Opinion Dynamics of Skeptical Agents

Alan Tsang
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
akhtsang@uwaterloo.ca

Kate Larson
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
klarson@uwaterloo.ca

ABSTRACT

How does skepticism affect opinion formation in networks? In many settings, agents exhibit skepticism in the presence of people whose beliefs radically different from their own, and they are reluctant to be persuaded by such individuals. We present a model of opinion dynamics where agents are receptive toward other agents that have similar opinions, but remain skeptical of agents holding disparate opinions. We analyze how agents with extreme opinions affect the general population, using simulations on Barabási-Albert random graphs, and modified Erdös-Rényi random graphs that incorporate homophily. Finally, we show that even skeptical agents are able to come to an early consensus and take coordinated action to reach a final opinion in most settings; but, agents in homophilic networks may fail to converge to a single opinion. Paradoxically, this happens when agents are *least* skeptical, and are able to stabilize themselves by balancing influence from extremists from opposing camps.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Experimentation, Theory

Keywords

Agent-Based Models, Social Simulation, Social Networks, Information Propagation, Innovation Diffusion, Opinion Dynamics, Cognitive Convergence, Degree-based Voter Model, Homophily

1. INTRODUCTION

The field of opinion dynamics draws its early roots from the study of innovation diffusion. Under these model, agents within a community individually choose whether or not to adopt a novel trait based on the actions of their neighbours, in a repeated coordination game. Early studies focused on the decision to adopt new technologies such as antibiotics [4] and hybrid corn [23]. These decisions are naturally modelled by binary variables, and the model can just as easily be applied to study operating system and social media adoptation today.

While binary variables are appropriate for modelling such decisions, they lack the richness necessary to capture more gradated opinions such as political leanings, socioeconomic standings, or various fashions and fads. The field of opinion dynamics generalizes the innovation model by interpreting opinions as continuous values in the interval $[0, 1]$. Agents' opinions are swayed by each other through repeated interactions, and the opinions of the community gradually converge to an equilibrium.

The analogous problem to innovation adoption in the continuous domain is the study of the effects of extremism in a community. In the discrete model, "early adopters" are represented as agents whose opinions are fixed to a certain value. In the continuous model, agents with fixed (or merely steadfast) opinions at the ends of the spectrum are akin to extremists in a population. The pitfall is that most mathematical models in this domain tend to focus on convergence of opinions [13]. The challenge then is to devise a model that allows fractions of a population to disagree with each other, even at equilibrium.

A class of phenomena known to cognitive scientists as *cognitive bias* motivates our approach. When subjects experience cognitive bias, they arrive at skewed or irrational conclusions based on an inaccurate and subjective reconstruction of reality [1]. One particular type of cognitive bias is *motivated cognition*, where observations are evaluated in ways most beneficial to the individual or compatible with the individual's beliefs.[1] In one experiment [17], when asked to rate the attractiveness and personality of a confederate, participants who were led to believe they must go on a date with the confederate consistently gave more favorable ratings. In a study of a more everyday phenomenon, after observing a sports event containing a minor but questionable call, fans of the losing team were more likely to attribute the outcome to referee error over qualities of the teams, when compared to fans of the winning team; however, in games where no such a questionable call is evident, there is no such bias.

The second study is very telling. Two groups of people were exposed to the same evidence, but their opinions (on

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

[1] A competing theory called *cognitive dissonance* explains the same behavior through a different set of mechanisms. The specific mechanics of these behaviors are unimportant to us, as we are only concerned with the fact that these behaviors *do* occur regularly in humans and other animals.

the competitive merits of the respective teams) did not converge. This seems to fly in the face of belief updates via Bayes' rule. Jaynes provides some insight on this by allowing agents to consider the possibility that the evidence is unreliable. The further away the evidence is from an agent's expectations, the more likely the agent is to believe that it is flawed, and therefore the less persuasive the evidence [14]. Laplace summarizes this idea nicely in his essay on probability [18], that outlandish claims "decrease rather than augment the belief which they wish to inspire; for the those recitals render very probable the error or the falsehood of their authors."

This idea of motivated cognition is central to our model. Agents are skeptical of another agent when their opinions diverge, but are more receptive to persuasion when their opinions better align. In the rest of this paper, we detail work on related models in opinion dynamics, then we formalize this concept of skepticism and trust[2] in our model of opinion dynamics, and explore its effects in simulated social networks.

2. RELATED WORK

Numerous researchers in the artificial intelligence community have explored how ideas diffuse through social networks. Recent works that emphasize the convergence of opinions include a model for how language features emerge, evolve and expire [24] and how opinions can be efficiently diffused in large communities [21]; Parunak also coins the term "collective cognitive convergence" in his study of the phenomenon, which also includes a more comprehensive review of literature [20].

