Understanding the Dynamics of Contemporary Political Polarization Through Agent-Based Models

Brad Ross, Rishi Bedi, Kenny Leung
{bross35, rbedi, kenleung} @stanford.edu

December 11, 2017

I Introduction

According to recent popular narrative, Americans and the politicians they support are developing increasingly extreme opinions across a wide range of issues, and, as a result, citizens and their representatives are much less likely to interact and collaborate with those with whom they disagree. Many have gone so far as to vilify others who identify with the opposite party, further hampering social cohesion and meaningful policy progress.

In the aftermath of the 2016 election, Professor Matthew Gentzkow wrote a summary of much of the contemporary literature on the nature of political polarization, describing several key phenomena that characterize the current state of polarization in the US. First, the distribution of individuals' self-identified positions along the liberal-conservative spectrum has not actually changed in the last 40 years. In fact, the overall distributions of beliefs about specific issues have remained singly-peaked rather than becoming bimodal as a result of people developing more extreme opinions. However, the median Democrat is much more liberal and the median Republican is much more conservative than they both were 20 years ago. How is this observation possible if the distributions of beliefs have not changed? Gentzkow points to the increase in the correlation between beliefs and the correlation between any one belief and identifying with a certain political party as the primary causes [1].

Political processes occur primarily in a social context, meaning that social structure, individual behaviors, and the landscape of political beliefs are intimately intertwined in a way that is not very well understood. In this paper, we study the bidirectional relationship between network structure and agent beliefs by defining and simulating several theoretical models to answer the following questions:

1. How do certain network structures and agent behavior schemes lead to belief polarization, particularly the correlation of beliefs?
2. How does correlation across beliefs affect network structure?
3. How are these two effects coupled?

Our core hypothesis is that particular assumptions about network structure and agent behavior could lead to an structural explanation of Gentzkow's observation of increased belief correlation despite continued unimodality of each individual belief.

Our work in this project is based on simulated networks - examining real network data is outside the scope of our work. First, we develop a model of belief propagation on fixed networks (Model 1). We proceed to develop a model of edge formation based on belief distributions (Model 2). Finally, we integrate the two models into a third unified approach, alternating belief and network updates using the mechanisms from Model 1 and Model 2. We then run repeated simulations of these models over time to determine whether or not the belief distributions and network structures they generate mimic the nature of contemporary political polarization.

II Model 1: Belief Propagation on Static Networks

Objective

The goal of developing this model is to understand how different agent behaviors dictating belief propagation on networks affect the distributions of beliefs on the network over time.

Related Work

In a 2003 paper, DeMarzo et al. [2] develop a model to show how a phenomenon they call persuasion bias affects the concurrent propagation of multiple beliefs through a static network. Persuasion bias describes the tendency of people to change a belief through repetitive exposure to a certain perspective, even if each new encounter provides no new information. In particular, the model first randomly assigns each agent in the network a set of prior beliefs. Then, over many discrete time steps, each agent takes a weighted average over the beliefs of its neighbors and assimilates those averaged beliefs into their own. These weights are initialized according to some prior, and then can vary together in relation to the weight agents place on their own beliefs over time. The authors prove

several important facts about the model. First, beliefs tend to become "unidimensional," meaning that beliefs that were initially uncorrelated tend to become highly correlated with each other over time, just as Gentzkow [1] describes. In the long run however, the beliefs of all agents in the network converge to the same set of stationary beliefs, i.e. all agents have the same set of beliefs. The paper also shows that the rate at which beliefs become unidimensional and then converge is determined by the structure of the network and the weights agents assign to the beliefs of their neighbors.

The Baseline Model

In this model inspired by the work of DeMarzo et al. [2], there are n agents that exchange information over a directed social network $G(V, E)$ between these agents. This network between agents is defined by the adjacency matrix A, where $A_{ij} = 1$ if $(i, j) \in E$, i.e. i "listens to" j (to borrow nomenclature from DeMarzo et al.), and 0 otherwise. Note that the graph is directed, so A is not guaranteed to be symmetric. We also assume that every agent listens to themself, i.e. they always factor in their current beliefs when considering how to update their beliefs based on new information. Each agent also has a set of ℓ beliefs, which are initialized randomly according to some distribution–for now, we will assume each belief is sampled independently from the standard normal distribution. The agents then exchange information and update their beliefs over a series of discrete rounds. Let $x_i^t \in R^\ell$ be a vector representing agent i's beliefs after round t. Then in round $t+1$, agent i determines x_i^{t+1} by integrating its own beliefs and the beliefs of all agents j such that $A_{ij} = 1$.

Before formally defining this belief updating process, it is important to describe qualitatively the desired properties of the updating process. First, it is important to represent persuasion bias, which is described by DeMarzo et al. as the tendency of people to not discount repeated information when updating their beliefs. Thus, the more times someone is exposed to some other belief value, the more likely that person is to integrate it into their own belief. Second, it is also important to incorporate confirmation bias into the model. Confirmation bias suggests that people tend to discount the opinions of those who do not share their current belief when updating their beliefs. Thus, in our model we want agents to weight the opinions of agents whose beliefs differ less than the opinions of agents whose beliefs are similar to their own.

