APPM 5515 Final Project

Brandon D. Finley
Applied Mathematics
University of Colorado, Boulder
Email: brfi3983@colorado.edu

Eugene Miller
Statistics and Data Science
University of Colorado, Boulder
Email: eumi4289@colorado.edu

February 28, 2021

Abstract

Graphs provide a useful abstraction of real life social networks, and the spread of communicable diseases. In our project, we analysed several real data sets representing social networks around the globe. Using this data we implemented the SIR model for epidemics and performed simulations of different scenarios. From these simulations we were able to draw important conclusions about how spread is affected by the strength of the infection and by vaccination.

1 Introduction

We were tasked with the analysis of several graphs from the perspective of epidemiologists. The graphs were collected from the SocioPatterns website. Downloaded MATLAB data files were converted to csv for use in Python 3. The most important libraries used in our project were NumPy for numerical calculation, pandas for data manipulation, and matplotlib.pyplot for plots. All code and results are available on GitHub.

2 Background

2.1 Network Statistics

Let $G = G(V, E)$ be the unweighted graph with vertex set V and edge set E. Then, we can define the following characteristics

- size = number of vertices $= n \overset{def}{=} |V|$
- number of edges = $m \overset{def}{=} |E|$
- volume = sum of weights on all edges = $\sum_{e \in E} w(e)$

Additionally, we also wish to define additional statistics that share insight into the geometry of our given network.

1. Density can be defined as the number of edges over the maximum possible number of edges (given n)

$$\text{density} \overset{def}{=} \frac{2m}{n(n-1)} \tag{1}$$

2. The dominant eigenvalue is the eigenvalue defined as the maximum absolute value. Thus, we can find λ by

$$\text{dominant eigenvalue} = \lambda \overset{def}{=} max\{|\lambda_1|, |\lambda_n|\} \tag{2}$$

with λ_1 being the largest (positive) and λ_n being the smallest (negative) eigenvalue where

$$\lambda_1 \geq \cdots \geq \lambda_n \tag{3}$$

3. The degree of a given vertex, v, is given by
$$\sum_{i=1}^{n} x_i \tag{4}$$
4. Additionally, we can find the clustering coefficient for a given vertex, v, with the following
$$\text{cc}(v) \overset{def}{=} \frac{2|\Delta(v)|}{\text{deg}(v)(\text{deg}(v)-1)} \tag{5}$$
where
$$\Delta(v) = \{(u,w)|u \in N(v), w \in N(v), (u,w) \in E\} \tag{6}$$
5. Using equation 5 and taking its mean value for a network encapsulates what is called the "transitivity" of the network. This is by considering mutual connections (explained more later on)
$$\bar{\text{cc}} \overset{def}{=} \frac{1}{n}\sum_{v \in V} \text{cc}(v) \tag{7}$$

Given these four statistics above, we can ultimately paint a picture of the network we are working with.

2.2 SIR Model

The model used was the Kermack-McKendrick epidemic model also known as the SIR model. This model considers the nodes n of V split into three subsets, where the variable t measures time elapsed since patient zero got infected.
$$n = S(t) + I(t) + R(t) \tag{8}$$
with

1. $S(t)$ is the number of nodes that are susceptible and can potentially be infected at time t (these nodes are not and have never been infected)
2. $I(t)$ measures the number of nodes currently infected and capable of spreading the infection
3. $R(t)$ is the number of recovered nodes that have been infected and are no longer able to spread of contract the infection

In order for the next section to read easier, we define the following variables and nomenclature for the SIR model

- β = infection rate
- μ = recovery rate
- Δt = size of time steps
- T = time interval for simulation
- A = adjacency matrix

Using the variables above, we take note of the following facts that have been previously derived

- If v is susceptible at time t and at least one of the neighbors of v (call it u) is currently infected, v is infected at time $t + \Delta t$ with a probability proportional to
$$\beta A_{uv} \Delta t, \tag{9}$$
where in our case $A_{uv} = 0$ or $A_{uv} = 1$ since we are working with an unweighted graph that does not take into account how many times a node is repeatably visited.
- If v is infected at time t, then v recovers at time $t + \Delta t$ with a probability proportional to
$$\mu \Delta t \tag{10}$$
- Once v is recovered it remains static and cannot be infected again

Now, if β represents the infection rate for a given epidemic, then taking its reciprocal should represent the mean time for a neighbor of an infected node to become infected. Specifically, for each vertex, v, we have

$$\frac{1}{\beta} \tag{11}$$

The mean time an infected node takes to recover can be modeled as

$$\frac{1}{\mu} \tag{12}$$

Therefore, the strongest epidemics will have large β and small μ.

We can also find the reproduction rate, that ultimately determines the extensive effect from the epidemic. This is done with the number

$$\rho_0 = \frac{\beta}{\mu}\bar{d}, \tag{13}$$

where $\bar{d}$ is the average degree of vertices in A. Also note that if $\rho_0 > 1$, then the epidemic will grow exponentially, if $\rho_0 < 1$, then the infection will die out, and if $\rho_0 = 1$, then it is called an *endemic*.

3 Results

3.1 Network Statistics

Below, we will focus on the first four evaluations defined above to model our network in a statistical sense. Most of these techniques are by conventional and signify various attributes of a network, such as connectivity, transitivity, overall size, etc. In section 3.2, we will do a numerical simulation as well.

3.1.1 Density

Next, we can simply take note of the density and dominant eigenvalue of each network as they are scalars. First, we have density for the contact networks in table 1.