Our skeptical paradigm places emphasis on limiting interactions between agents whose opinions diverge significantly. Many researchers in the 20th century have explored various linear models for opinion formation [13]. Krause [15] was the first in the field to incorporate nonlinear systems, formulating the *bounded confidence* model. In this scenario, a panel of experts must arrive at a consensus about the evaluation of a piece of work. Each begins with a private opinion and a level of confidence on the accuracy of that opinion. As they interact with each other, they allow their opinions to be swayed by only those experts who hold opinions within a certain interval of theirs. The more confident the expert, the smaller their interval. The more confident the other expert is, the larger the sway.[3]

In the bounded confidence model, the ability of agents to influence each other is cut off sharply at a certain limit. Deffuant [5] refines this model by incorporating a Gaussian kernel with bandwidth equal to the confidence level. This allows influence to be dropped off in a smooth, continuous manner. The initial motivation for this model was to study the emergence of "mob mentality", where sensible individuals are driven to extreme actions when present in a crowd containing only a small fraction of radicals [6]. Interestingly, while this avalanche effect sits as a counterpoint to the skeptical behavior motivating our model, it is emergent in our experiments. In his paper, Deffuant explores the ramifications of this model on Erdös-Rényi random graphs, while a variant of his model is explored in small-world social networks [11].

The idea of skepticism arising in social networks, between agents with different opinions, has also been explored more recently by Cho, Ver Steeg and Galstyan, and verified on data from the U.S. Senate [3]. In their paper, they consider co-membership in groups as being a surrogate for trust and a driver for evolution of network structure. Salzarulo [22] also investigates a similar phenomenon based on exogenously defined "in-group" and "out-group" mentalities.

Carvalho and Larson [2] explore the role skepticism plays in expert panels. In their model, a group of experts with initially different opinions revise their evaluations, with less weight given to experts whose opinions differ greatly from their own. They show that such a panel always reaches consensus, and such a model works efficiently on real world data.

The concepts of trust and persuasion have also been explored from different perspectives. Fang, Zhang and Thalmann [10] proposed a model for uniting the concepts of trust and innovation diffusion by allowing trust itself to be diffused through a network; Hazon, Lin and Kraus [12] considered how group decisions may be altered by appealing to self-interested individual to change their preference ballots.

Finally, Martins [19] proposes a model bridging the continuous and discrete domains, where agents maintain an internal (continuous) probability about which of two actions is more profitable, but is only able to communicate with each other through taking (discrete) actions. His simulations on a grid lattice show that a population eventually reaches stable equilibrium configurations of actions, where certain agents can become extremely confident of their choices.

3. OPINION DYNAMICS MODEL

In our model, agents $\{1, 2, \ldots n\}$ are embedded in a social network represented by a simple, undirected graph $G = (V, E)$. Each agent i has an *opinion* $x_i \in [0, 1]$ and is influenced by neighbours $N(i) = \{v \in V | \{i, v\} \in E\}$. For each neighbour j, i maintains a *trust* value $w_{i,j} > 0$ representing the weight i gives to j's opinions.

We define a trust function T, based on the distance between any particular opinions x and x' via the Gaussian kernel, described in Equation (1). The bandwidth parameter h represents the *empathy* of the population; a higher empathy reflects a population more willing to be persuaded by someone with a more different opinion.

$$T(x, x') = exp(-\frac{(x - x')^2}{h}) \tag{1}$$

Equations (2) and (3) describe the trust and opinion updates performed at each time step: each agent i updates its opinion x_i and trust values $w_{i,j}$ via a weighted average. A lower $w_{i,j}$ indicates i is more skeptical of j, and therefore, less influenced by j's opinions. We include a parameter $w_{i,i}$ as the inertia of i's trust and opinions, to be weighted against those of its neighbours. We also define a parameter r representing the *learning rate* of the population; a higher learning rate reflects a more judgemental population that more quickly distrusts someone with a different opinion. Note also that the opinion update (2) is performed before the trust update

[2] We use "trust" only in its plain, nontechnical sense, and not in reference to mathematical trust models within the multi-agent systems community.

[3] This description is based on subsequent work in [16], as [15] is written in German. A subsequent model by Deffuant [7] also weights the amount of influence exerted by the degree of overlap between the intervals.

(3) in each iteration.

$$x_i \leftarrow \frac{w_{i,i}x_i + \sum_{j\in N(i)} w_{i,j}x_j}{w_{i,i} + \sum_{j\in N(i)} w_{i,j}} \quad (2)$$

$$w_{i,j} \leftarrow \frac{w_{i,j} + r\ T(x_i, x_j)}{1 + r} \quad (3)$$

The majority of nodes in each network will represent *moderate* agents, with randomly chosen initial opinions which update as described above. The reminder of the vertices will represent *extremists*. Extremists have polarized opinions x_i fixed at one extreme of the spectrum (either 0 or 1), which is equivalent to setting their empathy to 0. They do not updated according to equations (2) and (3).

The use of the Gaussian kernel is reminiscent of the smoothed bounded confidence (SBC) model in [5]. Our model differs by replacing each agent's personal confidence value, with dynamically updated trust values between every pair of agents. This allows agents to remain receptive to some of their neighbours while becoming more skeptical of others, and also for trust to be gradually lost or recovered over time. The notion of equating confidence with persuasiveness is appropriate in a cooperative setting such an expert panel, but seems less suitable in a setting where agents are skeptical in their interactions.

