyond the model's validity region, and thus define the BEBA model as:

$$y_i(t+1) = \begin{cases} \operatorname{sgn}(y_i(t)) & \text{if } w_{ii} + \sum_{j \in N(i)} w_{ij}(t) \leq 0, \\ \frac{w_{ii} y_i(t) + \sum_{j \in N(i)} w_{ij}(t) y_j(t)}{w_{ii} + \sum_{j \in N(i)} w_{ij}(t)} & \text{otherwise.} \end{cases}$$

Moreover, for a small denominator the resulting opinions may fall outside the range $[-1, 1]$. To address this, we additionally clip negative values at -1 and positive values at 1.

3.3 Comparison of the BEBA and BOF models

There is a similarity between the BOF and our BEBA model, in that both alter the weights of the DeGroot's. Consider a simple star graph of five nodes where node 1 is in the center, and focus on one iteration of updating on node 1. In this case, we can observe how the two models update the opinion of a single node, given the opinions of her neighborhood.

First, we deal with the fact that BOF model assumes only positive opinion values, while our model assumes opinions being both positive and negative. Note that the value range of opinions is important in both models, since the BOF model weights the opinion values, while our model exploits the disagreement in the sign. To compare the models, we assume positive opinion values $x_i(t) \in [0, 1]$ on all nodes in the graph, and use them to implement an update of the BOF model. For our model, we transform opinions to the range $[-1, 1]$ by setting $y_i(t) = 2x_i(t) - 1$. Then we compute the value $y_1(t+1)$ as defined in BEBA, and rescale back.

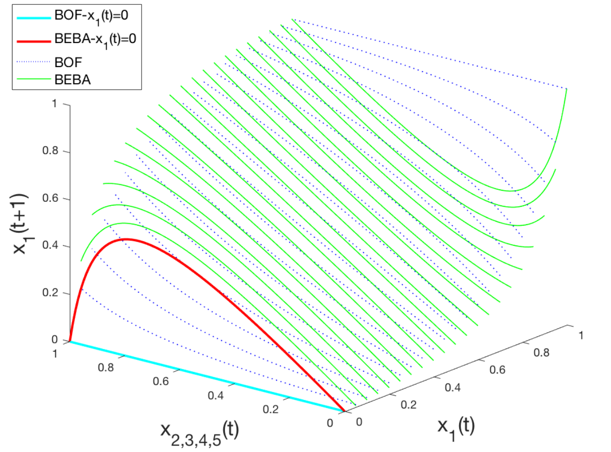

Figure 1: Opinion Formation on the Star Graph

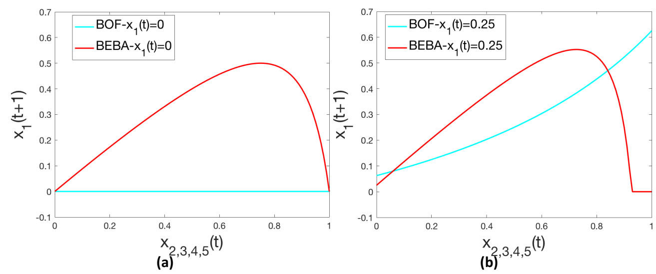

Figure 2: $x_1(t+1)$ as a function of $x_i(t)$, (a) $\beta_1 = 1$, $b_1 = 1$, $x_1(t) = 0$; (b) $\beta_1 = 2.5$, $b_1 = 1$, $x_1(t) = 0.25$.

In our experiment we assume $x_i(t)$ identical for all $i = 2, \ldots, 5$, and $x_i(t) \in [0, 1]$ for all nodes. We set $w_{11} = 1$ for both models, $b_1 = 1$ for BOF, and consider the values of 1 and 2.5 for β_1 in BEBA model. The opinion value $x_1(t+1)$ for both models, as a function of $x_{2,3,4,5}(t)$ and $x_1(t)$ is shown in Figure 1. The difference between the two models becomes clear when $x_1(t)$ takes extreme values (i.e., 0 or 1).

Figure 2(a) shows the curves for the two models when $x_1(t) = 0$. In BOF, the opinion $x_1(t+1)$ remains unchanged at value 0. This is true regardless of the value of b_1. Thus, extreme nodes never change their opinions, even a little, even when they are not biased at all. However, according to the biased assimilation, unbiased individuals should be influenced by similar opinions, while even extreme nodes assimilate opinions that are close to their own. In contrast, our model better captures the biased assimilation in this case. In Figure 2(a), for $\beta_1 = 1$, which corresponds to a mildly biased node, the opinion of node 1 can be moderated by that of her neighbors to different extents, while $x_1(t+1)$ never exceeds 0.5. Therefore, extreme nodes are not stuck in the extremes.

To better understand the backfire effect, we increase β_1 to 2.5, and set $x_1(t) = 0.25$ as shown in Figure 2(b). We observe that when the disagreement between node 1 and her neighbors becomes large (i.e., > 0.9), $x_1(t+1)$ drops under 0.25, until it becomes completely extreme with value 0.

From the plots in Figure 2 we also observe that for the different combinations of β_1 and $x_1(t)$, there exists a value of the neighboring opinions that causes the largest change in $x_1(t+1)$. For example, when $\beta_1 = 1$ and $x_1(t) = 0$, neighboring opinion of around 0.75 is the most influential as shown in Figure 2(a); for $\beta_1 = 2.5$ and $x_1(t) = 0.25$, opinion around 0.7 is the most influential according to Figure 2(b).

4 Theoretical Analysis

This section contains theoretical analysis of the BEBA model for two settings[1]. First we investigate the dynamics of opinions for a single agent in a fixed environment, and secondly we study the dynamics of polarization for all nodes in a connected social network.

4.1 A single agent in a fixed environment

Here we theoretically analyze the limit behavior of a single agent's opinion in an environment with a fixed opinion. An analysis of this type has been done for the BOF model [Dandekar *et al.*, 2013]. The setup is admittedly somewhat artificial but helps to gain a better understanding of the model. It has been deemed realistic in cases where the fixed environment consists of the news media, billboards, etc. [Dandekar *et al.*, 2013]. It also models the situation where the single agent is connected to a network that is large enough such that adding it will not meaningfully affect the network.