Name	Density
InVS13	0.180
InVS15	0.182
LH10	0.406
LyonSchool	0.285
SFHH	0.118
Thiers13	0.109

Table 1: Contact Densities

Doing the same for the presence networks, we have table 2.

Name	Density
InVS13	0.877
InVS15	0.701
LH10	0.525
LyonSchool	0.912
SFHH	0.908
Thiers13	0.811

Table 2: Presence Densities

You will notice that the values in the contact network have much lower densities than the presence network. This generally makes sense since in the presence networks, there should naturally be more connections as the criteria for an “edge” is less restrictive. This gives rise to more edges and so by definition increases the values for densities.

3.1.2 Spectral Distribution

Now, we begin by looking at the spectral distribution between the two networks. Firstly, however, we wish to characterize the largest eigenvalue between the contact and presence networks. Looking at the contact networks first, we have that the dominant eigenvalues are given by table 3.

Name	Dominant Eigenvalue
InVS13	19.889
InVS15	46.110
LH10	37.711
LyonSchool	80.248
SFHH	65.601
Thiers13	41.232

Table 3: Dominant Contact Eigenvalues

Next, we look at the presence networks in table 4.

Name	Dominant Eigenvalue
InVS13	84.719
InVS15	161.774
LH10	44.577
LyonSchool	221.549
SFHH	371.806
Thiers13	269.388

Table 4: Dominant Presence Eigenvalues

Similar to the densities, you will notice that the dominant eigenvalues of the presence networks are much larger, on average, when compared to the contact networks. This, again, makes intuitive sense as with presence networks, it appears that the overall structure should be larger than contact networks. This is due to the larger number of edges created where not only certain nodes might have a higher number of edges, but also each node might have a larger possible walk, and so, the operator norm of the presence networks appears to be larger.

Next, we will be looking at the total distribution for cluster analysis. Namely, for the contact networks, you will notice in figure 1 that there is clearly some noise that collects around an eigenvalue of magnitude zero. However, being the circle defined by Wigner's semicircle law, we see that there exists clusters for a given network.

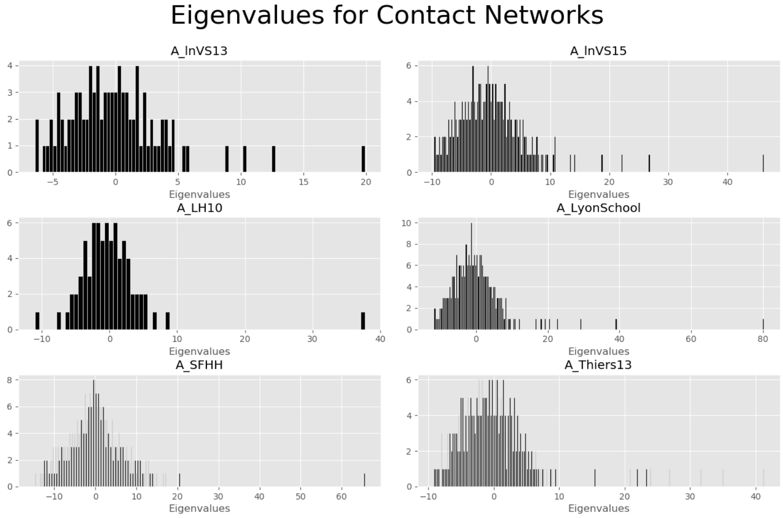


Figure 1: Spectral Distribution for Contact Networks

Looking at the distributions in figure 1, you will notice that the number of eigenvalues outside the defined circle of noise might relate to some physical clusters within the community. For instance, take A_LyonSchool, a primary school in Lyon, France. The number of clusters outside the circle appear to be the number 10. This could possibly signify 10 classrooms within the primary school where diseases spread within the given communities before moving onto the other parts of the network.

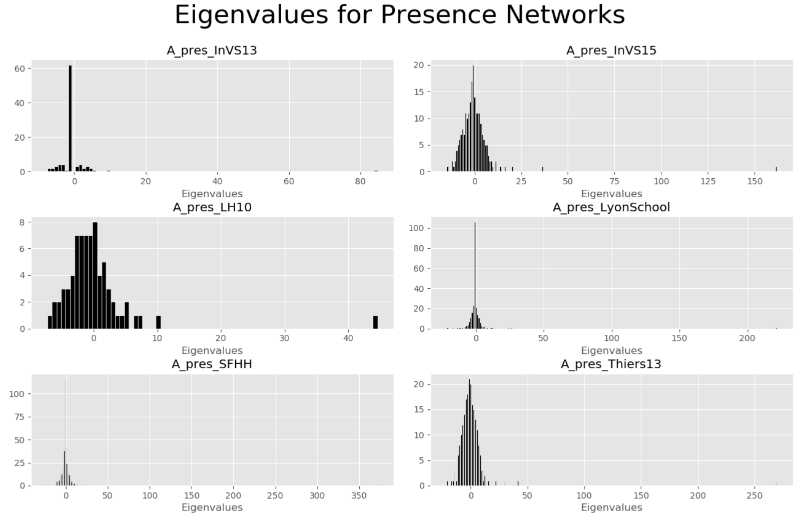


Figure 2: Spectral Distribution for Presence Networks

Next, we look at the presence network and compare its spectral distributions to that of those in the contact networks. Here, you will notice that the distributions are tightly concentrated around zero and that there does not appear to be any significant clusters. I believe this to be the case since in the presence networks, there is a form of "loose" connectivity. In other words, instead of defined communities where you clearly have to be in close proximity to be a part of them, the presence networks in a way "muddy" the connective tissue such that it is hard to distinguish when communities are present or just part of the overall structure of the graph.