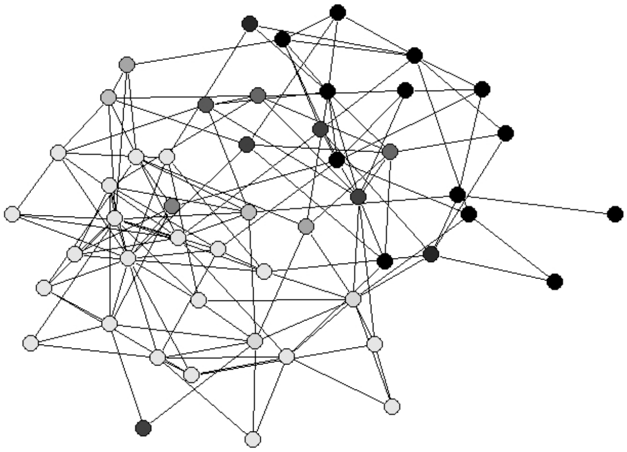

Figure 1: A Erdös-Rényi graph with homophily. Node colors indicate initial opinions, with progression from white (0) to black (1).

3.1 Graph Models

We consider two types of random graph models in our experiments: the classic Barabási-Albert random graph, and a homophily model based on Erdös-Rényi random graphs similar to that presented in [25].

A Barabási-Albert random graph with attachment parameter m is constructed by iteratively adding vertices, connecting them to m existing vertices with probability proportional to their respective degrees. It is often used to model the scale-free property of social networks where a relatively few number of vertices ("hubs") cover most of the edges.

An Erdös-Rényi random graph with connectivity probability p is constructed by considering every pair of vertices i and j, and connecting them with fixed probability p. We incorporate homophily in this model by reweighting the connection probability between i and j as $(1-d)p$, where $d = |x_i - x_j|$. This causes vertices with similar opinions to be joined with higher probability than those with disparate opinions. As with the classic Erdös-Rényi model, the resulting graph may be disconnected. If this is the case, we simply discard and regenerate the graph. A typical modified Erdös-Rényi graph on 50 vertices and $p = 0.2$ is shown in Figure 1; agent opinions were drawn from the distribution $Beta(0.5, 0.5)$.

3.2 A-priori Trust Models

The initial trust between the agents represent how much the agents trust each other prior to the start of the experiment. We utilize three different trust models:

First, we have the *uniform trust* model, where $w_{i,j} = 1, \forall\{i,j\} \in E$. We define $w_{i,i} = d_i$, where d_i is the degree of vertex i, which is consistent with a degree-based voter model where the interactions between an agent and its neighbours are modeled as a series of pairwise interactions. This model makes the fewest assumptions about how trust has been established.

Next, we have the *degree based trust* model, where more initial trust given to the opinions of well-connected ("popular") members of the community: $w_{i,j} = \frac{d_j}{d_i}, \forall\{i,j\} \in E$. Similarly by the logic above, we define $w_{i,i} = 1$.

Finally, we have the *kernel based trust* model. Here, we assume the vertices have interacted previously and their trust value have converged to equilibrium values specified by equation (1); that is, $w_{i,j} = T(x_i, x_j)$ and $w_{i,i} = 1$

4. EMPIRICAL SIMULATIONS

In this section, we describe two sets of experiments that explore the behavior of agents in our model. The first set of experiments operate only on Barabási-Albert random graphs, and aims to explore the ability of extremists to influence the moderate population on typical (i.e. scale-free) social networks.

In the second set of experiments, we explore the ability for extremists at both ends of spectrum to polarize the moderate population, with the ultimate goal of finding necessary conditions for the opinions of the moderates to stratify and stabilize at multiple, non-polarized levels. We introduce the modified Erdös-Rényi random graph model, and the kernel initial trust model in pursuit of this goal.

4.1 Experimental Design

For each experiment, we initialize the social network G with 200 nodes using the appropriate graph model, with varying parameters for graph construction and agent empathy. In the first set of experiments, 10% (20 nodes) of the population is chosen uniformly at random to be 1-extremists; we call this the 1-pole model. In the second set, the population contains 10% 0-extremists and 10% 1-extremists, also chosen uniformly at random; we call this the 2-pole model.

The remainder of the population comprise the moderates. They begin with opinions initialized to random values: either sampled uniformly from the interval $[0, 1)$, or from the partially polarized distribution $Beta(0.5, 0.5)$. Initial trust between them is set according to one of the models outlined in Section 3.2.

Once the instance is initialized, the variables are updated according to equations (1)-(3). We set $r = 1.5$ for all experiments, as preliminary tests did not find varying r changed our qualitative results.

The experiment terminates when no opinions changed by more than a small value ϵ, or a maximum number of iterations t_{max} has been reached. In our experiments, we set $\epsilon = 0.001$ and $t_{max} = 500$; t_{max} was rarely reached in practice. This model was implemented using Python 3.3.2. All results are averaged over 25 replicated trials.

4.2 Influence of Extremists

We begin by investigating the ability of extremists to affect the opinions of the moderates, and how that impact varies with graph structure and parameters of the agents. Figure 2 shows the evolution of opinions over the course of an experiment. To measure this impact, we measure the mean opinion of the moderates at the end of each experiment. If the moderates were completely unaffected by the extremists, the mean would hover around 0.5. If the extremists were completely successful at persuading the moderates, the mean would near 1.0.