For the agent i, we denote $y(t) \in [-1, 1]$ its opinion at time t, $\beta > 0$ its entrenchment parameter, and y its converged opinion (i.e., $\lim_{t\to\infty} y(t)$). We assume the agent weighs its own opinion with $w_{ii} = w$. For simplicity, we only consider the situation where the environment contains one node, but it should be noted that the analysis below can be easily generalized to several nodes. Let $p \in [-1, 1]$ be the fixed environmental opinion. Then, according to BEBA, the agent updates its opinion as follows:

$$y(t+1) = \begin{cases} \operatorname{sgn}(y(t)) & \text{if } w + \beta p y(t) + 1 \le 0, \\ \frac{wy(t)+\beta p^2 y(t)+p}{w+\beta p y(t)+1} & \text{otherwise.} \end{cases}$$

Before stating a theorem that quantitatively characterizes the limit y, we consider the behavior. [Case 1:] For sufficiently small entrenchment β (i.e., not biased), the fixed environment's opinion p will be sufficiently attracting such that $y = p$ regardless of $y(t)$. The same is true when $p = 0$: the neutral opinion is never polarizing and thus always attracting. [Case 2:] On the other hand, for sufficiently large entrenchment β (i.e., biased), the limit y will depend on the similarity of initial opinion $y(t)$ with the environment's opinion p: [Case 2a:] if $y(t)$ is similar to p, p should have an attracting effect on $y(t)$ such that its limit $y = p$; [Case 2b:] if $y(t)$ is very different from p, however, the backfire effect will cause the agent's opinion to diverge from p, such that $y = \operatorname{sgn}(y(t))$. [Case 2c:] Between Case 2a and Case 2b, there will be a 'sweet spot' where $y(t)$ is neither sufficiently similar to p for $y(t)$ to converge to p, nor sufficiently different for it to diverge to $\operatorname{sgn}(y(t))$. This is an unstable equilibrium where $y(t)$ remains constant through time, i.e. $y = y(t)$.

[1] Supplemental materials including theoretical proofs, datasets information, and more experimental results available in the Appendix.

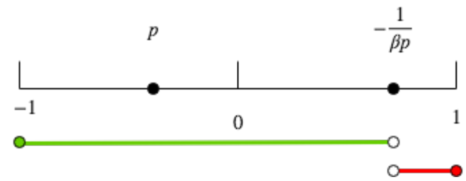

Figure 3: Graphical illustration of Case 2 from Theorem 1 (i.e. $p < 0$ and $\beta \ge -1/p$). [Case 2a:] For values of $y(t)$ in the green range, $y(t)$ will converge to $y = p$. [Case 2b:] For values of $y(t)$ in the red range, $y(t)$ will diverge to $y = 1$. [Case 2c:] For $y(t) = -\frac{1}{\beta p}$, $y(t)$ will not change such that $y = -\frac{1}{\beta p}$.

This intuition is formalized in the following theorem. For conciseness and transparency, we state it for the situation where $p \le 0$. It is trivial to adapt the theorem for $p \ge 0$.

Theorem 1. *Depending on the value of β relative to p:*

Case 1: *When $p = 0$ or $\beta < -1/p$, the agent's opinion always converges to p, i.e., $y = p$.*

Case 2: *When $p < 0$ and $\beta \ge -1/p$, there are three possibilities depending on how similar $y(t)$ is to p. (This situation is illustrated in Figure 3.)*

- **a:** *If $y(t) < -\frac{1}{\beta p}$, $y(t)$ will be sufficiently attracted to p such that $y = p$.*
- **b:** *If $y(t) > -\frac{1}{\beta p}$, $y(t)$ will diverge away from p such that $y = \operatorname{sgn}(y(t)) = 1$.*
- **c:** *If $y(t) = -\frac{1}{\beta p}$, $y(t)$ will remain constant through time, such that $y = -\frac{1}{\beta p}$.*

Theorem 1 already suggests that opinions under the BEBA model evolve to one of three possible states: consensus (Case 1 and Case 2a), polarization (Case 2b), and an unstable state of persistent disagreement (Case 2c).

4.2 Polarization and consensus for general networks and initial opinions

Here we extend from the single agent to a group of individuals that can update their opinions at any time step t. The dynamics of polarization are investigated theoretically with respect to different values of the entrenchment parameter. It was argued by the authors of the BOF model that homophily alone, without biased assimilation was not sufficient for polarization [Dandekar *et al.*, 2013]. In our BEBA model, the backfire effect and biased assimilation, without homophily, are sufficient to lead to polarization or consensus, depending on the parameters and the initial opinions. The theorem below makes this clear, by providing easy-to-realize sufficient conditions for polarization or consensus to occur.

Theorem 2. *Let $G = (V, E)$ be any connected unweighted undirected network. For all $i \in V$, $y_i(t) \in (-1, 0) \cup (0, 1)$ is the opinion of node i at time t, let $w_{ii} = 1$ and $\beta_i = \beta > 0$ for all $i \in V$. Denote $\mathbf{y}(t)$ the opinion vector of G at time t, $|\mathbf{y}(t)|$ is the vector with the absolute values of all opinions, and $\min(\mathbf{y}(t))$ is the minimum element in $\mathbf{y}(t)$. Then,*

1. *Polarization: If $\beta > \frac{1}{[\min(|\mathbf{y}(0)|)]^2}$, $\forall i \in V$, $|y_i| = 1$.*

2. *Consensus: If* $\beta < \frac{1}{[\max(|\mathbf{y}(0)|)]^2}$, *there exists a unique* $y^* \in [-\max(|\mathbf{y}(0)|), \max(|\mathbf{y}(0)|)]$ *such that* $y_i = y^*$, $\forall i \in V$.

A special case of particular theoretical interest is when $\min(|\mathbf{y}(0)|) = \max(|\mathbf{y}(0)|)$. Then there are only two different initial opinions in the network, with the same absolute value but opposite signs (i.e. they could represent 'for' and 'against' an issue of interest). In this case, the sufficient conditions also become necessary conditions, and a borderline situation emerges to which we refer as *persistent disagreement*. It can be proved concisely by relying on Theorem 2, and thus we state it as a Corollary:

Corollary 1. *Let* $G = (V_1, V_2, E)$ *be any connected unweighted undirected network. For all* $i \in V = V_1 \cup V_2$, *let* $w_{ii} = 1$ *and* $\beta_i = \beta > 0$. *Assume for all* $i \in V_1$, $y_i(0) = y_0$, *where* $0 < y_0 < 1$; *while for all* $i \in V_2$, $y_i(0) = -y_0$. *Then,*

1. *Polarization: If* $\beta > \frac{1}{y_0^2}$, $\forall i \in V_1 \cup V_2$, $|y_i| = 1$.
2. *Persistent disagreement: If* $\beta = \frac{1}{y_0^2}$ *(i.e., when* $w_{ij} = 0$ *if* $i \in V_1$ *and* $j \in V_2$*),* $\forall i \in V_1$, $y_i(t') = y_0$ *for all* $t' \geq 0$, *and* $\forall i \in V_2$, $y_i(t') = -y_0$ *for all* $t' \geq 0$.
3. *Consensus: If* $\beta < \frac{1}{y_0^2}$, *then there exists a unique* $y^* \in (-y_0, y_0)$ *such that* $\forall i \in V$, $y_i = y^*$.

