Stochastic Modeling of Epidemic Diseases Considering Dynamic Contact Networks and Genealogy Information

Haleh Ashki
STOCHASTIC MODELING OF EPIDEMIC DISEASES
CONSIDERING DYNAMIC CONTACT NETWORKS AND GENEALOGY INFORMATION.

By

HALEH ASHKI

A Dissertation submitted to the
Department of Scientific Computing
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2015

Copyright © 2015 Haleh Ashki. All Rights Reserved.
Haleh Ashki defended this dissertation on July 17, 2015.

The members of the supervisory committee were:

Peter Beerli  
Professor Directing Dissertation

Christopher Coutts  
University Representative

Sachin Shanbhag  
Committee Member

Dennis Slice  
Committee Member

Alan Lemmon  
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.
ACKNOWLEDGMENTS

This work would not have been possible without the endless support, guidance and efforts of a lot of people. Foremost, I would like to express my profound gratitude to my advisor Prof. Peter Beerli, who has been a constant source of guidance, patience and support over the last four years. I am deeply grateful for his patience during years of my research and for his trust that made me through PhD program in computational science. For everything you’ve done for me, Prof. Beerli, I thank you and I am indebted. You are more than an advisor to me.

I would also like to thank Dr. Jim Wilgenbusch, my former committee member and co-advisor for his guidance, encouragement and support.

I am also extremely thankful to my committee members, Prof. Sachin Shanbhag, Prof. Dennis Slice, Prof. Alan Lemmon and Prof. Christopher Coutts for their continued faith and encouragement of my efforts.

I would like to thank all the Scientific Computing staff at Florida State University for all their administrative help through these years. I would like to thank Scientific Computing faculties, especially those I had class with and have been their teaching assistant to help me increase my knowledge.

I want to express my sincere gratitude to the Beerli and Wilgenbusch families whom by their love and kindness made me feel home here.

I am indebted to many friends and family for their unconditional love and support. I am sincerely thankful to Dr. Shahriar Rouhani, my role model who I always look up to. Arsia Takeh, Hoda Mirafzal, Asal Mohamadi who supported me during these years, and incented me to strive towards my goal.

Last but not least, thank you to my lovely parents and sister, Abdollah, Fahimeh, and Yassi, for their everlasting love and encouragement, standing there for me all these years and supporting
me through every single challenge of my life. Words alone cannot express the thanks I owe them.

I would like to thank my beloved husband, Mehdi for his love, patient and support in last twelve years, and my little bundle of joy, Hannan.
# TABLE OF CONTENTS

List of Tables .................................................. viii
List of Figures .................................................. ix
Abstract ......................................................... xi

1 Background ...................................................... 1
   1.1 Epidemic Modeling ........................................ 2
      1.1.1 Definition of Important Terms in Epidemiology and Public Health .......... 3
   1.2 Compartmental Models .................................... 4
   1.3 Dynamic Network Models .................................. 4
      1.3.1 Probability Generating Function ....................... 5
   1.4 Overview of Some Well-Known Models ..................... 9
   1.5 Dynamics of SIR Model ................................... 11
   1.6 Dynamic Network .......................................... 12

2 Stochastic modeling of Dynamic Networks of Epidemic Diseases ............. 15
   2.1 States in Epidemic Diseases ............................. 15
   2.2 Heterogeneity ............................................ 18
   2.3 Population Structure .................................... 18
      2.3.1 Graphs ............................................... 19
      2.3.2 Random Networks ..................................... 19
   2.4 Dynamic Networks ...................................... 22
   2.5 Stochastic Models ...................................... 24
2.6 Current Research ......................................................... 25
  2.6.1 Method ............................................................... 25

2.7 Results ................................................................. 29
  2.7.1 Simulated Data ..................................................... 29
  2.7.2 Dynamic Versus Static Networks ............................... 31
  2.7.3 Real Data .......................................................... 32

3 Intervention Methods .................................................... 36
  3.1 Background .......................................................... 36
  3.2 My Model .............................................................. 38
    3.2.1 Eigenvector Centrality for Unweighted Adjacency Matrices ..... 38
    3.2.2 Eigenvector Centrality for Weighted Adjacency Matrices ........ 42
  3.3 Result ................................................................. 45
    3.3.1 Dynamic Network: Degree Decreased ........................ 46
    3.3.2 Intervention Methods .......................................... 46
  3.4 Discussion ............................................................ 49
  3.5 Conclusion ............................................................ 51

4 Basic Reproduction Number Estimation ............................... 54
  4.1 Background .......................................................... 54
  4.2 Traditional Formulation of $R_0$ .................................. 55
  4.3 $R_0$ Formulation Using Contact Network ......................... 56
    4.3.1 Degree Distribution ........................................... 56
    4.3.2 Coefficient of Variation ..................................... 57
    4.3.3 Clustering Coefficient ....................................... 57
  4.4 Comparison of $R_0$ Estimators for Erdős-Rényi and Watts-Strogatz Networks .................. 60
LIST OF TABLES

3.1 Eigenvalue calculation time by regular vs power method ................. 41
3.2 Different measures on the network ........................................ 42
3.3 Comparing the PageRank method with the centrality measurement for a weighted network .................................................. 45
4.1 Different studies finding on clustering effect on $R_0$ ................. 59
4.2 Comparison between network (a) and (b) ................................. 61
4.3 Epidemic characteristic and network summary statistics. All parameter values are average of 500 simulations in network with 1000 nodes, $\beta = 0.005$ and $\gamma = 0.05$ .................. 63
4.4 Comparison of $R_0$ based on clustering coefficient .................. 64
4.5 Epidemic characteristic and network summary statistics. ............ 67
5.1 Emission Probability Matrix .................................................. 82
5.2 Best state sequence path .................................................... 82
5.3 Emission Probability Matrix .................................................. 84
5.4 Best state sequence path .................................................... 85
# LIST OF FIGURES

1.1 Complexity of epidemic models. .................................................. 2

1.2 Example of a network with 7 nodes and 6 edges. ............................. 6

2.1 Graphical representation of the SIR model .................................... 16

2.2 (a) BA network with 100 nodes. (b) Node degree distribution ............. 21

2.3 (a) WS network with 100 nodes. (b) Node degree distribution ............. 22

2.4 (a) ER network with 100 nodes. (b) Node degree distribution ............. 23

2.5 Infection and Recovery Processes. .............................................. 27

2.6 Course of an epidemic in static networks ..................................... 30

2.7 Characteristics of three random networks (BA: a1,a2; ER: b1,b2; WS: c1,c2), a1,b1,c1 show the effect of different mean degrees \( \langle k \rangle \); a2,b2,c2 show the effect of different transmission rates \( \beta \). .............................. 33

2.8 Course of an epidemic in a population of size 1000 over 1000 time steps. Gray bands reflect the maximal range of the epidemic trajectory in the 500 simulations. Cumulative number of infected individuals averaged over 500 simulations, comparison for (a) **Day-Night** (b) **Gradually Decreasing Degree** among BA, ER and WS networks. .............................. 34

2.9 Static-Dynamic network comparison ............................................ 34

2.10 Cumulative numbers of infected individuals in epidemic simulations with real contact network. (a) Transition rate \( \beta \) vary. (b) Course of 100 simulated epidemic with \( \beta=0.05 \). ........................................ 35

3.1 Example of network composed of 13 nodes .................................. 41
ABSTRACT

Human life and diseases are inseparable. For millions of years, humans and their ancestors suffered from diseases, caused by infectious pathogens (e.g., bacteria, viruses, parasites) and caused by our own bodies as they age and degenerate. Within the last century, with the advent of public health measures, improved nutrition and medicine, such as antibiotics, some of the infectious diseases have been controlled. However, infectious diseases still lead to most of the non-age related deaths in the world, especially in nations with insufficient health support.

My research has taken the complex and dynamic contact networks as well as heterogeneity in disease transmission and recovery into account. Real social networks among individuals were used to generate an adjacency matrix in my formulae. Both, transition and recovery rates have been used as unique variables for each individual. I have used the forward Kolmogorov equation to solve the system.

To control and prevent the infectious diseases such as influenza and sexually transmitted diseases, we have to model the dynamics of a particular disease, estimate the parameters, and forecast the behavior of the disease over time. The estimated parameters help us to design and implement interventions, such as vaccination or closure of public places, to limit the spread of diseases.

