Predictive Modeling of Opinion and Connectivity Dynamics in Social Networks

Ajay Saini

Abstract

Social networks have been extensively studied in recent years with the aim of understanding how the connectivity of different societies and their subgroups influences the spread of innovations and opinions through human networks. Using data collected from real-world social networks, researchers are able to gain a better understanding of the dynamics of such networks and subsequently model the changes that occur in these networks over time. In our work, we use data from the Social Evolution dataset of the MIT Human Dynamics Lab to develop a data-driven model capable of predicting the trends and long term changes observed in a real-world social network. We demonstrate the effectiveness of the model by predicting changes in both opinion spread and connectivity that reflect the changes observed in our dataset. After validating the model, we use it to understand how different types of social networks behave over time by varying the conditions governing the change of opinions and connectivity. We conclude with a study of opinion propagation under different conditions in which we use the structure and opinion distribution of various networks to identify sets of agents capable of propagating their opinion throughout an entire network. Our results demonstrate the effectiveness of the proposed modeling approach in predicting the future state of social networks and provide further insight into the dynamics of interactions between agents in real-world social networks.

1 Introduction

The goal of this research is to develop a comprehensive model for the dynamics of social networks that reflects the changes observed in a real social network. Social networks are modeled as graphical structures of connected agents along with their opinions. The agents of a network interact with each other stochastically and change both their opinions and connections. Based on certain rules for interactions between agents, the change in opinions and connections leads to long term opinion propagation in the network as well as other notable changes such as the formation of clusters, or close-knit groups of agents. In order to predict how social networks change over time, we formulate a model that governs the dynamics of interactions between agents and use the model to simulate such interactions over a long period of time. Using our simulations results, we uncover the primary forces driving changes in social networks such as the propagation of opinions and the formation of clusters.

The literature dedicated to modeling social networks well represents the various dynamics used to simulate agent interaction in networks of theoretical structures [1, 2, 3, 5, 13]. However, such models are limited in their practical application because it has not been verified that they reflect the dynamics of real-world social networks. Also, much of the literature implements models that either focus solely on the update of opinions [1, 3, 5, 13] or solely on the update of connections [6, 8, 10]. However, in real-world networks it is often the case that both opinions and connections of agents change over time. In our work, we implement a data-driven modeling approach [6, 8, 10, 14, 18] in which we base our model on observations from a dataset in order to accurately model the dynamics of a real social network. Using the data, we show that social networks exhibit both opinion and connection change over time and accordingly develop a model that accounts for both [2].

The changing of connections over time in social networks gives rise to another prominent property of social networks: the formation of clusters, or close-knit groups of agents [6, 8, 10]. Several studies have accounted for clustering in social networks as a result of varied strength of social ties [2, 10]. In our work, we develop a model that follows this trend by using clusters as a measure of the strengths of connections between agents.

In order to understand how the opinion and connectivity of a real social network changes over time, we use data from the Social Evolution dataset of the MIT Human Dynamics Lab [16].

Using observations from the data along with social network theory, we formulate an agent-based model for the dynamics of the social network studied in the data [4, 10]. We then validate the model by comparing its prediction of the changes in opinions and connectivity of agents in the network with the changes observed in our dataset. We also perform a comprehensive evaluation of the model's parameters in which we analyze the long term opinion and connectivity dynamics of social networks representing different types of societies defined by the model's parameters. We conclude by introducing an optimization approach for opinion spread in the network which involves understanding how to change the opinion of the greatest proportion of agents in a network in a desired direction.

This work allows us to not only model and predict the long term changes that will occur in the networks defined by available data, but also to gain further understanding of how various theoretical networks change over time. The model, along with the theories and implications behind it, provides insight into the dynamics of interactions between agents as well as a set of tools with which one can investigate opinion propagation strategies for networks of various structures and opinion distributions. Such insights have applications in a diverse array of fields that involve manipulation of agents in social networks such as marketing [9], rumor propagation [12], and panic spread prevention [14].

The paper is organized as follows. In section 2 we detail the set of observations on which we base our model. In section 3 we formulate an agent-based model based off of the observations in section 2. In section 4, we validate the model by comparing its prediction to the data. In section 5, we evaluate the model's parameters in order to understand how opinion and connectivity change over time under different sets of dynamics for agent interaction. In section 6, we introduce the opinion propagation problem and detail our findings regarding maximal opinion propagation in social networks. In section 7, we discuss the overall conclusions of our work, and in section 8, we provide recommendations for further work. Lastly, in section 9, we state acknowledgements.

2 Data Analysis

2.1 The Available Data

We use the Social Evolution Dataset of the MIT Human Dynamics Lab [16] which details a study of students living in a Harvard dormitory. The students were surveyed at various times over the course of eight months and asked to self-report their opinions on several subjects as well as to indicate their close friends in the dormitory. The survey times we use are: September 2008, October 2008, December 2008, March 2009, and April 2009; which are denoted as: 2008.09, 2008.10, 2008.12, 2009.03, and 2009.04 respectively. From the surveys, we use each student's opinion of his or her own health and self-reported set of close friends.

Each student's opinion of his or her own health can take one of five values: 2 (healthy), 1 (average), 0 (below average), -1 (unhealthy), and -2 (very unhealthy).

Figure 1 shows the change in both opinion spread and connectivity between the first and last survey times of the data[1]. The connectivity in the graphs corresponds to the friendships between agents in the network as indicated in the survey. The edges are directed because in the data, friendship is not necessarily reciprocal. For example, if agent A considers agent B a friend, agent B does not necessarily consider agent A a friend. The colors in the graphs correspond to the health opinions of the agents.

Note that agents tend to change their opinions and friendships over time and form distinct clusters in the network. Our observation of clusters in the network supports previous studies that agents in social networks exhibit a natural tendency to form clusters [6, 10, 8].

2.2 Data Observations

In this section, we formulate observations from the data that serve as a basis for our modeling approach.

Observation 1: The opinions tend to move away from the neutral opinion value of 0 (yellow in Figure 1), and towards the more radical opinion values of -1, 1, and 2. This is demonstrated by the fact that the number of yellow nodes decreases from 20 in 2008.09 to 7 in 2009.04 and the

[1] When analyzing the data, we noticed that one agent's (agent 28 in the data) opinion and friendships were significant outliers. For this reason, we removed the agent in our analysis.