Now, we formalize the update process as follows. Let the degree to which agent i "trusts" or "has faith in" the beliefs of agent j, $f^t(i, j)$, be defined as follows:

$$f^t(i,j) = \exp\Big(-\frac{\left\|x_j^t - x_i^t\right\|_2^2}{c}\Big)$$

where the constant c in the expression represents agents' confidence in their own beliefs or their "narrow-mindedness." Note that as the distance between the beliefs of agent i and agent j increases, $f^t(i, j)$ decreases sharply, and that $f^t(i, j)$ is maximized when the two agents i and j have exactly the same beliefs (most notably when $i = j$), as desired. Observe also that as agents' narrow-mindedness c increases, the weight agents give to beliefs that differ significantly from their own decreases. Of course, it is one could use any function $f^t(i, j)$ to represent the "trust" between agent i and agent j, but for the purposes of the project we will not generalize this model further. We then define the weight that agent i assigns to the beliefs of agent j at time t as the following:

$$w_{ij}^t = \frac{A_{ij} f^t(i,j)}{\sum_k A_{ik} f^t(i,k)}$$

Note that if agent i does not listen to agent j, then $A_{ij} = 0$, so $w_{ij}^t = 0$. Also note that by construction, $\sum_j w_{ij}^t = 1$, since the denominator is simply the sum of the numerators for all other agents k to whom agent i may or may not listen. Finally, we define the update rule for agent i's beliefs from round t to round $t+1$:

$$x_i^{t+1} = \sum_j w_{ij}^t x_j^t = \sum_j \frac{A_{ij} f^t(i,j)}{\sum_k A_{ik} f^t(i,k)} x_j^t$$

To express the updates to all agents beliefs simultaneously, we first define X^t to be the matrix whose ith row is x_i^t. Thus, the ith row of X^t corresponds to the set of beliefs held by agent i at time t, and the ℓth column of X^t corresponds to the set of belief values held by each agent for belief ℓ. We then specify the transition matrix W^t between rounds by setting $W_{ij}^t = w_{ij}^t$ as defined above. Because the update rule for an individual agent can just be expressed as a linear combination of rows of X^t with weights specified by the ith row of W^t, the update rule for the entire system can be expressed simply as:

$$X^{t+1} = W^t X^t$$

The Backfire Effect

A natural modification of the baseline model is to incorporate another commonly observed phenomenon in real-world belief propagation: the backfire effect [3]. The backfire effect describes situations in which a person's current beliefs are reinforced, rather than undermined, when exposed to other information that challenges that person's beliefs. One could describe the phenomenon as "negative trust"–being exposed to beliefs that are significantly different from yours causes your beliefs to become more extreme in the opposing direction. One way to model this phenomenon is to subtract a constant b from the previous definition of $f^t(i, j)$:

$$f^t(i,j) = \exp\Big(-\frac{\left\|x_j^t - x_i^t\right\|_2^2}{c}\Big) - b$$

Now, if the distance between two agents' beliefs is sufficiently large, $f^t(i,j)$ will be negative. Besides this addition, the previous model of belief propagation remains unchanged.

Model Evaluation

For this model of belief propagation to accurately model contemporary political polarization, it must generate belief distributions that satisfy two main properties. First, beliefs must become more correlated with each other over time as observed by Gentzkow, i.e. they must converge to a "uni-dimensional" subspace of belief space. Second, the distribution of beliefs must remain somewhat normally distributed over time and not converge to a single point in belief space, unlike the behavior of the model proposed by DeMarzo et. al. [2]. Beyond visually inspecting nodes' beliefs plotted in belief space, we use the following two methods to assess the uni-dimensionality and normality of beliefs over time.

To determine empirically whether beliefs converge to a uni-dimensional distribution, we standardize each column of X^t for every t and compute the principal components of each X^t. To do so, we compute the Singular Value Decomposition of each X^t, $X^t = U^t \Sigma^t {V^t}^T$. Intuitively, the fraction of the sum of the squared singular values provided by the square of the first singular value:

$$\frac{{\sigma^{(t)}}_1^2}{\sum_{i=1}^{\ell} {\sigma^{(t)}}_i^2}$$

represents the proportion of variance in the data that occurs along the first principal component of X^t (the first right singular vector $v_1 \in \mathbb{R}^\ell$), which lies in belief space. If the proportion is closer to 1, then the variance in beliefs along the first right singular vector captures most of the total variance in beliefs, meaning that beliefs lie in a uni-dimensional subspace of belief space.

To assess the normality of beliefs over time, we first project the distributions of beliefs over time X^t onto the first principal components over time $u^{(t)}$ by computing $p_{u_1}(X^t) = u_1^{(t)} \sigma_1^{(t)}$. Then, we compute two statistics used in the statistics literature to compare empirical distributions to the standard normal distribution: *skewness* and *Fisher kurtosis*.

Skewness is a measure of distribution asymmetry, i.e. how much of its mass is to the left and right of its mean. It is computed by standardizing the third central moment of a distribution, like so:

$$\text{Skew}(X) = \mathbf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

For a standard normal distribution, the skewness is 0 because the distribution is symmetric.

Vanilla kurtosis is a measure of how little mass is located in the tails of a distribution. Higher kurtosis values indicate more probability mass towards the center of the distribution, and lower kurtosis values indicate more probability mass in the tails of the distribution. It is computed by standardizing the fourth central moment of a distribution, like so:

$$\text{Kurt}(X) = \mathbf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

Any normal distribution X has $\text{Kurt}(X) = 3$, so the Fisher kurtosis $\text{Kurt}_F(x) = \text{Kurt}(x) - 3$. Fisher kurtosis values slightly above 0 indicate distributions with kurtosis similar to that of normal distributions. Fisher kurtosis values higher than 0 indicate distributions with less mass in their tails than the normal distribution.