R0, the basic reproduction number, is an important parameter in epidemiology. R0 is the average number of secondary infections produced by a primary infection. If R0 is larger than one, an epidemic will most likely happen. An R0 smaller than one suggests that the disease outbreak is local and will die out. In this study I have shown that R0 estimators that only use the the number of contacts and some network features such as covariance of coefficient and clustering coefficient are not enough to estimate the epidemic threshold. I have formulated R0 to consider both node degree distribution as well as the spectral gap in the eigenvalue of a weighted adjacency matrix of a contact network.

Only recently, researchers have developed theoretical approaches that can take into account
dynamic networks and, independently, that can use genomic data of the pathogen, sampled from infected persons, to reconstruct the path of an epidemic. By considering the location and time of the sampled pathogen sequence data, we can combine the sampled infection network and the mutational history of the pathogen to reconstruct a more accurate contact network. We can reconstruct this dynamic contact network using genetic data and epidemic parameters via a Hidden Markov Model. Sampled genome sequenced data of the pathogen are the observation and a set of dynamic networks are the hidden states in our HMM framework. The system switches between the set of dynamic contact networks to fit the best pattern to observation data. The outcome of such an analysis is the accurate dynamic network among samples of the pathogen. This dynamic network captures the dynamics of the social contact network of the infected people. My model will most likely enable earlier detection of infectious disease spread in dynamic social networks than currently available methods.
CHAPTER 1

BACKGROUND

For hundred thousand of years, humans lived in widely dispersed, nomadic, small populations that minimized the effect of infectious diseases. With the agricultural revolution about 10,000 years ago, increasing sedentism and larger population groupings resulted in the first epidemiological transition in which infectious and nutritional diseases increased. Within the last century, with the advent of public health measures, improved nutrition and medicine such as antibiotics, several of diseases that cause epidemics have been controlled but still, infectious diseases are among the top 10 causes of human death (World Health Organization, 2012).

An infection results when a pathogen invades and begins growing within a host. The first step is the transmission into the host. For example for influenza, a close contact of an infected individual to a susceptible individual for a long enough period of time will allow the pathogen to be transmitted to the susceptible individual.

Since the pathogen is transmissible to other individuals, infectious diseases have a chance to lead to an outbreak or even epidemic. An outbreak means that a disease emerges, but by chance or because the transmission probability is low, dies out before spreading to the population at large. In an epidemic, on the other hand, the infection escapes the initial group of cases into the community and results in population-wide incidence of the disease (Meyers et al., 2005). Outbreak and epidemic are defined with specific meanings and are not interchangeable, as often is done in the common press. The Spanish flu in 1918 is an example of epidemic that spread around the world.

When a new infectious disease emerges, policy makers and doctors would like to have the answer to questions such as: Is it possible for an epidemic to occur? What is the probability of having a large epidemic by observing few cases? How can we predict the number of infected and the time-course of the epidemic?
Answers to these questions are crucial to predict the epidemic behavior, design an effective intervention method, and control the epidemic disease.

Over the years, many statistical methods were used in the field of public health. Statistical method are good to answer questions such as the probability of being sick while someone is exposed to some toxin, but they are not able to answer "why" is that happening. That is why a statistical approach does not work for infectious disease, and we need mechanistic/mathematical modeling to model epidemic diseases to answer the earlier posted questions.

1.1 Epidemic Modeling

We hope that mathematical models can help us to explain observed patterns and may be useful to generate forecasts. More complex models are considered to be more realistic. Epidemiology models are no exceptions to this rule. Figure 1.1 shows a relationship between complexity and realism in different epidemic models. Increasing the complexity doesn’t necessarily improve the model; we should only add as much detail to a model as it can affect the decision of policy makers. There are three different approaches to model the spread of an infectious disease:

![Figure 1.1: Complexity of epidemic models.](image-url)
• **Compartmental models**: Grouping individuals into compartments. For example, Susceptible, Infected, and Recovered (SIR) and calculate the number of individuals over time in each group solving a system of Ordinary Differential Equations (ODE).

• **Networks modes**: Considering the contact network among individuals in a population. The network can be either static or dynamic.

• **Agent-based**: Agent-based models are computer simulations where the agents are programmed to behave and interact with other agents and the environment in certain ways. In agent based models, each individuals is a variable in the model. The code store and analyze all the possible events according to their probability, thus the simulation is fully stochastic, for example, simulations of individuals contact within a city or country. Meyers et al. (2005) have simulated the individual contacts within the city of Vancouver; Barrett et al. (2005) have simulated movement of all individuals through Portland, Oregon; and Germann et al. (2006) have simulated the movement of individuals throughout the United States.

### 1.1.1 Definition of Important Terms in Epidemiology and Public Health

• **Basic Reproduction Ratio**

  The growth of a disease is usually parameterized by the basic reproduction ratio, $R_0$, which is the average number of additional infected individuals caused by a newly infected individual (Karrer and Newman, 2011). The basic reproductive number is also defined as the average number of infections caused by a individual infected early in an epidemic (Miller et al., 2011).

• **Epidemic Threshold**

  The epidemic threshold is the separation point where a disease dies out or where it grows exponentially, it is the time point at which $R_0 = 1$.

• **Transmissibility**

  Transmissibility or transition rate is an average probability that an infectious individual will transmit the disease to a susceptible individual.

• **Recovery Rate**

  Recovery rate is a rate at which an infected individual recovers from the disease or dies.
1.2 Compartmental Models

Miller (2014) has modeled the epidemic disease using a mass-action model. The mass-action model is a SIR model in which each individual is equally likely connected to others. Each infected individual transmits at rate $\beta k$, and $\gamma$ is a recovery rate per individual; $k$ is the number of contacts. Therefore, the ODE for this model is formulated as:

\[
\begin{align*}
\dot{S} &= -\beta k IS \\
\dot{I} &= \beta k IS - \gamma I \\
\dot{R} &= \gamma I
\end{align*}
\]

Epidemic probability is the probability that an epidemic occurs and the first infection has many "descendants". Mathematically it would be easier to calculate the probability of not starting an epidemic: $1 - P$. This probability for the mass-action model is:

\[
1 - P = \frac{\gamma}{\gamma + \beta P}
\]

And the final size can be solved numerically as:

\[
R(\infty) = 1 - e^{-R_0 R(\infty)}
\]

1.3 Dynamic Network Models

Meyers et al. (2005) presented a review on the Severe Acute Respiratory Syndrome (SARS) in 2002 and 2003 in different countries. They have compared the predicted $R_0$ and outbreak size estimated by mathematical epidemiologists with estimates using the real data reported from different countries. For instance, the predicted $R_0$ was in the range of 2.2 to 3.6, much higher that 1.0 (the epidemic threshold). Despite this, SARS has not emerged as a global pandemic. The discrepancy between the estimated $R_0$ and the observed epidemiology might come from early and effective intervention methods. Yet, even during the three and half months of the SARS outbreak in China, case counts were much less than expected given the predicted $R_0$. The models assumed fully mixed populations in which all individuals are equally likely to become infected, but this assumption did not hold. The rate of connection is not the same for every individual.
$R_0$ from simulated data often does not match $R_0$ from real data. This discrepancy suggests to model the population as a network with heterogeneous contact patterns to increase the accuracy of estimated parameters. This leads to a new class of epidemiology models taking into account the contact networks.

Before giving an overview on network models, I review some important parameters and mathematical properties that are used to formulate epidemic diseases considering contact networks.

### 1.3.1 Probability Generating Function

Figure 1.2 is an example of a network with 7 nodes and 6 edges. The number of edges attached to each node is the degree. The degree distribution is a distribution of frequencies of each node degree in a whole network. The probability generating function (PGF) for a random variable is a polynomial whose coefficients are the probabilities. It is an alternative representation for the probability distribution. PGF is a powerful method in network theory to calculate properties of a random network with known degree distribution function.

$$G_0(x) = \sum_{k=0}^{\infty} P_k x^k; \quad (1.6)$$

where $G_0(x)$ is a generating function for the probability distribution of node degree $k$ and $p_k$ is the probability that a randomly chosen node on the network has degree $k$. In other words, $G_0$ correctly generates the distribution of the degree of randomly chosen node in a network.

**Properties of Generating Function**

$$G_0(1) = \sum_{k} P(X = k) = 1. \quad (1.7)$$

$$G'_0(x) = \sum_{k=0}^{\infty} k P_k x^{k-1}. \quad (1.8)$$

$$G'_0(1) = \sum_{k} k P(X = k) = E(X) = \langle k \rangle. \quad (1.9)$$
Figure 1.2: Example of a network with 7 nodes and 6 edges.