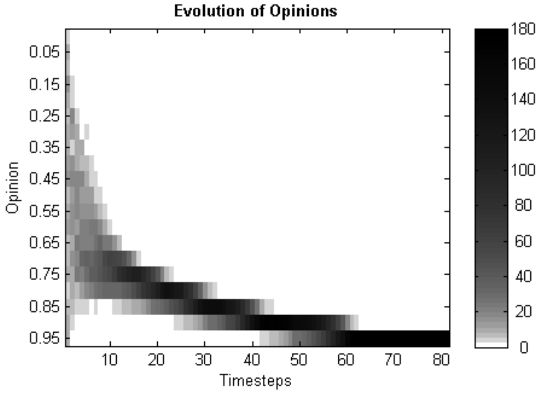

Figure 2: Opinions of moderates over the course of an experiment. Note the color scale is logarithmic.

Figure 3 shows how the average opinion at convergence changes as we adjust the empathy bandwidth parameter h, and the attachment parameter m. As expected, increasing empathy increases the impact of the extremists on the population. However, aside from the special case when $m = 1$, which imposes a tree structure on the network, increasing connectivity does not significantly impact the mean at convergence. This is likely due to the small-world property of these graphs, allowing influence to propagate quickly through the network.

Figure 3 also contrasts the effects of initializing using uniform trust (left) and degree-based trust (right). Adopting initial degree based trust introduces more degrees of freedom in the experiment, in the form of the portion of hubs becoming extremists. This accounts for the higher variability in our results.

One critique of Deffuant's SBC model is its sensitivity to noise [9]. We introduce a similar level of noise to our model, allowing each moderate agent to change their opinion by a small value drawn from a Gaussian distribution, with small probability at each update. More formally, each agent at each iteration has a 0.01 probability of using the following equation in place of equation (2) for their opinion update. Note that the resulting opinion is bounded within $[0, 1]$.

$$x_i \leftarrow \frac{w_{i,i}x_i + \sum\limits_{j\in N(i)} w_{i,j}x_j}{w_{i,i} + \sum\limits_{j\in N(i)} w_{i,j}} + \Delta, \Delta \sim \mathcal{N}(0, 0.15) \qquad (4)$$

Figure 4 shows the effect of introducing this degree of noise into the update process. Aside from the $m = 1$ case, there is little qualitative difference compared to Figure 3. Equation (3) controls the trust dynamics within our network and enables agents to react to sudden deviations in opinions. This gives our model the robustness to absorb noisy signals.

Examining the evolution of opinions in Figure 2, we see that in the initial stage of the experiment, the moderates rapidly converge toward a common opinion. Even agents near the pole are drawn in due to the initial trust conditions. This effect is amplified by the small-worlds property of these graphs. Once an early consensus is reached, the moderate opinion may slowly migrate to the extreme through gradual influence from extremists (as the case in Figure 2), or may successfully insulate the extremists from influencing the general opinion.

4.3 Opinion Polarization

In our second set of experiments, we incorporate two sets of extremists competing for the opinions of the moderate population. We initialize a randomly selected 10% of the population as 1-extremists and another 10% as 0-extremists. Deffuant [5] characterized 4 types of convergence in these two-pole situations: the moderates may converge to a single opinion that is either (I) moderate or (II) polarized, (III) the population may split in two with a portion converging at each pole, or (IV) the population may fragment, with fractions that retain non-extreme opinions.

The polarization of each agent's opinion is the absolute difference of their final opinion from the middle ground of 0.5. The higher the polarization, the more influence felt from the extremists. This allows us to differentiate non-polarized outcomes (types I and IV) from polarized outcomes (types II and III). To detect whether or not moderates have stratified opinions, we examine the final distribution of their opinions to see if they are unimodal (types I and II), or multimodal (types III and IV).

To eliminate false positives due to noise, we use the following procedure for identifying multimodality. First, we form a histogram of opinions, dividing the $[0, 1)$ interval into 20 buckets $b_1, \dots b_{20}$, each of width 0.05. A distribution is multimodal if there exist three buckets b_i, b_j, b_k $(i < j < k)$ such that $b_j < min(b_i, b_k)/2$, and $min(b_i, b_k) \geq T$. We arbitrary choose the threshold $T = 20$, which represents 10% of the agents.

Figure 5 shows the average polarization for our experiments. As before, higher empathy h is correlated with increased influence from extremists. However, now network structure plays a role as well, with impact from extremists being mitigated in more highly connected networks. We examine the final opinions and find that, in all cases with $m > 1$, the moderates converge to a unimodal distribution that drifts toward one of the extremes, reaching a type I or type II convergence. As before, we verify that these results

