CONTROLLING THE SPREAD OF DISEASE WITH NETWORK-BASED MODELS OF INFLUENCE

by

Matthew Jehrio

May 2022

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfilment of the requirements for the
degree of

Master of Science

Department of Biostatistics

Contents

Abstract

During epidemics, the ability of public health professionals and decision makers to minimize the impact of the contagion, depends, in large part, on their ability to quickly and efficiently deploy limited public health resources. This in turn, relies on the ability to quickly and accurately predict which areas and which sub-populations are most vulnerable in the context of an emerging epidemic. To this end, this research examines methodologies to identify such populations by leveraging existing algorithms that simulate a theoretical outbreak using individual scale resolution real-world data sets, as well as the PRINCE algorithm that predict the relative influence of nodes within a network.

1 Introduction

Methods for the control of infectious disease outbreaks are dependent on both effective disease surveillance mechanisms and accurate predictive models. Most recently, even with the advent of vaccines for the novel coronavirus, genetic mutations are still evolving (Mahase). This, combined with the fact that current production and distribution capabilities are still far from able to make vaccine doses readily available on the global scale (Asundi et al) means that classical methods of infection control are still critically important in both preventing infection transmission, as well as curtailing the spread of the virus to minimize the chance of newer, possibly more virulent and vaccine resistant mutant strains of the coronavirus (Greenhalgh et al.).

Networks can be used to model disease spread through a population.Vanunu et al. [2010] Networks allow simulation of contagious disease under different conditions by representing individuals in a population as nodes and the connections between them as links of transmission. Optimizing the parameters through simulating outbreaks allows researchers and policymakers to be better prepared to make evidence based decisions on policies protect the health and safety of the public.

One methodology involves modelling the transmission of infectious disease as a network of nodes and edges connecting them. The idea is to model transmission by simulating model outbreaks on this network and then see how various parameters change and react under differing initial conditions and constraints, and evaluating the extent of the resulting outbreak. This provides estimates for metrics that can ultimate be used to determine policy efficacy.

One of the most well known and widely used models is known as the Susceptible, Infected, Removed (SIR) model. Kuhl [2021] In the SIR framework, members of the population fall into one of three groups, susceptible, infected, or recovered/removed. Further, there is an assumption that, at a given rate, members of the population will transition from the susceptible state to the infected state; and with another probability members in the infected group will transition to the recovered/removed group. Different parameters of these transition probabilities and initial group sizes yield different dynamics for the overall behavior of the population as a whole throughout the course of the outbreak. This particular model can be generalized to include an entire class of models that might incorporate different classes for members of the population to fall into and/or different interactions between these classes.

While there has been previous work done on combining network analysis methods with SIR type models through simulation, there are still limitations to these methods. Some of the biggest gaps in understanding of these methods lie both in the design of the model and at the interpretation phase. Firstly, these models, as of yet, have difficulty in accommodating many of the different types of features that exist as part of the data collection process. The failure to capitalize on this unused information means that there are potentially many pieces of the pandemic puzzle that are being left out of the equation. Secondly, for the purposes of ultimately impacting public health policy, there needs to be an effective way of estimating the relative contribution that different nodes contribute to the disease burden so that public health officials can more optimally allocate resources such as testing.

In this paper, we combine both network simulation methodology with a network analysis algorithm known as PRINCE (PRIoritizatioN and Complex Elucidation).Vanunu et al. [2010] Additionally, we leverage the results of Firth et al. in "Using a real-world network to model localized COVID-19 control strategies" that created a network simulation model that utilizes geolocation data collected from residents in the town of Haslemere.

In the context of network science, SIR type models can be integrated by assigning each individual node as the susceptible, infected, or removed/recovered. The structure of the network itself gives rise to more interesting dynamics within the population throughout the course of the outbreak on account of the fact that there are now specific edges that connect the individual nodes as compared to either model separately. The advantage here is that now there is a structure to observe in finer detail the interactions between individual nodes.

Consequently, researchers can also investigate the differential effects of population structures within the community and investigate what effects, if any these structures have in the overall course of the outbreak.

During an outbreak such as the current coronavirus pandemic, there are many potential groups that the members of the population could fall in, susceptible, infected, quarantined, recovered etc. There has been work by previous studies to identify local, person level transmission dynamics. Vanunu et al. [2010] Petra Klepac [2018] In this work members of the community were recruited to have their movements traced by geolocation software, and their data and interactions recorded, as part of the BBC program "Contagion". Petra Klepac [2018] The aim was for this granular data to better predict the behavior of the infection on a local level, and that this fine grained level of specificity would yield a better understanding of the actionable steps that can be taken to reduce the potential for transmission. Further, the work done by Vanunu et al. constructs the transmission models themselves. These models were network based and utilized probability functions to calculate and adjust the risk of disease transmission.