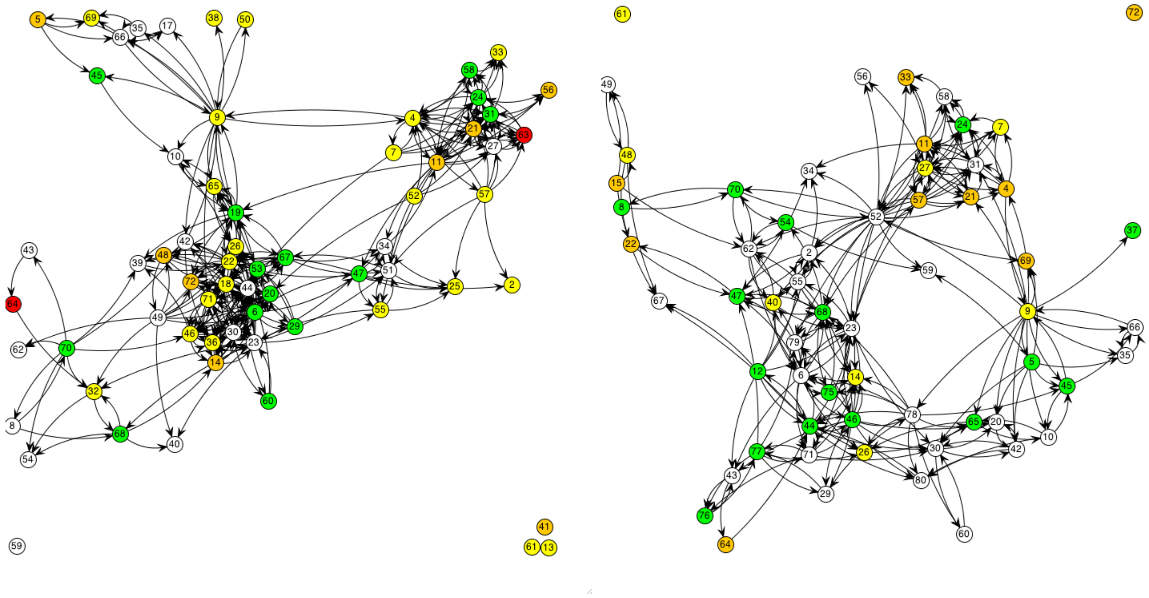

Figure 1: The network for the 2008.09 survey (left) and the 2009.04 survey (right). Colors represent opinion: healthy (green), average (white), below average (yellow), unhealthy (orange), very unhealthy (red). Edges represent friendships between nodes. Note the prominent changes in opinion distribution and connectivity of the network over time. Our goal in this research is to formulate a model capable of predicting these changes. Visuals created with the assistance of code from the JUNG library. [17]

number of orange (opinion $= -1$), white (opinion $= 1$), and green (opinion $= 2$), nodes increases from 43 in 2008.09 to 53 in 2009.04.

Our further data analysis involves analyzing trends in the average opinion and connectivity of the clusters in the network. However, to do so we must first define a method for clustering the network. We analytically determine clusters in the network[2] using a variant of the algorithm proposed by Girvan and Newman [7] as implemented by the JUNG library [17]. We determine the optimal set of clusters based on a metric proposed by Schaeffer which computes the quality of clusters based off of the connectivity of the nodes within the cluster and the the connectivity of the cluster to the rest of the network [15, pg. 38]. A high quality cluster is defined as a cluster whose nodes share many connections with each other and share few connections with the rest of the network [15, pg. 38]. In further analysis, we only use the optimal set of clusters[3].

In order to study how the clusters in the network change over time, we introduce several metrics: cluster opinion, opinion spread, and inner connectivity, which are defined as follows.

Let the nodes in cluster C of size k be $c_1, c_2, \ldots, c_k$ and have opinions $o(c_1), o(c_2), \ldots, o(c_k)$

[2] When clustering the network, we use an undirected version of the graph generated as follows. If i is friends with j and j is friends with i, i and j are connected by an undirected edge. Otherwise, there is no edge between i and j. This is done in order to increase the performance of the clustering algorithm and produce more optimal clusters as defined by our metric.

respectively.

Cluster Opinion: We define cluster opinion, $O(C)$, as the average opinion of the nodes in the cluster:

$$O(C) = \frac{\sum_{1 \leq i \leq k} \; o(c_i)}{k}. \tag{1}$$

Opinion Spread: We define opinion spread, $S(C)$, as a measure of how far, on average, the opinions of the nodes in the cluster are from the cluster opinion:

$$S(C) = \frac{\sum_{1 \leq i \leq k} \; |o(c_i) - O(C)|}{k}. \tag{2}$$

Inner Connectivity: We define the inner connectivity, $I(C)$, of a cluster as the number of edges in the cluster divided by the number of possible edges:

$$I(C) = \frac{|E|}{k(k-1)}. \tag{3}$$

Where $|E|$ is the number of directed edges in the cluster.

We cluster the network at each of the five survey times. For each survey time, we compute the average value of the proposed cluster metrics. The results are depicted in Table 1.

	2008.09	2008.10	2008.12	2009.03	2009.04
Cluster Opinion	.57	.67	.68	1.03	1.02
Opinion Spread	.87	.77	.66	.59	.52
Inner Connectivity	.48	.60	.63	.65	.66

Table 1: The cluster metrics for each survey time. At each time, we clustered the network and computed the value of the cluster metrics for each cluster. The values in the table are the averages of each metric over all clusters for each time set.

We observe a monotonic increase in both cluster opinion and inner connectivity over time and a monotonic decrease in opinion spread over time. In the following observations, we confirm these trends with a one-tailed linear regression t-test. The t-test uses an α-level of .05 with time as the independent variable and the cluster metric as the dependent variable.

Observation 2: The average cluster opinion significantly increases over time. The linear regression t-test gave a p-value of .00402.

Observation 3: The average opinion spread of the clusters decreases over time. The linear regression t-test gave a p-value of .00176.

Observation 4: The inner connectivity increases over time while the number of clusters decreases over time. The linear regression t-test gave a p-value of .03986.

3 Agent-Based Model of the Dynamic Network

Using observations from data, we develop an agent-based model [4, 10] capable of predicting the trends in both cluster opinion and connectivity observed in subsection 2.2. The model implements the changing of opinions and connections in a discretized way, where the state of the network changes from time t to time $t+1$. In this section we formally define the structure of the network, the rules governing the change of opinions, and the rules governing the change of connections (friendships between agents) that were suggested by data observation. In the later sections we will present a validation of the model using available data.

3.1 Structure of the Network

Define the network at time t as a directed graph $G(N_t, A_t)$ with set of nodes N_t and adjacency matrix ${A_t}^4$. Each node i in the network represents an agent having an opinion $o(i)_t$ at time t. The adjacency matrix A_t defines the edges in the graph (friendships between agents).

Studies have demonstrated that social ties often have unequal strength [2, 10]. To define the strengths of connections between nodes[5], we use the existence of clusters so that connections between nodes within a cluster are stronger than those between nodes from different clusters. Using this concept, we define A_t as

$$a(i,j)_t = \begin{cases} 0 & \text{if no edge connects node } i \text{ to node } j, \\ 1 & \text{if } i \text{ is connected to } j \text{ from a different cluster,} \\ w\ (w \geq 1) & \text{if } i \text{ is connected to } j \text{ from the same cluster,} \end{cases} \tag{4}$$

where $a(i,j)_t$ denotes row i, column j of A_t. The parameter w represents how much stronger

[3] We define a cluster as a set of nodes with size greater than 2. Throughout the paper, only clusters satisfying this property are considered in our modeling and analysis.

[4] Throughout the paper, we use the subscript t to denote a state at time t.

[5] Throughout the paper, we use the terms agent and node interchangeably.

friendships are between agents in the same cluster versus between agents in different clusters.

3.2 Opinion Update

When defining rules governing the update of opinions from time t to time $t+1$, we make the following assumptions from examples in existing literature:

1. If agent i considers agent j a friend, then agent j can change agent i's opinion but agent i cannot change agent j's opinion [3, 5].