III Model 2: Network Evolution with Static Beliefs

Objective

The objective of this model is to investigate a network formation algorithm that encodes homophily, the tendency for people to associate with those with whom they are similar. We begin with a random network and, over time, expect to form new clusters of nodes with similar beliefs, or isolated "echo chambers", i.e. connected components. We investigate the relationship between correlation of beliefs and the rate at which nodes form such clusters.

Related Work

Papadopoulos et al. [5] identify similarity as an important factor that should be included when modeling real-world networks using preferential attachment. In their paper, they describe a geometric interpretation of preferential attachment: when a node joins the network at time t, it is placed on a polar coordinate system with distance to the origin given by $ln(t)$, and angle given by a measure of the node's features (the belief vector, in our case). A new node connects to the m nodes (related to the desired average degree distribution) with which it has least hyperbolic distance, thus accounting for both popularity and similarity. Their simulated network exhibits high clustering that is characteristic of real-world networks.

Lee, Zaheer, et al. define a similar measure called "vertex affinity" [3]. Whether a new node connects to a particular existing node in the graph is governed by a probability distribution dependent on the degree of the receiving node and the computed affinity between the two nodes. Their metric of affinity is computed using logistic regression.

In our project, we depart from the preferential attachment model used by Papadopoulos, instead incorporating similarity of beliefs into an edge *rewiring* process on a randomly initialized graph (Erdos Renyi, Small-World) where degree distribution is more evenly spread across nodes in the network. We use Euclidean (2-norm) distance between belief vectors to influence the probability of deleting and creating new edges.

Homophilous Edge Rewiring Model

In this model, we initialize our network as either an Erdos-Renyi or Small World (Watts-Strogatz) undirected graph with nodes N and edges E without self-edges. Each node in the graph is associated with *fixed* a ℓ-dimensional belief vector $x_i \in \mathbb{R}^\ell$ drawn from a multivariate normal distribution with given covariance $C \in R^{\ell \times \ell}$. The covariance matrix is a symmetric matrix with ones in the diagonal and k in all other cells of the matrix, where k represents correlation between beliefs. Node beliefs remain fixed as the network evolves, but edges in the graph are removed and added over time.

We perform one instance of edge rewiring on the graph by first randomly sampling an edge (a, b) to delete, based the following probability distribution P_{del} where nodes a, b hold respective beliefs x_a, x_b:

$$P_{del}(a,b) = \frac{\|x_a - x_b\|_2}{\sum_{(i,j)\in E} \|x_i - x_j\|_2}$$

Next, we introduce an edge to the graph. At random, we pick a node $c \in \{a, b\}$ to be fixed as an endpoint of the new edge. We then sample a node d to serve as the edge's other endpoint. Node d cannot equal node c, because we do not permit self-edges, nor can it be a node that is already connected to c.

The probability of adding a new edge between two nodes depends on two factors: their belief similarity, and their proximity in the graph. In order to measure their proximity, we introduce a function $D(x, y)$, which evaluates to the shortest number of hops between nodes x and y in the network. If x and y are part of different connected components, this function returns $1 +$ the diameter of x's connected component. Using this function, we favor introducing an edge between nodes that are closer to one another or belong to the same connected component. Note that $D(x, y) \neq D(y, x)$, iff x and y belong to different connected components.

Then, we randomly sample a node d from the following probability distribution, parameterized by real-values k_H (termed the "homophily parameter") and k_C (termed the "community parameter"):

$$P_{add}(d|c) = \frac{\|x_c - x_d\|_2^{-k_H} \cdot D(c,d)^{-k_C}}{Z(c)}$$

The normalization value $Z(c)$ is computed by summing the numerator of P_{add} for all nodes not equal to and not already connected to c:

$$Z(c) = \sum_{i:i\neq c;(c,i)\notin E} \|x_c - x_i\|_2^{-k_H} \cdot D(c,i)$$

Model Evaluation

In theory, we would expect nodes to form clusters more quickly as the correlation between beliefs increases. In a setting where agents are more likely to form connections with like-minded peers, increased correlation initializes nodes that are more consistently conservative or liberal across beliefs. We use the average clustering coefficient of the network and the progressive number of connected components to gauge the rate at which cluster formation occurs.

IV Model 3

Objective

Model 1 describes belief evolution over time given a fixed graph. Model 2 describes graph evolution over time given a fixed belief distribution. In Model 3, we unify these approaches in an alternating update process to interrogate their coupled behavior.

Model Description

We initialize an undirected Small World (Watts-Strogatz) random graph with nodes N and edges E. Node beliefs are initialized as samples from a multivariate normal distribution. At every iteration, we perform the following updates:

1. Holding edges E fixed, update beliefs x according to the following update rule:

$$f^t(i,j) = \exp\Big(-\frac{\|x_j^t - x_i^t\|_2^2}{c}\Big) - b$$

$$x_i^{t+1} = \sum_j w_{ij}^t x_j^t = \sum_j \frac{A_{ij} f^t(i,j)}{\sum_k A_{ik} f^t(i,k)} x_j^t$$

   We vary the backfire parameter b in our experiments, but hold the confirmation bias (narrow-mindedness) parameter c fixed at 1.

2. Holding beliefs x fixed, update edges according to the edge updating algorithm described below.

for $i := 1$ *to EdgeFlips* **do**
  Sample edge $e = (a, b) \sim P_{del}$;
  Delete edge e;
  Sample node $b' \sim ||x_a - x_b'||_2{}^{-k_H}$;
  Add edge (a, b')
end

Algorithm 1: Edge updating procedure

We vary the *EdgeFlips* parameter, but hold k_H constant. More *EdgeFlips* per iteration means more edges are rewired given the belief distribution - this parameter controls the relative rates of belief updating and edge rewiring.