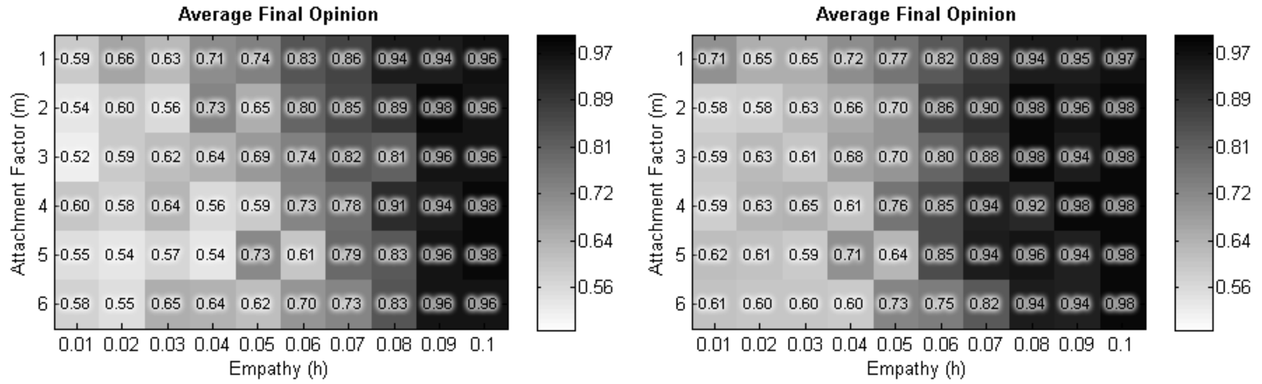

Figure 3: The convergence mean opinion of moderates, in the presence of 10% 1-extremists. The model on the left is initialized using uniform trust (95% confidence interval within ±0.11 for all sets), and the right, using degree based trust (95% C.I. within ±0.10).

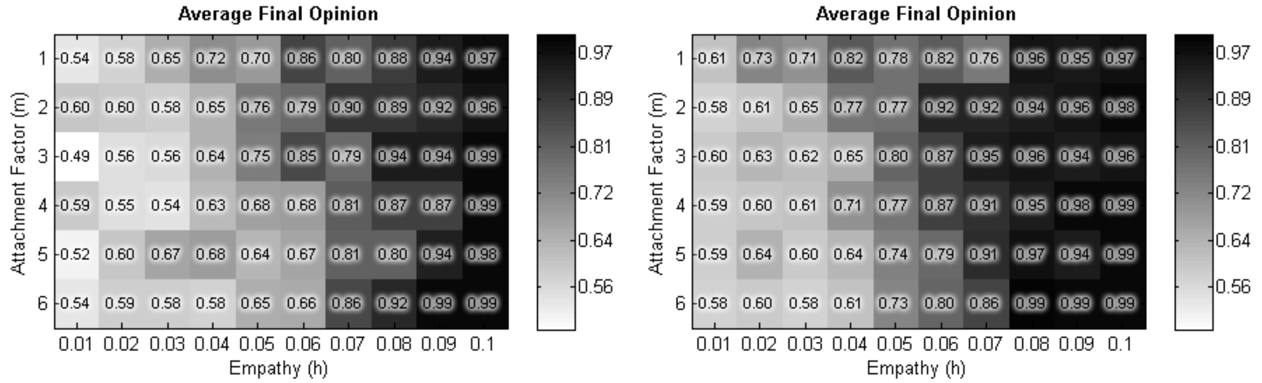

Figure 4: Effects of introducing noise to the model of Figure 3. Uniform trust (left, 95% C.I. within ±0.11) and degree based trust (right, 95% C.I. within ±0.10).

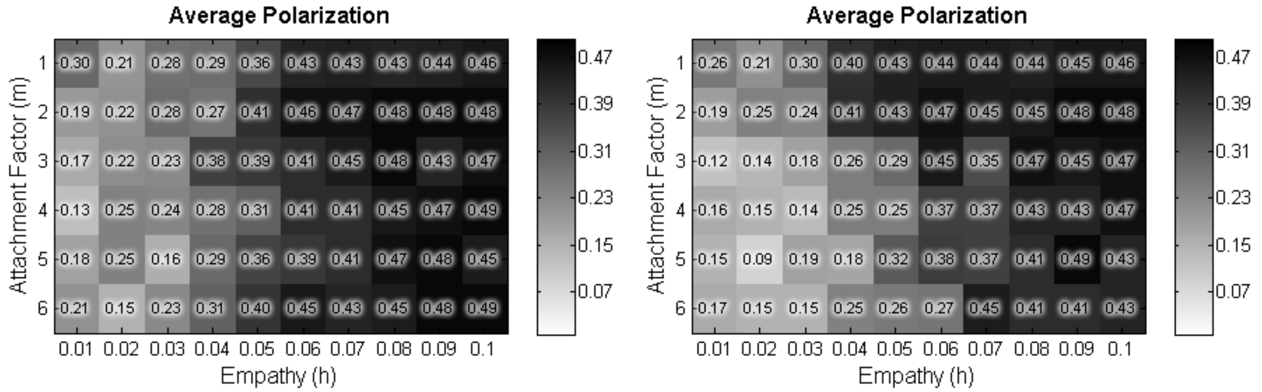

Figure 5: The average polarization of moderates when exposed to extremists of opposing camps. The model on the left is initialized using uniform trust (95% C.I. within ±0.09), and the right, using degree based trust (95% C.I. within ±0.09).