2. When agent j influences agent i, agent i's opinion changes by some fraction of the difference between its opinion and j's opinion [3].

3. The degree to which agent i is influenced by agent j depends on how strongly agent i considers agent j a friend [3, 5].

Each node i's opinion at time $t+1$ is a function of its opinion and its friends' opinions at time t and is defined as

$$o(i)_{t+1} = s(o(i)_t + q(i, N_t)_t), \tag{5}$$

where $q(i, N_t)_t$ is the opinion change caused by node i's friends and $s(x)$ is the result of opinion amplification, used to account for data Observation 1 in subsection 2.2.

By Assumption 2, we define $q(i, N_t)_t$ as

$$q(i, N_t)_t = \sum_{j \in N_t} (\alpha)(\Delta x(i, j, N_t)_t)(o(j)_t - o(i)_t), \tag{6}$$

where $\Delta x(i, j, N_t)_t$ is the degree to which node j changes node i's opinion and $\alpha \in (0, 1]$ is a scale factor introduced to reduce the degree to which node j changes node i's opinion thereby slowing opinion change and implementing a slow learning model [13, Part II]. By Assumption 3, we define $\Delta x(i, j, N_t)_t$ using edge weights, which are indicative of how strongly node i considers node j a friend. We let $\Delta x(i, j, N_t)_t$ be the weight that node j exerts on node i divided by the sum of the weights exerted by all of i's friends. Formally,

$$\Delta x(i, j, N_t)_t = \frac{a(i, j)_t}{\sum_{1 \le k \le |N_t|} a(i, k)_t}.$$

Note that if j is not a friend of i, $\Delta x(i, j, N_t)_t = 0$, meaning that node j has no influence on node

i's opinion in accordance with Assumption 1.

In Observation 1 of subsection 2.2, we observed that opinions in the network tend to radicalize by moving toward more extreme values such as -1 and 2. In order to radicalize node i's opinion, we introduce an **opinion amplification function** $s(x)$ defined as

$$s(x) = \begin{cases} x & \text{if } x \in [-2,-1] \cup [0,1.5], \\ kx \ (k \geq 1) & \text{if } x \in (-1,0) \cup (1.5,2), \end{cases} \tag{7}$$

where k is a parameter representing the extent to which agents tend to radicalize their opinions. The bounds on the cases are empirically determined based off the data. We will show that these bounds are accurate in modeling the data in section 4

Using the definitions of $q(i,N_t)_t$ and $s(x)$ defined in (6) and (7) respectively, we update the opinion of every node i in the network using the rule in (5).

3.3 Connection Update

When defining the rules governing the update of connections from time t to time $t+1$, we make the following four assumptions from examples in existing literature:

1. Agent i can form friendships with friends of its friends (agents k such that for some j, $a(i,j) > 0$ and $a(j,k) > 0$) [8, 10]

2. Agent i can form friendships with agents that already consider it a friend (agents j such that $a(j,i) > 0$) [6, 8, 18].

3. The stronger the potential connection between two disconnected agents (the higher the value of $a(i,j)$ or $a(j,i)$ would be if disconnected agents i and j were to form a connection), the more likely the connection is to form [2, 10].

4. The stronger the connection between two connected agents (the higher the value of $a(i,j)$ if agent i is connected to agent j), the less likely the connection is to break [11].

In the proposed model, after the update of opinions occurs at time t, each node will probabilistically form one connection and probabilistically break one connection.

For each node i in the network, we use Assumptions 1 and 2 to randomly select one agent j with which node i can potentially form a connection and one agent k (such that $a(i,k)_t > 0$) with which node i can potentially break a connection. Let set S_f contain all ordered pairs of nodes (m_f, n_f) that participate in probabilistic connection forming and let set S_b contain all ordered pairs of nodes (m_b, n_b) that participate in probabilistic connection breaking. Note that $S_f \cap S_b = \emptyset$.

For all ordered node pairs (i, j) such that $i, j \in N_t$, we define probabilities $f(i, j)_t$, the probability that a directed connection from i to j forms, and $b(i, j)_t$, the probability that a directed connection from i to j breaks as

$$f(i,j)_t = \begin{cases} \beta p(i,j)_t & \text{if } (i,j) \in S_f, \\ 0 & \text{if } (i,j) \notin S_f, \end{cases} \quad \text{and } b(i,j)_t = \begin{cases} \beta(1 - p(i,k)_t) & \text{if } (i,j) \in S_b. \\ 0 & \text{if } (i,j) \notin S_b, \end{cases}$$

respectively. The scalar $\beta \in (0, 1]$ is used to scale down the probability of connection change in order to slow the rate of connection change over time similar to the way we use α in subsection 3.2 to slow the rate of opinion change over time.

The function $p(x, y)_t$ is used to determine the probabilities of connections from a node x to a node y forming and breaking and is defined as

$$p(x,y)_t = \begin{cases} (.5 - c) & \text{if } x \text{ and } y \text{ are not in the same cluster,} \\ (.5 + c) & \text{if } x \text{ and } y \text{ are in the same cluster,} \\ (.5) & \text{if } x \text{ is in not in a cluster,} \end{cases}$$

where the parameter $c \in [0, .5)$ represents the extent to which agents restrict their friendships to those within their own cluster.

We update the adjacency matrix from A_t to A_{t+1} via an algorithm in which we implement the following cases[6] for all ordered node pairs (i, j) such that $i, j \in N_t$:

1. If $f(i, j)_t = 0$ and $b(i, j)_t = 0$, then $a(i, j)_{t+1} = a(i, j)_t$ (there is no edge change).

2. If $f(i, j)_t \neq 0$, then $a(i, j)_{t+1} = 1$ (an edge forms) with probability $f(i, j)_t$ and $a(i, j)_{t+1} = 0$ (an edge fails to form) with probability $1 - f(i, j)_t$.

3. If $b(i, j)_t \neq 0$, then $a(i, j)_{t+1} = 0$ (an edge breaks) with probability $b(i, j)_t$ and $a(i, j)_{t+1} = 1$ (an edge fails to break) with probability $1 - b(i, j)_t$.

[6] Note that we omit the case $f(i, j)_t \neq 0$ and $b(i, j)_t \neq 0$ because by our definitions of $f(i, j)_t$ and $b(i, j)_t$, such a case is impossible.

After probabilistic connection update, A_{t+1} contains information about the new set of edges at time $t + 1$. The network is then re-clustered using the algorithm in subsection 2.2 and A_{t+1} is updated again using its cluster-based definition described in (4).

4 Modeling the Data

The model has three parameters: w (tendency to follow the opinion of those from the same cluster versus those from a different cluster), k (tendency to radicalize opinion), and c (tendency to form more friendships with those from the same cluster versus with those from a different cluster), along with two scalars, α and β, which regulate the rate of opinion and connection change respectively. Using the network from the first survey time in our data (2008.09) as an initial condition ($t = 0$), we identified values for the three parameters w, k, and c and two scalars α and β in the model that best describe the social network dynamics of the data.