3.1.3 Degree Distribution

Now, to further understand our network, it is useful to look at the degree distribution, that is, the number of edges each vertex has. This will give us a sense of connectivity in the graph as there might be select vertices that might have an enormous amount of friends or rather everyone who has about the same number of friends. Note that this only takes into account immediate connections and not mutual connections (that is the clustering coefficient).

We begin by looking at the degree distribution for the contact networks.

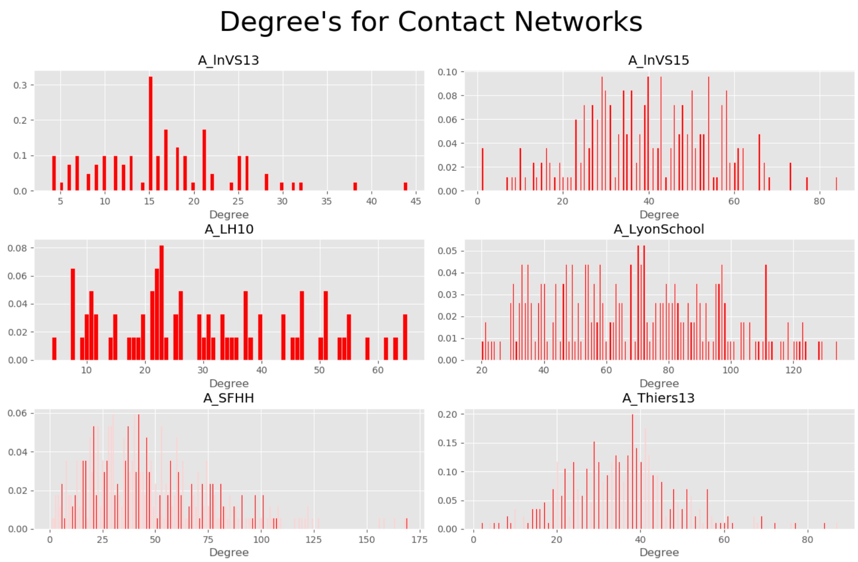


Figure 3: Degree Distribution for Contact Networks

In the contact networks, you will notice that there is some form of uniformity in the degree distributions. That is, neither does there exist “hubs” of high degree with most of the network with low connectivity nor large amounts of hubs of high degrees.

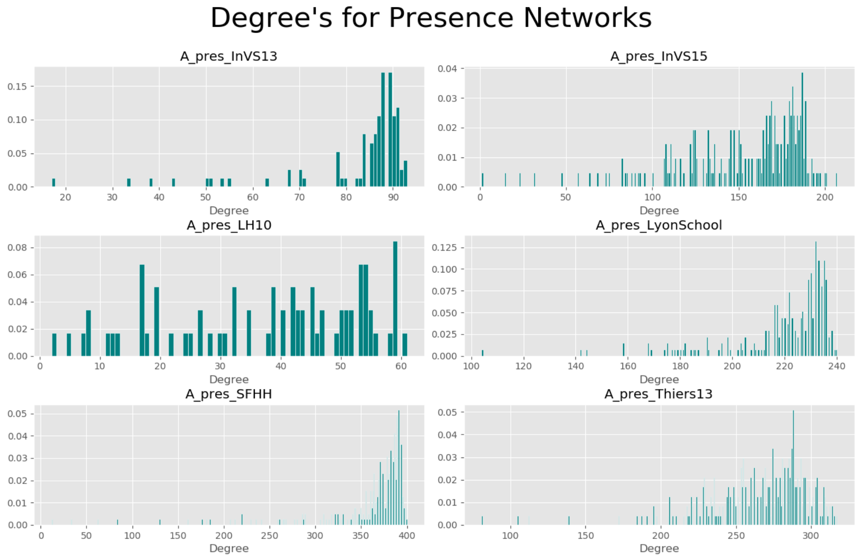


Figure 4: Degree Distribution for Presence Networks

Unlike the contact networks, there is clear skewness involved in these networks. And so, since the skew is to the left, it shows that there exist many hubs of high degree that would cause a propagation of the virus. This would cause an exponential increase in cases as the cases would multiply upon themselves if starting in a community of high degree. The exception to this observation is the network A_pres_LH10, where a uniform distribution is seen.

3.1.4 Clustering Coefficient Distribution

Finally, we have the clustering coefficient distribution. Although it sounds fairly similar to the degree distribution, the difference is that this takes into account mutual connections. Namely, the way our set is defined in the clustering coefficient formula, it takes into account what is known as *triangles* within a graph. This cardinality of the set can be calculated by raising our adjacency matrix to the third power. Specifically,

Definition 1. *Given an adjacency matrix A, one can find the number of walks of length k from node i to node j by looking at the i, j entry of A^k.*

Thus, in our case, we care about the triangles of node i and so we can look along the diagonal of the A^3 matrix. Using this fact and equation 5, we can find the clustering coefficient for each vertex v. Graphing the contact networks first yield the six distribution plots in figure 5.