However, a problem with this approach is that there is a lot of potential information that is left unused. In a simple network, there is no information about the structure of the network. Ideally, a network based model would be able to incorporate this kind of information and become more informative and effective in understanding intervention. PRIoritizatioN and Complex Elucidation (PRINCE) is an algorithm for probing a network to see the relative importance of all the nodes in the network, and see which nodes appear to be most influential. PRINCE was originally developed to study prioritization of proteins based on previously available information. This was accomplished by repeatedly calculating and updating a prioritization function with respect to a matrix of preexisting similarity metrics. Vanunu et al. [2010] It is this point that allows f or the flexibility in incorporating multiple features in the network model.

In this work, we leverage the methods of both the PRINCE algorithm by Vanunu et al. and the Haslemere network data and methodology of Firth et al to explore strategies for controlling outbreak spread, as well as minimizing the amount of necessary restrictions on social life. To do this, we utilized the Prince algorithm to determine which nodes in the Haslemere network structure were likely to be most influential in the ultimate propagation of the contagion over the course of the outbreak simulation. Then, we considered the effects of isolating the individual nodes that are present in the Haslemere outbreak sim-

ulation sequentially in order of their predicted influence from the PRINCE algorithm. By examining the degree to which the epidemic is contained by the removal of each number of influential nodes in the Haslemere network and data set model, we ascertain the degree to which the PRINCE algorithm is able to identify the top priority nodes, an approximation of the distribution of the relative importance of the entire set of nodes present in the simulation, the properties of the nodes in the network that are associated with being a high priority and thus a risk of becoming a *super spreader*, and an estimate of the number of high priority nodes/ what percentage of high priority nodes are necessary to isolate to contain an epidemic. Results show that PRINCE is effective in identifying high priority nodes within the network. Further, prophylactic isolation of the highest priority nodes as selected by PRINCE is shown to have a substantial ability to stem the transmission rates and overall disease burden, especially in the case of isolating nodes that are not yet infected.

2 Materials and Methods

Networks can be represented as an $N \times N$ adjacency matrix. While these network models can certainly be useful in and of themselves for analyzing the structure of a population, a further advancement in their application is dynamic network modelling. In these models, the networks do not represent just a static state, but a population at a single point in time. In these models, whether or not any individual node will transition is governed by probabilistic models that rely on, in part, the values of the nodes most adjacent to the individual node in question.

A central question in network analysis is predicting which nodes have the greatest influence. Contrary to popular belief, these nodes are not always central hub nodes that are highly connected. The PRioritization and Complex Elucidation (PRINCE) algorithm is a propagation based method to score and rank nodes based on their relative influence in a given network. Firth [2020] In our approaches, the adjacency matrix is the Haslemere network. Petra Klepac [2018] This particular network consists of 468 nodes and was constructed using GPS tracking of participants in a study conducted in in Haslemere England, where the edges of the network represent close contact encounters between participants.

2.1 SIR Model

The SIR model of infectious disease propagation is a well known and widely used tool to simulate disease spread and ultimately control outbreaks. In the SIR model, individuals are categorized as being in the Susceptible (S), Infected(I), or Recovered/Removed (R) category. In the model, Susceptible individuals will, at a given rate, be converted from their original susceptible status to having an infected status. Similarly, people in this model that are currently in the infected group will, at a different constant rate, be converted into the recovered/removed group.

Figure 1: Diagram of SIR model state transitions

The rates of conversion between the susceptible and infected groups, as well as the rate of conversion from the infected group to the recovered/removed group are both determined by the individual characteristics of the contagion itself and its ability to spread within a population, as well as the current size of the population itself. Given a set of initial populations, as well as the relative rates at which people in the model convert from each group for a specific contagion, the SIR model then predicts the behavior and ultimate dynamics of the system and the people in them.

2.2 Node Prioritization

In this method, the authors propose a new method for determining which nodes in a network are most likely to be influential in a dynamic system such as a dynamic network model. Although the paper itself originally focused on protein-protein interaction networks, the framework is easily adapted to other networks, such as the Haslemere, England COVID-19 propagation network reviewed here.