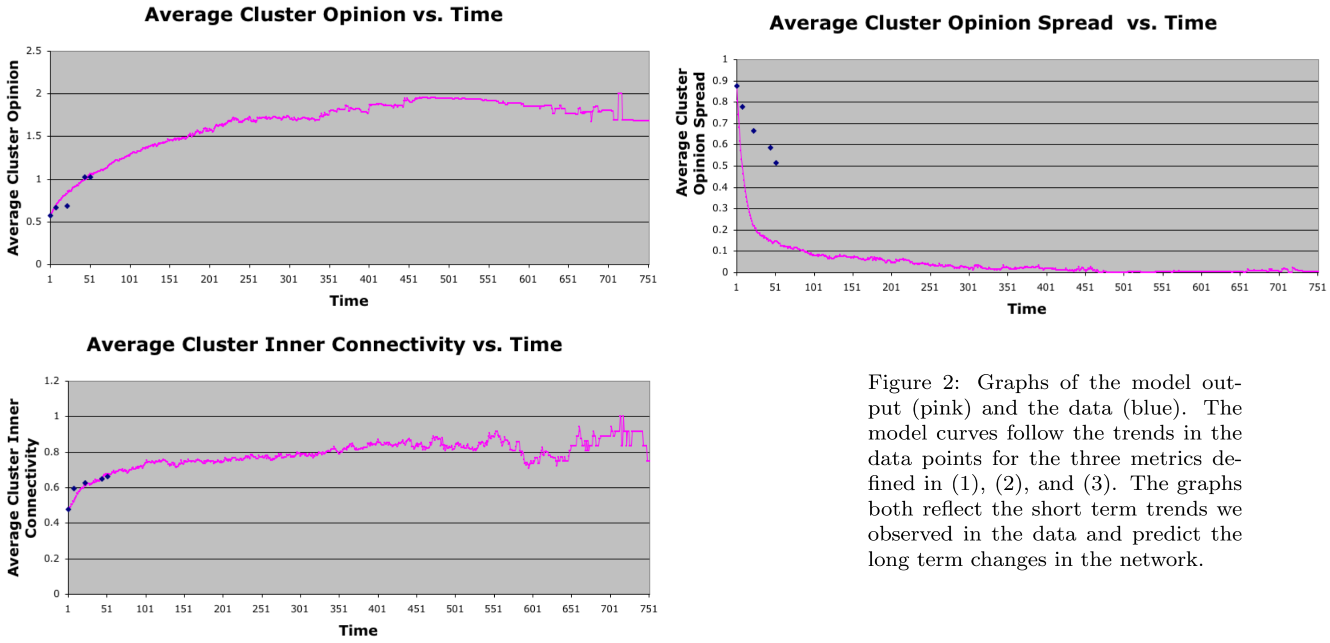

Figure 2: Graphs of the model output (pink) and the data (blue). The model curves follow the trends in the data points for the three metrics defined in (1), (2), and (3). The graphs both reflect the short term trends we observed in the data and predict the long term changes in the network.

Figure 2 shows the results of simulations[7] using the model (the pink curves) with parameter values $w = 5$, $k = 1.05$, $c = .245$ and scalar values $\alpha = .10$, $\beta = .15$. These values were determined to yield the best fit to the data (shown as the blue points along the curves).

[7] To generate the graphs, we ran 50 simulations averaged the curves generated by each simulation at each time t to produce the curves shown throughout this section and throughout section 5.

In order to determine the accuracy of the model, we average the percent errors[8] at each data point after $t = 0$. Cluster opinion had an average percent error of 7.07% and inner connectivity had an average percent error of 5.22%, thus indicating that the model is reasonably accurate in predicting the trends in these two metrics. As shown in Figure 2, opinion spread generally followed the observed trend, though the percent error was significantly larger.

Also, note that after $t = 400$, cluster opinion, inner connectivity, and opinion spread stabilize around certain values. As time increases, the average cluster opinion spread approaches 0, which suggests a convergence to the same opinion among the nodes in each cluster. The idea that the opinion of the network stabilizes around a given value is critical in our analysis regarding how to use the model to optimize long term opinion propagation in a network. In section 6, we will use this idea to study how opinions converge in networks of various structures and initial opinion distributions.

From the graphs, we can see that around $t = 50$, the average opinion in the model is close to 1 and the average inner connectivity is close to .65. These are the approximately the same average cluster opinion and inner connectivity values as seen in 2009.04, the last survey time in the data.

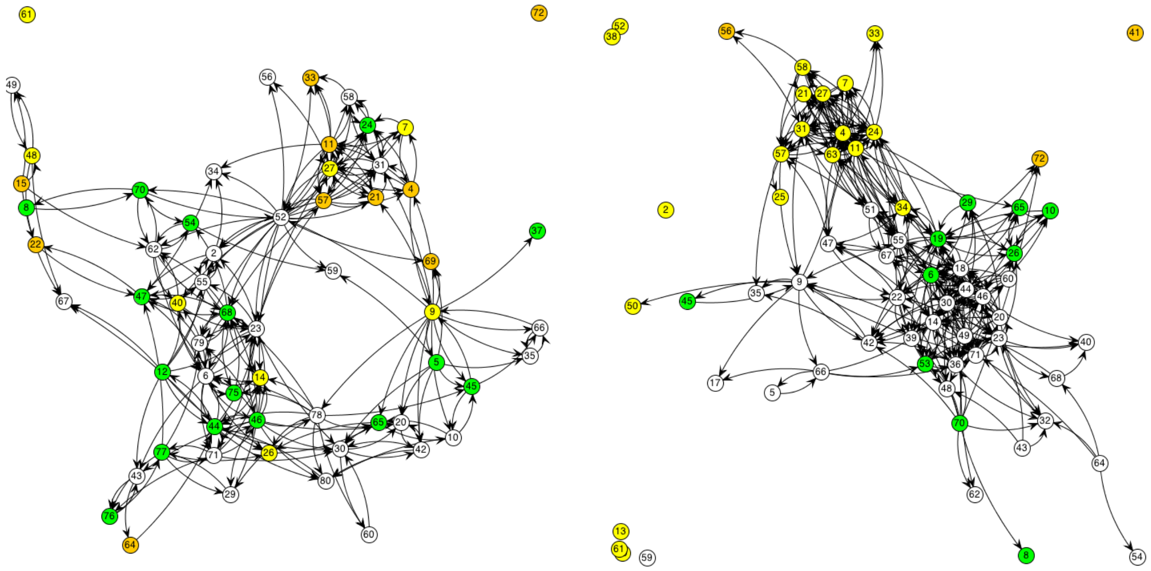

Figure 3: The data at 2009.04 (left) and the model at $t = 50$ (right). Note the similarity between the two networks in the existence of a larger cluster that is mostly green (opinion value 2) and white (opinion value 1) and a smaller cluster that is mostly yellow (opinion value 0) and orange (opinion value -1).Visuals created with assistance of code from the JUNG library. [17]

[8] percent error = $\frac{|\text{model value}-\text{data value}|}{\text{data value}} * 100\%$

Figure 3 further demonstrates the similarity between the network at $t = 50$ and the last survey time of the data.