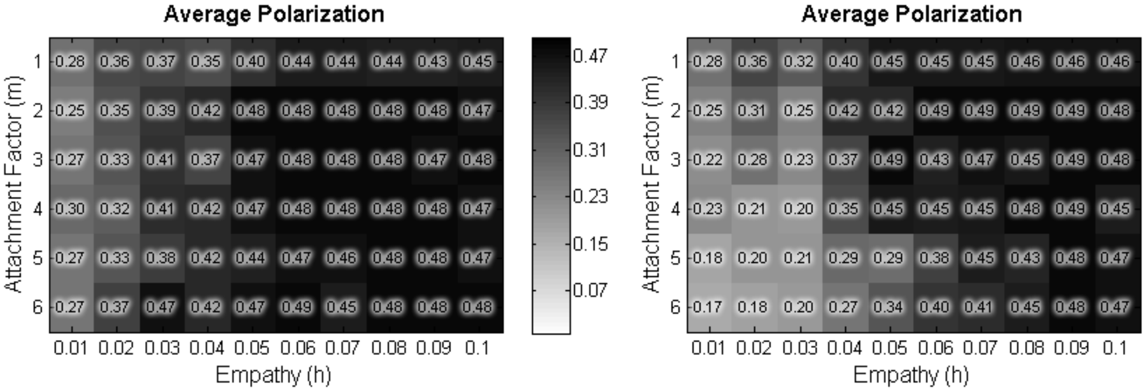

Figure 6: The average polarization of moderates with initial opinions drawn from $\beta(0.5, 0.5)$. The model on the left is initialized using uniform trust (95% C.I. within ± 0.07), and the right, using degree based trust (95% C.I. within ± 0.09).

are robust against noise (data not shown).

One might wonder whether a population that is initially divided can produce type III or type IV convergences. We investigate this possibility by drawing initial opinions x_i from $Beta(0.5, 0.5)$. Figure 7 shows a run under these parameters. We observe a behavior similar to that of Figure 2 – the population converges toward an early consensus, and gradually shift to a unimodal distribution near one of the extremes. This behavior is consistent across all trials, with no multimodal distributions arising when $m > 1$.

Figure 6 shows the average polarization of the general population using the two initial trust models. On the left, we observe that uniform initial trust allows polarization to occur rapidly, regardless of network structure, with nearly complete polarization occurring at $h > 0.04$. This is a surprising result, since in order for moderates to polarize at one extreme, a large portion of the population must be converted from their initial opinions set on the other end of the spectrum. We also observe this trend when the network is initialized using degree-based trust (Figure 6) (right), but it is not as obvious as with uniform initial trust.

Thus, there appear to be two main factors preventing opinions from stratifying. The initial trust given to agents of significantly different opinions, and the lack of homophily in the graph structure. To remedy the first issue, we implement the kernel trust model, where agents are inoculated with skepticism right from the start, modeling a situation where agents have previously interacted and trust dynamics have reached an equilibrium between them. To combat the second issue, we define the modified Erdös-Rényi random graph to capture the homophily property of social networks.

Interestingly, in our simulations of this new model, stratification does not occur until empathy $h > 0.3$, far above the point at which opinions normally become polarized. Figure 8 shows the evolution of opinions in a run that ends in a Type IV convergence. Notice the concentration of opinions migrate gradually from the poles, but do not converge. As shown in Figure 9 (left), the amount of polarization actually *decreases* as empathy increases beyond 0.3. Figure 9 (right) shows the fraction of runs that converge to multimodal distributions. As empathy exceeds 0.3, the likelihood of a type III or type IV outcome increases, and the presence of type IV convergences necessarily lowers the average polarization.

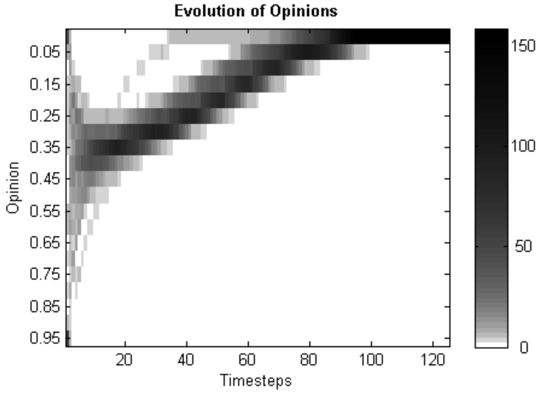

Figure 7: Evolution of opinions in moderates, with partially polarized initial opinions. Note the color scale is logarithmic.

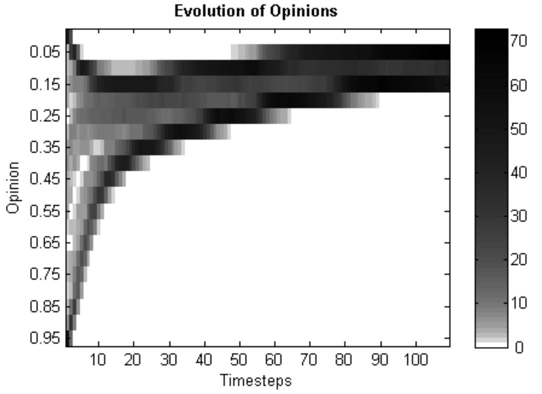

Figure 8: Evolution of opinions in moderates, on a modified ER-graph with homophily, with partially polarized initial opinions. Note the color scale is logarithmic.