The PRINCE algorithm works by constructing an adjacency matrix of all the nodes in a given system. In the original paper, Vanunu et al. use the reliability of a given node interaction as weights for this adjacency matrix. The PRINCE algorithm itself normalizes the weights to values in the interval [0,1]. The algorithm relies on iteratively computing the value of the equation:

$$F' = \alpha W' F^{t-1} + (1 - \alpha)Y$$

where F^{t-1} is the value of the prioritizations at the previous step, Y represents an apriori estimate of known interactivity, W' represents the adjacency matrix.

2.3 Simulations

The primary objective of our simulations is to investigate the abilities of PRINCE in predicting influential nodes to preemptively isolate. As such, the tuning parameters of the outbreak model were set so as to minimize compounding factors. As such, the parameter controlling for testing was turned off, the parameter that controls for the probability that a node will become infected at a given step as a result of infection from outside of the model itself was set to a low value at .001. The proportion of cases that are asymptomatic was set to .94 and the proportion of cases that were presymptomatic, that is, displaying symptoms before the period in which the individual is contagious was .2. We did assume that a certain proportion of people that are positive enter quarantine, and another proportion enter isolation. In this model, neither quarantine nor isolation completely removes the possibility that the infected individual will transmit the infection, but the probability that transmission will occur is reduced. This is meant to simulate the fact that people in isolation are oftentimes still physically close to members of their own households etc.

The parameters of the baseline model wereset such that nodes were isolated, but not quarantined upon testing positive. The probability of a node becoming infected from a source outside of the network is .002. Three scenarios were considered using the Haslemere network:

1. Isolation nodes that were initially infected based on the node's score from the PRINCE algorithm.

2. Isolated the susceptible nodes on a prophylactic basis.

3. Isolate all of the susceptible nodes as well as an increasing number of susceptible nodes.

For each scenario, a random model that represents an isolation strategy that is uninformed by the underlying network structure was used as a baseline for comparison. These random models isolated the same number of nodes from the same population subset, i.e. infected/susceptible, but with the exception that in the null model the selection process is random. As an additional point of comparison, we also computed a baseline model that represents the condition of no prophylactic isolation.

100 simulations were performed for each combination of parameters in each scenario. We rerun the simulation and isolate a sequentially increasing number of nodes that are initially infected for each case. For example, on the first simulation the highest priority infected node in the scenario group is isolated, on the second iteration the two highest priority nodes are isolated. Nodes were isolated as a percentage of the total number of initial infections in increments of 5 from 10 to 100 for the models isolating infected nodes, and 10 to 150 for models isolating susceptible nodes. Throughout each iteration of the simulation, the number of new infections was recorded, and the total number of new infections at the end of the simulation was stored. This process was repeated for each scenario.

2.4 Setup

The models we run here start with the network structure as defined by the data collected in the Haslemere data set. Petra Klepac [2018]. From here, an outbreak is simulated according to the methods prescribed by Vanunu et al. . In addition to the probability of transmission to nodes that are adjacent to infected nodes, this model also takes into account a variety of other factors. Once a node is denoted infected, there is an incubation period of a certain number of days during which the node is infectious, but not symptomatic. There is a parameter in the model that allows for adjusting the fraction of nodes that are in this incubation period will be presymptomatic. Once a node becomes symptomatic there is an option to trace first degree contacts of the infected node, as well as a parameter to contact trace the second degree contacts of the infected node. During contact tracing, there is a parameter to control the false negative rate. Upon contact tracing, there is the option to either isolate or quarantine infected nodes. For this model, isolation and quarantine both correspond to lowering the probability that the infection will spread between the infected node to the adjacent nodes. Quarantine reduces the probability of transmission lower than isolation, but

neither state entire eliminates the probability of transmission. In a real world context this could easily be seen as living in a common household, among other explanations.

The probability of transmission to nodes adjacent to a previously infected nodes is governed by a Weibull distribution which in these experiments had fixed shape and scale parameters of 1 and .5. If the returned value was greater than .5, transmission between nodes occurred. The remaining tuning parameters are either scalars such that $0 \leq x \leq 1$, or Boolean values.

Following Klepec, the number of new infections after each step was recorded. In the cases of running the model under the PRINCE and random isolation strategies, the nodes that are selected for preemptive isolation are put into isolation before the model commences. Isolation in the PRINCE cases preemptively isolates the top n ranked nodes in the specified population using PRINCE, and the random preemptive isolation strategy isolates nodes within the same subpopulation with equal probability.