5 Parameter Dependent Model Variability

The parameters in the model represent a number of aspects of society such as tendency to follow the opinion of those from the same cluster versus those from a different cluster (parameter w), tendency to radicalize opinion (parameter k), and tendency to form more friendships with those from the same cluster versus with those from a different cluster (parameter c). In this section, we vary the parameters in order to gain an understanding of how the network dynamics vary under different sets of conditions governing interactions between agents.[9]

We use the 2008.09 network as the condition for the model at time $t = 0$. Each time we vary a parameter, we hold the other two parameters constant at their data-confirmed values in order to avoid the confounding effects of changing multiple parameter values.

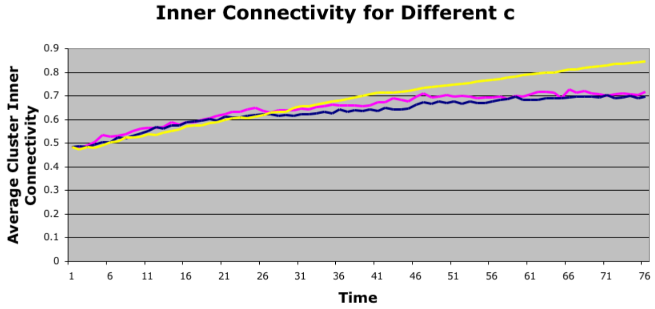

Figure 4: Average cluster inner connectivity for $c = 0$ (pink), $c = .245$ (blue), and $c = .49$ with $w = 5$ and $k = 1.05$ held constant. Note that the curves for $c = 0$ and $c = .245$ level off long term while the curve for $c = .49$ tends to increase linearly.

We vary the parameter c as shown in Figure 4 and find that $c = .49$ yields the greatest increase in inner connectivity. This is due to the fact that $c = .49$ causes agents to have a very high tendency to increase their number of friends within their own clusters, thus increasing the inner connectivity of the clusters.

When varying w as shown in Figure 5, we observe that for values of t less than 450, the average cluster opinion curves for $w = 5$ and $w = 100$ are essentially the same while the curve for $w = 1$ noticeably differs. Since high values of w indicate that agents allow those within their cluster to influence them to a greater degree than those outside their cluster, our results suggest that the

[9] Note however, that we do not consider the constants α and β parameters. Instead, they are data driven constants used to control the speed at which opinion and connection update respectively occur for which the values $\alpha = .10$ and $\beta = .15$ were confirmed in section 4.

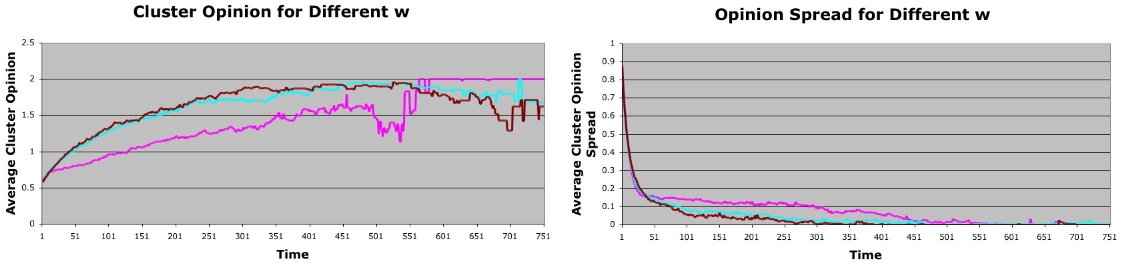

Figure 5: Average cluster opinion (left) and average cluster opinion spread (right) for $w = 1$ (pink), $w = 5$ (blue), and $w = 100$ (brown). We hold $c = .245$ and $k = 1.05$ constant. Note that as w increases to high values, there is a negligible difference in the long term average cluster opinion, thus suggesting that weighted friendships only affect long term opinion propagation up to a certain point.

degree to which agents follow the opinions of friends within their own cluster only impacts the average cluster opinion up to a certain point. Beyond that point, no matter how much more agents are influenced by those within their cluster over those outside their cluster, the average cluster opinion will not be significantly affected. We observe the same for the average cluster opinion spread over time as w increases.

Furthermore, note that the deviation created when $w = 1$ is one that signifies a slower rate of change in both cluster opinion and opinion spread, though the difference is much more significant in cluster opinion. To explain this, we note that when $w = 1$, all agents in the network have an equal influence on any given agent's opinion, regardless of the clusters they belong to. In both Figure 1 and Table 1 in subsection 2.2, we found that initially, the average cluster opinion of the network is positive and that there are significantly more agents with initial opinion ≥ 0 than those with initial opinion < 0. As a result, if $w > 1$, then the agents with positive initial opinions will have a much greater influence on the network as a whole than those with negative opinions. This is because the greater number of agents with a positive initial opinion means that they will share a cluster with a greater number of agents than those with a negative initial opinion, thus allowing the agents with a positive initial opinion to exert their influence, which is increased by the parameter w, on a greater number of agents and cause a rapid increase in average cluster opinion. However, if $w = 1$, then the agents with a positive initial opinion do not experience an increase in influence on other agent's opinions. Instead, all agents exert an equal influence on each other. Therefore, although the agents with a positive initial opinion still have a greater influence on the network as a whole than the agents with a negative initial opinion because of their greater number, the difference is not as great due to the lack of increase in influence based off of clusters by the parameter w. As a result, the average cluster opinion still increases, but at a much slower rate.

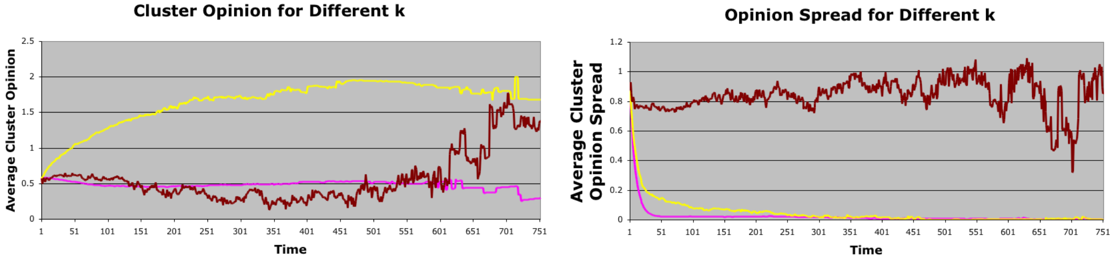

Figure 6: Average cluster opinion (left) and average cluster opinion spread (right) for $k = 1$ (pink), $k = 1.05$ (yellow), and $k = 2.00$ (brown) while holding $c = .245$ and $w = 5$ constant. Note the long term instability of the curves for $k = 2.00$ which results from very large opinion amplification.