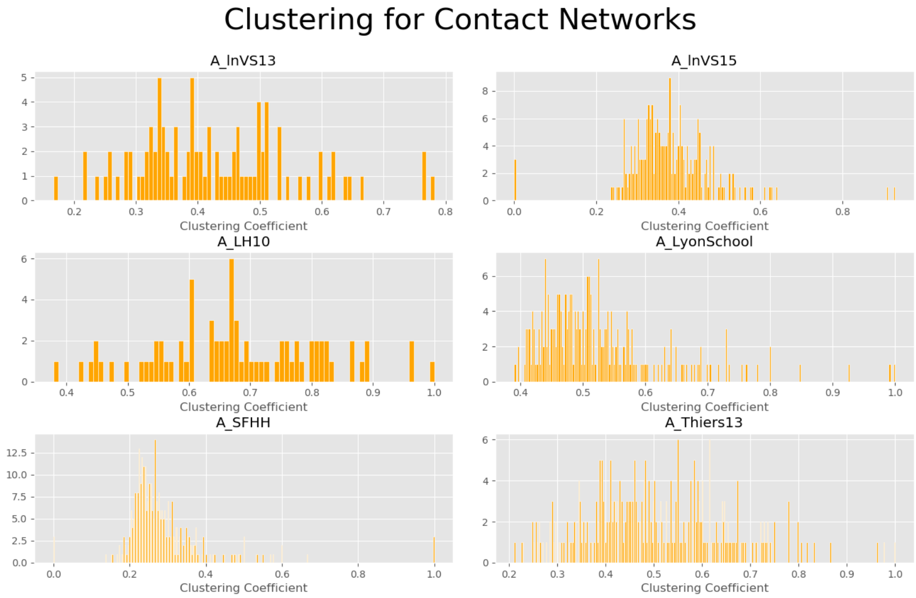


Figure 5: Cluster Coefficient Distribution for Contact Networks

From figure 5 you will notice that the distributions are fairly Poisson-like with some networks have more uniform distributions. It does appear that some are skewed right, and so, have various “hubs” with high transitivity, where most represent a lower number.

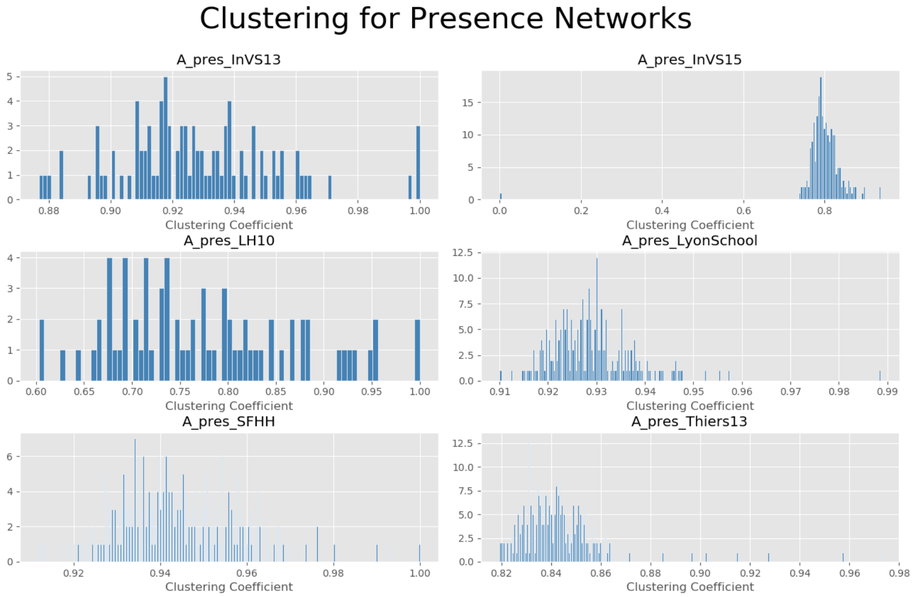


Figure 6: Cluster Coefficient Distribution for Presence Networks

At first glance, it appears that regardless of contact vs presence networks, the distributions look fairly comparable. However, upon further inspection, we see that the distribution in the presence networks are essentially shifted over to the right. In a way, this demonstrates that the global structure of the two networks appear to be the same; however, with the presence networks, there is an overall growth of clustering since their magnitudes of the clustering coefficients are higher.

3.1.5 Erdos-Renyi vs. Preferential Attachment

In this section, we will determine which graph model, Erdos-Renyi G(n,p) or Preferential Attachment G(n), to choose to model particular networks, both contact and presence. While some seem fairly obvious, others might not be a great fit for either models.

Additionally, if the case is that the best fit model is the Erdos-Renyi, then we have in the limit of large n, that is nodes, the clustering coefficient is p. For this we will take equation 7 as our estimate on p.

As mentioned above, a model for Erdos-Renyi is best used for when the degree distribution resembles a Poisson distribution. However, we note that this is the case for large n. In the case of smaller n's, we notice that the distributions are more uniform. This contrasts to the idea of the preferential attachment graph, where there is a skewed distribution, causing the growth of the virus to unfold at a higher rate. Using this as the backbone of our decision as well as 7, we will choose which model is appropriate for each network in table 5 and 6.

Name (Contact)	Model	p
InVS13	Erdos-Renyi	0.928
InVS15	Erdos-Renyi	0.799
LH10	Erdos-Renyi	0.776
LyonSchool	Erdos-Renyi	0.929
SFHH	Erdos-Renyi	0.944
Thiers13	Erdos-Renyi	0.843

Table 5: Modeling Contact Networks

Next, we model the presence networks. Notice here that for simplicity, I used the shorthand PA for Preferential Attachment. Apart from that, all other parameters are the same.