$G'_0(1)$ is the mean of the probability distribution and I use $\langle k \rangle$ for that.

Higher moments of the distribution can be calculated from higher derivatives.

$$\langle k^n \rangle = \sum_k k^n P_k = \left[ \left( x \frac{d}{dx} \right)^n G_0(x) \right]_{x=1}.$$  \hspace{1cm} (1.10)

If the distribution of a property $k$ of an object is generated by a given generating function, then the distribution of the total of $k$ summed over $m$ independent realizations of the object is generated by the $m$th power of the generating function. For example, if we choose $m$ nodes in a network at random, then the distribution of the sum of the degrees of those nodes is generated by:

$$[G_0(x)]^m = \left[ \sum_k P_k x^k \right]^m,$$  \hspace{1cm} (1.11)
Therefore the equation for the second derivative evaluated at 1 would be:

\[ G''_0(1) = \sum_k k(k - 1)P(X = k). \] (1.12)

### 1.3.1.1 Excess degree.

Consider a node in a network, all connections lead to neighboring nodes. The excess degree is the number of neighbors that can be infected. It is one less than the total number of neighbors, because the node has received the infection from one of the neighbors and therefore that neighbor cannot be infected again. The excess degree calculation is important for the \( R_0 \) calculation.

The degree distribution of a node reached by following a randomly chosen edge in a graph is proportional to \( kp_k \); \( q_k \) denote this degree distribution, after normalizing we obtain:

\[ q_{k-1} = \frac{kp_k}{\sum_k kp_k}. \] (1.13)

To get the \( q_k \), we adjust the index

\[ q_k = \frac{(k + 1)p_{k+1}}{\sum_k kp_k}. \] (1.14)

and hence the generating function is:

\[ \frac{\sum k p_k x^k}{\sum_k kp_k} = \frac{xG'_0(x)}{G'_0(1)}. \] (1.15)

Excess degree of a node is the number of neighbors excluding the one we reach that node by, which is the degree minus 1. we simply divide the function above by one power of \( x \), thus arriving at a new generating function:

\[ G_1(1) = \frac{G'_0(x)}{G'_0(1)}. \] (1.16)

The average of the excess degree is calculated as:

\[
\sum_{k=0}^{\infty} (kq_k) = \frac{\sum_{k=0}^{\infty} (k(k + 1)p_{k+1})}{\sum_k kp_k} \\
= \frac{\sum_{k=0}^{\infty} ((k - 1)kp_k)}{\sum_k kp_k} \\
= \frac{<k^2> - <k>}{<k>},
\] (1.17)
expressed as PGF:

\[ G'_1(1) = \frac{\sum_{k=0}^{\infty}((k-1)kp_k)}{\sum_k kp_k} \]  
\[ = \frac{G''_0(1)}{G_0(1)} \]  
\( (1.20) \)

1.3.1.2 Some example of specific graphs with known distribution function and their corresponding \( G_0(x) \).

(a) Poisson-distributed graph

\[ P_k = \frac{Z^k e^{-z}}{k!} \]  
\[ G_0(x) = e^{x(x-1)} \]  
\( (1.22) \)

(b) Power-law distributed graph

\[ P_k = C k^{-\tau} e^{-\frac{k}{\kappa}} \]  
\( (1.24) \)

where \( C, \tau \) and \( \kappa \) are constant and

\[ C = [Li_{\tau}(e^{-\frac{1}{\kappa}})]^{-1} \]  
\( (1.25) \)

\[ P_k = \frac{k^{-\tau} e^{-\frac{k}{\kappa}}}{Li_{\tau}(e^{-\frac{1}{\kappa}})} \] for \( k \geq 1 \)  
\( (1.26) \)

Where \( Li_{\tau}(x) \) is the \( \tau \)th polylogarithm of \( x \).

\[ G_0(x) = \frac{Li_{\tau}(xe^{-\frac{1}{\kappa}})}{Li_{\tau}(e^{-\frac{1}{\kappa}})} \]  
\( (1.28) \)

(c) Graph with arbitrary specified degree distribution. For this kind of graphs, we know the exact number \( n_k \) of vertices having degree \( k \)

\[ G_0(x) = \frac{\sum_k n_k x^k}{\sum_k n_k} \]  
\( (1.29) \)
1.4 Overview of Some Well-Known Models

The complement of the epidemic probability, \((1 - P)\), has been formulated by Miller (2014) as:

\[
\Omega(\tau) = 1 - T(\tau) + T(\tau) \int_0^\infty \gamma e^{-\gamma \frac{\psi'(\Omega(\tau))}{\psi'(1)}} d(\hat{\tau})
\]  

(1.30)

Where \(\psi\) is a probability generating function, \(T\) is transmissibility, and \(\hat{\tau}\) is the infection duration.

The final size has been formulated as:

\[
\theta = 1 - T + T \frac{\psi'(\theta)}{\psi'(1)}
\]

(1.31)

Newman (2010) has done an extensive study on networks and formulated the epidemic parameters using PGF and network properties. He reviews and discusses different network models in detail.

Karrer and Newman (2011) defined \(R_0\) as the average number of additional infections caused by a newly infected individual. There are two factors causing the spread of infection: transmissibility \(T\) and the number of contacts of each individual. He calculated the \(R_0\) as:

\[
R_0 = T \sum_{k=0}^{\infty} (k q_k) = T \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}
\]

(1.32)

Using the epidemic threshold definition, gives us:

\[
T_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}
\]

(1.33)

Using equation 1.19 and 1.21 we can show the epidemic threshold as:

\[
T_c = \frac{1}{G_1'(1)} = \frac{G_0'(1)}{G_0''(1)}
\]

(1.34)

In other words, the epidemic thresholds, \(T_c\), is the minimum transmissibility (T) required for an outbreak to become a large-scale epidemic.

The average size of an outbreak is calculated by Newman (2002) by nesting PGF’s of the number of new infections emanating from an infected vertex as:

\[
\langle s \rangle = 1 + \frac{T G_0'(1)}{1 - T G_1'(1)}
\]

(1.35)
Equations 1.23 to 1.28 show the $G_0(x)$ of some well-known networks. The $G_1(x)$ and $G_1(1)$, as well as their first and second derivative, are defined. Knowing the property of the graph, calculating the epidemic threshold and $R_0$ is straightforward. $R_0$ and epidemic threshold are also simple to calculate for networks with arbitrary specified degree distribution as shown in equation 1.29. $R_0$ and outbreak size have been calculated by Newman et al. for many other networks.

Erik Volz (2007) formulated the epidemic threshold as:

$$\epsilon \left( r \frac{g''(\theta)}{g'(\theta)} - r - \mu \right) = 0 \quad (1.36)$$

where we have an initial condition as $\epsilon << 1$. Rearranging the formula yields the critical ratio $r/\mu$ in terms of the PGF:

$$(r/\mu)^* = \frac{g'(1)}{g''(1) - g'(1)}, \quad (1.37)$$

which is the epidemic threshold. It also can be put in terms of transmissibility $\tau$ as:

$$\tau = \int_0^\infty (1 - e^{-rT})(\mu e^{-\mu T})dT = \frac{r}{r + \mu}, \quad (1.38)$$

where the probability that an infectious individual will transfer an infection to a neighbor for a duration $T$ is

$$1 - e^{-rT}. \quad (1.39)$$

Then rearranging the equation 1.37 yields the epidemic threshold in terms of $\tau$:

$$\tau = \frac{g'(1)}{g''(1)}. \quad (1.40)$$

That is exactly the same as equation 1.34.

Keeling and Rohani (2007) discuss a variety of infectious disease modeling: approaches that are deterministic or stochastic. They formulated the SIR and SIS model for multiple pathogens, sexually transmitted diseases, and spatial models. There are different derivation of simple SIR model formulation based on the specific parameters for each model. $R_0$ and dynamic of the disease has been formulated for all compartmental and network models.
1.5 Dynamics of SIR Model

The SIR model was originally developed without assuming an underlying contact network that will bias the dynamics of the system, for example, the number disease incidences. In the early 2000, researchers had started to meld the SIR model with network models.

