# Modeling the Spread of Information

Andrew Nemec

Carthage College

anemec@carthage.edu

March 8, 2017

**Abstract**

With advances in technology we can study interesting problems involving large networks. Such networks include websites such as Facebook, which is used by millions of people everyday. In particular, we will study the spread of information in a network. We will use the Barabasi-Albert model and techniques for studying networks in order to make predictions about the spread of information. We will use the programming language Python and a library called NetworkX to create simulations and analyze our results.

## 1 Introduction

Graph theory is the study of the properties of structures containing vertices connected by edges. Graph theory is used extensively in the study of networks by assigning attributes to vertices and edges in order to model a network. Some examples of this include power grids, social networks, disease spread, and others. In this paper we will be examining the spread of information in social networks using these techniques.

We can learn a lot through the study of social structures. We can examine the process by which information flows through a network. A study published in 1950 examined the adoption rate of hybrid corn seeds in two communities in Iowa. The study showed that how influential farmers sped up the adoption rate of these disease resistant crops. With the advent of more powerful computers we can study larger scale problems. Large scale internet access has provided many with the opportunity to connect to each other. This is seen through the popularity of online social networks. Facebook being an example of one the most popular. Its estimated daily users are a 1.23 billion. Studying these networks can help identify the reach of a users post or the audience of an advertisement.

We will focus on the spread of information in a network and analyze the relationships that cause this. To do this we will be using the programming language Python. We will be using a powerful library for Python called NetworkX to help in creating and analyzing our network.

## 2 Definitions and Development

Our goal is to study the spread of information in a society. To do this we will create a network and a simulation to be run on this network. Our network will use a node to represent a person. The relationship between these nodes will be represented by an edge. This edge represents a

mutual relationship that these people share, that is this network is not a directed network. To help us in modelling a real world network we will use the Barabasi-Albert model. This model uses a process called preferential attachment to grow the network. At each step in building the network a new node is added and 1 or more edges are attached from the new node to existing nodes. A node with a higher degree is more likely to receive this connection. This process creates a network with nodes with varying degrees, mimicking the communities formed in social networks. Furthermore, it guarantees that our network is connected. We will use techniques from graph theory to help us understand our network. The degree of a node is the number of connections the given node has to other nodes. This will give us an understanding of how important a node is locally. The betweenness centrality of a node is the number of shortest paths that pass through the given node. This does not include paths starting or ending at the given node. Betweenness centrality tells us how important a node is in the entire network based on how likely information is to pass through a given node. Eigenvector centrality is a measure of importance of a node in a network. A node receives a higher score based on the number of important nodes its connected to. Eigenvector centrality will help us understand how important a node is overall.

**Simulation**

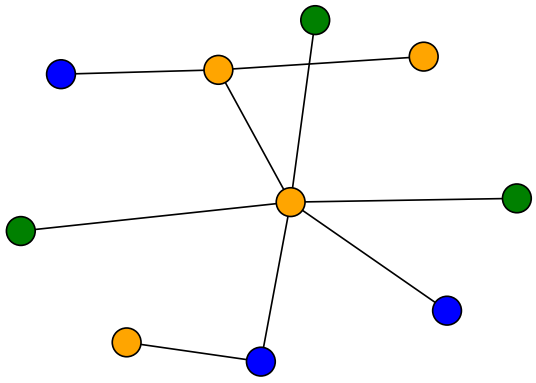


Figure 1: Example of simulation

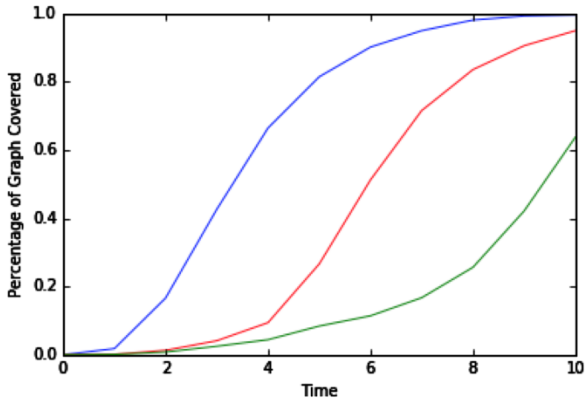
We will have two simulations. Our simulations will use a network containing 100 nodes and be created using the Barabasi-Albert model with a minimum degree of one. We seed our model to keep the network constant in each simulation. The first simulation is meant to be a basic demonstration of the importance of a node in a network. We accomplish this by having 2 opinions and simulating the time for one opinion to cover the entire network. We will select a node based on importance measured by eigenvector centrality. The node will then spread its information to its neighbors until the entire graph has been covered.