Name (Presence)	Model	p
InVS13	PA	NA
InVS15	PA	NA
LH10	Erdos-Renyi	0.776
LyonSchool	PA	NA
SFHH	PA	NA
Thiers13	PA	NA

Table 6: Modeling Presence Networks

3.2 SIR Model

Using the SIR model defined above, for each dataset, we will run 100 simulations with fixed parameters that measure the "spreadability" of a virus. This will be done for all contact networks, not presence networks. Additionally, note that patient zero will be selected at random to start the spread of the virus. We also have the following parameters.

- $\beta = 4 \cdot 10^{-4}$
- $\mu = 100 \cdot \frac{\beta}{k}, \quad k \in \{1, 2, 3, 4, 5\}$
- $\Delta t = 10^{-3} \cdot \frac{1}{\beta}$

Noting equation 13, we will plot the distribution of recovered nodes for each value of ρ_0. Notice that ρ_0 is dependent on our parameters, β, μ, and $\bar{d}$. While β and $\bar{d}$ are fixed, μ varies with the parameter k. Thus, for each dataset, we should have five different distributions.

In the order of the datasets given, we can begin our plotting.

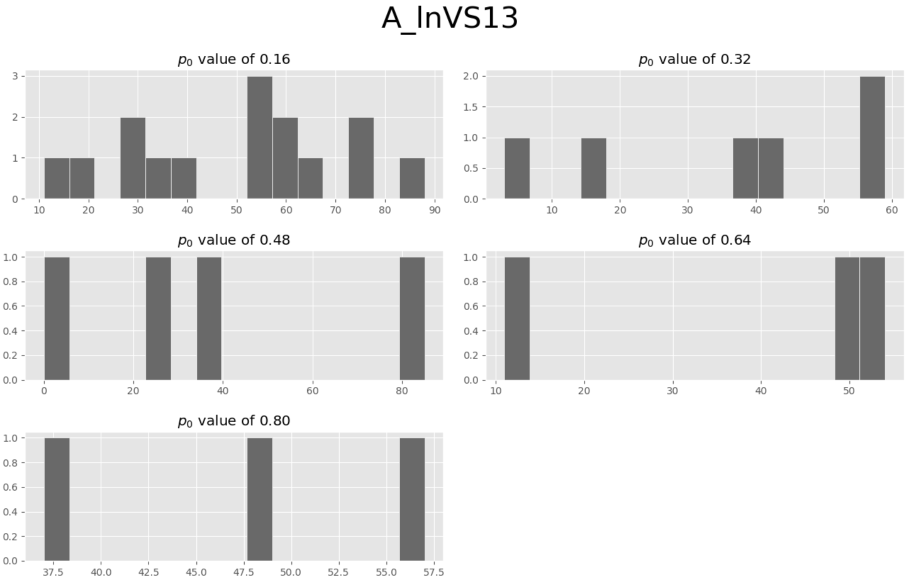


Figure 7: Recovered Distributions for InVS13

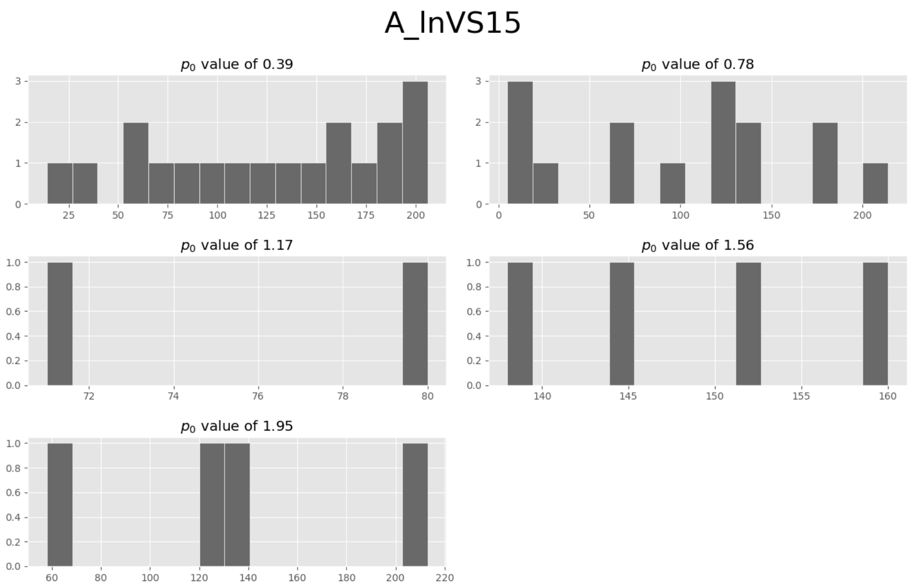


Figure 8: Recovered Distributions for InVS15