Eames and Keeling (2002); Miller et al. (2011); Lindquist et al. (2011); Hébert-Dufresne et al. (2013); Ball and Neal (2008) have formulated the dynamics of system as:

\[ \theta(t) = P(\text{v not yet transmitted to u}) \]  \hspace{1cm} (1.41)

to get the dynamic of disease, he has formulated how the \( \theta \) evolve as:

\[ \frac{d\theta}{dt} = -\beta \theta + \beta \frac{\psi'(\theta)}{\psi(1)} + \gamma (1 - \theta) \]  \hspace{1cm} (1.42)
\[ \frac{dR}{dt} = \gamma I, S = \psi(\theta), I = 1 - S - R \]  \hspace{1cm} (1.43)

Volz (2007) has modeled the dynamics of a SIR-type epidemic using ODEs and PGF to represent the degree distribution on a network. He has used the SIR model ODE equations 2.2, 2.3, 2.4 along with the PGF equation 1.6. The fraction of connection between nodes belonging to each compartment in the SIR model is shown by \( M_{ij} \), where \( i \) and \( j \) represent the compartments. For example, the fraction of contacts between susceptible and infected individuals is \( M_{SI} \), and the probability that an edge is connecting a susceptible to infectious node is \( PI = M_{SI}/M_S \). He has used the concept of derivation to calculate the dynamics of the system and has formulated them as ODEs. The following formulae, assuming that transition rates and recovery rate are constant, summarize the method of Volz (2007).

\[ PI = \frac{M_{SI}}{M_S} \]  \hspace{1cm} (1.44)
\[ PS = \frac{M_{SS}}{M_S} \]  \hspace{1cm} (1.45)
\[ \frac{d\theta}{dt} = -r PI \theta \]  \hspace{1cm} (1.46)
\[ \frac{dS}{dt} = -r PI \theta g' (\theta) \]  \hspace{1cm} (1.47)
\[ \frac{dI}{dt} = -r PI \theta g' (\theta) - \mu I \]  \hspace{1cm} (1.48)
where $\theta$ is the fraction of degree of nodes that remain susceptible at time $t$. $PI$ is the probability that a susceptible $a$, the first element in the ordered pair $(a,b)$ in graph, has an infectious neighbor $b$; $PS$ is the probability that susceptible $a$, the first element in the ordered pair $(a,b)$ in graph has a susceptible neighbor $b$; $r$ is the constant infectious rate and $\mu$ is the constant recovery rate.

Solving the above ODEs (formulae 1.46 - 1.48) give us the number of individuals in each epidemic state (susceptible, infected, and recovered) at each time $t$. The initial conditions are defined as:

$$\theta(t = 0) = 1 - \epsilon$$  \hspace{1cm} (1.49)

$$PI(t = 0) = M_{SI}/M_S = \frac{\epsilon}{(1 - \epsilon)}$$  \hspace{1cm} (1.50)

$$PS(t = 0) = M_{SS}/M_S = \frac{(1 - 2\epsilon)}{(1 - \epsilon)}$$  \hspace{1cm} (1.51)

where $\epsilon$ is a small fraction of the nodes in the network that are selected uniformly at random and initially infected.

As in the previous methods, the PGF and its derivative are predefined for most of the networks and are used as functions that need to be replaced by values, for example equations 1.23, 1.28, and 1.29.

### 1.6 Dynamic Network

All previously discussed models and equations are designed for static networks where the number of contacts can be heterogeneous among nodes but the numbers and properties of contacts remain constant over time. But in actual contact networks the number as well as the properties of contacts among individuals in a population changes over time.

Volz and Meyers (2007) have modeled such a dynamic network using a Neighbor Exchange (NE) model. They have calculated the dynamics of an epidemic disease as well as $R_0$ and epidemic threshold.

The NE model assumes that the neighbors of a node will continually change, while the total numbers of the current neighbors remain constant. The rate of change or swap between edges is a constant value $\rho$. 
By taking this new parameter into account, the dynamics of the system can be calculated. The dynamics equations are almost the same as equations 1.46 to 1.48 except for having an extra parameter $\rho$.

$$\frac{dPI}{dt} = rPSPIg''(\theta) - rPI(1 - PI) - PI\mu + \rho(MI - PI) \tag{1.52}$$

$$\frac{dPS}{dt} = rPSPI(1 - \theta g''(\theta) + \rho(\frac{g(1)}{g(1)} - PS) \tag{1.53}$$

Volz and Meyers (2009) have calculated reproduction ratio as:

$$R_0 = \frac{\tau\rho}{\mu} + \tau\frac{\mu + \rho g''(1)}{g'(1)} \tag{1.54}$$

using $\tau$ as probability of transmission from $a$ to $b$ during their infectious contact period. By using the theory of Markov process and a Poisson process with rate $\mu + \rho$

$$\tau = \frac{r}{r + \mu + \rho} \tag{1.55}$$

and the epidemic threshold is

$$\frac{\mu g'(1)}{g''(1)(\mu + \rho) + \rho g'(1)} \tag{1.56}$$

Miller et al. 2009b describe a SIR epidemics in a dynamic network. The partnerships are assumed to be for a very short period of time. Multiple transmission is neglected in the same partnership. The probability of not causing an epidemic has been formulated as:

$$1 - M = \int_0^\infty P(k)\frac{\gamma}{k\beta(1 - \alpha) + \gamma} \tag{1.57}$$

$$\alpha = \int_0^\infty kP(k)\frac{\gamma}{<K > k\beta(1 - \alpha) + \gamma} \tag{1.58}$$

which can be solved numerically.

The dynamics of diseases considering dynamic network in modeling has been formulated as:

$$\frac{d\theta}{dt} = -\beta\theta + \beta\frac{\theta^2\psi'(\theta)}{\psi'(1)} - \theta\gamma\ln(\theta) \tag{1.59}$$

$$\frac{dR}{dt} = \gamma I, S = \psi(\theta), I = 1 - S - R \tag{1.60}$$

Balcan et al. (2009) have modeled epidemic diseases taking mobility and migration between cities into account. They have both population and mobility layer in their modeling. The GLEAMviz
software is been developed by this group (Balcan et al., 2009). The global epidemic and mobility model, GLEAM, forecasts an epidemic by combining real-world data of social contact network among individuals and stochastic models of disease transition. This model has been used successfully to estimate epidemic parameter and disease spread for many epidemics (Bajardi et al., 2011; Balcan et al., 2010; Tizzoni et al., 2012).

Stadler et al. (2011); Stadler (2008); Stadler et al. (2013); Stadler and Bonhoeffer (2013); Leventhal et al. (2014); Kühnert et al. (2014); Stadler (2010, 2011) have estimated the epidemic parameters and spread of a disease taking the viral sequence data into account. Their formulation is mainly based on the genealogy data. They infer epidemiological parameters directly from genome data of pathogens. Their method is based on phylogenetic analysis using a birth-death model.

My study is mathematically melding of infectious disease models focusing on dynamic networks with genealogical data. My model is intermediate between a network model and an agent-based model. I use contact networks that are dynamic and change over time. The dynamics of the system is stochastic; calculations are based on ODE and Markov chain property. An extension of the model takes the genealogy of the pathogens into account. A Hidden Markov Model estimates the dynamic network structures that represent the most likely connectivity and mobility among individuals in a population.
CHAPTER 2

STOCHASTIC MODELING OF DYNAMIC NETWORKS OF EPIDEMIC DISEASES

Human life and diseases are inseparable. Diseases can be caused by our own bodies as they age and degenerate or by infectious pathogens. It is not commonly thought that mathematics can save lives and improve quality of life. My research is about infectious diseases, such as flu or sexually transmitted diseases. There are several important factors on modeling epidemic diseases such as the population structure representing the possible contact among individuals. Understanding the contact network among individuals is essential for forecasting how an epidemic will spread. States in epidemic diseases are a fundamental concept of modeling. The simple model of progress of an epidemic in a large population divides the population into three different compartments: Susceptible, Infected, and Recovered (SIR). Transition and recovery rate are two important parameters in epidemiology explaining the likelihood that a susceptible individual can be infected by contacting an infectious one and the rate an individual can recover from the disease. Considering the Heterogeneity in these parameters make the model more accurate.

2.1 States in Epidemic Diseases

To model the progress of an epidemic in a large population, individuals have been classified in three or more groups based on the behavior of the disease. A very simple model groups the population into three different groups: Susceptible, Infected, and Recovered. Susceptible are individuals who are not yet infected and are susceptible to disease. Infected are those individuals who have been infected with the disease and can spread the disease. Recovered are those individuals who have been recovered from the disease and are not able to be infected again. This model is named the SIR model, it was invented in 1927 by Kermack and McKendrick (Figure 2.1). The change of
the system over time has been modeled by differential equations. This model considers a closed population (no birth, death or migration) with three groups. Susceptible $S(t)$, infected $I(t)$ and recovered $R(t)$ represent the number of individuals in each class at time $t$.