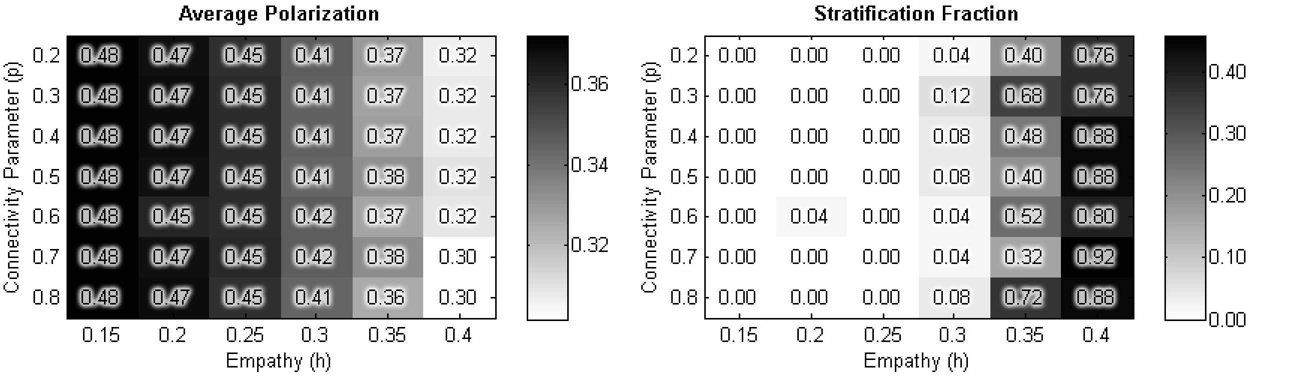

Figure 9: Average polarization of moderates on a modified ER-graph with homophily, with partially polarized initial opinions (left, 95% C.I. within ± 0.03), and the frequency of stratification (right).

The notion that agents with higher empathy, and therefore "listen" to, and are influenced by, a wider range of opinions, is a necessary ingredient for opinions to stratify is very surprising. We hypothesize that this is because agents with such high empathy values are simultaneously affected by extremists from both poles, stabilizing their opinions in a bimodal configuration. Similar stratification is not observed in the Barabási-Albert or the unmodified Erdös-Rényi models, even when employing kernel trust; nor is it observed in the modified Erdös-Rényi model without kernel trust (data not shown).

5. DISCUSSION

One natural question to ask is how the conversion of half the population from one end of the opinion spectrum to the other occurs in Barabási-Albert graphs. The answer may be found by approximating the amount of influence that can be exerted on a densely connected community, even when they have already reached a unified opinion (this is a best case scenario that lower bounds the amount of influence that can be exerted on it). To do this, we extend the concept of *cluster densities* from innovation diffusion. We define a *cluster of density* p as a set of nodes in G such that no node in the cluster has more than fraction p of its neighbours outside the cluster [8].

Now, suppose A is a cluster of density p, $B = G \setminus A$, and all agents in A have opinion x, while all agents in B have opinion $x + \Delta$.

Consider a node i in A with degree d. According to Equation (2), x_i will be updated according to

$$\begin{aligned} x_i \quad \leftarrow \quad & \frac{dx_i + \sum\limits_{j \in N(i)} w_{i,j} x_j}{d + \sum\limits_{j \in N(i)} w_{i,j}} \\ = \quad & \frac{dx_i + \sum\limits_{j \in N(i) \cap A} x_i + \sum\limits_{j \in N(i) \cap B} w_{i,j} x_j}{d + |N(i) \cap A| + \sum\limits_{j \in N(i) \cap B} w_{i,j}} \\ = \quad & \frac{d(1+p)x_i + \sum\limits_{j \in N(i) \cap B} w_{i,j} x_j}{d(1+p) + \sum\limits_{j \in N(i) \cap B} w_{i,j}} \end{aligned}$$

Now if we approximate the weights $w_{i,j}$ with the target trust function $T(x, x + \Delta)$, written as $T(\Delta)$ for brevity,

$$\begin{aligned} = \quad & \frac{d(1+p)x_i + d(1-p)\ T(\Delta)(x_i + \Delta)}{d(1+p) + d(1-p)\ T(\Delta)} \\ = \quad & x_i + \frac{((1-p)\ T(\Delta))\Delta}{(1+p) + (1-p)\ T(\Delta)} \end{aligned}$$

Finally, if we assume $(1+p) >> (1-p)\ T(\Delta)$, then,

$$\begin{aligned} \cong \quad & x_i + \frac{1-p}{1+p} T(\Delta)\Delta \\ = \quad & x_i + \frac{1-p}{1+p} exp(-\frac{\Delta^2}{h})\Delta \qquad (5) \end{aligned}$$

If the right hand side of this expression is bounded within ϵ of x_i, then the simulation will terminate. By comparison, let us modify the above setup by allowing x_i to have a very small fraction p'' of its neighbours that are bridge vertices, with an intermediate opinion $x_i + \Delta/2$. x_i still has fraction p of its neighbours in A, and p' of its neighbours in B with opinion $x_i + \Delta$ $(p + p' + p'' = 1)$. By a similar analysis, approximating the weights $w_{i,j}$ with T yields:

$$\begin{aligned} x_i \quad \leftarrow \quad & \frac{(1+p)x_i + p'\ T(\Delta)(x_i + \Delta) + p''\ T(\frac{\Delta}{2})(x_i + \frac{\Delta}{2})}{(1+p) + p'\ T(\Delta) + p''\ T(\frac{\Delta}{2})} \\ = \quad & x_i + \frac{(p'\ T(\Delta))\Delta + p''\ T(\frac{\Delta}{2}))\frac{\Delta}{2}}{(1+p) + p'\ T(\Delta) + p''\ T(\frac{\Delta}{2})} \end{aligned}$$

And if we assume $(1+p) >> p'\ T(\Delta) + p''\ T(\frac{\Delta}{2})$, then,

$$\begin{aligned} \cong \quad & x_i + \frac{1}{1+p}\left[p'\ T(\Delta)\Delta + p''\ T(\frac{\Delta}{2})\frac{\Delta}{2}\right] \\ = \quad & x_i + \frac{1}{1+p}\left[exp(-\frac{\Delta^2}{h})\Delta + exp(-\frac{\Delta^2}{4h})\frac{\Delta}{2}\right] \qquad (6) \end{aligned}$$

By comparing equation (5) with (6), we see that the amount of influence effected on x_i is greater in the presence of bridge vertices if $T(\Delta) < 2/3$, which is certainly true if we expect the simulation to halt in the bridgeless case.