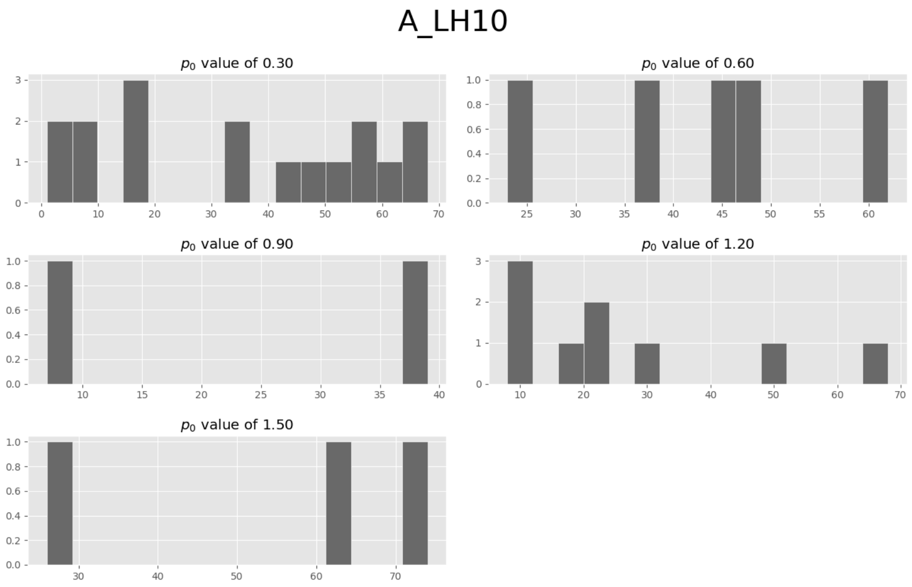


Figure 9: Recovered Distributions for A_LH10

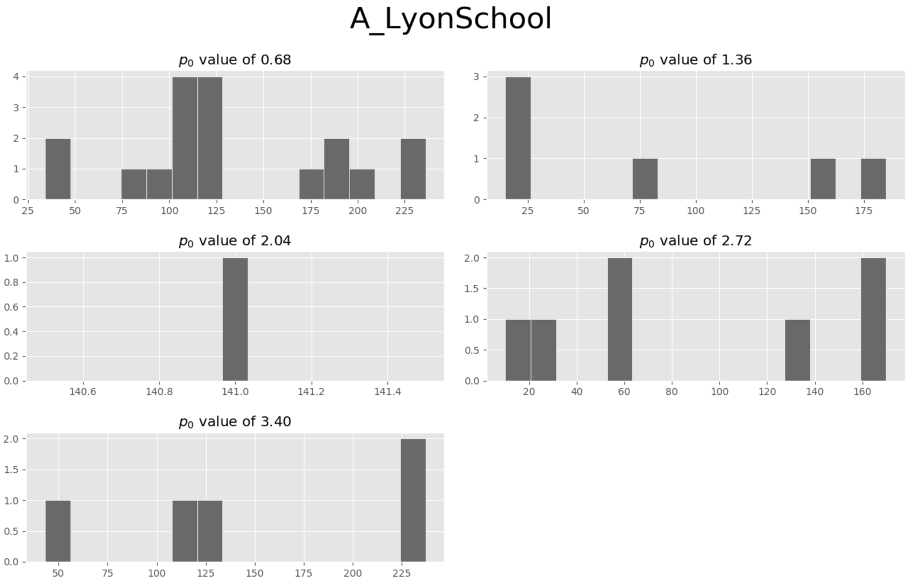


Figure 10: Recovered Distributions for LyonSchool

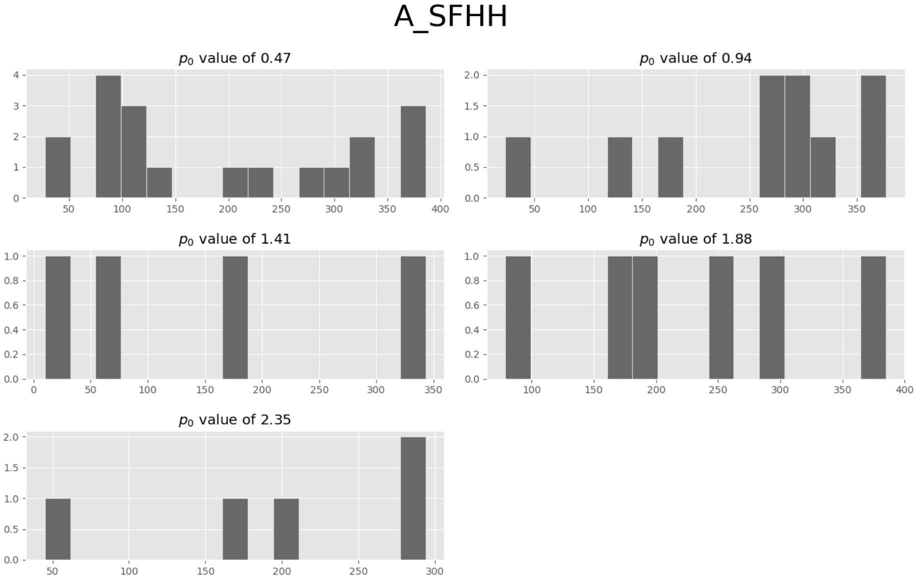


Figure 11: Recovered Distributions for A_SFHH

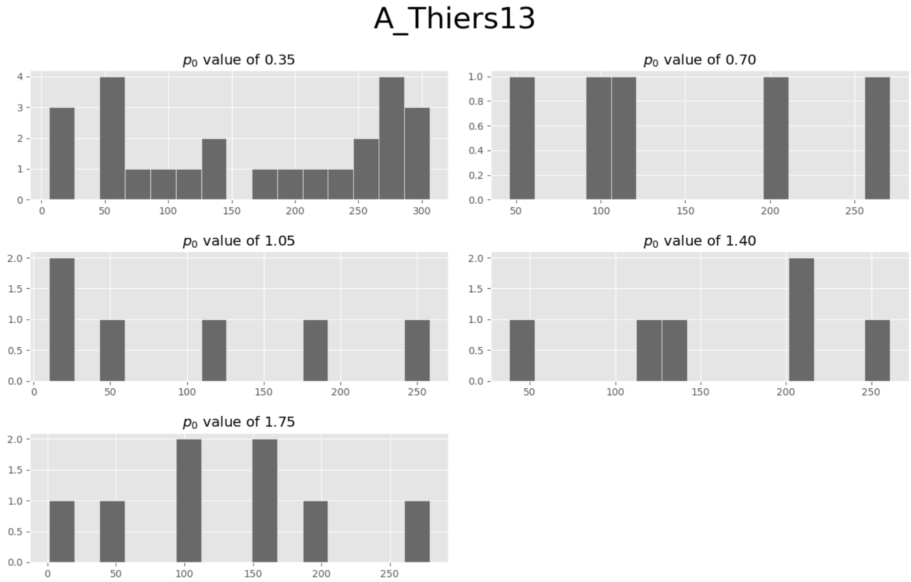


Figure 12: Recovered Distributions for A_Thiers13

From the graphs (7, 8, 9, 10, 11, 12), it appears that as we stipulated before (with ρ_0's relationship to the value 1), that for $\rho_0 > 1$, the number of recovered nodes were most skewed towards extreme values. This was not the case for when $\rho_0 < 1$, where the distribution seems to be fairly uniform across the board. This makes sense, as I will touch upon this later, since the more the growth of the virus, the more infected people and so if those people do not die out, then the more recovered people you have. Likewise, if it reaches herd immunity quickly, then there will be multiple nodes that never got infected and so never recovered. In short, with higher ρ_0 values, we see a skew to the extreme values while with lower ρ_0 values we see more uniformity.