We can set up a model that has this basic relationship

$$N = S + I + R \quad (2.1)$$

where the variables are defined as:

- $S$ is the number of susceptible.
- $I$ is the number of infected.
- $R$ is the number of recovered.
- $N$ is the population size.

Each individual in the population has an equal probability of contracting the disease with a rate of $\beta$. Each individual makes contact and is able to transmit the disease with $\beta(N - 1)$ others. For a large $N$ it would approximate to $\beta N$. Fraction of contact by an infected with a susceptible is $S/N$, therefore the effective transmission rate is $\beta N(S/N) = \beta S$. Each infected will recover at rate $\gamma$ and will not be susceptible to the disease again.

There are $I$ infected at time $t$ in a population, therefore the fraction of newly infected individuals will be $I\beta S$ (Brauer and Castillo-Chávez, 2001). These processes occur simultaneously. They are referred to as the Law of Mass Action, a widely accepted idea that the rate of contact between two groups in a population is proportional to the size of each of the groups concerned (Daley and Gani, 2005). We can say that $S(t + 1) = S(t)I\beta$ and transform this difference equation to a differential

Figure 2.1: Graphical representation of the SIR model

![SIR Model Diagram](image)
\[
\begin{align*}
\frac{dS}{dt} &= -\beta SI, \\
\frac{dI}{dt} &= \beta SI - \gamma I, \\
\frac{dR}{dt} &= \gamma I.
\end{align*}
\] (2.2) (2.3) (2.4)

Each infectious disease has its own characteristic, and therefore different models with more parameters or compartments are needed to model them accurately. Some other important epidemic models are:

**SIS Model:** Some infectious diseases, such as the common cold, do not induce any long lasting immunity. Infected individuals recover with no immunity to the disease and will be susceptible again. It can be expressed as:

\[ S \rightarrow I \rightarrow S \]

**SEIR and SEIS Models:** Many diseases have another phase called latent or exposed phase, during which an individual is infected but not infectious. This model considers the exposed phase by adding a compartment \( E \). The immunity can be permanent or temporary and therefore the infected individual can recover or be susceptible again. It can be expressed as:

\[ S \rightarrow E \rightarrow I \rightarrow R \quad \text{or} \quad S \rightarrow E \rightarrow I \rightarrow S \]

**MSIR Model:** For many infectious diseases, including measles, babies are born with immunity to the disease for the first few months of their life protected by maternal antibodies. The compartment \( M \) represents this:

\[ M \rightarrow S \rightarrow I \rightarrow R \]

Birth and death rates are ignored in all previously described models. Adding these two parameters does change the differential equations of each model. For the SIR model the equations are
augmented to

\[
\frac{dS}{dt} = -\beta SI + \mu - \mu S, \tag{2.5}
\]

\[
\frac{dI}{dt} = \beta SI - \gamma I - \mu I, \tag{2.6}
\]

\[
\frac{dR}{dt} = \gamma I - \mu R. \tag{2.7}
\]

where \( \mu \) is the per capita death rate, and the population level birth rate.

I will focus in the SIR model for my work because it fits the characteristics of the majority of infectious diseases.

### 2.2 Heterogeneity

In many epidemic models, each individual contracts the disease with the same probability as others, but that is not true in real life. For example, in the real world people have different connectivity patterns. They also have different rates of being infected or recovered. Heterogeneity can take many forms: heterogeneity with respect to contact rate, susceptibility, infectiousness, and network structure. To develop better models, we have to take heterogeneity into account.

### 2.3 Population Structure

Traditional theories of disease propagation ignore network effects (Anderson and May, 1991a; Hethcote, 2000). In these models everyone has an equal probability of infecting others and the population is assumed to be fully mixed. This along with the assumption of large population size allows us to formulate the epidemic as ordinary differential equations. In the real world, however, people contact others in different pattern, for example the number of connections and contact duration are different. Individual contacts are also not static and change over time. In recent years a significant amount of work has investigated the role of population structure in epidemiology. Network epidemiology uses the network of individuals as a representation of population structure. Graph theory as a tool has been used to represent these networks.
2.3.1 Graphs

A graph is a mathematical representation of a set of objects. Some or all of those objects are connected. The interconnected objects are represented by mathematical abstractions called vertices, and the links that connect some pairs of vertices are called edges. We can formulate the graph in terms of vertices or nodes $V$ and edges or lines $E$ as

$$G = (V, E).$$

Instead of using a list of nodes and edges, we can represent the network as an adjacency matrix. The matrix element $P(i,j)$ is one if node $i$ and node $j$ are connected and zero otherwise. For undirected graph this adjacency matrix is symmetric. And for large and low-density graph, the matrix is sparse, too.

2.3.2 Random Networks

The Internet, the World Wide Web, Facebook, and eBay are examples of some of the myriad types of networks that are a part of our daily life. Each of us is also part of a bigger network, the social contact network. For example, there are social networks where nodes can represent individuals and links can represent a friendship, a sexual relationship, a past communication, a coauthorship, or a citation.

Networks depicting social interactions between people have been studied for decades (Scott, 2000; Wasserman and Faust, 1994). Recently the use of social contact network structure has improved epidemic disease modeling. Many researches have modeled the social contact network using graph theory. Contact networks can be studied as graphs, and graph analysis has become crucial to understand the features of these systems. The origin of graph theory dates back to Euler’s solution of the puzzle of the bridges of Königsbergs in 1736 (Euler, 1736). In the 20th century they have also become extremely useful as representation of a wide variety of systems in different areas such social science, computer science, and epidemiology. Watts and Strogatz (1998); Barábasi and Albert (1999); Barrat et al. (2008); Boccaletti et al. (2006); Dorogovtsev and Mendes (2013); Newman (2003c); Pastor-Satorras and Vespignani (2007) have produced networks by new approaches that can handle networks with millions of nodes.
Three well known networks that have been widely used in epidemiology as social contact network modeling are the graph described by Barabási-Albert (BA; Barabási and Albert, 1999), Erdős-Rényi (ER; Erdős et al., 1959) and Watts-Strogatz (WS; Watts and Strogatz, 1998).

2.3.2.1 Barabási-Albert Model. Many of the natural or human-made networks such as the internet, the world wide web, and some social networks such as the network of friends in Facebook fall in the class of scale-free networks. Scale-free networks have power law degree distributions; nodes that have already many connections will accumulate more new connections than others, for example in social media such as Facebook or Twitter.

The BA model, named for its inventors Albert-László Barabási and Réka Albert, is one of the best known models in this class. It has been widely used in study of epidemic models where the contact network has a power law distribution.

**Algorithm**

The generating BA network starts with an initial connected network with $m_0$ nodes. Then a new node is added and is been connected to one of the existing nodes with probability:

$$p_i = \frac{k_i}{\sum_j k_j}$$  \hspace{1cm} (2.9)

where $k_i$ is the degree of node $i$ and the sum is made over all pre-existing nodes $j$. The new nodes have a preference to link to already heavily liked nodes. Thus, the hub nodes quickly accumulate more nodes.

The degree distribution follows the power law:

$$P(k) \sim k^{-3}.$$  \hspace{1cm} (2.10)

Figure 2.2 (a) is an example of a BA network with 100 nodes and (b) shows the degree distribution of this network.

2.3.2.2 Watts-Strogatz Model. The WS model, named after its inventors Duncan J. Watts and Steven Strogatz in 1998, is one of the random graphs in the class of small-world networks.
A small-world network is defined to be a network where the typical distance $L$ between two randomly chosen nodes (the number of steps required) grows proportionally to the logarithm of the number of nodes $N$ in the network, that is:

$$L \propto \log N$$

Therefore, the degree distribution in the case of the ring lattice is just a Dirac delta function centered at $k$, where $k$ is mean node degree. The average shortest path length is small, meaning that most nodes can be reached from every other by a small number of steps.