Intriguingly, these conditions in the Theorem and Corollary are independent of the network structure and depend only on the entrenchment parameter β and the opinion vector at time 0. Yet, it should be noted that the value of the consensus and the eventual polarized state do depend on the network structure. Moreover, the network structure, and the distribution of the opinions over it, do determine whether polarization or consensus will arise when neither of the sufficient conditions of Theorem 2 are satisfied. These claims are confirmed in experiments in the next section.

5 Experimental Analysis

In Section 4 we provided sufficient conditions for our model to reach consensus or polarization. In this section we perform an experimental analysis of how these two phenomena manifest themselves on real and synthetic networks. Our goal is to answer the following questions:

- In the case that the network reaches consensus, what is the value of the consensus opinion, and how does the network structure, β, and the initial opinion vector affect this value?
- In the case that the opinions polarize, what is the state of the polarization and how is it affected by the initial opinions, β, and the structure of the networks?

We use both real-world and synthetic networks in our experiments. The real datasets include Zachary's Karate Club network [Zachary, 1977] and six Twitter networks with given opinions for different events ranging from political elections to sports activities [Zarezade *et al.*, 2017; De *et al.*, 2016]. See the Appendix for details. The synthetic networks are:

- Erdős-Rényi (ER) networks $G(n, \rho)$ have binomial degree distributions, where ρ is the edge connection probability between nodes [Bollobás, 2001].
- Watts-Strogatz (WS) networks $G(n, K, 1)$ have the small world property of with K being the average degree, and we fix the rewiring probability to be 1 (i.e., random graph), thus only refer to K [Watts and Strogatz, 1998].
- Barabási-Albert (BA) networks $G(n, M_0, M)$ are scale-free, where M_0 is the number of initial nodes and M the number of nodes that a new node is connected to [Albert and Barabási, 2002].

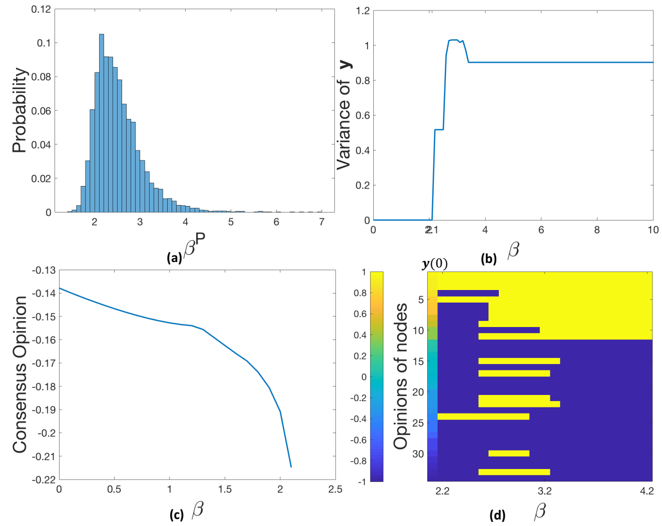

Figure 4: For the Karate network: (a) the distribution of β^P (i.e., the smallest β that results in polarization) for 10000 random opinion vectors (uniform on $[-1, 1]$); for one opinion vector, (b) the variance of all converged $\mathbf{y}$ as β increases from 0 to 10; (c) consensus opinion values for $\beta \in [0, 2.1]$; (d) final opinions for each of the nodes.

5.1 The influence of the entrenchment β

From Theorem 2, we know the stationary opinion vector $\mathbf{y}$ of our model polarizes when $\beta > \frac{1}{[\min(|\mathbf{y}(0)|)]^2}$, and reaches consensus when $\beta < \frac{1}{[\max(|\mathbf{y}(0)|)]^2}$. However, these limits are far away from each other and polarization may occur at much lower values of β in practice, similarly consensus for higher β. We now take the Karate network as an example and examine the relation between β and polarization experimentally using random initial opinion vectors.

Let β^P denote the threshold between consensus and polarization for any pair of network and opinion vector. Figure 4(a) shows the distribution of the empirical β^P values for 10000 different random opinion vectors, where $y_i(0)$ is uniform within $[-1, 1]$. We observe that the threshold for polarization is much smaller than the theoretical value, which should be lager than 10^4. However, the empirical value of β^P is below 5 for most of the $\mathbf{y}(0)$, and never exceeds 7.

In Figure 4(b), the variance of the stationary opinion vector is plotted as a function of β, for one of the opinion vectors. When there is consensus the variance is zero, while when the variance is greater than zero, polarization is obtained (i.e., different variances correspond to different polarized states). We observe that as β increases, the opinion vector converges from consensus to polarized states. Empirically, no persistent disagreement is achieved. For this $\mathbf{y}(0)$, polarization is shown if $\beta > 2.1$ such that $\beta^P = 2.1$.

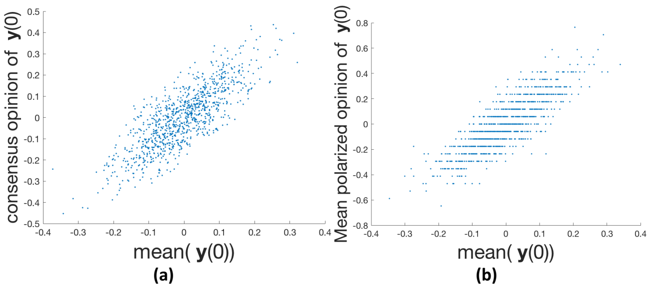

Figure 5: For 1000 random $\mathbf{y}(0)$ on Karate network: (a) consensus opinion when $\beta = 1$; (b) mean polarized opinion when $\beta = 10$.

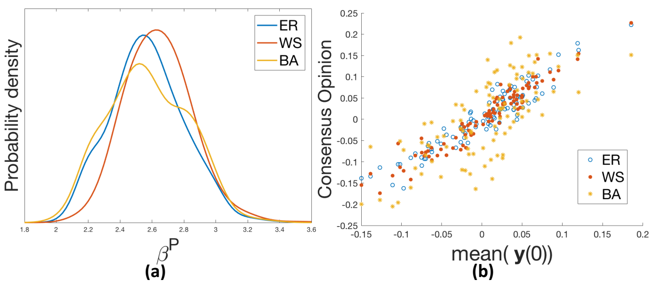

Figure 6: Based on one ER ($n = 100, \rho = 0.0606$), one WS ($n = 100, K = 3$), and one BA network ($n = 100, M_0 = 4, M = 3$): (a) distribution of β^P for 1000 random opinion vectors; (b) for 100 opinion vectors, mean $\mathbf{y}(0)$ vs. the consensus value ($\beta = 1$).

