Rose-Hulman Undergraduate Mathematics Journal

Volume 17
Issue 1

Article 12

SI Dynamics of Disease Spread

Catherine Northrup
Carthage College

Elisabeth Rutter
Carthage College

Kerry Stapf
Carthage College

Rose-Hulman Undergraduate Mathematics Journal

Network Science: SI Dynamics of Disease Spread

Catherine Northrup[a] Elisabeth Rutter [b]
Kerry Stapf [c]

Volume 17, No. 1, Spring 2016

Sponsored by

Rose-Hulman Institute of Technology
Department of Mathematics
Terre Haute, IN 47803
Email: mathjournal@rose-hulman.edu
http://www.rose-hulman.edu/mathjournal

[a]Carthage College
[b]Carthage College
[c]Carthage College

Rose-Hulman Undergraduate Mathematics Journal
Volume 17, No. 1, Spring 2016

Network Science: SI Dynamics of Disease Spread

Catherine Northrup Elisabeth Rutter Kerry Stapf

Abstract. Imagine you are walking down a crowded hallway. You aren't in contact with everyone all at once. You talk to or simply pass by different people at different times as you walk down the hall. These connections would best be represented using a temporal network. In this work, we examine temporal networks to determine the behavior of disease spread across these networks and how it differs from the behavior of static networks. We use differential equations for mean field approximations to theoretically model how infection spreads throughout a temporal network. We extend our model to incorporate network structure by deriving a degree-based mean field theory. We then validate our theories with simulations in Mathematica. We also look into including multiple rounds of infections to see how it affects the spreading behavior. From our results we are able to determine how the temporal aspect affects the rate of spread of the disease and the overall size of the infected population.

Acknowledgements: We would like to thank Carthage College for funding our research and Dr. Yaple for being our mentor.

1 Introduction

Network science is a fascinating interdisciplinary field and as such has a varied array of applications. It draws ideas from mathematics, statistical physics, computer science, and even sociology. Networks can be used to model anything from the spreading of internet memes across social networks to mapping connections between genes to figure out the origins of various diseases. One of the most common applications is using networks to model epidemiological processes. This is the type of modeling we use in order to study the characteristics and dynamics of networks, specifically the SI compartmental model. Our aim in this research is to determine how those characteristics differ when dealing with a network that changes over time.

Some of the earliest work modeling networks was done by Paul Erdős and Alfréd Rényi, who developed random graph models using probabilistic theory [3]. Another model, the small-world model, was created by Duncan Watts and Steven Strogatz [5] while studying the six degrees of separation in celebrity social networks. This type of random graph differs from the Erdős–Rényi graphs in that the nodes are highly clustered and any two nodes are connected by a short path. With the advent of the internet, physicist Albert-László Barabási used networks to map the internet and a new network structure emerged: the scale-free network [1]. In scale-free networks, there are a handful of highly connected hubs while the rest of the nodes have only a small number of connections. Together with Réka Albert, Barabási found that most real world networks (natural, technological and social) form scale-free networks.

We use a graph to represent a network, where each node or vertex is a person. The two types of graphs we most often use to represent a network are Erdős–Rényi graphs, also known as Bernoulli graphs, and scale-free graphs, also known as Barabási–Albert graphs. Nodes in a Bernoulli graph have very similar degrees, or number of connections. The graph is built randomly, with any two nodes having the same independent probability of being connected. The degrees of the resulting graph then follow a binomial distribution. On the other hand, the Barabási–Albert graphs are built using preferential attachment, meaning a high degree node is more likely to gain more edges. Therefore, in Barabási–Albert graphs there are nodes of very high degree, and nodes of very low degree.

After deciding to use Bernoulli and Barabási–Albert graphs, we chose to look into the SI model to see how a temporal network structure affects the spread of infection. The SI model is the simplest type of epidemic spread model. The S is the fraction of susceptible people out of the entire number of nodes, which means they can become infected with the disease being modeled. We also use S to denote that a node is part of the fraction of susceptible people. The I is the fraction of infected people that can spread the disease by contact with others out of the entire number of nodes, but I also indicates that a person is part of the fraction of infected. This model follows the fully-mixed idea, where anyone can come into contact (meaning it disregards geographic location and community groups). Because this model does not include any sort of recovery time, the infection has the potential to spread indefinitely. Mathematically, this model can be represented using differential equations.