Model Evaluation

We hypothesize that coupled belief and edge dynamics will result in uni-dimensionality of beliefs without belief convergence in the limit. Specifically, we believe the variance of the belief distribution will be along a single axis, and that the distribution of beliefs when projected on this axis will be normal. We also hypothesize an increase in the global clustering coefficient, reflecting the emergence of "echo chambers" of belief.

We evaluate uni-dimensionality of beliefs by using PCA on the belief matrix, as explained in the Model Evaluation section of model 1. The percentage of variance explained by the top principal component reveals the uni-dimensionality of beliefs. We evaluate non-convergence of beliefs by computing the Fisher kurtosis of the belief distribution (high kurtosis means convergence) as done for model 1, and normality by computing the skew and Fisher kurtosis of the distribution of beliefs projected on the top principal component (both should be close to 0 if the belief distribution is approximately normal).

V Simulations and Results

A Model 1

As described above, if our model were an accurate representation of belief propagation over time, we would expect to see agents' beliefs collapse to a single axis (uni-dimensionality), not collapse to a single point in belief space, and remain symmetrically and somewhat normally distributed.

Displayed below is a sample simulation of the baseline model (backfire effect constant $b = 0$) using 200 agents, each with 2 beliefs, interacting on a Small World graph with starting out-degree 2 and rewire probability 0.5:

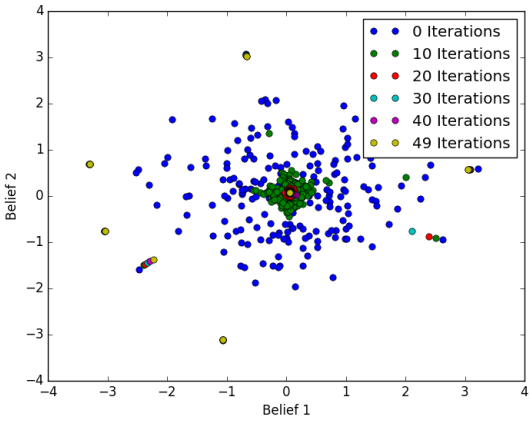


Figure 1: Beliefs of agents in a Small World graph after various iterations of belief propagation.

Unfortunately, as can be seen in the graph, after just 40 iterations, most nodes beliefs converge to roughly the same beliefs, as predicted by DeMarzo et. al [2]. Repeating this simulation 100 times yields the same behavior. Clearly then, the introduction of varying edge weights based on the similarity of connected nodes' beliefs is not sufficient to induce the belief polarization characteristic of our era, i.e. beliefs that are distributed with a single peak along some axis in belief space. Note that several nodes do remain roughly in their original belief positions over the course of the simulation. This phenomenon occurs because nodes that are not connected to the rest of the graph, cannot integrate the beliefs of the rest of the nodes in the graph into their own.

Below are the results of a sample simulation with backfire constant $b = 0.05$ over the same Small World graph.

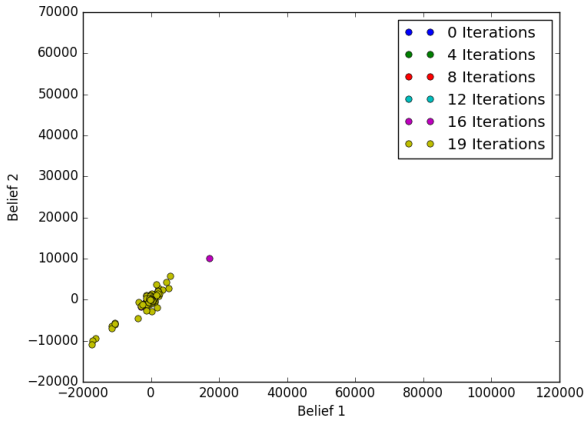


Figure 2: Beliefs of agents in a Small World graph after various iterations of belief propagation, with the backfire effect.

Oddly enough, we see that beliefs tend towards extreme values after just 20 iterations, completely obscuring the initial positions of beliefs. However, agents' beliefs do appear to dissipate outwards along a single axis. Again, this behavior is consistent across 100 simulations. It appears then that the introduction of the backfire effect causes agents' beliefs to become uni-dimensional and much more extreme over time. Intuitively, this observation makes sense: if an agent is exposed to another agent with very different beliefs, the two agents' beliefs will be pushed in opposing directions. Unfortunately, the extreme rate at which beliefs diverge does not appear realistic at all.

To determine empirically that the model with nonzero backfire converges to a single axis of beliefs and to determine that axis, we employ the principal component computation methodology explained above. To conduct this analysis, we run 20 simulations of belief propagation on the same 200-node Small World graph as before with 5 beliefs and backfire constant $b = 0.05$. We then plot the average proportions of each squared singular value to the sum of squared singular values across all 20 simulations over time:

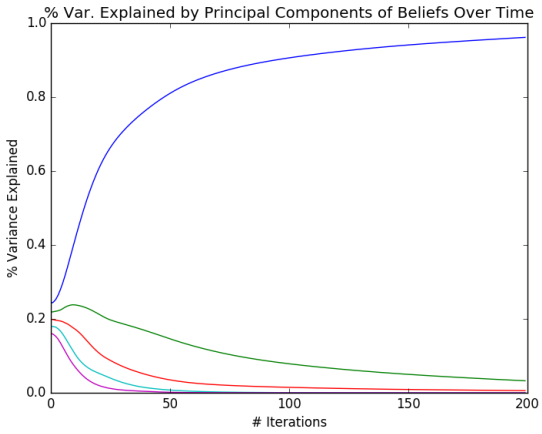


Figure 3: Average percentage of variance explained over 20 trials by principal components of agent beliefs after various iterations of belief propagation.