When reaching consensus, Figure 4(c) shows that the consensus value becomes less neutral as β increases. This is true for 78.74% of the 10000 vectors on Karate network. Meanwhile, different βs do not necessarily result in the same polarized state (see Figure 4(d)). The heatmap shows different polarized states for different values of β for this $\mathbf{y}(0)$.

5.2 The influence of the opinion vector $\mathbf{y}(0)$

In this experiment, we investigate the influence $\mathbf{y}(0)$ on the consensus opinion value and the mean polarized opinion. Figure 5 shows that the consensus value and the mean polarized opinion are strongly correlated to the mean of $\mathbf{y}(0)$. Meanwhile, Figure 5(b) shows that in the case of polarization, opinion vectors with similar initial means may result in quite different polarized states because the placements of the opinions on nodes differ. Also, $\mathbf{y}(0)$ with different means may result in similar polarization (i.e., mean polarized opinion).

Then we analyze two real datasets Tw:Club (i.e., Barcelona getting the first place in La-liga 2016) and Tw: Sport (Champions League final in 2015 between Juventus and Real Madrid), which have the same network but different initial opinion vectors. It is found that the β^P is 11.7 for Tw:Club and 3.3 for Tw:Sport, which indicates the Champions League final gets polarized more easily than the other event.

5.3 The influence of the network topology G

In this experiment, we study how the topology affects the β^P for the same (set of) $\mathbf{y}(0)$, as well as the stationary opinion vectors of our model. To this end, we generated networks with the three random network models, with the same number of nodes, intialized with the same opinion vectors.

We observe that for networks with the same number of nodes and similar numbers of edges, different network properties result in different dynamics of polarization. Figure 6(a) shows that for the same set of $\mathbf{y}(0)$, the distributions of the β^P value for the three models. It shows that the β^P has a larger mean in the WS model, indicating networks with this structure may be more robust against polarization. We also observe the standard deviation of the β^P values for the BA distribution is larger, which appears to be due to 'hub' nodes, whose opinions strongly affect the value of β^P.

Figure 6(b) plots the consensus values reached by a set of 100 random opinion vectors on the three networks. The shapes of scatter plots become increasingly compact from the BA model, the ER model, to the WS model, corroborating the larger variance in the opinion dynamics on BA networks.

The parameters in each model also affect the dynamics, see the supplement in the Appendix. For example, when the edge probability ρ in the ER model increases from a small number, which guarantees a connected network, to 1, β^P varies less for ER models with similar ρ. The experimental results are similar for the consensus value, and the polarized opinion. Not only the number of edges has an influence on the dynamics of polarization, but also the placement of the edges.

Table 1: β^P for real-world twitter datasets

Network	β^P	Network	β^P	Network	β^P
Tw:GoT	2.9	Tw:Club	3.3	Tw:US	4.9
Tw:UK	7.5	Tw:Delhi	7.7	Tw:Sport	11.7

5.4 Real-world dataset analysis

Based on the six real-world twitter datasets [Zarezade *et al.*, 2017; De *et al.*, 2016], we investigate how easily each event gets polarized opinions, namely the value of β^P. It is shown in Table 1 that political events are apparently less likely to polarize, except the US one. While the sport or TV events are more likely get polarized, except when people had to bet instead of supporting (i.e., Tw:Club).

6 Conclusion and Future Work

Modeling how opinions evolve when individuals interact in social networks is an important computational social science challenge that has received renewed attention recently. The availability of realistic models of this type may have substantial real-life impact on a variety of applications, from political campaigns design, to conflict prevention and mitigation.

A large number of models have been proposed in the literature. To the best of our knowledge, however, none of them model the so-called Backfire Effect: the fact that individuals, when exposed to a strongly opposing view, will not be moderated, but rather become more entrenched in their opinion.

Here we proposed the BEBA model, which models both Biased Assimilation and Backfire Effect. It is governed by one parameter (which can vary over the individuals), called the entrenchment parameter, determining the strength of both.

The BEBA model naturally generates different behaviors: from convergence to a consensus, to polarization.

Theoretical and empirical analyses demonstrate that the resulting model is not only realistic, its behavior also provides an interesting view on the interplay between network structure, the entrenchment parameter, and the opinions.

These properties make the BEBA model a useful tool for simulating the effect of interventions, such as editing the network (e.g. by facilitating communication between particular pairs of individuals), altering the initial opinions (e.g. through targeted information campaigns), or affecting the entrenchment of particular individuals (e.g. through education).

Acknowledgements. This work was supported by the ERC under the EU's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agr. no. 615517, FWO (project no. G091017N, G0F9816N), the EU's Horizon 2020 research and innovation programme and the FWO under the Marie Sklodowska-Curie Grant Agr. no. 665501.

References

[Akoglu, 2014] L. Akoglu. Quantifying political polarity based on bipartite opinion networks. In *Proc. of ICWSM*, 2014.

[Albert and Barabási, 2002] R. Albert and A. Barabási. Statistical mechanics of complex networks. *Rev Mod Phys*, 74(1):47, 2002.

[Allahverdyan and Galstyan, 2014] A. E. Allahverdyan and A. Galstyan. Opinion dynamics with confirmation bias. *PloS One*, 9(7):e99557, 2014.

[Baron *et al.*, 1996] R. S. Baron, S. I. Hoppe, C. F. Kao, B. Brunsman, B. Linneweh, and D. Rogers. Social corroboration and opinion extremity. *J Exp Soc Psychol*, 32(6):537–560, 1996.

[Bollobás, 2001] B. Bollobás. *Random graphs*. Cambridge University Press, 2001.

[Castellano *et al.*, 2009] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646, May 2009.

[Chong and Druckman, 2007] D. Chong and J. N. Druckman. Framing public opinion in competitive democracies. *Am Polit Sci Rev*, 101(4):637–655, 2007.

[Clifford and Sudbury, 1973] P. Clifford and A. Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973.

[Conover *et al.*, 2011] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proc. of ICWSM*, pages 89–96, 2011.

[Corner *et al.*, 2012] A. Corner, L. Whitmarsh, and D. Xenias. Uncertainty, scepticism and attitudes towards climate change: biased assimilation and attitude polarisation. *Climatic change*, 114(3-4):463–478, 2012.