3 Results

3.1 Histograms

Isolation of Infected Nodes Model

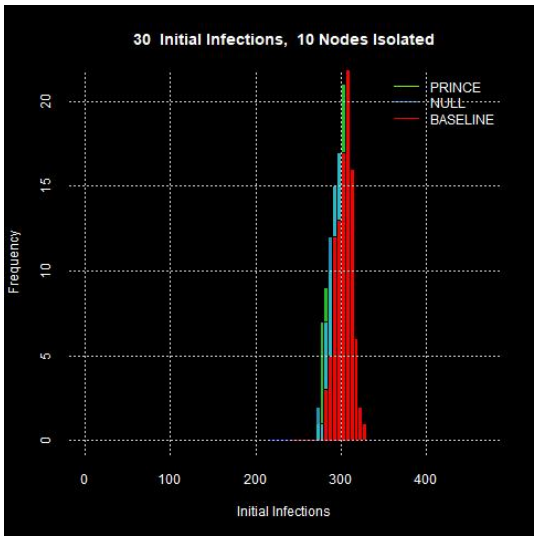

Figure 2: 30 initial infections, with 10 Infected Nodes Isolated

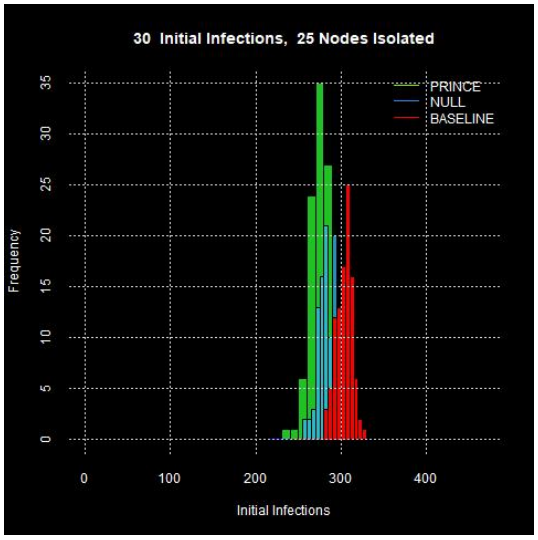

Figure 3: 30 initial infections, with 25 Infected Nodes Isolated

To assess, as a rough estimate of the scale of variation with respect to the change in input parameters, this entire simulation was run multiple times with different initial conditions corresponding to relatively high and relatively low values of initial infections and prophylactic isolations. Figures 2 - 5 show the histograms for the model corresponding to different numbers of initially infected nodes being isolated.

Isolation of Susceptible Nodes

Figures 6- 8 show the histograms of the results where susceptible nodes were isolated. We can see further that when we increase the number of initial infections, that the model using PRINCE appears to perform differentially well against not only the baseline model with no nodes isolated, but also against a null model in which random nodes are isolated. In addition to a reduction in the mean number of cases, there is also a skewness that is added to the data under the PRINCE model isolation strategy.

3.2 Multiple Point Model

The simulation was run across consecutive initial infection values. The resultant number of infections was recorded after each simulation. Firstly, we examined the figures that plot the total number of infections after the model has run against the percentage of the initially infected nodes that were isolated.

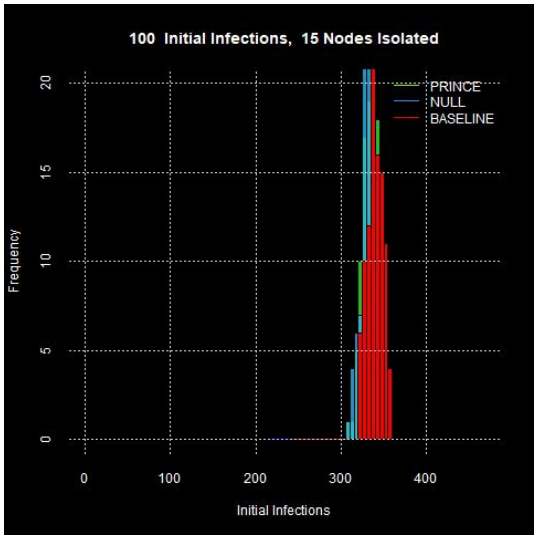

Figure 4: 100 initial infections, with 15 Infected Nodes Isolated

3.3 Total Infections by Initial Infections and Number of Isolations

The next show the average number of resultant infections over 100 iterations of he simulated outbreak for given number of initial infections. The error bars on the figures represent one standard deviation from the point estimate of the mean for all the iterations.