We begin in Section 2 with the derivation of mean field theory, used for the Bernoulli graphs, and then discuss how the simulation compares to the theoretical equation found. In Section 3, we discuss how we simulated multiple rounds of infection. Section 4 includes the degree-based mean field theory equation for the Barabási–Albert graphs, along with a comparison graph of mean field theory to degree based mean field theory to a simulation. Section 5 discusses how we tested our theories for error and convergence within simulations.

2 Mean Field Theory

First, we derive an equation that would represent how an infection would theoretically spread throughout a network. In order to generate a theoretical equation for a temporal SI network, we look at the relationship between S and I, which is $S + I = 1$. Then we calculate the probability for each edge combination, which could be SI, SS, or II. After the probabilities of each possible combination are found, the expected value for the number of infected nodes at the next time step is determined. Since the SS and II combinations of nodes will not result in an increase in the fraction of infected, the combinations are classified as a 0, because 0 nodes become infected. The last combination, SI, will result in one more node becoming infected, so the proportion of infected nodes increases with $\frac{1}{n}$. The rate of change of the number of infected nodes is then written as

$$\frac{dI}{dt} = \begin{cases} \frac{0}{n} & \text{with probability } (1-I)^2 + I^2 \\ \frac{1}{n} & \text{with probability } 2I(1-I) \end{cases}$$

From here, we write the expected value function for the number of infected nodes, which comes from the idea that the expected value is equivalent to the sum of the possible outcomes multiplied by their probabilities. The expected change of infected nodes simplifies to:

$$\frac{dI}{dt} = \frac{2}{n} I(1-I). \tag{1}$$

We then go through a similar derivation for activating two edges at one time, and then three, and design a general mean field theory based on how many edges are activated at one time (w):

$$\frac{dI}{dt} = \frac{2w}{n} I(1-I). \tag{2}$$

In this derivation, edges that share a common node are ignored.

We then simulate the spread of infection on a temporal network under the same assumptions as the model and compare it to our theoretical equation.

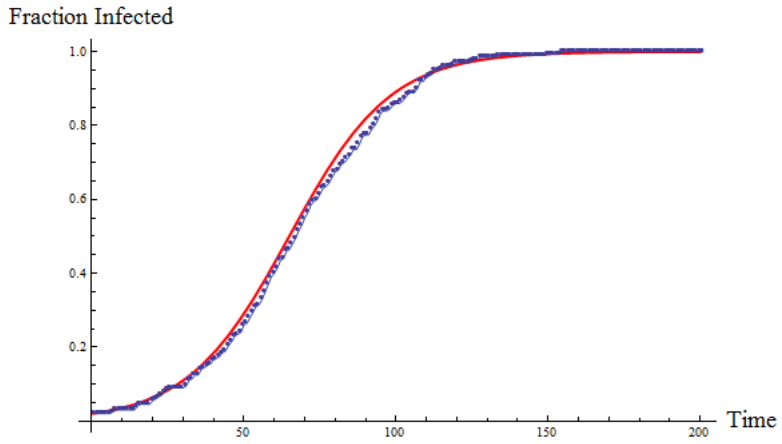

Figure 1: The theoretical mean field curve from Equation 2 is in red and the blue dotted line is the median curve of different experimental curves based on simulations that were run. Note that the two match well.

As can be seen in Figure 1, the median of the experimental curves and the theoretical curve are quite similar. We use the median instead of the mean because there are outliers that would skew the mean either higher or lower than what is expected. This shows that the temporal network closely matches our theoretical prediction of how the disease would spread through the temporal network.

After looking into the temporal mean field theory, we notice that some edge combinations could share nodes. This could then allow the neighbors of the neighbors of an infected node to become infected in a single time step, if the infection was allowed to spread from an infected node to a susceptible node and across another connection to a different susceptible node. Thus, we will look into multiple rounds of infection for a network.

3 Multiple Rounds of Infection

One thing that we overlook in our mean field model is the scenario in which there is an overlap of the nodes connected by multiple edges. This gives rise to our multiple rounds of infection model.