[Dandekar *et al.*, 2013] P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *PNAS*, 110(15):5791–5796, 2013.

[Das *et al.*, 2014] A. Das, S. Gollapudi, and K. Munagala. Modeling opinion dynamics in social networks. In *Proc. of WSDM*, pages 403–412, 2014.

[De *et al.*, 2016] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. G. Rodriguez. Learning and forecasting opinion dynamics in social networks. In *Proc. of NIPS*, pages 397–405, 2016.

[Deffuant *et al.*, 2000] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Adv. Compl. Syst.*, 3(1-4):87–98, 2000.

[Deffuant *et al.*, 2002] G. Deffuant, F. Amblard, G. Weisbuch, and T. Faure. How can extremism prevail? a study based on the relative agreement interaction model. *JASSS*, 5(4), 2002.

[DeGroot, 1974] M. H. DeGroot. Reaching a consensus. *JASA*, 69(345):118–121, 1974.

[Friedkin and Johnsen, 1990] N. E. Friedkin and E. C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

[Gilbert *et al.*, 2009] E. Gilbert, T. Bergstrom, and K. Karahalios. Blogs are echo chambers: Blogs are echo chambers. In *Proc. of HICSS*, pages 1–10, 2009.

[Gionis *et al.*, 2013] A. Gionis, E. Terzi, and P. Tsaparas. Opinion maximization in social networks. In *Proc. of SDM*, pages 387–395, 2013.

[Hegselmann and Krause, 2002] R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *JASSS*, 5(3), 2002.

[Herr, 1986] P. M. Herr. Consequences of priming: Judgment and behavior. *J Pers Soc Psychol*, 51(6):1106, 1986.

[Holley and Liggett, 1975] R. A. Holley and T. M. Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The Annals of Probability*, 3(4):643–663, 1975.

[Jackson, 2008] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[Kempe *et al.*, 2016] D. Kempe, J. Kleinberg, S. Oren, and A. Slivkins. Selection and influence in cultural dynamics. *Network Science*, 4(1):1–27, 2016.

[Krause, 2000] U. Krause. A discrete nonlinear and nonautonomous model of consensus formation. In *Proc. of Difference Equations*, pages 227–236, 2000.

[Lord and Taylor, 2009] C. G. Lord and C. A. Taylor. Biased assimilation: Effects of assumptions and expectations on the interpretation of new evidence. *Soc Personal Psychol Compass*, 3(5):827–841, 2009.

[Lord *et al.*, 1979] C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol*, 37(11):2098, 1979.

[Nyhan and Reifler, 2010] B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[Watts and Strogatz, 1998] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[Zachary, 1977] W. W. Zachary. An information flow model for conflict and fission in small groups. *J Anthropol Res*, 33(4):452–473, 1977.

[Zarezade *et al.*, 2017] A. Zarezade, A. De, H. Rabiee, and M. G. Rodriguez. Cheshire: An online algorithm for activity maximization in social networks. *arXiv:1703.02059*, 2017.

A Proof of Theorem 1

A.1 Only one node in the environment

Recall that there is one node with a fixed opinion $p \in [-1, 1]$ in the environment. The opinion of the agent is updated as mentioned in Section 4.

Lemma 3. *If $w + \beta p y(t) + 1 \leq 0$, the opinion of the agent stays at $\mathrm{sgn}(y(t))$ for all $t' > t$.*

Proof. As shown in the updating rule that when $w+\beta p y(t)+1 \leq 0$, $y(t+1) = \mathrm{sgn}(y(t))$. $w+\beta p y(t)+1 \leq 0$ is equivalent to $\beta p y(t) \leq -w - 1 < 0$. Knowing that $|y(t + 1)| = 1 \geq |y(t)|$,

$$\beta p y(t + 1) \leq -w - 1$$

Therefore, $y(t') = \mathrm{sgn}(y(t + 1)) = \mathrm{sgn}(y(t))$ for all $t' > t$. □

Lemma 4. *If $w+\beta p y(t)+1 > 0$, there exist two fixed points where $y(t+1) = y(t)$: p and $-\frac{1}{\beta p}$. p is attracting while $-\frac{1}{\beta p}$ is repelling.*

Proof. The converged opinion y of the agent should satisfy

$$f(y) = \frac{wy + \beta p^2 y + p}{w + \beta p y + 1}$$

$$f(y) - y = \frac{-\beta p y^2 + (\beta p^2 - 1)y + p}{w + \beta p y + 1} = \frac{u(y)}{v(y)} = 0 \quad (3)$$

where

$$u(y) = -\beta p y^2 + (\beta p^2 - 1)y + p$$
$$v(y) = \beta p y + w + 1$$

By solving $u(y) = 0$, which is equivalent to $f(y) - y = 0$ since $u(y) > 0$, the two fixed points of $f(y)$ are: p and $-\frac{1}{\beta p}$.

Next, we prove that p is attracting and $-\frac{1}{\beta p}$ is repelling.

$$f'(y) = \frac{w(w + \beta p^2)}{(w + \beta p y + 1)^2} \geq 0$$

$|f'(y)| = f'(y)$, then $f'(p) = \frac{w(w+\beta p^2)}{(w+\beta p^2+1)^2} < 1$, thus attracting; while $f'(y - \frac{1}{\beta p}) = \frac{w(w+\beta p^2)}{w^2} > 1$, thus repelling. □

Lemma 5. *If $w + \beta p y(t) + 1 > 0$ and $py(t) \geq 0$, $y = p$.*

Proof. If $p = 0$, $y(t + 1) = \frac{w}{w+1} y(t)$, as the iteration goes, $\lim_{t\to\infty} y(t) = 0$;

If $py(t) > 0$, e.g., they are both positive

- when $0 < y(t) < p$, $y(t + 1) - y(t) = \frac{u(y(t))}{v(y(t))} > 0$, thus $y(t + 1) > y(t)$, the agent's opinion increases until it reaches p;
- when $p < y(t) < 1$, $y(t + 1) - y(t) < 0$, the agent's opinion decreases to p.