Next, we wish to capture the fraction of epidemics where the final fraction of recovered nodes n_r is greater than 20% as a function of ρ_0. Unlike before, where we needed a histogram for each variation within the dataset, we can simply plot the scalar value, so that each dataset has its own distribution. Doing so, we have figure 13.

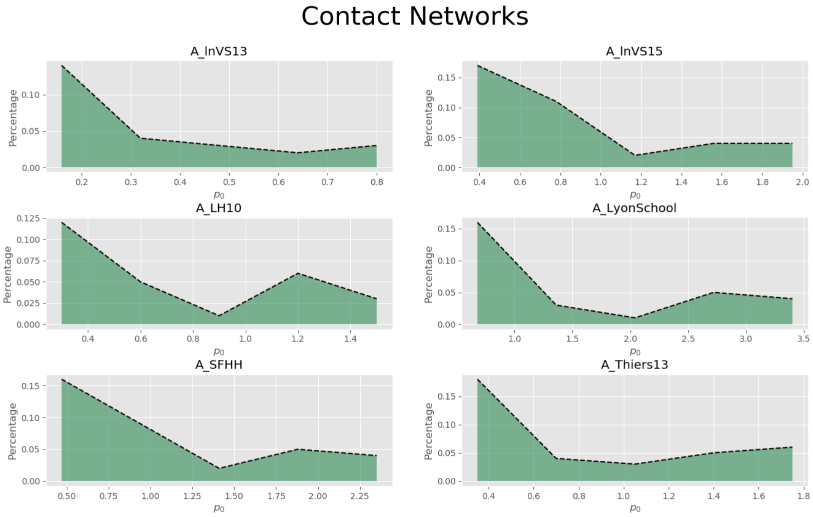


Figure 13: Fraction of epidemics where $n_r > 20\%$

Similar to figure 13, we also wish to compute the average number of recovered nodes for those epidemics that had more than 20% of their population infected. And so, we have figure 14.

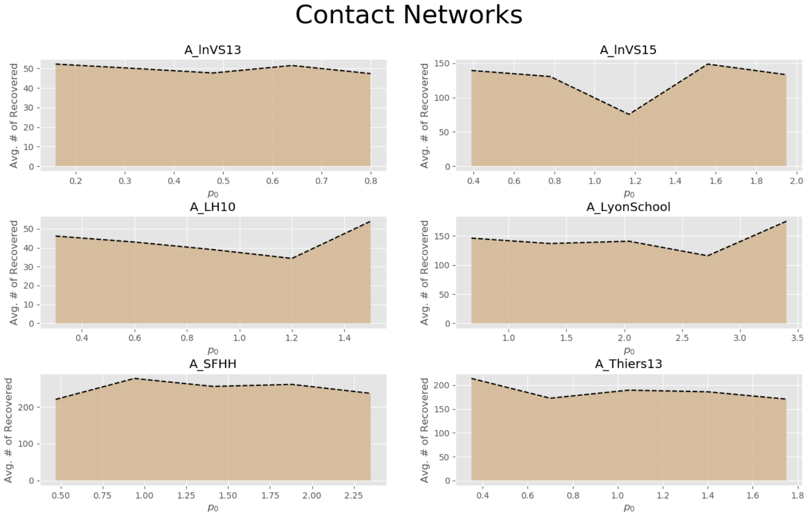


Figure 14: Average Number of Recovered Nodes for when $n_r > 20\%$

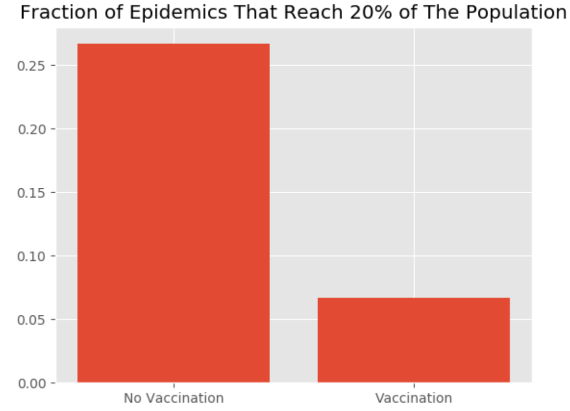
In general, we see that figure 13 has an overall downward trend, meaning that more epidemics that gave rise to $n_r > 20\%$ occurred when ρ_0 was small. Given that β is fixed, and that $\bar{d}$ was fixed for a given network, that when μ, the recovery rate was large, that more of the network was left recovered. This makes sense as we want our bodies to heal quickly before spreading the virus; I suspect this to have an even great effect in the presence networks.