In our previous simulations, the infection could only spread from the infected node to its direct neighbors. We want to test how the infection will spread if there are multiple "rounds" of infection. For example, if node i is infected and is connected to susceptible node j, and susceptible node k is connected to node j but not i, the infection could still reach node k in the same time step by way of node j.

In order to model this mathematically, we first need to list out all of the possible combinations in which pairs of susceptible or infected nodes could connect and the probabilities with which these cases can happen. The setup for our temporal networks follow the same implementation as in our previous work, where at each time step only certain edges are

"active" and can be used to advance the infection. We start by calculating the probabilities for when two edges at a time are active, then repeat the process for three edges. A simple visual of multiple rounds with two edges being activated is shown below in Figure 2.

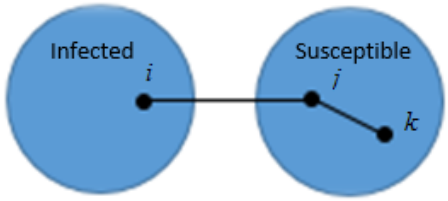

Figure 2: The SI connection and the SS connection share node j, and infected node i is infecting nodes j and k in a single time step.

When determining the probabilities of each scenario of I and S combinations, there are two different types: the original probability, which is what we used before, and the exception to the original probability. The original probabilities correspond to the combinations in which each connection has its own nodes; no nodes overlap between two connections. The exceptions are when some of the nodes overlap between connections and result in either a lower or higher change in the number of infected individuals at each time step. Once the process of determining the probabilities of the exceptions, the probabilities were then organized into groups depending on the number of nodes that became infected. For all of the probabilities that resulted in zero nodes becoming infected, the probabilities were added together and multiplied by I because there was no change in the number of infected individuals. If the probabilities represented one node becoming infected, they were added together and multiplied by $I + \frac{1}{n}$, and if two nodes became infected, the probabilities were added together and multiplied by $I + \frac{2}{n}$. Once this was done, all of the probabilities were added together and simplified to get an equation representing multiple rounds of infection. The final equations for this process for two and three edges added at each time step are shown below along with the general equation for one round of infection.

w	Equation
1	$\frac{dI}{dt} = \frac{2wI}{n}(1-I)$
2	$\frac{dI}{dt} = \frac{2wI(1-I)}{n}(1+\frac{1-2I}{n})$
3	$\frac{dI}{dt} = \frac{2wI(1-I)}{n}\left[1-\frac{2(3-3I+\frac{2I^2}{3}-\frac{2I^3}{3})}{n}+\frac{16(-1+2I-\frac{2I^2}{3})}{n^2}\right]$

Table 1: Multiple Rounds of Infection with w being the number of rounds

The equations in Table 1 all have the same initial coefficient of $\frac{2w}{n}I(1-I)$, which is the mean field model from Equation (2). However, the two and three rounds of infection

equations have additional correction terms. These terms are in the form of the truncated series $\sum_{k=1}^{w} a_k(I)n^{1-k}$, where the a_k coefficient is interpreted as a correction, accounting for the likelihood that an edge is shared by more than one infected node. This means that as the number of rounds of infection increase, the extra terms will continue until $\frac{1}{n^{w-1}}$, where w is the number of edges activated at each time step. This behavior leads us to believe that correction terms become less important for large networks. After observing the equations, we validate our theory by plotting the results with our simulation of a temporal network.