Isolating Infected Nodes

The plots in figures 9- 11 represent the condition that the prophylactic isolations were made among the population of nodes that were infected at the beginning of the outbreak, with the PRINCE model, Null model, and baseline with no isolations being represented.

As would be expected, for relatively small initial infection values, there is hardly any difference at all between the PRINCE model and the Null model. Also consistent with intuition is the fact that the PRINCE model and Null model perform very close to the baseline results.

Although there appears to be the beginnings of separation between the PRINCE model and the Null model with an increasing number of initial infections present, in even up to 100 initial cases,the models have similar performance with overlapping error bars.

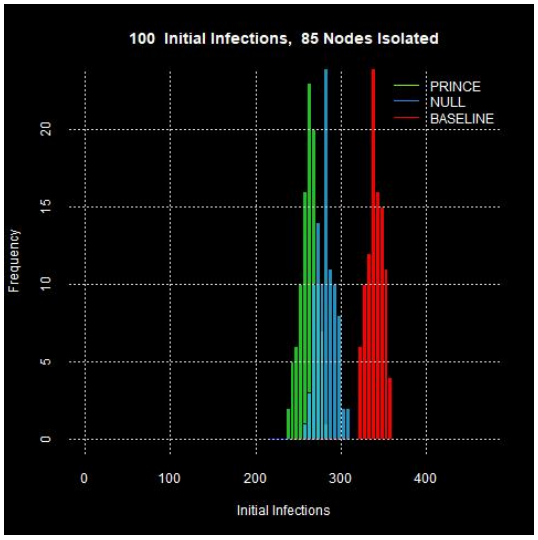

Figure 5: 100 initial infections, with 85 Infected Nodes Isolated

Isolating Susceptible Nodes

The plots in figures 12 - 14 correspond to a model in which nodes are proactively isolated from the population that is initially in a susceptible state at the start of the simulation. This includes both the PRINCE and Null methods of selection, as well as, again, a baseline model for the sake of comparison. As a note, while the x-axes for the previous case were variable with the magnitude of the number of initial infections, such a restriction is not relevant when isolating susceptible nodes. In the following figures, the x-axes are both fixed and greater than the number of initially infected nodes.

Here, we begin to see results that are interesting. Initially, we can see by visual inspection that, in this case, selecting nodes via the PRINCE algorithm yields results there are considerably more favorable over the results of not just the baseline model, but the Null model as well.

There are also differences with respect to both the number of initial infections and with respect to the number of initial isolations.

In the first figure, there is an increasing difference in efficacy between the PRINCE model and the Null model as the number of initial isolations increases. This is suggestive of PRINCE's effectiveness at identifying high priority nodes, as well as the cumulative effect of isolating a group of higher priority nodes as opposed to random nodes.

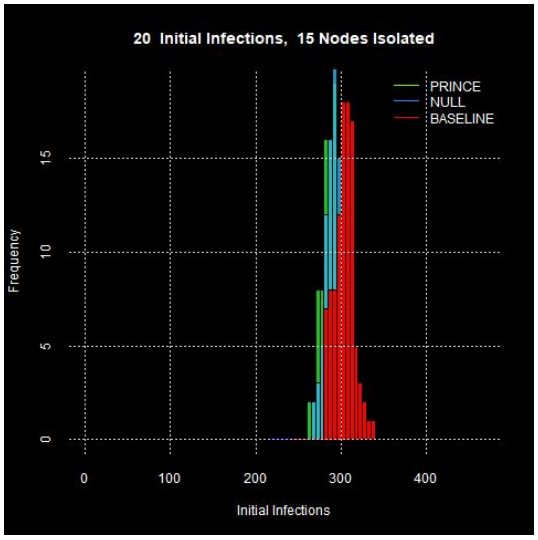

Figure 6: 20 initial infections, with 15 Infected Nodes Isolated

In the other two figures, we can see that the differential benefit of selecting nodes to isolate via the PRINCE algorithm does decrease to a certain extent, but there is still a clear value in isolating nodes based on PRINCE prioritization. While there is a relative decrease in potency associated with an increased number of initial infections, it is important to bear in mind that population size of this model was 468 nodes, that this model simulated 70 days of a pandemic, and that people can contract the virus more than once. When applying this model to a real world public health situation, there would be far more nodes involved. Depending on when these strategies are applied in a pandemic, there is the possibility that the behavior under public health guidance will more closely resemble that of the first figure, given the ratio of susceptible to infected people is sufficiently high. Additionally, because the usefulness of prioritizing nodes via PRINCE is dependent on the current state of the system, and because this model simulated 70 days of an outbreak, there is potential for an increased differential with an outbreak that lasts for a longer time, as would be the case in a real world pandemic setting.

