Epidemiology Case Study

Margaret-Ann Seger, Boris Taratutin, Eun Abe Kim
December 16, 2011

Each fall, millions of college kids go back to school. And each fall, numerous colleges experience the fabled 'Freshman Plague' where a large percentage of the incoming students contract an unknown disease (usually some variety of influenza) that sweeps through the class like a plague.

University health offices often urge students to follow stringent health and safety guidelines to fight the spread of sickness. Occasionally, for illnesses such as norovirus or the H1N1 'Swine Flu' outbreak of 2009, students have been quarantined or entire campuses shut down. Quarantines and school closures are traumatic and disruptive, as well as expensive.

Imagine however, the university could predict how the disease would spread and take preventative actions with those individuals most likely to perpetuate the spread of the disease. These pre-emptive actions might take the form of immunization, quarantine, or extra health precautions.

In this chapter, we develop a way to pinpoint the most crucial students to target for such preventative measures, whose removal or vaccination minimize the spread of disease.

Disease Spread Models

In *The impacts of network topology on disease spread*, Mark Shirley describes a particular implementation of a disease spread model in an arbitrary network[1]. In the real world, when two people come into contact with each other, there is the potential of infecting one another. To model this, Shirley represents individuals as nodes, and represents the links between them as potential points for disease transmission. In Shirley's model, nodes can be in one of four states: 'healthy', 'infected', 'infectious', or 'immune'. For the purposes of the model, 'immune' includes both deceased individuals and individuals who have been infected but recovered from the disease and cannot contract it again.

These states, coupled with the following state transitions, are the basis for our implemen-

[1] To read more about Shirley's model, go to http://www.sciencedirect.com/science/article/pii/S1476945X05000280

tation of Shirley's model:

- 'Healthy' nodes adjacent to an infected node have a chance of becoming infected by infectious nodes (with probability $p_infection$)
- Infected nodes will eventually become infectious (after $latent_period$)
- Infectious nodes will eventually become immune (after $infectious_period$)
- Immune nodes can no longer infect other nodes or be infected themselves

Shirley tested his model on random, small-world, and scale-free networks. We chose to use a scale-free network as this topology most closely mimics the structure of a human network.

In his 1999 paper, *Emergence of Scaling in Random Networks*, Albert Barabasi introduces a method of preferential attachment for creating scale-free graphs and describes it as a network that is grown, where new nodes prefer to attach to nodes that already have a large number of connections. Barabasi argues that scale-free network topologies are useful for representing abstract systems like the World Wide Web and human social interactions, hence making scale-free graphs the logical choice as a base for our disease-spread model.

Simulation

Figure 1 shows Shirley's disease-transmission model on Barabasi's scale-free network.

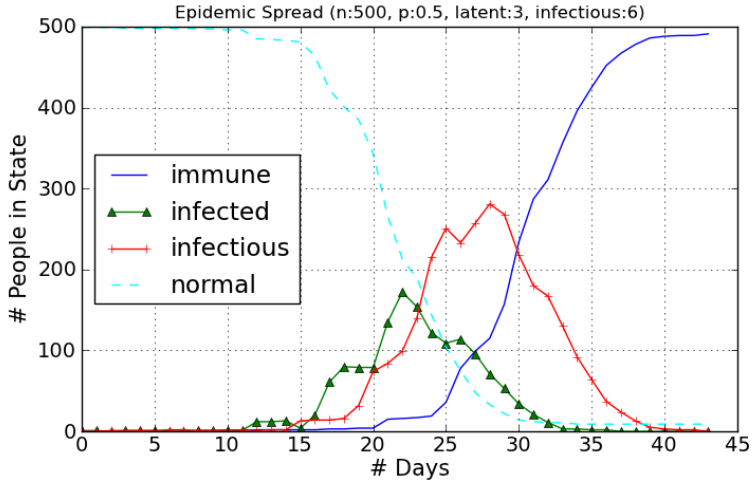

Figure 1: A plot of the number of people in each state of Shirley's disease spread model (normal', infected', infectious', and immune') over the course of an epidemic. The simulation parameters were as follows: $n = 500$, $p_infection = 0.5$, $latent_period = 3$, $infectious_period = 6$). Each line represents the number of people in that given state at any point in time.

The lines represent the number of normal, infected, infectious, and immune people at any given point in time. Looking at Figure 1, we see that the epidemic seems to grow as a logistic function. This is consistent with what happens in the real world where hubs' - very well-connected nodes - act to expedite the spread of disease.