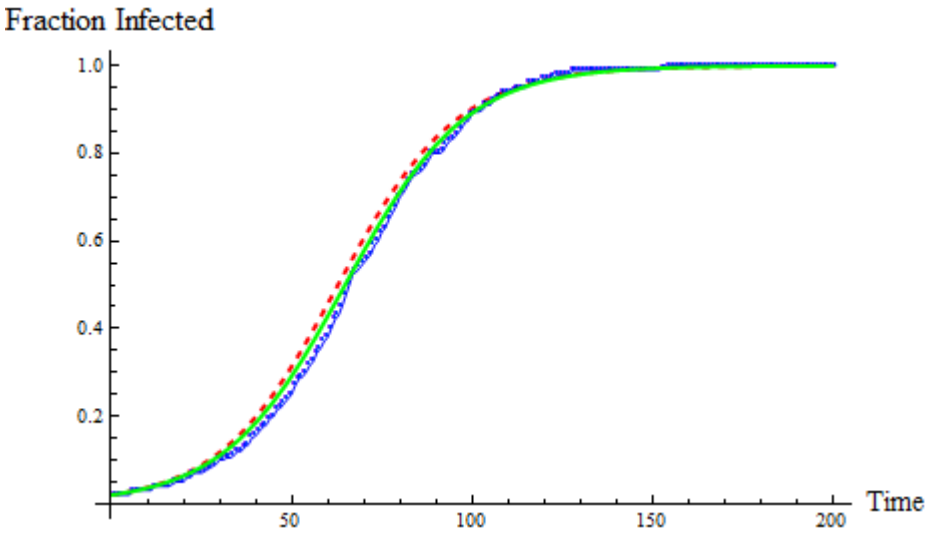

Figure 3: The blue, dotted curve is the simulation. The red, dashed curve is the mean field theoretical equation, which is one round of infection. The solid green curve is the theoretical equation for three rounds of infection.

After observing Figure 3, it is clear that the two theoretical curves of one round of infection and three rounds of infection match almost exactly to the experimental curve. This indicates that the number of rounds of infection does not affect the rate of infection. Thinking about it theoretically, allowing for multiple rounds of infection should cause the infection to spread through the network faster than if there were only one round of infection. However, as seen in this section, incorporating multiple rounds of infection into our simulation does not seem to affect the rate of infection as much as expected, so it can be ignored when running the simulations. Since the multiple rounds does not seem to show any significant changes in the rate of infection compared to the mean field model, we will turn our focus to the dynamics on a network of a different structure.

4 Degree-Based Mean Field

In our previous models, we found that while they fit well for Bernoulli graphs, they did not match well for Barabási–Albert graphs. This is due to the fact that Barabási–Albert graphs

follow the principle of preferential attachment (the rich get richer) and as such are much less homogeneous in regards to the distribution of edge connections compared to Bernoulli graphs, where all of the edge connections are equally probable. The differing edge distribution then alters the spreading behavior. Because our mean-field model in Equation 2 treats each node identically, it does not match well with simulations on networks where nodes are heterogeneous. To remedy this, we look into degree-based mean field approximations that will allow us to incorporate the network structure into the model and simplify matters by assuming that nodes of the same degree (same number of edges branching to other nodes) the same dynamical behavior, which allows us to consider the dynamics on a network with an exactly prescribed degree distribution. Nodes of the same degree spread the infection at the same rate and they will also infect the same number of nodes. The approximation [4] we use is given by the system of equations

$$\frac{d\rho_k}{dt} = \lambda(1-\rho_k)kv \tag{3}$$

where v is

$$v(t) = \sum_k P_k \rho_k \frac{k}{z} \ . \tag{4}$$

In this system of equations, there is a separate equation for each degree k in the network. The λ is the average infection probability, $\rho_k(t)$ represents the fraction of degree k nodes that are infected at a time t, and $v(t)$ is the overall probability that, given an edge, it's attached to an infected node with degree k. This probability is calculated by, for each degree k, multiplying the likelihood of having this degree P_k by the number of neighboring nodes (k), then by the fraction of these neighbors who are infected (ρ_k). The factor of $\frac{1}{z}$ is a normalization constant, where z is the mean degree.

We then use a modified version of Equation (3) to take into account the temporal aspect. The naive guess would be to add the same coefficient as before, $\frac{2w}{n}$, however this would be incorrect. The proper coefficient also includes a factor of z, the mean degree. This gives us the coefficient $\frac{2w}{nz}$, which represents the probability of choosing either an SI or an IS edge, since there are two edge combinations multiplied by w edges chosen, all over the total number of edges nz in the network. The full version of our new temporal degree-based mean field equation is as follows:

$$\frac{d\rho_k}{dt} = \frac{2w}{nz}\lambda(1-\rho_k)kv. \tag{5}$$

Our theory is then plotted against the median of several simulations run with Barabási–Albert graphs as well as our original mean field equation. In Figure 4, the median of the 10 graphs fits more closely to the degree-based mean field theory, while the mean field theory dips quite low, not matching the experimental curve at all. Note that the theoretical curve in red has the same general shape as the experimental curve, with the only major difference being at the peak of the curve. This slight disparity is due to the fact that the high degree

hubs in Barabási–Albert graphs cause the infection to spread more quickly than predicted by the theory. Overall, this plot shows that the degree-based mean field approximation fits much better than the regular mean field for Barabási–Albert graphs.