**Algorithm**

Generating the WS network requires the number of nodes, $N$, mean degree of nodes, $\langle k \rangle$, the rewiring probability, $\beta$, as an input. First, a regular ring lattice with $N$ nodes has been constructed. Each node is connected to $K$ neighbors, $\frac{K}{2}$ on each side. Then for every node $n_i$ take all edges $(n_i, n_j)$ where $i < j$ and rewire it with rewiring probability $\beta$. Rewiring is done by replacing $(n_i, n_j)$ with $(n_i, n_k)$ where $k$ is chosen with uniform probability from all possible values that avoid self-loops. Low rewiring probability leads to a network that looks like the original ring lattice. In contrast, high rewiring probability leads to a network that is equivalent to an ER network.

Figure 2.3 (a) gives an example of WS network.
2.3.2.3 Erdős-Rényi model. There are two closely related variants of the ER model that have been generated by Paul Erdős and Alfréd Rényi in 1959 and 2001. The ER model generates a random graph with $n$ nodes where the edges are connecting two nodes with probability $p$. The distribution of the degree of any particular node is binomial.

$$P(deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$  \hspace{1cm} (2.12)

For large $n$ the distribution would tend to be Poisson distributed. The $p$ value plays an important role in the behavior of the ER graph. By having lower or larger $p$ value the graph would have more or fewer connected components.

Algorithm

A graph is constructed by connecting nodes randomly. Each edge is included in the graph with probability $p$ independent of other edges, calculated using formula 2.12.

Figure 2.4 (a) gives an example of an ER network.

2.4 Dynamic Networks

Dynamic networks are a class of networks where the components, nodes and edges, have a dynamic structure; nodes and edges may appear or disappear over time. These changes can happen
simultaneously. A dynamic network can have a fixed number of nodes and mean degree while the edges appear and disappear by connecting new sets of nodes; or a network can grow or shrink by having more or fewer nodes over time. A population generated with a birth and death process is a good example for this class of networks. The number of edges and mean degree of the whole network can change over time as well. For example, an attempt to change the behavior of an epidemic by closing schools and other public places may decrease the mean degree and number of edges in the contact network.

To simulate data, I constructed dynamic random networks using two scenarios:

- **First scenario: Day and Night Network** I combined two static networks with different connectivities and repeat this pattern for some time. This would represent a standard social network, where we go to work during the day and interact with others, and during the night we only interact with our friend and families.

- **Second scenario: Gradually Decreasing Degree Network** The connectivity of random nodes is gradually reduced over time. This represents a social network during the course of an epidemic, when people reduce their connectivity as much as possible. To generate this network, an arbitrary number of nodes are chosen randomly. Then their degree is reduced by an arbitrary value, for example 40%. The edges that need to be removed to achieve the node degree reduction are also chosen randomly.
2.5 Stochastic Models

There are different stochastic modeling processes, such as a discrete time Markov chain (DTMC) model, a continuous time Markov chain (CTMC) model, and a stochastic differential equation (SDE) model. These stochastic processes differ in the underlying assumptions regarding the time and the state variables. In a DTMC model, the time and the state variables are discrete. In a CTMC model, time is continuous, but the state variable is discrete. Finally, the SDE model is based on a diffusion process, where both the time and the state variables are continuous.

I will outline the CTMC because it will be useful for my own explorations of the SIR model.

The CTMC epidemic processes are defined on a continuous time scale, \( t \in [0, \infty) \), but the states \( S(t), I(t), R(t) \) are discrete random variables, i.e., \( S(t), I(t), R(t) \in (0, 1, 2, ..., N) \).

**Markov chain**

A Markov chain is a stochastic process with the Markov property. The Markov property is named after the Russian mathematician Andrey Markov who said that the conditional probability distribution of the next future state of the process depends only upon the present state, not on the sequence of events that preceded it. It can be formulated as follow:

\[
P(X_n | X_{n-1}, X_{n-2}, ...) = P(X_n | X_{n-1})
\] (2.13)

Here \( X_n \) is the state of the Markov chain and \( n \) is an integer time or index. In general a Markov model consist of a transition matrix and a starting state probability vector to calculate the probability of being in the next state. The transition matrix is a stochastic matrix that describes a Markov chain \( X_n \) over a finite state space. In other words:

\[
P(j | i) = P_{i,j}.
\] (2.14)

\( P_{i,j} \) is the probability of movement (transition) from state \( i \) to state \( j \). By having the transition matrix \( P \) we can say:

\[
X_{n+1} = PX_n
\] (2.15)
Where $X_{n+1}$ and $X_n$ are two consecutive state vectors of a Markov chain.

The transition matrix is constructed by using the property of the model and parameters such as recovery rate, transition rate, birth and death rate. The matrix differs from model to model but will still have the same structure. By applying the Markov property and infinitesimal transitional probabilities, a continuous time analogue of the transition matrix can be defined. Instead of a system of difference equations, a system of differential equations is obtained.

In the context of a continuous-time Markov process, the Kolmogorov equations, forward and backward, are a pair of systems of differential equations that describe the time-evolution of the probability. The Kolmogorov forward equations describe an initial value problem for finding the probabilities of the process, given the quantities $A_{jk}(t)$.

$$\frac{dP_k}{dt}(t) = \sum_j A_{jk}P_j(t) \tag{2.16}$$

This can be rewritten in matrix form, where $A$ is a transition probability matrix as:

$$\frac{dP}{dt} = AP \tag{2.17}$$

2.6 Current Research

Instead of assuming to have a static network, I allow changes in the network structure at different times to solve the dynamics of an epidemic. I have used a Markov property-based approach because it gives me the ability to use the adjacency matrix of the social network, representing the population interaction.

2.6.1 Method

The most commonly used epidemiological model is the SIR model (Kermack and McKendrick, 1927), which divides the population into three different groups: Susceptible, Infected, and Recovered (S,I,R). This model is deterministic and is usually formulated as a system of differential equations. It assumes that the population size is large and is differentiable in terms of time using
the equations

\[
\frac{dS}{dt} = -\beta SI, \\
\frac{dI}{dt} = \beta SI - \gamma I, \\
\frac{dR}{dt} = \gamma I,
\]

(2.18)

where \(\beta\) is the transition rate at which susceptible individuals become infected and \(\gamma\) is the recovery rate at which infected individuals either die or recover and therefore are not available for the infection process anymore. This most basic model implies an unstructured network in which each individual is connected to all other individuals.

I extend this model to allow complex contact networks that can change over time. The extension is based on the continuous time Markov process concept as well as a system of forward Kolmogorov equations.

I calculate the probability of each individual being in each state (S, I, R) at each time step. There are \(N\) individuals, each representing a node in the contact network. Each individual can only be in one state at a given time. \(P_S(t)\) is a probability vector of size \(n \times 1\) indicating the probability of each individual being in state \(S\) at time \(t\). \(P_I(t)\) is a probability vector of size \(n \times 1\) indicating the probability of each individual being in state \(I\) at time \(t\). \(P_R(t)\) is a probability vector of size \(n \times 1\) indicating the probability of each individual being in state \(R\) at time \(t\).

Each of the three state vectors contains the probability of all individuals in the population:

\[
P_I = [P_{I,1}, P_{I,2}, ..., P_{I,N}] \\
P_S = [P_{S,1}, P_{S,2}, ..., P_{S,N}] \\
P_R = [P_{R,1}, P_{R,2}, ..., P_{R,N}] \\
P_{R,i} = 1 - (P_{I,i} + P_{S,i}).
\]

(2.19)  (2.20)  (2.21)  (2.22)

The initial states having one or more infected individuals and all others are susceptible; the \(P_I(0) = (0,0,1,0,0,1,0,...)\) has few ones and the rest are zeros, and \(P_S(0) = (1,1,0,1,1,0,1,...)\) has the value 0 for those who are initially infected and the rest are ones. There is no recovered individual at time zero, therefore \(P_R(0) = (0,0,0,...)\) is a vector of size \(n\) of zeros.
Since there are two independent processes for each individual (Figure 2.5) and within each process only two states are allowed (infection process: S, I; recovery process I, R), in each process the two probabilities sum to 1.0. Therefore in the infection process, we only need to calculate the $P_I; P_S = 1 - P_I$. For the recovery process we calculate $P_R; P_I = 1 - P_R$.

A susceptible individual can become infected by contacting infected individuals. An infected individual can recover after some time lag. These two processes: Infection and Recovery are independent of each other (Figure 2.5).

![Infection and Recovery Processes](image.png)

The probability that a person is infected in the next time step depends on both the number of neighbors, their epidemic state of being infected, recovered or susceptible, and the rate at which that particular person can become sick.