□

Lemma 6. *If $w + \beta p y(t) + 1 > 0$ and $py(t) < 0$,*

1. *If $\left|\frac{1}{\beta p}\right| > 1$, $\lim_{t\to\infty} y(t) = y^e$.*
2. *If $\left|\frac{1}{\beta p}\right| \leq 1$,*
   (a) *If $|y(t)| < \left|\frac{1}{\beta p}\right|$, $y = p$.*
   (b) *If $y(t) = -\frac{1}{\beta p}$, $y(t') = -\frac{1}{\beta p}$ for all $t' \geq t$.*
   (c) *If $\left|\frac{1}{\beta p}\right| < |y(t)| \leq 1$, $y = \mathrm{sgn}(y(t))$.*

Proof. Assume $y(t) \in (0, 1]$ and $p \in (-1, 0)$,

- if $\left|\frac{1}{\beta p}\right| > 1$, all $y(t) \in (0, 1] < -\frac{1}{\beta p}$, $y(t)$ is attracted to p as the updating goes;
- if $\left|\frac{1}{\beta p}\right| = 1$, $y(t)$ is repelled by the extreme point and goes to the attracting one unless it starts with $-\frac{1}{\beta p}$ at time t;
- if $\left|\frac{1}{\beta p}\right| < 1$, when $0 < y(t) < -\frac{1}{\beta p}$, $y(t + 1) - y(t) = \frac{u(y(t))}{v(y(t))} < 0$, $y(t+1) < y(t)$, the agent's opinion decreases to p; when $y(t) = -\frac{1}{\beta p}$, $y(t)$ stays there; when $y(t) > -\frac{1}{\beta p}$, $y(t + 1) > y(t)$, the agent's opinion increases to the extreme value on its side.

□

A.2 A group of nodes in the environment

Assume there is a set of m neighbour with different fixed opinions, $\mathbf{p} = (p_1, p_2, ..., p_m)$, $m > 1$. We denote

- $q = \sum_j p_j^2$ the sum of the squares of the fixed opinions.
- $s = \sum_j p_j$ the sum of the fixed opinions.
- $m = \sum_j 1$ the number of nodes in the environment.

Lemma 7. *$mq - s^2 \geq 0$, which is $m \sum_j p_j^2 \geq (\sum_j p_j)^2$.*

Proof.

$$m \sum_j p_j^2 - (\sum_j x_j)^2 = \frac{1}{2} \sum_i \sum_j (p_i - p_j)^2 \geq 0$$

□

The agent's opinion is updated by

$$y(t + 1) = \begin{cases} \mathrm{sgn}(y(t)) & \text{if } w + \beta s y(t) + m \leq 0, \\ \frac{wy(t)+\beta q y(t)+s}{w+\beta s y(t)+m} & \text{otherwise.} \end{cases} \quad (4)$$

Lemma 8. *If $w+\beta sy(t)+m>0$, there exist two fixed points where $y(t+1)=y(t)$:*

$$y^a=\frac{\beta q-m+\sqrt{\Delta}}{2\beta s}\qquad y^r=\frac{\beta q-m-\sqrt{\Delta}}{2\beta s}$$

where $\Delta=(\beta q-m)^2+4\beta s^2$. y^a is attracting while y^r is repelling.

Proof. The function is $f(y)=\frac{wy+\beta qy+s}{w+\beta sy+m}$. The two fixed points satisfy $f(y)=y$. $|f'(y)|=f'(y)$ since

$$\begin{aligned}f'(y)=&\frac{(w+\beta q)(w+m)-\beta s^2}{(\beta sy+w+m)^2}\\=&\frac{w(w+m)+\beta qw+\beta(qm-s^2)}{(\beta sy+w+m)^2}>0\end{aligned}$$

For $y^a=\frac{\beta q-m+\sqrt{\Delta}}{2\beta s}$, $f'(y^a)<1$ because

$$\begin{aligned}f'(y^a)-1=&-\frac{1}{2}\frac{(m-\beta q)^2+4\beta s^2+(2w+m+\beta q)\sqrt{\Delta}}{(\beta sy^a+w+m)^2}\\&<0\end{aligned}$$

For $y^r=\frac{\beta q-m-\sqrt{\Delta}}{2\beta s}$, $f'(y^r)>1$ because

$$\begin{aligned}f'(y^r)-1=&-\frac{1}{2}\frac{(m-\beta q)^2+4\beta s^2-(2w+m+\beta q)\sqrt{\Delta}}{(\beta sy^r+w+m)^2}\\=&-\frac{1}{2}\frac{A}{B}\end{aligned}$$

$\frac{A}{B}<0$ since $B>0$ and it can be proved as below that $A<0$.

$$\begin{aligned}&\left[(m-\beta q)^2+4\beta s^2\right]^2-\left[(2w+m+\beta q)\sqrt{\Delta}\right]^2\\=&4\left[(m-\beta q)^2+4\beta s^2\right]\left[\beta(s^2-qm)-w(m+w+\beta q)\right]\\<&0\end{aligned}$$

Therefore, y^a is attracting and y^r is repelling. □

B Proof of Theorem 2

Recall that $y_i(t)\in(-1,0)\cup(0,1)$. Given any opinion vector $\mathbf{y}(0)$ of a given connected network $G=(V,E)$, the opinions can be divided into two groups V_1 and V_2 at any time t: a) $\forall i\in V_1$, $y_i(t)>0$; b)$\forall i\in V_2$, $y_i(t)<0$, and $V=V_1\cup V_2$. Denote $n_i^s(t)$ the number of node i's neighbors node that are in the **s**ame group with i at time t, and $n_i^d(t)$ the number of neighbors in the **d**ifferent group. Specifically, they are denoted as

$$\begin{aligned}n_i^s(t)=&|N(i)^s|, N(i)^s=\{j|j\in N(i),\text{ and }y_i(t)y_j(t)>0\}\\n_i^d(t)=&|N(i)^d|, N(i)^d=\{k|k\in N(i),\text{ and }y_i(t)y_k(t)<0\}\end{aligned}$$

Lemma 9. *For node $i\in V$ fix $\beta_i=\beta>0$, if $\beta>\frac{1}{[\min(|\mathbf{y}(0)|)]^2}$, $\lim_{t\to\infty}|y_i(t)|=1$.*

Proof. For node $i\in V$, the opinion is updated with BEBA. If $\gamma=1+\sum_{j\in N(i)}w_{ij}\le 0$, $y_i(t+1)$ reaches the extreme value in one iteration due to strong backfire effect.

While when $\gamma>0$, for any $t>0$, $y_i(t+1)$ is updated as

$$y_i(t)\frac{1+\sum_{j\in N(i)^s}w_{ij}\frac{y_j(t)}{y_i(t)}+\sum_{k\in N(i)^d}w_{ik}\frac{y_k(t)}{y_i(t)}}{1+\sum_{j\in N(i)^s}w_{ij}+\sum_{k\in N(i)^d}w_{ik}}=y_i(t)\frac{C}{D}\tag{5}$$

When $\beta>\frac{1}{[\min(|\mathbf{y}(t)|)]^2}$, for all $k\in N(i)^d$, $w_{ik}=\beta y_i(t)y_k(t)+1<0$. The sums in Equation (5) satisfy: $\sum_{j\in N(i)^s}w_{ij}\frac{y_j(t)}{y_i(t)}$, $\sum_{j\in N(i)^s}w_{ij}$, $\sum_{k\in N(i)^d}w_{ik}\frac{y_k(t)}{y_i(t)}>0$, and $\sum_{k\in N(i)^d}w_{ik}<0$.