The second simulation attempts to model what the spread of information in a network might look like. Specifically, we want to study the effect that changing the likelihood to spread an opinion has on a network. We will have 5 opinions in this simulation and a variable that will represent a chance to spread opinion. Each node in this simulation will start with one of five opinions

randomly distributed over the network. Each node will start with one opinion but can add more to its list of opinions. At each step in the simulation we randomly choose a node to spread its opinion. This node will form a list of its neighbors. We use a variable to control the chance to spread its opinion. If the probability to spread is met, we will spread the opinion. If the nodes have no opinion in common the neighboring node will add the spreading nodes opinion. If each nodes list of opinions has an opinion in common, the nodes will delete all opinions except the similar one. For example, we have an initial node and its neighbor. The initial node has the yellow opinion and the neighbor has the green opinion. If the chance to spread is met, then the neighbor will add the yellow opinion to its list. The neighboring nodes list now contains the green opinion and the yellow opinion. Now let us say that we have an initial node with the red and yellow opinion. We also have a neighbor with the red opinion and the green opinion. If the chance to spread is met, the initial node and the neighbor will reset their lists to only contain the red opinion since the nodes share that opinion.

## 3 Results

In this section we will discuss the results of our simulations. We will use the Barabasi-Albert model with a minimum degree of one and a network size of 100 nodes.




Changed the wording here to make it clearer that we're talking about the first simulation

Figure 2: The blue line represents the highest eigenvector centrality. The red line represents an average eigenvector centrality. The green line represents the lowest eigenvector centrality.

In our first simulation we will demonstrate the importance of being an influential node, that is, a node with a higher eigenvector centrality. In this simulation one node is chosen as the initial node to spread information. In each step of the simulation it will spread information to its neighbors. The blue line represents an initial node with the highest eigenvector centrality. The red line represents an average eigenvector centrality and the green line represents a low centrality. As we can see the higher the eigenvector centrality, the faster the spread of information. This is a result of the initial node having a larger influence so its ability to spread information to many nodes at once results in the information being able to pass through the network quickly. On the other hand, an unimportant node will take much longer to spread

Removed a couple sentences here talking about probability to make it clearer

information leading to a much slower initial growth and a longer time for the information to spread across he entire network. This leads us to our next, more interesting simulation.

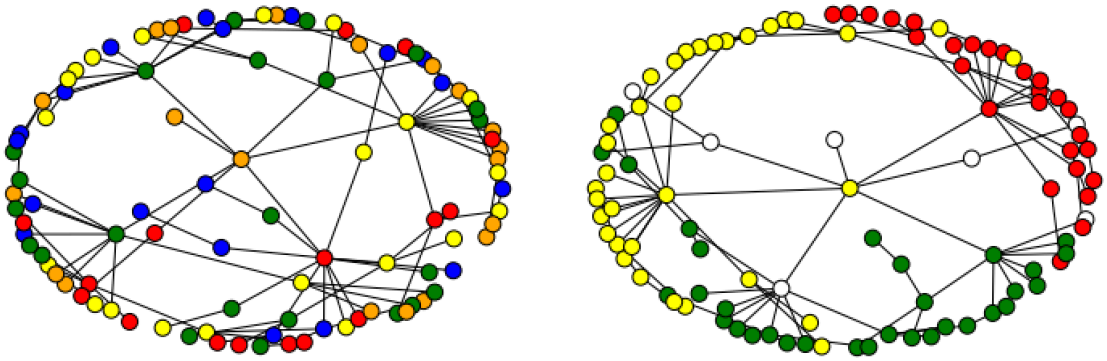


Figure 3: Networks before and after simulation. Each color represents a different opinion. The white nodes represent nodes with multiple opinions.