Exercise 1: *Try using the Python file DiseaseSpreadModel.py to run the simulation with different parameters. How do different parameters affect the spread of disease? Note that each time the simulation is run, the generated graph has different shape and form from the one before. How does this randomness in the simulation affect the results? Run the simulation an appropriate number of times to see some convergence in results: how many times do you have to run it?*

Identifying the 'key' Nodes in an Epidemic

Alex Ng and Janet Efstathiou, in *The Structural Robustness of Complex Networks*, describe a technique whereby each node in a network is systematically removed, the disease-spread simulation run, and the final size of the epidemic calculated. We incorporated the same technique in our simulation to measure the effect of quarantining any given node on the final epidemic size.

We ran the simulation 500 times to account for variability in the network topology due to randomness in scale-free network 'growth' and in the disease spread model. We averaged these results to calculate an epidemic size for each removed node, and plotted the results in Figure 2.

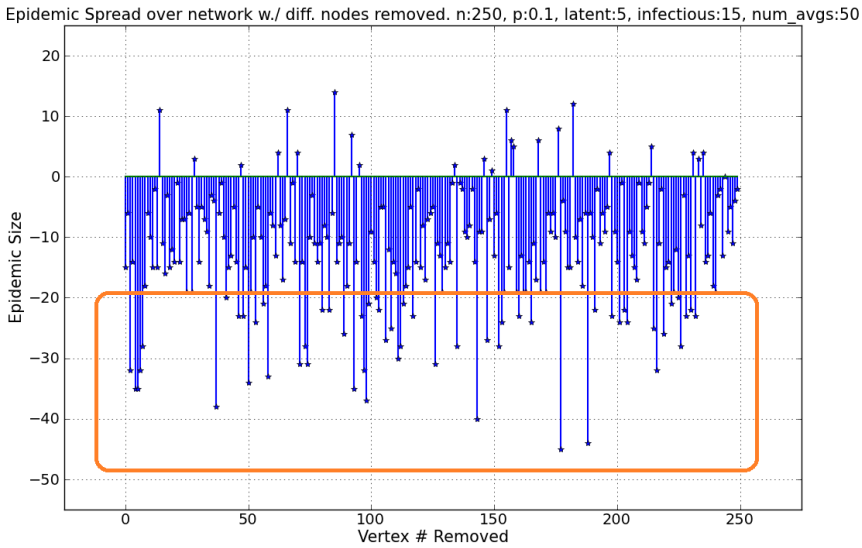

Figure 2: A plot of change in epidemic size for each node removed, for a randomly-generated scale-free graph topology. In this scenario, there are a significant number of nodes that could drastically reduce the size of the epidemic.

In Figure 2, the nodes highlighted in orange are the ones that, if removed, produce the largest reduction in epidemic size. We call these 'key' nodes.

Finding a More General Node-Removal Method

Up to this point, all of the results we have attained have used techniques that require perfect knowledge of the network. However, having a strategy that depends on a highly specific network is not the optimal solution for determining what will happen in the real world - as it is subject to small variances in network topology or simulation parameters, and runs the risk of being too much of an abstraction from the real world.

In a real-world scenario like a virus outbreak on a college campus, disease spreads through networks that are inherently complex and cannot be represented by a simple graph topology - making it very difficult, if not impossible, to discern individuals to vaccinate .

The natural extension then is to identify the general characteristics of these 'key' nodes, so that we can identify which nodes to vaccinate, without needing perfect knowledge of the underlying network structure. We concluded that a node's degree is a good and simple measure of 'importance'.

To evaluate the effectiveness of using node degree as a method for determining key nodes, we ran a test that plots epidemic size as a function of the number of nodes deleted from the network for three different node-removal techniques: a 'worst-case' random-selection method, a 'best case' exhaustive search method, and the node degree-based selection method we wanted to test. We hypothesized that degree-based selection would be more effective than random selection, but less effective than selection by exhaustive search.

In Figure 3, we ran a comparison that, one at a time, removed nodes from the network, using each of these three methods for choosing which nodes to remove, calculated the resulting epidemic size, and plotted the results.

Figure 3 shows that selecting nodes to vaccinate based on their degree is just as good as selecting them by exhaustive search, and far better than a random selection method. Furthermore, figure 3 shows that, with relatively few highly-connected nodes removed, the epidemic size drops dramatically. The implications of this are significant: in a college of 2,000 students, if an epidemic broke out, you could save 1000 students by vaccinating the top 100 most connected students.

Tying Back to the Real World

We now know that even if we don't have full knowledge of a network, we can accurately determine which nodes to vaccinate by selecting the ones that have the highest degree.

But what does having a 'high degree' mean in real life?