Now we focus on the node that has the most moderate opinion, namely the node with absolute value of opinion $\min|\mathbf{y}(t)|$ at each time step, starting from time 0. Knowing $C,D>0$,

$$C-D=\sum_{j\in N(i)^s}w_{ij}(\frac{y_j(t)}{y_i(t)}-1)+\sum_{k\in N(i)^d}w_{ik}(\frac{y_k(t)}{y_i(t)}-1)\tag{6}$$

Since $y_i(t)$ has the smallest absolute opinion value, for any $j\in N(i)^s$, $\frac{y_j(t)}{y_i(t)}\ge 1$, thus $C>D$, $\frac{C}{D}>1$, and $|y_i(t+1)|>|y_i(t)|$.

After every iteration from time t to $t+1$, the opinion of the most moderate node becomes more extreme, until it reaches the absolute value of 1, thus for any $i\in V$, $\lim_{t\to\infty}|y_i(t)|=1$. □

Lemma 10. *For node $i\in V$, if $\beta<\frac{1}{[\max(|\mathbf{y}(0)|)]^2}$, there exists a unique $y^*\in[-\max(|\mathbf{y}(0)|),\max(|\mathbf{y}(0)|)]$ such that $\lim_{t\to\infty}y_i(t)=y^*$ for all $i\in V$.*

Proof. When $\beta<\frac{1}{[\max(|\mathbf{y}(0)|)]^2}$, $\gamma=1+\sum_{j\in N(i)}w_{ij}>0$ because for any $j\in N(i)$, $w_{ij}=\beta y_i(t)y_j(t)+1>0$.

For any $t>1$, $y_i(t+1)$ is updated as in Equation (5), however, the sums have different values: $\sum_{j\in N(i)^s}w_{ij}\frac{y_j(t)}{y_i(t)}$, $\sum_{j\in N(i)^s}w_{ij}$, $\sum_{k\in N(i)^d}w_{ik}>0$, and $\sum_{k\in N(i)^d}w_{ik}\frac{y_k(t)}{y_i(t)}<0$.

Then we focus on the most opinionated node, which means the node has the largest absolution value of its opinion $\max|\mathbf{y}(t)|$, starting from time 0. Knowing $D>0$,

- when $C>0$, $C-D$ is shown in Equation (6). With i being the most opinionated node, $\frac{y_j(t)}{y_i(t)}\le 1$ for all $j\in N(i)^s$; $\frac{y_k(t)}{y_i(t)}<0$ for all $k\in N(i)^d$. Therefore, $C<D$, $0<\frac{C}{D}<1$ and $|y_i(t+1)|<|y_i(t)|$.
- when $C=0$, $y_i(t+1)=0$.
- when $C<0$, $-C-D$ is shown in Equation (7). As $-1\le\frac{y_k(t)}{y_i(t)}\le 0$ for $k\in N(i)^d$, $-C-D<0$, $0<\left|\frac{C}{D}\right|<1$, thus $|y_i(t+1)|<|y_i(t)|$.

$$-2-\sum_{j\in N(i)^s}w_{ij}(\frac{y_j(t)}{y_i(t)}+1)-\sum_{k\in N(i)^d}w_{ik}(\frac{y_k(t)}{y_i(t)}+1)\tag{7}$$

At every time step, the most opinionated node get moderated until they reach consensus - there is no such node and the updating process stops because consensus is reached. □

Lemma 11. *For node $i \in V_1$, $y_i(0) = y_0$, where $0 < y_0 < 1$; $\forall i \in V_2$, $y_i(0) = -y_0$. If $\beta = \frac{1}{y_0^2}$, $y_i(t) = y_i(0)$ for all $t \geq 0$.*

Proof. When $\beta = \frac{1}{y_0^2}$, $w_{ij} = \frac{1}{y_0^2} y_i(t) y_j(t)$. At time 1,

$$y_i(1) = \frac{y_i(0) + 2n_i^s(0) y_i(0)}{1 + 2n_i^s(0)} = y_i(0)$$

For any $t \geq 1$,

$$y_i(t+1) = \frac{y_i(t) + 2n_i^s(t) y_i(t)}{1 + 2n_i^s(t)} = y_i(t) = y_i(0)$$

. □

C Datasets and experimental results

C.1 Real-world datasets

Table 2: Real-world dataset summary

Network	$\lvert V \rvert$	$\lvert E \rvert$	Event
Karate	34	78	Friendship
Tw:Club	703	3322	Barcelona in La-liga 2016
Tw:Sport	703	3322	Juventus vs Real Madrid 2015
Tw:US	533	13564	US Presidential Election 2016
Tw:UK	231	905	British Election 2015
Tw:Delhi	548	3638	Delhi Assembly Election 2013
Tw:GoT	947	7922	GoT promotion 2015

C.2 Influence of the opinion vector $\mathbf{y}(0)$ and network topology G

This figure corresponds to Figure 5, and is used to investigate both the effects of the opinion vector and the network topology. Horizontal subfigures show the different consensus and polarization converging states for different $\mathbf{y}(0)$s, while the vertical subfigures show the differences between the three types of random networks of similar sizes. The finding of this experiment is consistent with that of Figure 6(b).

C.3 Influence of model parameters

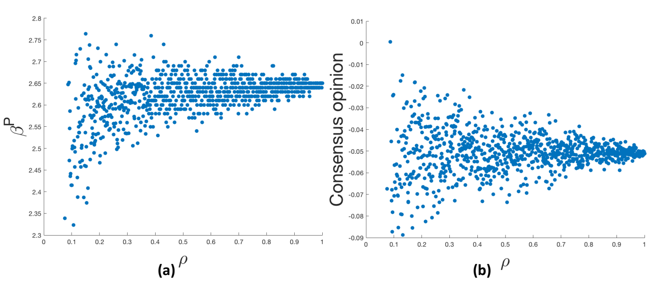

Figure 8: For an random opinion vector $\mathbf{y}(0)$, on ER models with $n = 100$ and $\rho \in (0, 1]$. (a) the value of β^P for the $\mathbf{y}(0)$; (b) the consensus opinion reach by $\mathbf{y}(0)$ when $\beta = 1$.