In the plot, we see that the proportion of variance along the first principal component increases towards 1, indicating that, in the limit, most, if not all the variance in the data occurs along a single principal axis.

We also assess the normality of the distribution by plotting the average skewness and Fisher kurtosis of beliefs projected onto their first principal components over time, as described above. We report the average statistics over 20 trials:

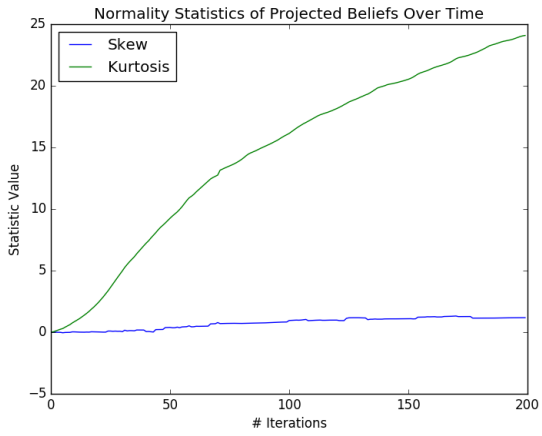


Figure 4: Average skewness and Fisher kurtosis of beliefs projected onto the first principal component of the belief matrix over time.

As can be seen in the plot, the skewness of the distribution remains close to 0, indicating that the distribution is symmetric. However, the Fisher kurtosis grows continually over time, indicating that somehow, the distribution of projected beliefs is becoming more and more peaked about its center. As we explain below, this is a sign that beliefs are actually converging with backfire, albeit on an odd scale.

Before going further, we try to rectify the divergence of beliefs with the backfire model. To do so, we standardize beliefs (subtract the mean of each belief from each agent's belief values and divide by standard deviation of each belief) after each iteration.

Below are the results of a sample simulation with backfire constant $b = 0.05$ over the same Small World graph with standardization after every iteration:

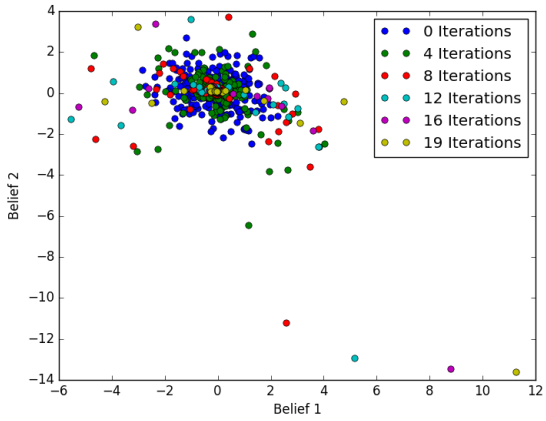


Figure 5: Beliefs of agents in a Small World graph after various iterations of belief propagation, with the backfire effect and belief standardization after every iteration.

Unfortunately, we again see belief convergence after just 20 iterations. The fact that standardizing beliefs after each iteration yields the same convergence pattern as that seen without the backfire effect gives us insight into what exactly is occurring when we simulate belief propagation with backfire but no standardization. By including the backfire effect in the model without standardizing, we allow beliefs to repel each other, causing more extreme beliefs to form over time. While the absolute values of these beliefs do increase over time, the fact that they still converge if we do standardize them indicates that they are actually getting closer together relative to their increasing magnitudes. In other words, the scale on which beliefs are measured increases, but the variance of beliefs measured in that scale still goes to zero over time. The continually increasing kurtosis in the plot of normality statistics for the model with backfire above only confirms this fact, since more and more of the mass of the distribution shifts to its center.

B Model 2

In our simulations, we initialize our network as:

- an Erdos-Renyi graph consisting of $n = 100$ nodes and $m = 100$ undirected edges
- a Small-World graph consisting of $n = 100$ nodes each initialized with 3 outgoing edges and a rewire probability of 0.5

Each node has a fixed 2-dimensional belief vector, which is drawn from a multivariate normal distribution with a covariance matrix constructed from correlation k, as described earlier in the model specification. Across

our simulations, we vary the correlation, measuring the average clustering coefficient and number of distinct connected components in the network changes over each iteration.

We fixed the homophily parameter $k_H = 1.0$. We alos fixed the community parameter to be $kC = 0$, essentially ignoring node path lengths during the rewiring step), since we found that the parameter did not impact the clustering of the graph (data not shown).

Our simulations demonstrate that both the clustering coefficient and number of connected components increase after iterations of edge rewiring, across a range of correlation values, and then begin to plateau. These values are averaged across ten independently-initialized trials.

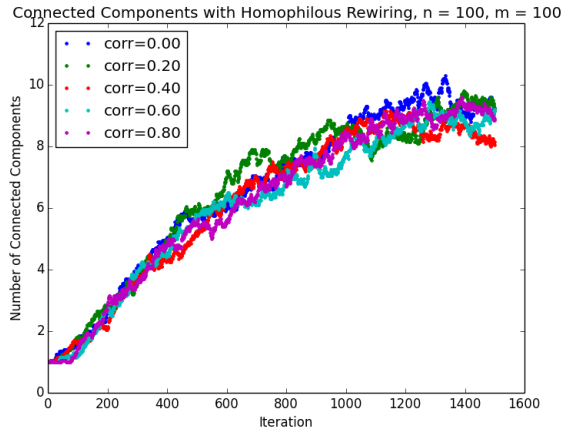


Figure 6: Number of connected components in an Small-World graph after iterations of homophilous edge rewiring.