Isolating Infected and Susceptible Nodes

Finally, we consider the case where all of the initially infected nodes are isolated, and a certain percentage of susceptible nodes. This is done for the purpose of testing PRINCE's ability to predict nodes that are influential above and beyond a random isolation strategy in a situation with a more tightly constrained environment.

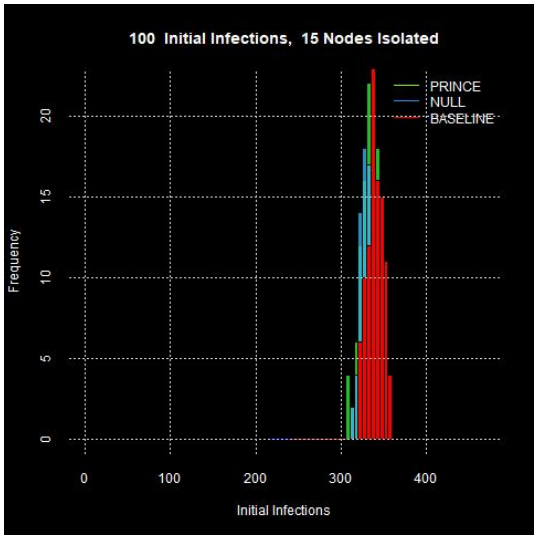

Figure 7: 100 initial infections, with 15 Susceptible Nodes Isolated

Figure 15 shows the results for running the model under the above specifications with 30 initial infections present. As can be anticipated, isolating all initial infections decreased the total number of resultant infections relative to the no isolation baseline case. However, there is also a decrease in the relative effect of using PRINCE as opposed to a random null model. This is consistent with the idea that a lower amount of total disease spread blunts the relative effect of isolation. However, there is still a decently large signal left from using PRINCE instead of a random null model.

Figure 16 shows the results of applying the same conditions as before, but with 100 initial infections. As shown in the figure, this represents the hardest test PRINCE was put through here. With so many nodes already isolated, there is less latitude for any preemptive isolation model to demonstrate a differential effect on the total number of resultant isolations. While there is a decrease in the differential, there still does appear to be a noticeable positive effect of using PRINCE, even in this ”Worst-Case” Scenario. As 100 nodes is also approximately a quarter of the simulated population, the implication is that PRINCE still provides an advantage even if you begin using it as an isolation strategy later on in the course of an outbreak.

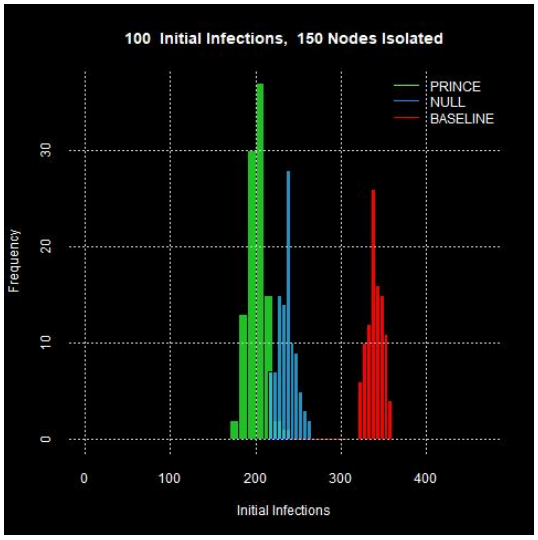

Figure 8: 100 initial infections, with 150 Susceptible Nodes Isolated

4 Discussion and Future Work

From the results presented above, we can see that the PRINCE algorithm has significant ability to predict nodes that are most influential in terms of the structure of the network itself, as well as on a process propagated across it. Further, this model was trained on real world geolocation data that represents a specific subset of the population's actual network structure. Training models on actual interaction data is critically important as it allows a more realistic view of how prevention and management strategies perform and compare in the real world. This is of utmost importance for public health officials, decision makers, and to a certain extent extent the general public in stemming the spread of disease. Furthermore, the approach shown here is useful not just in its efficacy at identifying nodes that are likely to become critical in an actual outbreak, but also in that it can be used to estimate a minimal set of nodes that need to be quarantined to reach certain public health benchmarks. In an outbreak in the real world, the reality is that public health resources are already limited, and further strain on these resources necessitates a carefully devised delegation of these resources. The results of this paper provide a framework with which to structure this delegation and how to best manage critical resources in a way that maximizes public safety and minimizes the ultimate impact of an outbreak.