As shown in Figure 6, as k increases from 1 to 1.05, the average cluster opinion and average opinion spread curves increase as well. However, note that the curves for $k = 2.00$ are very unstable. To explain the anomaly caused by $k = 2.00$, we note that the amplification function defined in (7) dramatically amplifies both positive and negative opinions for high values of k. At small values of t, when the network is close to its initial condition, there is a mix of both positive and negative opinions in the network. Therefore, as a result of interaction with their neighbors, the opinions of nodes will be pulled towards the positive direction at some times and towards the negative direction at other times. This pulling of opinions in both positive and negative directions is greatly amplified by the large value of k. As noted in Observation 1 of subsection 2.2, although the network contains both positive and negative opinions at $t = 0$, the overall opinion is fairly neutral (close to 0). Therefore, the oscillations in both cluster opinion and opinion spread we observe for small values of t are still fairly small despite the large value of k. However, as t increases and the opinions become increasingly radicalized over time due to the large value of k, the oscillations in both cluster opinion and connectivity increase as well, resulting in the dramatic fluctuations seen in Figure 6.

6 Opinion Propagation through Various Networks

One useful application of social network science is in understanding how to maximize the propagation of an opinion throughout a network [1, 3, 5, 9]. Since our data involves a study of the health of agents over time, we will study how to propagate good health throughout a network so that very unhealthy agents with a health opinion of -2 will become healthy through interaction with healthy agents which have a health opinion of 2. Starting with networks that have a homogeneous opinion[10] of -2 (very unhealthy), we change some proportion of the nodes to an opinion of 2 (healthy) at time $t = 0$ [1, 5]. From the graph of Opinion Spread vs. Time presented in section 4, we know that the opinions of nodes in the network tend to converge to a given value over a long period of time. Our goal in the following experiments is to gain an understanding of what proportion of nodes to change and which nodes to change based off the structure of the network so that the nodes changed to an opinion of 2 will cause the entire network to converge to an opinion of 2, an optimal opinion value, over a long period of time. The following three experiments will demonstrate our approach to this problem and the subsequent results. In order to reduce variability and the confounding effect of multiple parameter values in the results, our experiments only use the parameters values that accurately reflected the data ($w = 5$, $k = 1.05$, $c = .245$).

6.1 Experiment 1

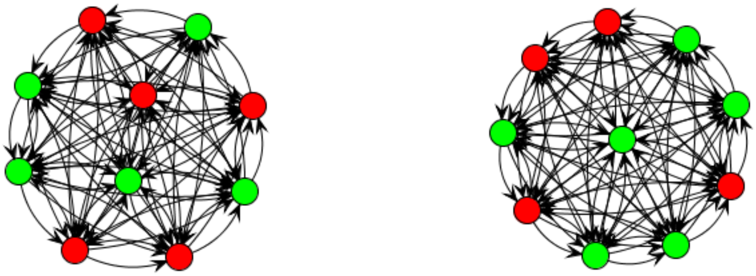

Figure 7: A complete network of size 10 with 50% of nodes changed to an opinion of 2 (left) and 60% of nodes changed to an opinion of 2 (right). These are sample initial conditions for the simulation in Experiment 1. In Experiment 1, we varied the size of the network, using networks of sized 10, 50, and 100. For each network size, we changed 50% of the nodes to an opinion of 2 and 60% of the nodes to an opinion of 2. The results of simulations with the model are shown in Figure 8.

In the first experiment, we analyze complete networks[11] of sizes 10, 50, and 100. Note that with this structure, the number of clusters is 1, implying that every node has equal influence on every other node. We ran simulations with different proportions of nodes in the network changed from an opinion of -2 to 2. Figure 7 depicts sample initial conditions for a network of size 10 with both 50% and 60% of the nodes changed to a healthy opinion. Figure 8 shows the results[12] of changing 50% and 60% of the network from an opinion of -2 to 2, with the two graphs demonstrating a

[10] We define a cluster as having a homogeneous opinion if all the nodes in the cluster have the same opinion.

[11] Networks in which the number of edges is equal to the number of possible edges.

[12] Throughout section 6, we averaged the curves produced by 30 simulations at each time step to produce the smooth curves shown in the graphs.

significant turning point in the average opinion trend.

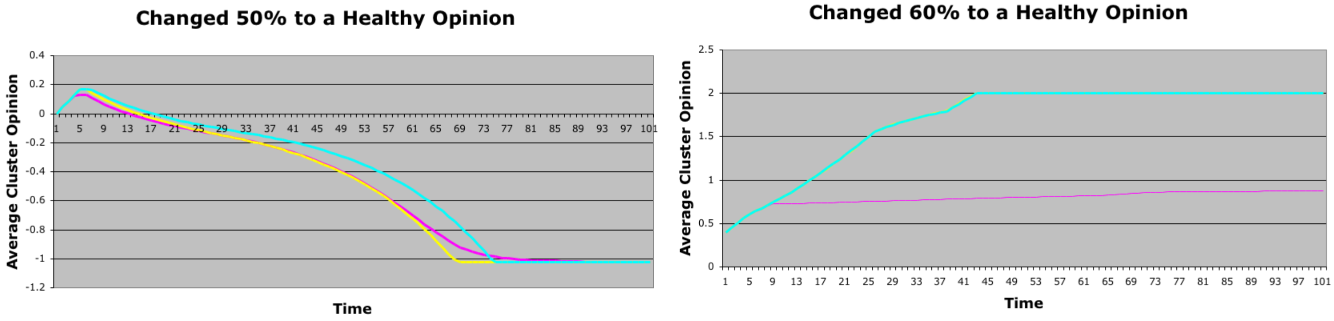

Figure 8: Average Cluster Opinion vs. Time for networks with 50% of nodes changed (left) and 60% of nodes changed (right). Each graph contains the Average Cluster Opinion vs. Time for networks with 10 nodes (pink), 50 nodes (yellow) and 100 nodes (blue). Being essentially the same, the blue and yellow curves overlap in the graph on the right. Note the positive association between the total number of nodes in a cluster and the proportion required to turn the network completely healthy. Also note that when 60% of nodes are changed, the clusters with 50 and 100 nodes follow nearly the same trend, thus suggesting that past a certain point, cluster size has no effect on opinion propagation.

If only 50% of the nodes are changed, the average cluster opinion stabilizes around -1 for all three networks. However, if 60% of the nodes are changed, the average cluster opinion for the networks of sizes 50 and 100 converges to 2 while the average cluster opinion for the network of size 10 does not. Note that this suggests the existence of a threshold between 50% and 60% such that beyond the threshold, the networks of sizes 50 and 100 will consistently converge to an opinion of 2. Furthermore, note that the networks of sizes 50 and 100 required a fewer proportion of nodes starting with an opinion of 2 in order to cause entire network to become healthy, thus suggesting that the larger a network is, the easier it is to propagate a healthy opinion.

6.2 Experiment 2

In the second experiment, we use two clusters with opposite opinions (2 and -2) in order to understand which opinion would prevail if the two clusters were to interact under the rules defined by the model.