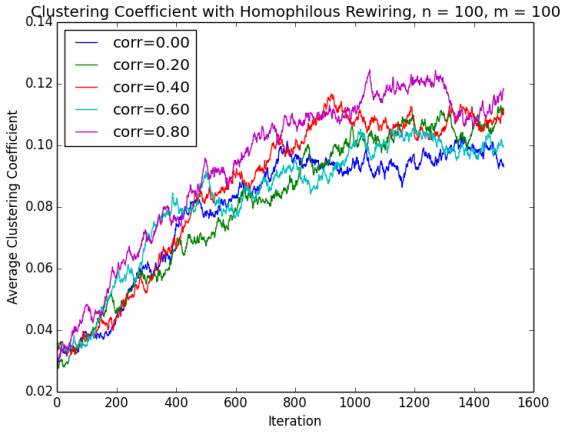


Figure 7: Average clustering coefficient of an Small-World graph after iterations of homophilous edge rewiring.

We performed the graph updates on an Erdos-Renyi graph as well, with similar results. Our results indicate that on average, the rate of clustering remains similar across different correlations and initial graph structures. This is contrary to what we expected to see, which was an increase in the rate of clustering when accompanied by a higher correlation.

It is possible that the correlation plays a more significant role in the clustering rate in graphs that are initialized with a particular combination of nodes and edges, homophily parameter value.

C Model 3

For this section, we use Watts-Strogatz Small-World graphs with 100 nodes and out-degree of 2. Unless otherwise specified, $EdgeFlips = 20$ per iteration, the rewire probability was set to 0.5, and simulations were run for 400 iterations. Note that simulations did not necessarily converge after this long, but were cut off nonetheless for consistency.

We observe that without the backfire effect ($b = 0$), beliefs converge invariant to other parameter settings, as in DeMarzo's work [2]. When we include the backfire effect, however, we observe uni-dimensionality without obvious convergence, even in the limit as we run longer simulations. This seemingly convergence-free unidimensionality, however, coincides with unconstrained belief drift and unconstrained increase in belief variance. Similar to the results described in Model 1, these beliefs are actually becoming closer together relative to their increasing magnitude. Thus, while our frame of reference of belief space has changed, the variance of beliefs in this scale still goes to zero over time. We can demonstrate this quantitatively by measuring the Fisher kurtosis of the belief distribution.

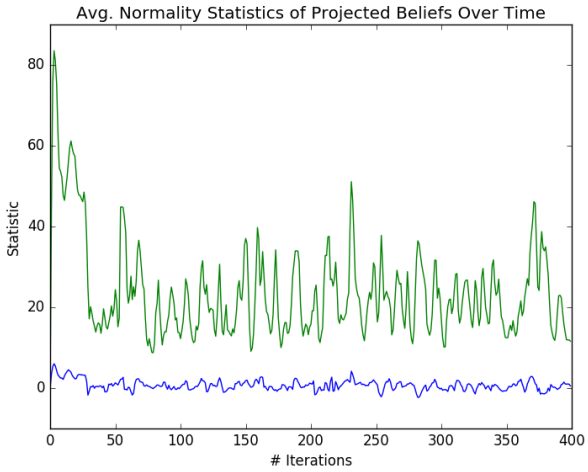


Figure 8: Skew and kurtosis of belief distribution of Model 3 with backfire and no belief rescaling.

We attempt to remedy this unconstrained drift by rescaling the belief values to always be between -1 and 1 ("belief rescaling"). This results in the following final statistics over the distribution of belief values, over time:

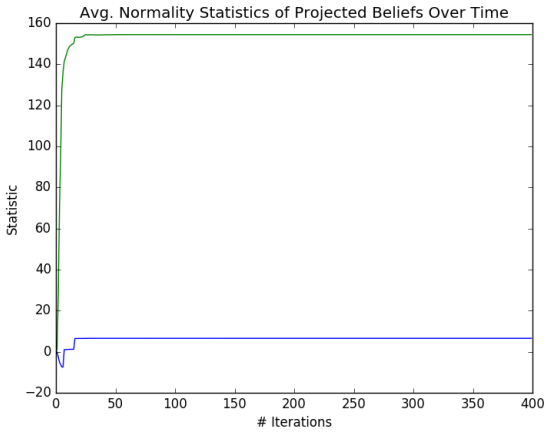


Figure 9: High kurtosis of belief distribution for Model 3 with backfire and belief rescaling indicates belief collapse.

We will address the plateauing Fisher kurtosis below.

Now we examine the actual belief distributions over time for the model with backfire and belief rescaling. It is not obvious that convergence is occurring from the plot of all beliefs, since many singletons are becoming isolated over the course of the simulation and thus not continuing to collapse with the remaining non-singleton connected components.

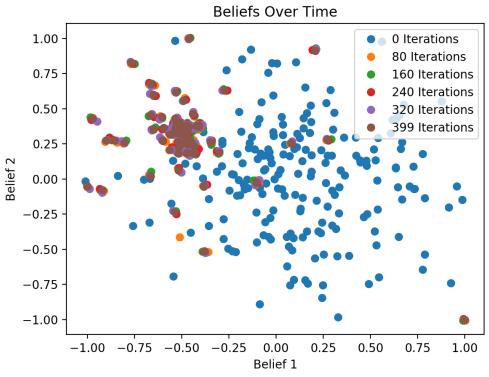


Figure 10: Beliefs over time for Model 3 with backfire and belief rescaling.