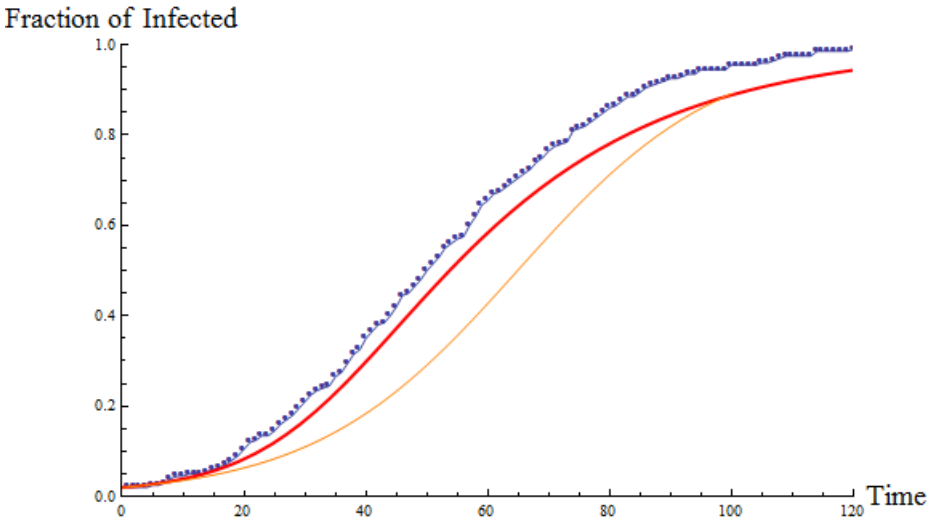

Figure 4: Comparison plot of two models versus simulation of a Barabási–Albert graph. The orange curve is the mean field theory, thick red curve is the degree-based mean field theory, dotted blue curve is the median of 10 simulations. Note how the degree-based mean field theory matches the simulation much better than the mean field theory.

5 Error Testing

Looking for patterns and testing for convergence is a way of assessing the validity of a model to see how well it actually fits a real world scenario or simulation. We began to look into ways to test our simulations for error compared to our mean field theory. We first tested if changing the number of simulations run affected the error by finding the difference in the fraction of infected between the theoretical model and simulation at a specific slice of time, all while running a loop to continuously increase the number of simulations aggregated. When we plotted the data, the error did not converge anywhere, so we determined that the number of simulations run did not have a significant impact on the accuracy of the fit. Therefore, we did not include the plots of error versus simulations.

We tested the convergence in a similar way for increasing the number of nodes, and then also looked at the root mean square of the error for nodes because we observed that the error tended towards zero.

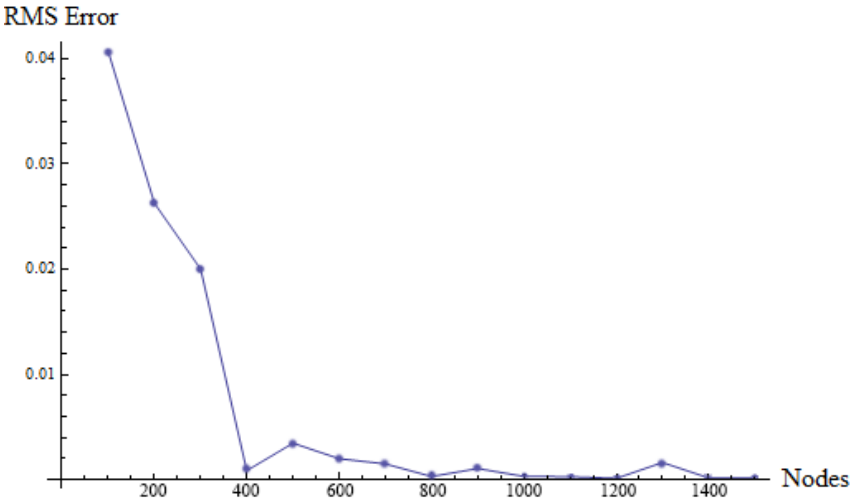

Figure 5: RMS versus Number of Nodes for a Bernoulli network.