For the initial conditions, we generate two complete clusters of size 10, one with a homogeneous opinion of 2 and the other with a homogeneous opinion of -2. In order to facilitate opinion update between the clusters, we connect the two clusters as follows: randomly connect a node in the positive-opinion cluster to a node in the negative-opinion cluster and randomly connect a node in the negative-opinion cluster to a node in the positive-opinion cluster. Using these initial conditions for the network state at $t = 0$, we run simulations using the model. Figure 9 depicts the initial

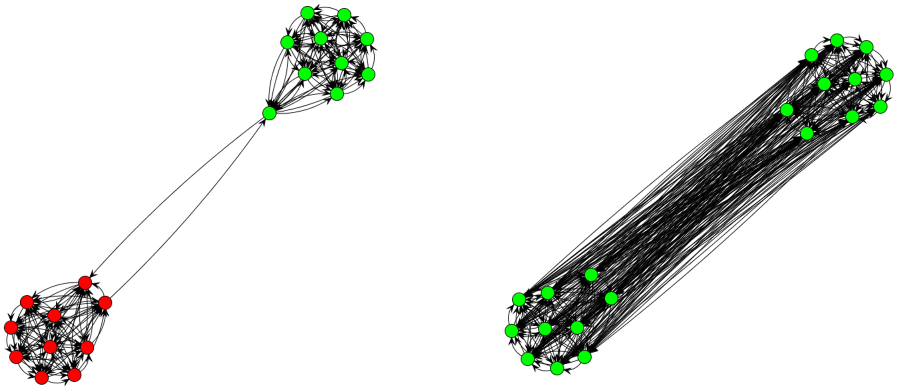

Figure 9: The initial condition for Experiment 2 (left) and the result of running simulations with the model on the initial condition (right). In Experiment 2, we found that when two complete clusters of opposite homogeneous opinion interact, the healthy opinion prevails and the the entire network adopts an opinion of 2 over a long period of time. This is due to the fact that the data-driven opinion amplification function defined in (7) is biased towards amplifying healthy opinions to a greater extent than unhealthy ones.

conditions and the simulation result. Figure 10 shows a graph of the average cluster opinion over time.

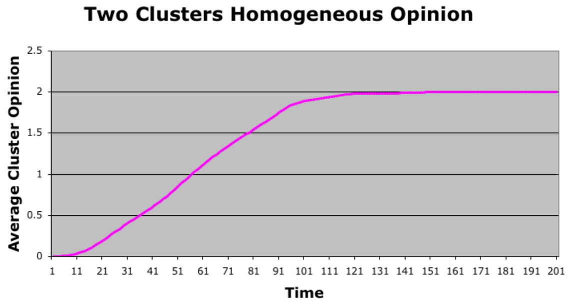

Figure 10: Average Opinion vs. Time for a network with two clusters of opposite opinions (2 and -2) as an initial condition. Note that over time, the average opinion increases such that eventually, the entire network has a healthy opinion of 2. Because the amplification function detailed in equation (7) in subsection 3.2 allows nodes to reach a more extreme positive opinion value (upper bound is 2) than a negative opinion value (lower bound is -1), the network consistently converges to the highest possible positive opinion: 2.

Note that as time progresses, all the nodes in the network eventually obtain an opinion of 2. This result is due to the amplification function introduced in subsection 3.2 in order to account for opinion radicalization in the positive direction as seen in data Observation 1 of subsection 2.2. Because we observed that the opinions in the data increased over time, we implemented the amplification function so that it was naturally biased towards creating a positive opinion in the network, with 2 being the upper bound for amplifying positive opinion and -1 being the lower bound for amplifying negative opinion. Since the amplification function favors the increasing of opinions, we can conclude that whenever a very unhealthy cluster competes against a healthy cluster, the opinion of the healthy cluster will dominate.

6.3 Experiment 3

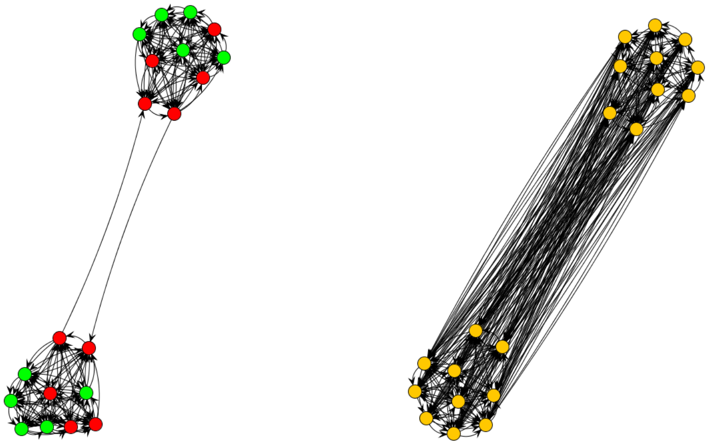

Figure 11: The initial condition for Experiment 3 (left) and the result of running simulations with the model on the initial condition (right). In Experiment 3, we found that when two complete clusters with each being 50% healthy and 50% very unhealthy interact, the network becomes entirely unhealthy over a long period of time. The fact that this scenario resulted in a healthy opinion failing to propagate suggests that when trying to maximize the spread of a healthy opinion, it is not advantageous to disperse the nodes of healthy opinion between clusters.

The goal of this experiment is to identify some subset of nodes in a homogeneous very unhealthy network to change to a healthy opinion at time $t = 0$ such that those nodes cause the rest of the network to adopt a healthy opinion. The only restriction is that no more than 50% of the nodes can be changed.

In subsection 6.1, we observed that we cannot cause the entire network to adopt a healthy opinion by only changing 50% of the nodes at $t = 0$ if the network is a complete graph. However, in subsection 6.2 we observed that if a network consists of two clusters of equal size, with one cluster having an opinion of 2 and another cluster having an opinion of -2 (thus 50% of the network is healthy) at $t = 0$, then the amplification function causes the entire network to become healthy over time. In this section, we use the same two cluster network structure used in subsection 6.2 to investigate whether it is more advantageous to have all the nodes in a single cluster be healthy at $t = 0$, as in subsection 6.2, or to have half the nodes in each cluster be healthy at $t = 0$.

For the initial condition, we generate a network with the same structure as the one in 6.2, the only difference being that instead of having opposite homogeneous opinions, each cluster is 50% healthy and 50% very unhealthy at $t = 0$. Using this initial condition, we run simulations using the model. Figure 11 shows the initial condition and the simulation result. Figure 12 shows a graph of the average cluster opinion over time.

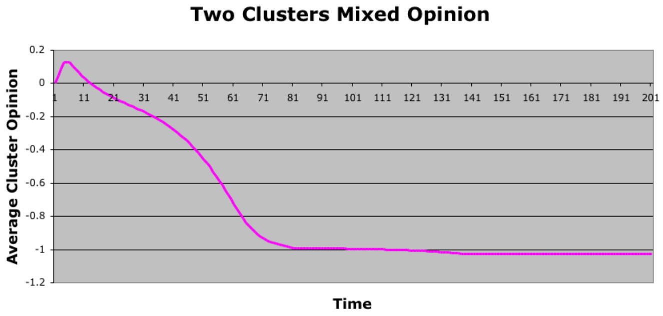

Figure 12: Average Opinion vs. Time for 50% of the nodes changed in each cluster in a network with two clusters. Note that when each cluster starts out as being 50% healthy, the opinion stabilizes at a negative value. Conversely, we saw in subsection 6.2 that when one cluster starts as completely healthy and the other starts as completely very unhealthy, the opinion converges at 2 (healthy). This difference suggests that certain subsets of nodes are more easily able to propagate a healthy opinion than others are.