The collapsing behavior of connected nodes is evident, however, from this plot of belief distribution over time for the largest connected component that remains at the end of our simulation.

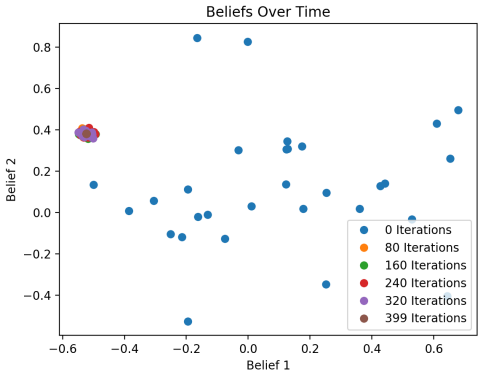


Figure 11: Beliefs over time for the nodes in the largest connected component at the end of the simulation; Model 3 with backfire and belief rescaling.

The singletons which become isolated over the course of simulation still generally fall in the same axis of variance as the collapsing components, as demonstrated by the percentage of variance explained by each principal component of the belief matrix.

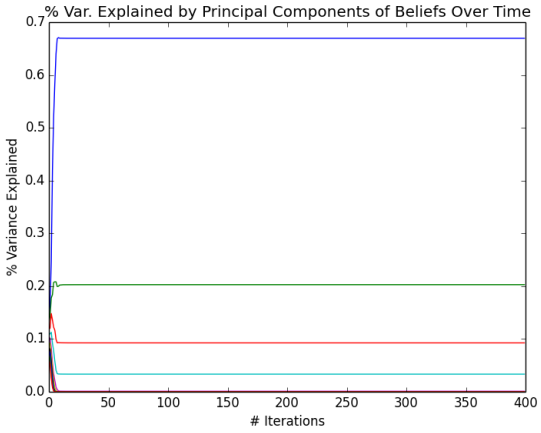


Figure 12: Average percentage of variance explained over 5 trials by principal components of agent beliefs after various iterations of belief propagation in model 3 with backfire and rescaling.

We observe that the number of "singleton" components (connected components of size 1) steadily increases over time, reflecting nodes who lose their only edges to other nodes in the network through the edge rewiring process. This naturally coincides with a steady decrease in size of the largest connected component. Since these orphaned nodes originated in the same connected component, they remain essentially along the axis of the principal component, as uni-dimensionality emerges before belief convergence due to the faster speed at which beliefs propagate compared to the speed at which nodes are shed. The increasing then decreasing behavior in the size of the second largest connected component is interesting and not intuitive, so we refrain from discussing it further. Looking

back at Figure 9, we note that the kurtosis likely plateaus because the isolated singletons no longer converge along with the remaining connected components, thus enforcing some non-zero spread in the belief distribution despite convergence of remaining connected nodes.

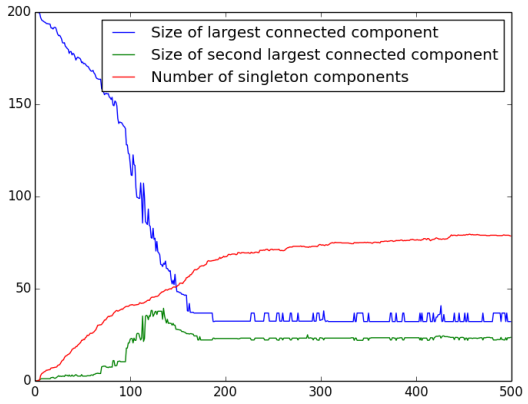


Figure 13: Sizes of the first and second largest components, as well as the total number of singleton components, across iterations and averaged over 5 trials.

Another way to demonstrate convergence in the non-isolated nodes is by looking at the average L2 distance of nodes in a given connected component to the center of mass of beliefs in that component. We find that this average distance collapses to near-zero very quickly, further supporting convergence of beliefs of connected nodes, despite backfire and edge rewiring.

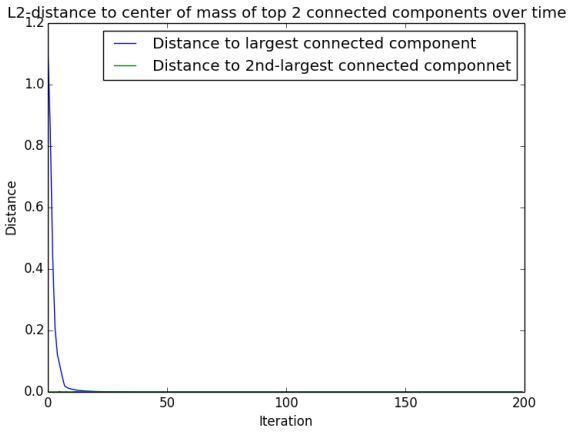
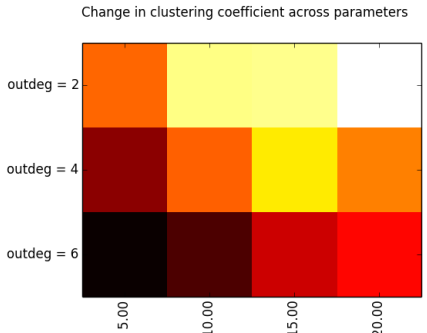


Figure 14: For the first and second largest connected components, the average Euclidean distance between each node in the component and the average belief of all nodes in the component, across iterations and averaged over 5 trials.