Thus, when there is a large gulf in opinions, influence is quite limited and skepticism is high. However, the presence of even a handful of unpolarized intermediaries will serve as a siphon through which influence will flow, starting an avalanche effect where the two clusters' opinions begin to converge with increasing speed. This is reminiscent of the "mob mentality" that inspired Deffuant's SBC model.

6. CONCLUSION

We have introduced a robust model of opinion dynamics that captures trust and skepticism between agents that changes over time based on the difference in opinions between the agents. We show that agents operating in a preferential attachment, small-world network will quickly converge to an early, loose consensus before taking coordinated action to migrate the collective opinion to the equilibrium. This equilibrium may be moderate or polar, with agent empathy being the primary factor influencing the final outcome. A secondary factor is connectivity, which has a significant moderating effect, but only in the two-pole model.

We have also modified the Erdös-Rényi graph to incorporate homophily. Only by combining this graph model and inoculating our agents with an equilibrium amount of skepticism for other agents, can we cause opinions to stratify away from extreme values. We hypothesize that this stratification can only exist when individual opinions are stabilized by more extreme opinions from both ends of the spectrum.

Future work include further exploration of the properties of homophilic Erdös-Rényi graphs. The model could be adjusted to include heterogeneous empathy and learning rates within the population. It could also be extended to a dynamic population, with agents entering and leaving the community over time, and with their opinions growing more confident (higher skepticism) as they interact with the community. Finally, our model could be adapted to the discrete action domain, where each agent possesses a private continuous opinion, takes discrete actions based on that private opinion, and only observes the discrete actions of their neighbours.

7. REFERENCES

[1] H. Bless, K. Fiedler, and F. Strack. *Social cognition: How individuals construct social reality.* Hove and New York: Psychology Press, 2004.

[2] A. Carvalho and K. Larson. A consensual linear opinion pool. *23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[3] Y.-S. Cho, G. V. Steeg, and A. Galstyan. Co-evolution of selection and influence in social networks. *Proc. of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, 2011.

[4] J. Coleman, H. Menzel, and E. Katz. *Medical Innovations: A Diffusion Study.* Bobbs Merrill, 1966.

[5] G. Deffuant. Comparing extremism propagation patterns in continuous opinion models. *Journal of Artificial Societies and Social Simulation*, 9(3), 2006.

[6] G. Deffuant, F. Amblard, and G. Weisbuch. Modelling group opinion shift to extreme: the smooth bounded confidence model. *2nd ESSA Conference (Valladolid, Spain)*, 2004.

[7] G. Deffuant, F. Amblard, G. Weisbuch, and T. Faure. How can extremism prevail? a study based on the relative agreement interaction model. *Journal of Artificial Societies and Social Simulation*, 5, 2002.

[8] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World.* Cambridge University Press, 2010.

[9] B. Edmonds. Assessing the safety of (numerical) representation in social simulation. *3rd European Social Simulation Association conference (ESSA)*, 2005.

[10] H. Fang, J. Zhang, and N. M. Thalmann. A trust model stemmed from diffusion theory for opinion evaluation. *12th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2013.

[11] D. W. Franks, J. Noble, P. Kaufmann, and S. Stagl. Extremism propagation in social networks with hubs. *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, 16(4), 2008.

[12] N. Hazon, R. Lin, and S. Kraus. How to change a group's collective decision. *23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[13] R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[14] E. T. Jaynes. *Probability Theory: the Logic of Science.* Cambridge University Press, 2003.

[15] U. Krause. Soziale dynamiken mit vielen interakteuren. eine problemskizze. *Modellierung und Simulation von Dynamiken mit vielen interagierenden Akteuren*, 1997.

[16] U. Krause. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in Difference Equations*, 2000.

[17] Z. Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990.

[18] P. S. Laplace. *Essai philosophique sur les probabilités.* Paris Bachelier, 1840.

[19] A. C. R. Martins. Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics*, 19, 2008.

[20] H. V. Parunak, T. C. Belding, R. Hilscher, and S. Brueckner. Modeling an managing collective cognitive convergence. *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2008.

[21] O. Pryymak, A. Rogers, and N. R. Jennings. Efficient opinion sharing in large decentralised teams. *11th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2012.

[22] L. Salzarulo. A continuous opinion dynamics model based on the principle of meta-contrast. *Journal of Artificial Societies and Social Simulation*, 9, 2006.

[23] D. Strang and S. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24:265–290, 1998.

[24] S. Swarup, A. Apolloni, and Z. Fagyal. A model of norm emergence and innovation in language change. *10th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011.

[25] L. H. Wong, P. Pattison, and G. Robins. A spatial model for social networks. *Physica A: Statistical Mechanics and its Applications*, 360, 2006.