While the analysis here provides evidence that the PRINCE algorithm has significant ability to predict which nodes will be the most critical in focusing public health resources in the context of an emerging pandemic, there are still many facets of both the model it-

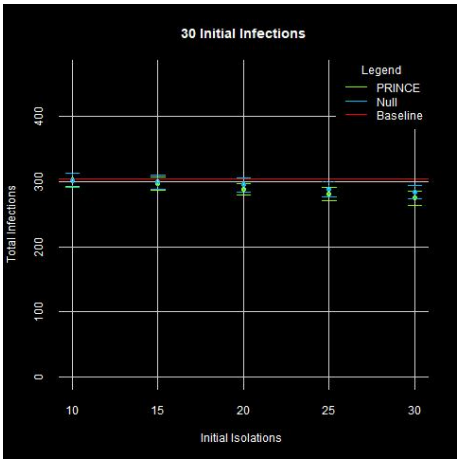

Figure 9: 30 initial infections; isolate infected nodes

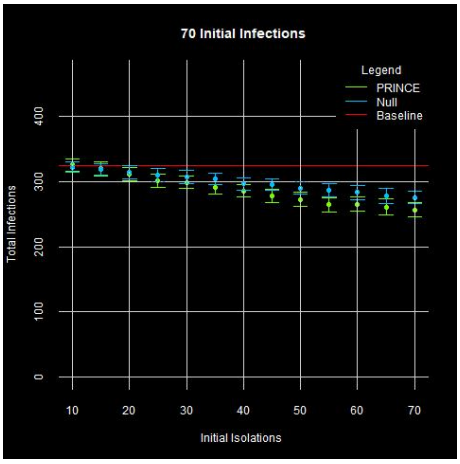

Figure 10: 70 initial infections; isolate susceptible nodes; isolate infected nodes