The figure below contains the increase in clustering coefficient measured after using Model 3 to update a Small-World network. Each cell represents an independent simulation with some configured initial value of node out-degree and number of edges rewired at each iteration. Clustering coefficient values range from 0 (dark) to 1 (light).



	rewire = 5.00	rewire = 10.00	rewire = 15.00	rewire = 20.00
outdeg = 2	0.0196	0.0802	0.0816	0.1008
outdeg = 4	-0.0352	0.0182	0.0532	0.0262
outdeg = 6	-0.0678	-0.051	-0.019	-0.0046

Figure 15: Heatmap and associated table displaying the average change in clustering coefficient of a network initialized with $n = 200, m = 400$, and rewire probability $= 0.5$, after 400 iterations of graph updates and averaged over 5 trials.

We observe that the average clustering coefficient increases as a result of the graph updates. This can be explained by homophilous rewiring coupled with belief propagation. As more edges rewire from connecting belief-divergent nodes to belief-similar ones, the network begins to splinter into disparate connected components.

On the other hand, the clustering coefficient increases less significantly when the network is initialized with a high out-degree. We notice that the change in clustering coefficient is negative for high out-degrees, meaning that nodes become less clustered over time. The reason for this is that an initial node out-degree of 6 is already high in proportion to the total number of edges in the graph, so the graph is already highly clustered to begin with.

VI Discussion

At the beginning of our project, we hypothesized that developing a coupled model of belief propagation and network evolution would generate belief distributions and network structures similar to those observed in today's highly polarized climate. In particular, we hoped that our models would generate belief distributions that converged over time to uni-dimensional subspaces of belief space. We also expected that these uni-dimensional distributions would not collapse to single points in belief space as they had for DeMarzo et. al. [2], instead remaining roughly normally distributed along their axis of principal variance. Regarding network evolution, we anticipated that increasing the correlation between beliefs would lead to increased network clustering and even the development of echo chambers isolated from other parts of the graph.

However, our results for Model 1 demonstrated that, although agents' beliefs became uni-dimensional over time, it was impossible to avoid convergence of their distribution to a single point in belief space. Despite appearing to guarantee non-convergence, including the backfire effect did little to stop the agents' beliefs from coming together.

In our analysis of Model 2, we observed that belief correlation did not have a significant impact on the resulting clustering of the network after iterations of homophilous rewiring.

In Model 3, which consisted of edge rewiring coupled with belief propagation, the beliefs of nodes in our network tended toward a uni-dimensional distribution in the short run, but converged in the long run. We observed that the largest connected component in our network consistently lost mass as the network continued to evolve. This occurred because outlying nodes were ousted from the larger connected components due to the forces of confirmation bias and the backfire.

Although our modeling choices were informed by the psychology literature and intuition about human behavior - confirmation bias and the backfire effect are both well-documented and intuitive assumptions - they appear insufficient for producing the sort of stable, non-convergent uni-dimensional belief distribution which would provide a mechanistic underpinning of Gentzkow's observations about the nature of political polarization [1].

VII Future Work

There is always possible future work in expanding the breadth and rigor of our modeling choices. For example, it would be interesting to run coupled simulations as in Model 3 on real-world networks, initialized with real belief distributions. Our results appear quite invariant to parameter choices and the random graph models we tried, but it is of course possible there are important characteristics of real-world networks missing from all our attempts.

Just as our modeling choices were influenced by priors on likely social behavior, it is instructive to consider how our model results might inform our understanding of social behavior. Many of our conclusions appear unrealistic, such as the belief drift observed in Model 3 without normalization and the progressive isolation of singletons with normalization. It would be interesting to consider what might be missing from our understanding of agent behavior, then, which if added could remedy these unrealistic simulation outcomes.

VIII Acknowledgements

We would like to thank Professor Matthew O. Jackson of Stanford Economics for his advice on how to approach the project and his suggestions of papers to read for inspiration. We would also like to thank Prof. Jure Leskovec and the rest of the teaching staff of CS224W for teaching us the fundamentals of network analysis.

IX Work Sharing

We would like to receive equal credit for our work on the project, as we mostly did our work jointly.

References

[1] M. Gentzkow. Polarization in 2016. *Toulouse Network of Information Technology white paper, 2016.*

[2] P. Demarzo, D. Vayanos, and J. Zwiebel. Persuasion Bias, Social Influence, and Unidimensional Opinions. *The Quarterly Journal of Economics, Volume 118.* Issue 3, 1 August 2003, Pages 909-968.

[3] B. Nyhan and J. Reifler. When corrections fail: The per- sistence of political misperceptions. Political Behavior, 32(2):303âĂŞ330, 2010.

[4] J.-Y. Lee, M. Zaheer, S. Günnemann, and A. Smola. Preferential Attachment in Graphs with Affinities. *Proceedings of Machine Learning Research.* PMLR 38:571-580, 2015.

[5] F. Papadopoulous, M. Kitsak, M. Serrano, M. Boguná, and D. Krioukov. Popularity versus similarity in growing networks. *arXiv preprint.* arXiv:1106.0286, 2011.

[6] M. Vicario, A. Scala, G. Caldarelli, H. Eugene Stanley, and W. Quattrociocchi. Modeling confirmation bias and polarization. *Scientific Reports 7*, Article number: 40391 (2017).