Our contact network is expressed as a weighted adjacency matrix representing connectivity and transition rates. The matrix is a symmetric adjacency matrix because there is a mutual connection between each two connected nodes. The matrix is sparse because in a real society an individual is usually not connected to many others; each individual is connected to a small number of other individuals compared to the total number of individuals in the population.

In addition to the adjacency matrix, the transition rate has to be taken into account. We can think of the transition rate as a probability or rate that someone become sick by contacting an infectious person. This is not constant over time and can vary from individual to individual. In
modeling the system we could define different parameters for each individual, which is not that accurate because we cannot estimate the transmission rate for each person. But by taking some parameters, such as age, race, biological background, and vaccination history into account we can apply some more precise arbitrary parameters to affect the disease transition rates. This immunity or vulnerability rate act as node weight. To map this weighted graph to a matrix, the rates have been transformed to edge rates, so the graph would be a directed graph with different in degree and out degree values. Therefore, the weighted adjacency matrix is not symmetric.

The system of the dynamics of an epidemic using equation 2.17 can be written as:

\[
\frac{d(P_I)}{dt} = CP_I
\]  \hspace{1cm} (2.23)

where \( C \) is the weighted adjacency matrix. Expanding \( C \) in (2.23) we get

\[
\frac{d(P_I)}{dt} = \left( \sum_i \sum_j \beta_i C_{ij}(P_I)_j \right),
\]  \hspace{1cm} (2.24)

where \( \beta_i \) is the transition rate of individual \( i \) and \( C_{ij} \) is the connection among individuals. The recovery of an infected individual is modeled as:

\[
\frac{d(P_R)}{dt} = -\gamma_i P_R,
\]  \hspace{1cm} (2.25)

where \( \gamma_i \) is the recovery rate of individual \( i \). By combining (2.24) and (2.25) we get

\[
\frac{d(P_I)}{dt} = \delta \left( \sum_i \sum_j \beta_i C_{ij}(P_I)_j \right) + (\delta - 1)(\gamma_i P_I),
\]  \hspace{1cm} (2.26)

\[
\delta = \begin{cases} 
1 & \text{if } \frac{P_I}{\tau_{SI}} \leq 1, \\
0 & \text{otherwise},
\end{cases}
\]  \hspace{1cm} (2.27)

where \( \tau_{SI} \) is an infection threshold or the duration of infectiousness. The range is between 0 and 1. The default value is set to 1.0. It can have different arbitrary values based on the life history of the virus or disease incubation time. The longer the incubation time, the more likely the virus spread. Thus by having the larger \( \tau_{SI} \) value, it took longer for the \( PI \) value to reach the threshold and the individual been labeled as infected. The smaller the \( \tau_{SI} \) value, meaning shorter incubation time and individual been considered as infected sooner.

My model is extendable to other disease models than SIR. For example, in SEIR type diseases, where an infected individual can spread the disease but does not show symptoms, our simple two-state framework is extended to include an additional parameter \( \tau_{SE} \) which is the latency threshold.
2.7 Results

I have tested my method on simulated and real data.

2.7.1 Simulated Data

I have tested the model on both static and dynamic network. For static network simulation, I have constructed three commonly used random networks, those developed by Barábasi and Albert (1999) [BA], Erdős et al. (1959) [ER], and Watts and Strogatz (1998) [WS]. The package NetworkX (Hagberg et al., 2008a) was used to generate these networks. All three network types have 1000 nodes each and mean degree is set to 6. The transition rate $\beta$ is 0.005 and the recovery rate $\gamma$ is 0.05 in all simulations. Each simulation was run for 1000 time steps. All results are averaged over 500 runs.

2.7.1.1 Static Network. As shown in Figure 2.6 for this specific parameter sets, the epidemics happen sooner in a BA network. Epidemics starts later on an ER network but have almost the same patterns as BA networks and the rate of increase of infected individuals is almost the same in these two networks. Since there are few nodes with high node degree in BA network, those nodes spread the disease quickly to their neighbors, that is why the epidemic happens sooner in BA network. The distribution of the node degree for the rest of the nodes is almost the same as the ER network, that is why after that peak, the course of an epidemic has almost the same pattern as ER. An epidemic starts much later on a WS network and its propagation in the network is also slower with a gentler slope, but the maximum number of infected individuals is larger than in ER and BA and almost all individuals in the population would be infected, if we run the simulation for long enough. Because there is no isolated nodes in the WS network, therefore, everyone would get infected but it would just take longer.

2.7.1.1 Testing the Different Parameter Sets. There are three key properties that define the course of an epidemic: the contact among individuals, transition rate, and recovery rate. To test the impact of each parameter, I have simulated the epidemic on static networks BA, ER,
and WS for different ranges of $\langle k \rangle$ and $\beta$. Figure 2.7 shows the result. Each row represents the result for each network. Each columns is based on different parameter sets.

In a population with more dense contact structure (a network with high $\langle k \rangle$), the disease is more likely to become an epidemic. In all three networks, Figure 2.7 (a1,b1,c1), there are more infected individuals for larger $\langle k \rangle$ equal to 20 than 6.

The second column shows results for different transition rates, $\beta$. A larger transition rate means that an individual is more likely to be infected while in contact with infected individuals. Therefore diseases with higher transition rates are more likely to become epidemics compared to those with smaller transition rates. Figure 2.7 (a2,b2,c2) clearly shows that there are more infected individuals for larger $\beta$ than smaller ones.

2.7.1.2 Dynamic Network. For the dynamic simulation, the two scenarios explained at 2.4 are used. To generate the dynamic network for the "Day and Night" network, I simulated a 'Day' network with mean degree of 9.95-10 and a 'Night' network with a reduced mean degree of 2.5-3. So the average mean degree for these would be 6. These two static networks were then combined so that every 50 time steps the network switches between day and night for a total of 1000 steps; this is equivalent to 10 full days. The mean degree overall is around 6. Figure 2.8(a)
shows the result for an epidemic simulation on the three different network types (BA, ER, WS) using my Day-Night dynamic network. The degree of nodes has been decreased compared to a static network but the degree distribution of each network is been intact. For example, BA has the power law distribution with a smaller mean degree. Thus, the total pattern of the course of an epidemic is same as Figure 2.6 and there is not a large difference in total number of infected individuals. But since individuals have fewer connections at night it is less probable that infected individuals infect others; the course of the epidemic is delayed.

For the "Gradually Decreasing Degree" network, I reduced the degree of 50 randomly picked nodes by 40 percent at at arbitrary steps 30, 60, 90, 120, and 200 over 1000 time steps. Therefore, at each of this time steps, the new network with less connectivity is used as a representation of the social contact network. Figure 2.8(b) shows that the cumulative number of infected in the gradually decreasing network will never rise as high as in the static network. But since the mean node degree does not decrease a lot, there is only a small change in the total number of infected individuals. The total number of infected is decreased in average of 5%.

2.7.2 Dynamic Versus Static Networks

Researchers have applied static networks using constant parameters that are averages over the whole population. Static contact networks have been often used to simulate epidemics. They are considered an average of a real dynamic network over time.

To show that averaging the parameters to simulate a network is not accurate, I have compared the epidemic simulations using a dynamic network with epidemics simulation using an average static network.

In our dynamic network simulation Figure 2.9(b) the averaged mean degree is around 6.5; the static network, Figure 2.9(a) has similar mean degree; \( \langle k \rangle = 6 \). Even though the two networks have a similar mean degree, the course of an epidemic is different in these two simulations. Results show that fewer people were infected in the dynamic version compared to the static network. The epidemic also peaks later and less abruptly in the dynamic network. My comparison suggests that averaging is a poor method that does not give accurate results.
2.7.3 Real Data

A potential challenge for testing the accuracy of the method is the scarcity of empirical data, specifically for dynamic contact networks. Recording each person’s contacts for a long period of time in a large society is a time consuming process. However, availability of devices that record such dynamic networks automatically are becoming more common. Wearable active Radio-Frequency Identification Devices (RFID) are able to detect face-to-face contact among individuals with a spatial resolution of about 1.5 meters, and a time resolution of 20 seconds (cf. Salathé et al., 2010). Simulations can be used to exhibit the dynamics of the epidemic in these networks.