Figure 3 shows an example of the starting opinions and the ending opinion in one run of the simulation. At the start of the simulation, opinion is spread evenly amongst all nodes, as seen in the network to the left. The network on the right shows the opinion that forms within communities by the end of a simulation. We can see there are three distinct opinions still remaining. In each community there is a node with a high degree that is influencing the rest of the community. In the next graphs we will further examine the results of one run of the simulation.

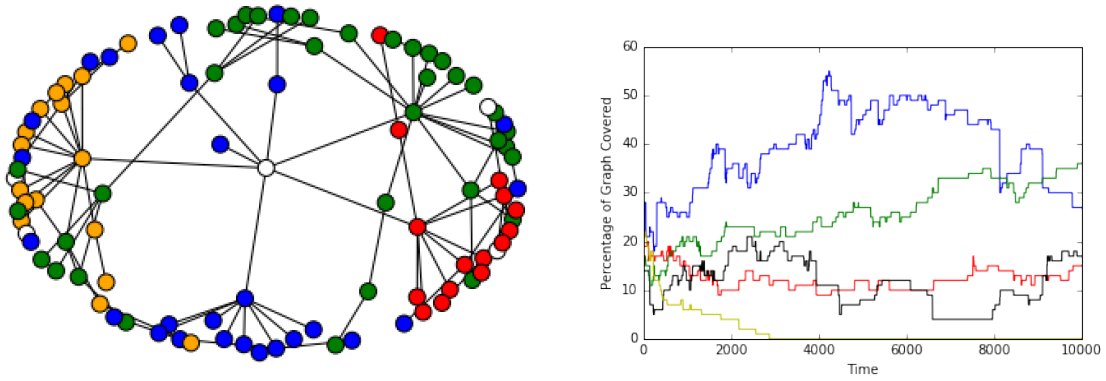



Figure 4: Network after simulation and graph of simulation at 10% chance

These graphs show one run of the simulation. The network on the left shows the opinion at the end of the simulation. The graph on the right shows the amount of the network covered by an opinion throughout the entire run of the simulation. This simulation was run with a chance to spread of 10%. We again see that each community starts to form one distinct opinion, although there isn't the same amount of agreement as we have seen in the previous example. From the graph of the simulation we see that there is variation in the size of each opinion. In fact, blue had the highest size for most of the simulation but was passed by green near the end. We

continue to see this trend in variance in the size of opinion through multiple simulations as well as if we continue the simulation past the 10000 iterations shown here. We don't begin to see more consistent convergence to a single opinion until we increase the chance to spread past 35%. This could be caused by a few things. One possibility is that a higher chance to spread makes it easier for an opinion to spread to a new community an establish itself in that community. A lower chance means that even if an opinion manages to spread to a new community it still has a difficult time establishing itself.