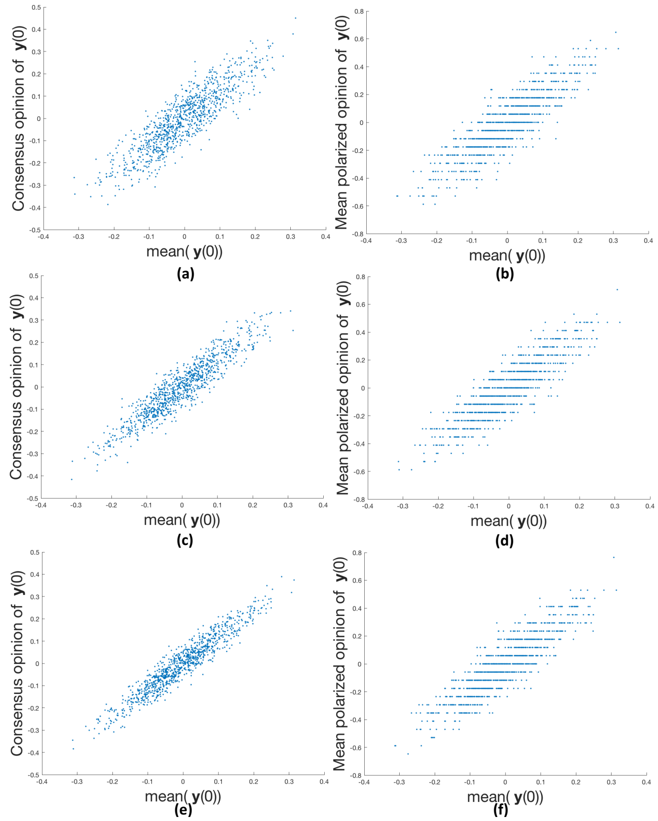

Figure 7: For 1000 random $\mathbf{y}(0)$. (a) and (b) on a BA model ($n = 34, M_0 = 3, M = 2$); (c) and (d) on an ER model ($n = 34, \rho = 0.139$); (e) and (f) on a WS model ($n = 34, K = 2$). The left column of (a), (c), (e) - the consensus opinion when $\beta = 1$; the right column of (b), (d), (f) - the mean polarized opinion when $\beta = 10$.

This experiment takes the ER model as an example and investigates the influence of the parameter ρ on the network topology, thus resulting in the influence on opinion dynamics.

C.4 Influence of edge placements

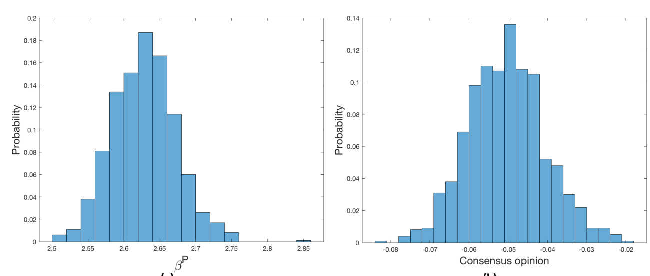

Figure 9: For an random opinion vector $\mathbf{y}(0)$ with mean -0.0395, on 1000 ER models with $n = 100$ and $\rho = 0.4$. (a) the value of β^P for the $\mathbf{y}(0)$; (b) the consensus opinion reach by $\mathbf{y}(0)$ when $\beta = 1$.

ER model is taken again as the example here for investigating the influence of the network edge placements on opinion dynamics. It shows that the network topology does have significant influence on the value of β^P and the consensus opinion value.

C.5 Influence of the edge addition/deletion in the network

We can also investigate the question: If someone wants to maximally increase/decrease the value of consensus opinion or the average polarized opinion, which edge should be removed/added?

Add One Edge - Consensus.

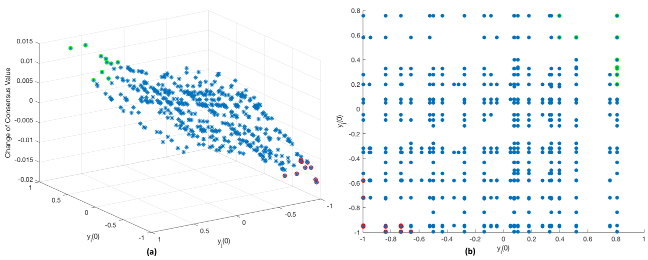

Figure 10: Add one edge on Karate network to change the consensus opinion - $\beta = 1$. Top 10 best choices are highlighted: green for increase and red for decrease.

Delete One Edge - Consensus.

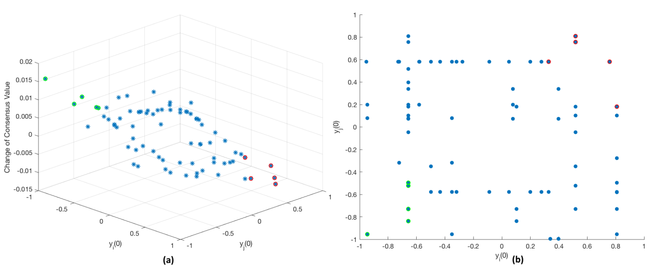

Figure 11: Delete one edge on Karate network to change the consensus opinion - $\beta = 1$. Top 5 best choices are highlighted: green for increase and red for decrease.

It shows in Figure 10 and 11 that in order to maximally increase the consensus value by editing one edge, adding the edge between the most opinionated disconnected negative nodes is the best choice when allowing only addition; while deleting the edge between the most opinionated connected positive nodes is the most effective way if allowing only deletion. A smaller consensus opinion value can be achieved by adding the edge between the most positive opinionated nodes or deleting the one between the most negatively opinionated nodes.

Edge edition that has almost no influence on consensus.

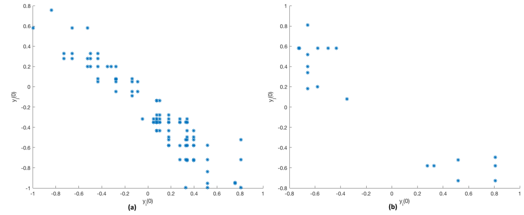

Figure 12: Additions - (a) and Deletions - (b) that cause minor change in consensus values on Karate network. ($|\text{change}| < 10^{-3}$)

However, the connections between nodes with equivalently (i.e., in terms of absolute opinion value) opposing opinions have almost no influence on the consensus value, as shown in Figure 12. In contrast, when the network gets polarized, the neighbors of the neutral nodes have more significant influence on the mean polarized opinions.