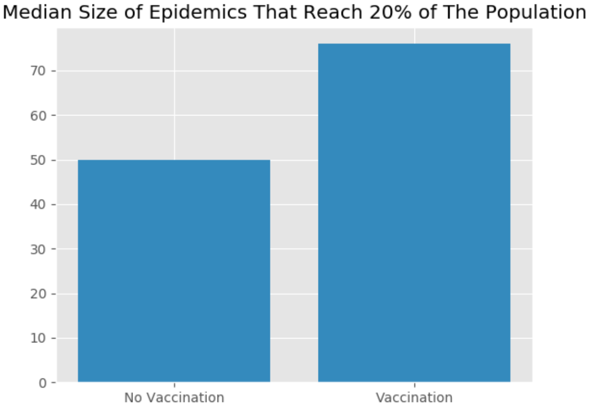
While a trend was noticed for the percentage of epidemics, a pattern within figure 14 appears to be a bit harder to recognize. Namely, the average number of recovered nodes appears to be independent to our value of ρ_0. For example, if the infection rate is larger, then a node would be less likely to infect other nodes, which mean that while the other nodes do not get infected and then heal afterwards, they could just never get infected. Thus, we end up with a larger amount of susceptible nodes that escape even having the virus. I suspect that this appears to be the case for each network.

3.3 SIR Model with Vaccination

To model the effect of a vaccination on the epidemic, the 20 nodes with highest degree were moved from S to R before the beginning of the simulation.
For each of the 6 data sets specified, 10 simulations were run using parameters:
$\beta = 4 \times 10^{-4}$
$\mu = 100 \times \beta$

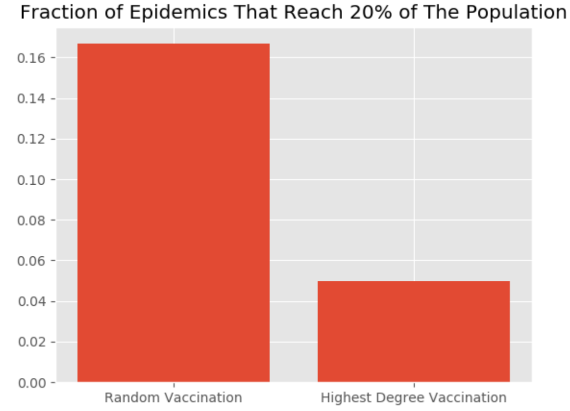
The number and largest size ($max_{t\in(0,T)}\{I(t)\}$) of epidemics that became large outbreaks that reached 20% of the population ($max_{t\in(0,T)}\{I(t)\} > 0.2 \times n$) was tracked across simulations. The following plots describe the results of the simulations:

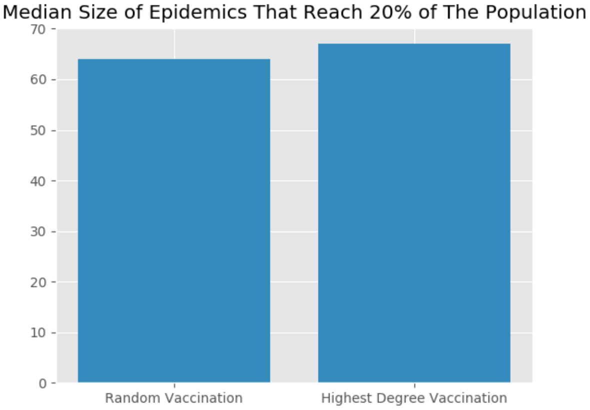




Although vaccination decreased the observed proportion of epidemics became large outbreaks, the median size of these outbreaks was larger for vaccinated populations than un-vaccinated ones.

An additional simulation was performed to compare the effect of highest degree vaccination to random vaccination. The same metrics were evaluated and are presented below:





The observed proportion of randomly vaccinated populations that reached large epidemic size was larger than the highest degree vaccinated populations. The median size of outbreaks in both simulations was about the same.

4 Conclusion

The effect of vaccination was demonstrated to have a negative effect on the observed number of epidemics that reach large size. Populations without vaccination were observed to be $\sim 5\times$ more likely to have large epidemics. Therefore, vaccination is preferable to no vaccination.

When considering whether to vaccinate randomly or by highest amount of contact, the simulation showed that random vaccination had $\sim 4\times$ the observed number of large epidemics compared to the highest degree vaccination. We can conclude that it is preferable to vaccinate according to individuals with highest contact.

It is important to note that vaccination did not prevent outbreaks from occurring. In fact, the large outbreaks that occurred in vaccinated populations had a larger median size than those in un-vaccinated ones. Despite this fact vaccination decreased the average number infected. In cases where diseases are potentially fatal, this means that vaccination decreases fatality.