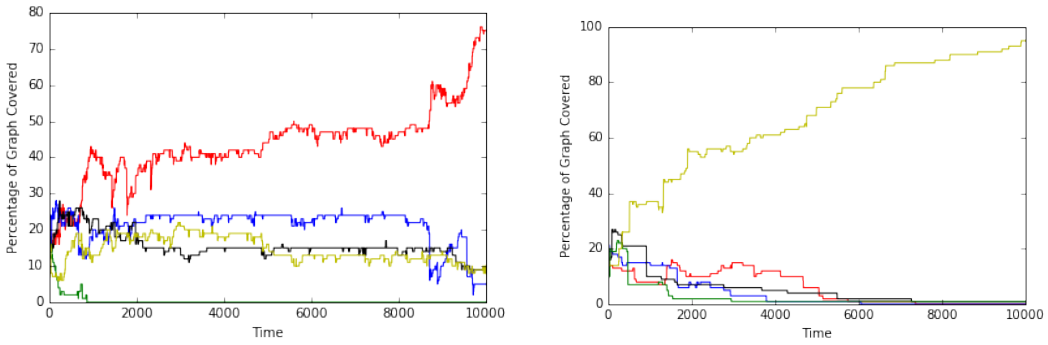



Figure 5: Graph of simulation at 50% chance and 90% chance

These graphs show what the simulation looks like after changing the chance to spread. The graph on the left shows a 50% chance to spread and the graph on the right shows a 90% chance to spread. Compared to the graph from before, we can see how changing the chance to spread affects how quickly a network converges to a single opinion. When we run the simulation with a 10% chance to spread we rarely see convergence to a single opinion, even when we run the simulation much longer than the 10000 iterations we have been using. However, the higher the chance the more quickly we see the network converge to a single opinion. One possibility could be that a lower chance to spread doesn't favor the majority opinion in the way a higher chance to spread does. With a higher chance, information spreads more easily. This means that once an opinion gains traction it is harder to slow it down.

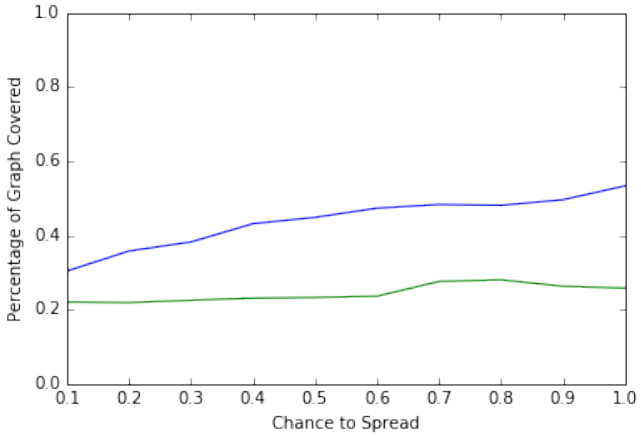


Figure 6: Graph of average amount of network covered vs change in probability to spread

This graph shows the average percentage of the graph covered by an opinion after 100 runs of the simulation. We do this for each change we make in chance to spread. The blue line represents the most held opinion at the end of a simulation and the green line represents the second most held opinion. We can see a trend that as an opinion is more likely to spread, the percentage of the graph it will cover at the end increases. However, the second most held opinion does not follow this trend, it stays relatively consistent. One possibility as to why this is happening is that the second opinion is relatively strong in its community and the larger opinion grows by converting the opinions that are less strongly held. We can further see this in the next graph where we include the other three opinions.

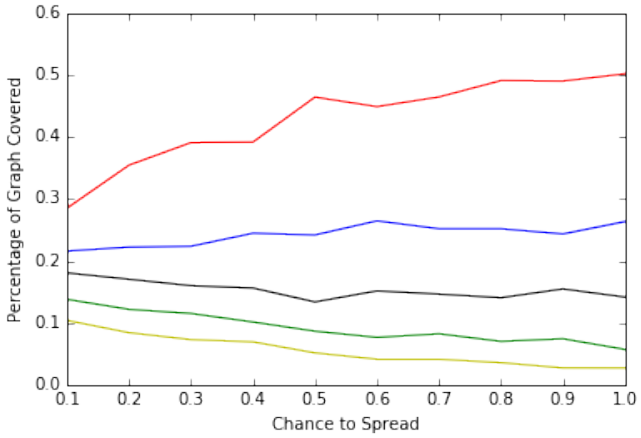



Figure 7: Graph of average amount of network covered vs change in probability to spread

We can see from this graph that the highest line starts at 28.5% and goes to 50.2% and the second goes from 21.6% to 26.3%. Further, the two least held opinions are decreasing as we increase the chance to spread. This supports our theory that the strongest opinion is affecting the least held opinions while the second most held opinion is able to remain steady.

Fixed the labels.