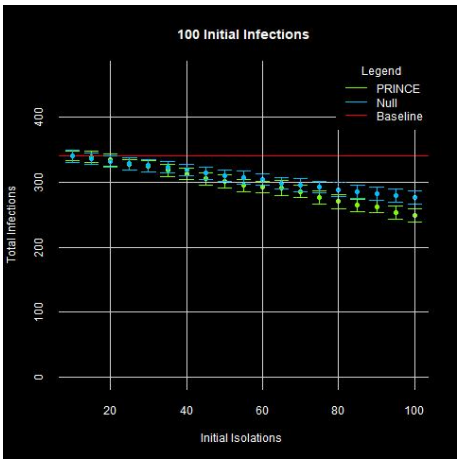

Figure 11: 100 initial infections ; isolate susceptible nodes

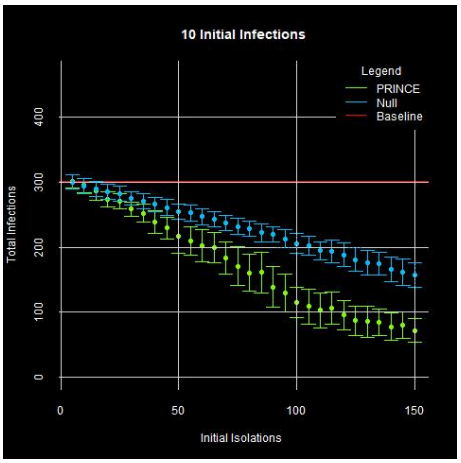

Figure 12: 10 initial infections; isolate susceptible nodes

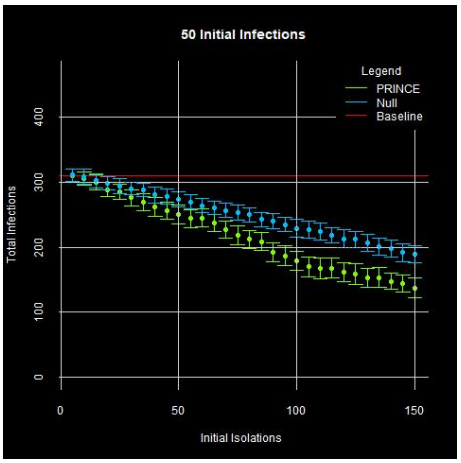

Figure 13: 50 initial infections; isolate susceptible nodes

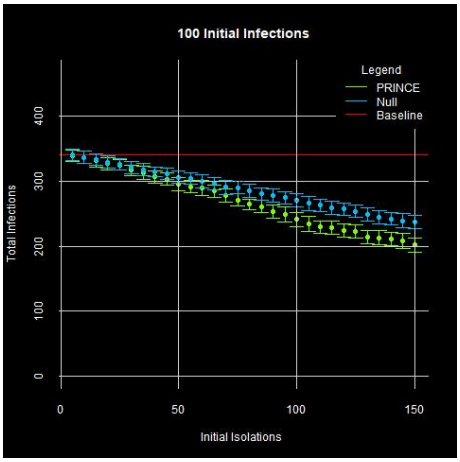

Figure 14: 100 initial infections; isolate susceptible nodes

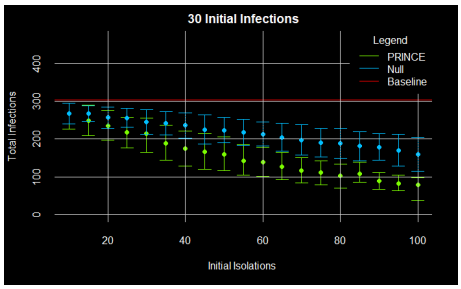

Figure 15: All Infected Nodes Isolated, 30 Susceptible Nodes Isolated

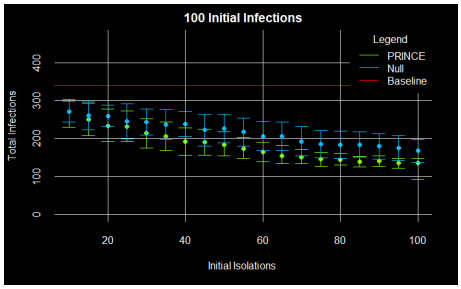

Figure 16: All Infected Nodes Isolated, 100 Initial Infections

self and the PRINCE algorithm that warrant further investigation. Due to the substantial computational resources required to run this simulation, only this particular set of parameters was investigated here. Future work may involve running the aforementioned models repeatedly to evaluate PRINCE's performance on the COVID-HM model using a wider variety of combinations of tuning parameters. This will give public health officials, policy makers, the general public, and other researchers a clearer picture as to what populations and individuals are most at risk for both transmitting and becoming infected by a novel pathogen by way of the structure of their social circle.

Additionally, the COVID-HM model that was used here to simulate the disease propagation was built on top of a network structure that was constructed using data collected from a small town in England at one point in time. Alterations in the structure of the underlying network, as well as changes to the underlying network's properties(i.e. total population, average degree, connectivity etc.) would ostensibly produce a corresponding change in the dynamics of the evolving outbreak. Hence, future work could also focus on the characterization of the behavior of the models specified with PRINCE among varying network sizes and structures, giving results that are applicable in a more general set of circumstances. Additionally, due to the modularity of the PRINCE algorithm and the dynamic network simulation, this combined methodology has the benefit of being highly extensible both in its ability to adapt to different underlying network structures and its ability to incorporate previous data regarding subpopulation transmissibility from other studies directly into the final modelling result.

Bibliography

Hellewell J. Klepac P. et al Firth, J.A. Using a real-world network to model localized covid-19 control strategies. *Nat Med*, 26, 2020.

Ellen Kuhl. *The classical SIR model*, pages 41–59. Springer International Publishing, Cham, 2021. ISBN 978-3-030-82890-5. doi: 10.1007/978-3-030-82890-5_3. URL https://doi.org/10.1007/978-3-030-82890-5_3.

Julia Gog Petra Klepac, Stephen Kissler. Contagion! the bbc four pandemic – the model behind the documentary. *Epidemics*, 24, 2018.

Vanunu, Oron, and et al. Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6, 2010.

Vanunu, Oron, et al. "Associating Genes and Protein Complexes with Disease via Network Propagation." PLoS Computational Biology, edited by Wyeth W. Wasserman, vol. 6, no. 1, 2010, p. e1000641. Crossref, https://doi.org/10.1371/journal.pcbi.1000641.

Robins, Garry, et al. "An Introduction to Exponential Random Graph (P*) Models for Social Networks." Social Networks, vol. 29, no. 2, 2007, pp. 173–91. Crossref, https://doi.org/10.1016/j.socnet.2006.08.002.

Santolini, Marc, and Albert-László Barabási. "Predicting Perturbation Patterns from the Topology of Biological Networks." Proceedings of the National Academy of Sciences, vol. 115, no. 27, 2018, pp. E6375–83. Crossref, https://doi.org/10.1073/pnas.1720589115.