Though campus social media gurus may have a thousand or more digital 'friends' and 'followers', they would not be the key students to vaccinate in case of an epidemic. Rather, those students who are in close physical proximity to large numbers of people on a day-to-day

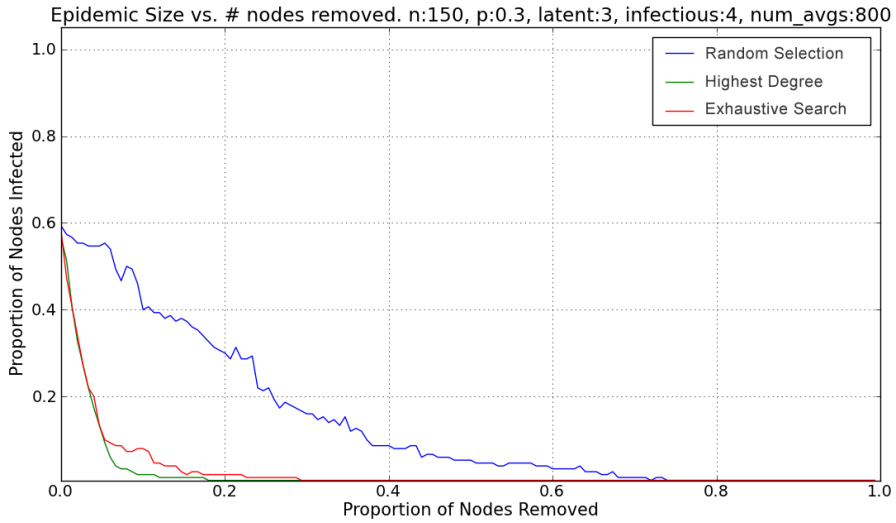

Figure 3: A plot of epidemic size as a function of how many nodes were removed. Three different techniques were used for selecting which order to remove nodes - random (worst-case), exhaustive search (best-case) and based on degree.

basis should be targeted.

This realization worsens the problem of pinpointing the correct students for immunization for college officials: almost all university students come into contact with a large number of other students on a day-to-day basis, whether this be in the dorms, communal kitchens, large cafeterias, or at after-hours frat parties.

Further Exploration and Exercises

In their research exploring the optimization of vaccination distribution, Dr. Jan Medlock and Dr. Alison Galvani present a very interesting theory about disease spread, which they present in their paper *Optimizing Influenza Vaccine Distribution*. Galvani and Medlock propose that schoolchildren, who are most responsible for transmission, should, not surprisingly, be prioritized for vaccination. However, they further suggest that parents of schoolchildren, aged 30 to 39, should also be prioritized, as they serve as the 'bridges' between their children and the rest of the population, which ultimately allow diseases to jump between schools and the workplace.

By looking for the bridges and most-connected nodes in a network, we can identify the most important types of nodes to vaccinate to minimize epidemic size, even if we don't have full knowledge of the network topology.

Exercise 2: *Read about the idea of bridges and see if you can find a data set that could help you identify what some real-world bridges might be. Some possible people you might think*

of are flight attendants or store clerks, who come into physical contact with a large number of people that may not normally interact with each other. Notice that you never talk about a specific individual as a bridge but rather a group of people, whether it be defined by their profession or the place they spend time in, etc.

Exercise 3: *When we made the decision to pursue node degree as a generalizable strategy for identifying key nodes, there were many other characteristics we could have explored. These include:*

1. **Node connectivity**: Node connectivity is defined to be, The minimum number of nodes which need to be removed to disconnect the remaining nodes from each other." (Wikipedia).
2. **Network diameter**: A network's diameter is the maximum of all of the shortest paths between any two nodes in the graph.
3. **Average shortest path length in network**: The average shortest path length is the average of all of the shortest path lengths between any two nodes in the network.
4. **Network clustering coefficient**: A network's clustering coefficient is, "a measure of degree to which nodes in a graph tend to cluster together." (Wikipedia).
5. **Link Load**: Link load is, "The amount of traffic carried by a link, or by the most heavily loaded link in the system (http://parallel.ru/info/reference/hpccgloss.html)

Many of these node and graph characteristics provide varying degrees of effectiveness in identifying key nodes for removal in disease spread networks. Experiment with using a few of these other characteristics to identify important nodes for quarantine. As a bonus, experiment with different graph topologies such as Watts-Strogatz, Erdos-Renyi, etc. and determine which characteristics are most effective for identifying key nodes.

Bibliography

Barabasi, Albert. *Emergence of Scaling in Random Networks.* 1999.
Downey, Allen. *Complexity and Computation.* 2011
Galvani and Medlock *Optimizing Influenza Vaccine Distribution.* 2009.
Ng and Efstathiou. *The Structural Robustness of Complex Networks.* 2006.
Shirley, Mark. *The impacts of network topology on disease spread.* 2005.