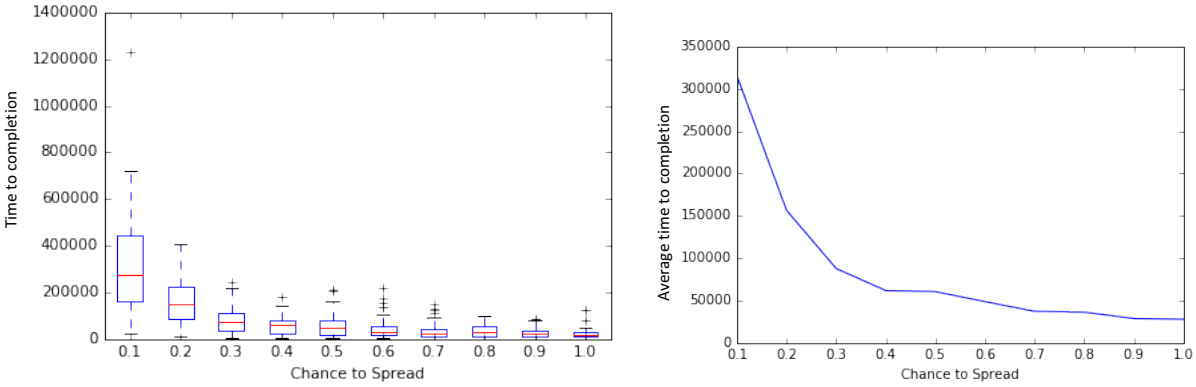



Figure 8: Box plot of time to complete simulation and graph of median time to completion

I attempted to work with these graphs but I couldn't it to work in python.

These graphs show the amount of time for a network to converge to one opinion. The simulation was run 50 times for each change in chance to spread. We then made a box plot of the time to completion. The graph on the right gives a better look at the median time to completion for each probability. We can see from the graph on the left that the lower the chance to spread the more variance in time to completion. With a 10% chance to spread the median is just above 300000 iterations but the interquartile range lies between 185000 and 415000. One interesting trend we observe is that the median time to completion remains steady after a 40% chance to spread. We also see less variation in the time to completion after a 20% chance to spread.

## 4 Conclusion and Directions for Further Research

One trend we can see is that communities tend to converge to one opinion. The higher the chance to spread the more likely it is that the entire network will converge to one opinion quickly. In addition, once an opinion becomes the majority opinion it becomes difficult for other opinions to spread at probabilities over 35%. However, below 35% and we see many graphs in which the majority opinion changes several times. We also see that important nodes are the ones driving the spread while less connected nodes help to reinforce the opinion in a community. It is difficult for opinions to bridge to new communities. We also see that the lower the probability of spreading information the more volatile a networks opinions are.

We don't often see convergence to a single opinion in real life. However, we do see streamlining of an opinion, that is similar opinions converging into one. This study might be more useful in studying an effect such as this where it is more likely that an opinion will spread and change than opinions that differ greatly. This study might also give us a deeper understanding of product trends. For example, we might be able to examine why a certain brand is more popular than its competitors. From our study we might be able to conclude that the more people that know of a certain brand the more likely they will be to talk about that brand with others or in terms of our study, a higher chance to spread information. Even if this conversation is negative towards the brand a higher chance to spread might mean a brand grows in popularity.

Access to the internet contributes to an interesting change in how information is spread. Over the past 30 years, access has gone from extremely limited to being readily available via mobile phones. This change has led to information being more easily obtainable. In addition, the past 10 years has seen a rise in online social networks. The fact that it is easier to connect to others could mean that information spreads at a higher rate. One factor that we did not study in this paper is stubbornness, or the willingness of a person to change their opinion. An interesting topic to explore further would be how stubbornness affects the spread of information in a network.

Our simulations aimed to study the effect that the probability of an opinion spreading has on a network. The simulation used the Barabasi-Albert model to create a network that we might see in the real world. There are many more factors that exist in these real life networks that could give us a better understanding of how information spreads. Further examining of variables such as these would lead to a greater understanding of how they affect a network. Another example of further research would be to gather and analyze data from a real network and compare it to the data created from simulations. Twitter could be used to map users who are connected to each other and the spread of a tweet or hashtag.

## References

[1] Newman, M., Networks: An Introduction, Oxford University Press, 2010.
[2] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008
[3] Jackson, M., Social and Economic Networks, Princeton University Press, 2008

## Appendices