We see that when each cluster starts as being 50% healthy, the long term opinion stabilizes at around -1, which is a long term result similar to the one from subsection 6.1 in which we simulated a network with a single cluster that was 50% healthy at $t = 0$. To explain this, we note that in this case, we have the same situation as in subsection 6.1, except that there are now two clusters that are 50% healthy instead of one. However, since the two clusters are very loosely connected for small t, having only two edges between them, they essentially act as two separate clusters for a time, thus acting in the same manner as the cluster in subsection 6.1 and stabilizing at an opinion around -1.

From the results in subsection 6.2 and in this section, we can conclude that when aiming to propagate a healthy opinion throughout a two-cluster network with only 50% of the network being healthy at $t = 0$, it is advantageous to have all the healthy nodes concentrated in a single cluster versus spread over two clusters. In subsection 6.2, we saw that having the healthy nodes concentrated in a single cluster at $t = 0$ causes the entire network to become healthy over a long time. In this section, however, we saw that when each cluster is a combination of healthy and very unhealthy nodes, the network, although it is still 50% healthy at $t = 0$, stabilizes around a negative opinion value instead of a positive one.

7 Conclusion

In this paper, we formulated an agent-based model [4, 10] with dynamics based off of observations from a real-world dataset containing surveys of a student population. The model contains a set of parameters w, k, and c which represent different aspects of society such as tendency to follow the opinion of those from the same cluster versus those from a different cluster (parameter w), tendency to radicalize opinion (parameter k), and tendency to restrict friendships to those from the same cluster (parameter c). We have demonstrated that the model with parameter values $w = 5$, $k = 1.05$, $c = .245$ accurately predicts the trends in the observed data. Furthermore, by using the parameter values that reflected the data, we were able to predict the long term dynamics of the network.

By varying the parameters of the model, we simulated opinion propagation and friendship changes in different types of social networks. When varying c, we observed that when agents have a high tendency to restrict their friendships to those within the same cluster, the average inner connectivity of the clusters increases. When varying w, the extent to which weighted friendships due to clustering affect opinion change, we observed that weighted friendships between agents only affect long term opinion propagation up to a certain maximum allowed strength, beyond which the long term opinion propagation changes negligibly. Lastly, when varying k, the extent to which opinions radicalize, we observed a long-term instability in both cluster opinion and opinion spread associated with a high tendency to radicalize.

By performing three experiments in which we varied the initial network structure and opinion distribution, we derived a set of tools that can be used to gain a more thorough understanding of how to maximize healthy opinion propagation in social networks. In the first experiment, we found that in networks consisting of a single cluster, healthy opinions tend to propagate more easily as the network gets larger. In the second experiment, we found that when two clusters of opinions 2 and -2 interact, the one with opinion 2 always succeeds in propagating its opinion due to the bias of the data-driven opinion amplification function defined in (7). In the third experiment, our results suggested that in order for a healthy opinion to spread throughout an unhealthy network, it is advantageous for the healthy nodes to be concentrated in the same cluster at $t = 0$.

The model created in our work yields not only a method for reflecting the dynamics of and predicting the changes in a real human society, but also several tools for studying the changes that occur in various theoretical networks, such as the ones in section 6.

8 Further Research

In our work, we successfully modeled the dynamics of a network in a dataset but did not verify that our model correctly reflects the changes in other real social networks. To improve upon the strength of our model, we will simulate other datasets with our model and compare the predictions made by the model to the actual change in data in order to verify the accuracy of the model in different types of networks.

In section 6, we provided an introduction to a study of health opinion propagation in the network by deriving conclusions regarding opinion propagation in networks containing one or two clusters. However, their are infinitely many theoretical social network structures as well as network structures yielded by analysis of data. Analyzing other network structures would yield more insight into the opinion propagation problem initiated by section 6.

9 Acknowledgements

First, I would like to thank my mentor, Dr. Natasha Markuzon, both for suggesting the idea of modeling social networks and for her guidance throughout the course of the research. I would also like to thank the MIT PRIMES program for providing me with the opportunity to perform this research. Lastly, I would like to thank my parents, whose constant support allowed me to get where I am today.

References

[1] Acemoglu, Daron, Asuman Ozdaglar, and Ercan Yildiz. ”Diffusion of innovations in social networks.” Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on. IEEE, 2011.

[2] Acemoglu, Daron, Mohamed Mostagir, and Asuman Ozdaglar. ”State-Dependent Opinion Dynamics.” (2012).

[3] Bradwick, M. E. (2012). Belief propagation analysis in two-player games for peer-influence social networks (Doctoral dissertation, Massachusetts Institute of Technology). http://dspace.mit.edu/bitstream/handle/1721.1/72645/807215820.pdf?sequence=1

[4] Carley, Kathleen M., et al. ”BioWar: scalable agent-based model of bioattacks.” Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 36.2 (2006): 252-265.

[5] David Kempe, Jon Kleinberg, va Tardos: Influential Nodes in a Diffusion Model for Social Networks. In Proceedings of ICALP 2005, Lisboa, Portugal.

[6] D. Rand, S. Arbesman, and N.A. Christakis, ”Dynamic Social Networks Promote Cooperation in Experiments with Humans,” PNAS: Proceedings of the National Academy of Sciences 108(48): 19193-19198 (November 2011); doi:10.1073/pnas.1108243108

[7] Girvan M and Newman MEJ, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 8271-8276 (2002).

[8] Handcock, Mark S., Adrian E. Raftery, and Jeremy M. Tantrum. ”Model?based clustering for social networks.” Journal of the Royal Statistical Society: Series A (Statistics in Society) 170.2 (2007): 301-354.

[9] Hartline, Jason, Vahab Mirrokni, and Mukund Sundararajan. ”Optimal marketing strategies over social networks.” Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

[10] Levine, S. S. & Kurzban, R. (2006). Explaining clustering within and between organizations: Towards an evolutionary theory of cascading benefits.Managerial and Decision Economics, 27, 173-187.

[11] McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. ”Birds of a feather: Homophily in social networks.” Annual review of sociology (2001): 415-444.

[12] Nekovee, Maziar, et al. ”Theory of rumour spreading in complex social networks.” Physica A: Statistical Mechanics and its Applications 374.1 (2007): 457-470.

[13] Norman, M. Frank. Markov processes and learning models. Vol. 84. New York: Academic Press, 1972.

[14] Read, Jonathan M., Ken TD Eames, and W. John Edmunds. ”Dynamic social networks and the implications for the spread of infectious disease.” Journal of the Royal Society Interface 5.26 (2008): 1001-1007.

[15] Schaeffer, Satu Elisa. ”Graph clustering.” Computer Science Review 1.1 (2007): 27-64.

[16] Sensing the Health State of a Community, A. Madan, M. Cebrian, S. Moturu, K. Farrahi, A. Pentland, Pervasive Computing, Vol. 11, No. 4, pp. 36-45 Oct 2012

[17] The JUNG (Java Universal Network/Graph) Framework No. UCI-ICS 03-17. (2003) by Joshua O'Madadhain, Danyel Fisher, Scott White, Yan B. Boey

[18] Vaquera, Elizabeth and Grace Kao (2008). ”Do You Like Me as Much as I Like You? Friendship Reciprocity and its Effects on School Outcomes among Adolescents.” Social Science Research. 37: 55-72.