In Figure 5, the root mean square error was plotted against the number of nodes until it reached 1500 nodes. With this large amount of nodes, the curve does come very close to zero error, and stays around zero. This showed that increasing the number of nodes in our simulation better matches our theoretical equation.

We also ran the code using a Barabási–Albert graph for both the mean field theory and the degree-based mean field theory. This was done to get quantitative evidence that the degree-based model is a better fit for Barabási–Albert graphs than the regular mean field theory. A comparison plot of the root mean square error for these two models is shown in Figure 6.

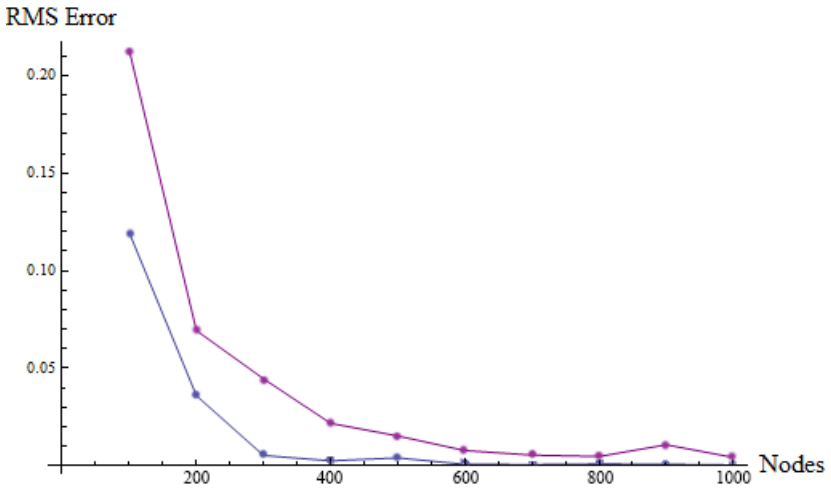

Figure 6: Comparison plot of root mean square error. The mean field is the top purple curve, while the degree-based mean field is the bottom blue curve.

As can be seen from Figure 6, the error for both models converges towards zero as the number of nodes increases. The error for the degree-based mean field model, however, is consistently lower than the mean field model, which makes it evident that the degree-based model is a better fit.

6 Summary

Throughout our research, we have come across many thought-provoking findings. We were able to successfully incorporate the temporal aspects of a network in our basic derivation of the mean field model and the degree-based mean field with SI dynamics. These derivations allowed us to compare them against simulations to see how well they predict how a disease spreads over a network. In most of our simulations, the experimental outcomes and the theoretical outcome matched very well. In the cases that did not match as well, we were able to come up with a different theory in order to compensate for those errors. Our convergence testing showed that as the number of nodes in a network increased, the error decreased. This indicates that to produce more realistic and accurate results, a larger network would need to be studied. The reason we did not use larger networks was because of the length of time necessary to run the simulations. We were also able to determine that temporal networks, as opposed to static networks, spread the disease much slower by a factor of $\frac{2w}{n}$. In both cases we found that as long as all the nodes were connected or if the infection started within a connected component, eventually all of the nodes in the network or in that connected component became infected. If given the time for further research, we would be interested in running the multiple rounds of infection code with Barabási–Albert graphs, as well as adding another layer of complexity by using SIR or SIS dynamics.

References

[1] Barabasi, Albert-László, Albert, Réka, Jeong, Hawoong "Mean-field theory for scale-free random networks." Physica A 272:173–187, 1999.

[2] Newman, M. E. J. Networks: An Introduction. Oxford: Oxford UP, 2010. Print.

[3] Erdős, P., Rényi, A. "On random graphs." Publicationes Mathematicae Debrecen 6:290-297, 1959.

[4] Porter, Mason A., and James P. Gleeson. "Dynamical Systems on Networks: A Tutorial."arXiv preprint arXiv:1403.7663(2014).

[5] Watts, D.J., Strogatz, S.H. "Collective dynamics of 'small-world' networks." Nature 393:440-442, 1998.