I tested my approach with the data recorded by Salathé et al. (2010). The data were collected in a high school of 789 individuals for four groups: students, teachers, staff, others. Every day, the data were collected from 0600 until the end of the school day, although the majority of the population was present only between 0800 and 1600. Salathé et al. (2010) applied a static network using averages of contact durations as edge weights. In contrast, I used our dynamic network implementation using network structure captures by their data at every 20 seconds. I assumed that no transmission occurred at night or on the weekend because infected individuals are disconnected from the network at these times. In contrast, recovery of infected individuals can continue at night and on the weekend. The initial infected individual was randomly chosen from the population during the first week. An infected individual will not infect any other participant of the network after hours because they are disconnected from the network. I tested a range of transition rates $\beta$. The highest number of infected individuals occurs in the simulations with higher transition rate (Figure 2.10a).

Figure 2.10b shows the cumulative number of infected individuals over 100 runs. Because initiation of an epidemic occurred randomly in the first week, the peak of an epidemic also varies by one week. The fact that the propagation of the epidemic is paused at night and on weekends explains the staircase shape of the curve.
Figure 2.7: Characteristics of three random networks (BA: a1,a2; ER: b1,b2; WS: c1,c2), a1,b1,c1 show the effect of different mean degrees \( \langle k \rangle \); a2,b2,c2 show the effect of different transmission rates \( \beta \).
Figure 2.8: Course of an epidemic in a population of size 1000 over 1000 time steps. Gray bands reflect the maximal range of the epidemic trajectory in the 500 simulations. Cumulative number of infected individuals averaged over 500 simulations, comparison for (a) Day-Night (b) Gradually Decreasing Degree among BA, ER and WS networks.

Figure 2.9: Static-Dynamic network comparison
Figure 2.10: Cumulative numbers of infected individuals in epidemic simulations with real contact network. (a) Transition rate $\beta$ vary. (b) Course of 100 simulated epidemic with $\beta=0.05$. 
CHAPTER 3

INTERVENTION METHODS

3.1 Background

Mitigating or preventing the spread of infectious diseases is the ultimate goal of infectious disease epidemiology. Understanding the dynamics of epidemics is an important tool to achieve this goal. Infectious disease epidemics in recent years, such as HIV 1981, foot-and-mouth in 2002, severe acute respiratory syndrome (SARS) in 2003, and swine flu (H1N1) in 2009, highlight the need to plan for controlling the spread of infectious diseases worldwide (Woolhouse et al., 2005; Jones et al., 2008; Tildesley et al., 2006; Donnelly et al., 2004; Germann et al., 2006; Hatchett et al., 2007; Fraser et al., 2009).

The structure of contact networks is crucial in explaining epidemiological patterns seen in the spread of directly transmissible diseases such as sexually transmitted diseases, for example HIV/AIDS (Anderson and May, 1991b; Anderson, 2000; Morris, 1993), the respiratory or close-contact route (e.g. SARS and influenza-like illnesses) (McLean et al., 2005; Lloyd-Smith et al., 2005), or influenza (Ferguson et al., 2005, 2006; Halloran et al., 2008; Longini et al., 2005).

Intervention methods are classified as: 1. Pharmaceutical (PI) vs Non-Pharmaceutical (NPI) and 2. Non-Adaptive/Adaptive. Pharmaceutical interventions are those using antivirals, antibiotics, and vaccinations. In contrast, NPI interventions are methods that change the social network structure in a population without administering any drugs, for example closure of public spaces, quarantine, and sequestration. Adaptive interventions are those that change the mobility in a population like school closures, and treatments that have temporary effects such as antiviral medications which are only effective when being taken. In contrast, Non-Adaptive strategies are limited to treatments that have a permanent effect, like vaccination and treatments that occur before the
start of an epidemic. They assume that the behavior of individuals does not change while the epidemic is happening.

The main problem for public health officials and policy makers is how to apply the intervention methods to minimize transmission and control the epidemics. The intervention sources such as vaccination are usually limited and can not be applied to all individuals. Other methods such as public space closure are costly. Therefore, the main question is how to choose a group of the population that should receive priority in getting the intervention to maximize the effect and reduce the epidemic risk.

Designing effective intervention methods needs the identification of important nodes in the network. One way to do this is to detect the communities. Communities, also called clusters or modules, are groups of vertices which probably share common properties and/or play similar roles within the network. A large body of literature has studied the community structure in a network for the past several years (Girvan and Newman, 2002; Newman, 2006a; Fortunato, 2010; Wu and Huberman, 2004; Gfeller et al., 2005; Newman, 2006b; Duch and Arenas, 2005; Hu et al., 2010; Lancichinetti et al., 2010; Danon et al., 2005; Fortunato, 2010). Community detection is a computationally difficult task, however, several algorithms have been developed and tested, for example Minimum-cut method, Hierarchical clustering, Girvan-Newman algorithm, Modularity maximization, Statistical inference, and Clique based methods. They mainly divide the network into communities by finding the edges between the communities. For example, the Girvan-Newman algorithm assigns a “betweenness” number to each edge. Betweenness of an edge is the number of shortest paths between pairs of nodes. Betweenness is large when the edge lies “between” many pairs of nodes. The edges with high values are then removed, leading to a graph that contains communities. Other widely used methods are ‘modularity maximization’ and clustering. These methods identify communities by searching over all possible divisions, or similarities, respectively.

For this study, I have used an alternative method that depends on the matrix property to detect the important nodes in a network for employing intervention methods. I have compared the accuracy of this method with other graph theory methods and have shown that this new method is as reliable as those other graph theory methods. Matrix property methods are easier to use on dynamic networks which represent time series data.
I present the details of my model, its accuracy, and computational cost. I show the results for different scenarios of modeling dynamic versus static networks.

3.2 My Model

3.2.1 Eigenvector Centrality for Unweighted Adjacency Matrices

The social contact network has been represented as an adjacency matrix in my model. The adjacency matrix is square and contains entries for every pair of nodes (individuals) in the network. If the individuals share a connection (edge) then the adjacency matrix contains a one otherwise a zero. Eigenvalues and eigenvectors are two important features of the adjacency matrix. To get the eigenvalues and eigenvectors of the matrix, we have to set and solve equation 3.1. For the square matrix $A$, we can write:

$$Av = \lambda v \quad (3.1)$$

where $\lambda$ is a scalar known as the eigenvalue or characteristic value associated with the eigenvector $v$. By rearranging the equation we get:

$$(A - \lambda I)v = 0 \quad (3.2)$$

which has a solution only when its determinant $|A - \lambda I|$ equals zero. Setting the determinant to zero allows one to obtain the polynomial equation:

$$P(\lambda) = |A - \lambda I| = 0 \quad (3.3)$$

known as the characteristic polynomial of the matrix $A$. The roots of this polynomial are the eigenvalues. Once the eigenvalues are found, solving the following linear system allows one to get the eigenvectors:

$$Ax = b \quad (3.4)$$

There are different factorization methods to solve this system such as $LU$ factorization; all have complexity of $O(n^3)$. 

38
The eigenvector of the largest eigenvalue contains information about the importance of a node in a network. The largest eigenvector contains a measure of the centrality of nodes in a graph or network. It is not just related to the degree of the node, but rather its connectivity to other important nodes in the network. The centrality of a node can be expressed as:

$$x_v = \frac{1}{\lambda} \sum_{t \in G} a_{ij} x_j$$  \hspace{1cm} (3.5)

With a rearrangement, this can be rewritten in vector notation as the eigenvector equation 3.1.

There is an eigenvector corresponding to the largest eigenvalue. If we sort this eigenvector, the index is associated with the importance of the nodes in the network. Thus, by having sorted eigenvectors we can find a set of important nodes by their priority.

Let $\lambda_1, \lambda_2, ..., \lambda_n$ be the eigenvalues of an $n \times n$ matrix A. $\lambda_1$ is called the dominant eigenvalue of A if:

$$|\lambda_1| > |\lambda_i|, i = 2, ..., n$$

The eigenvector corresponding to $\lambda_1$ is called the dominant eigenvector of A.

3.2.1.1 Power Methods. Matrix decomposition methods used to calculate the eigenvalue and eigenvectors are expensive numerical methods. The adjacency matrix representing the social contact network is a very large sparse matrix. So the eigenvector calculation $O(n^3)$ would be a costly method for this specific aim. But, for obtaining the targeted nodes in a network, we only need the eigenvector corresponding to the largest eigenvalue. We do not need to follow a costly matrix decomposition method to get all the eigenvectors, therefore a less costly calculation, such as the power method may be beneficial.

The Power Method is an iterative method for finding the eigenvector corresponding to the dominant eigenvalue, which simply takes powers of the original matrix A. In order for the method to work, I assume that A is symmetric so that it is guaranteed to have real eigenvalues and a complete set of linearly independent eigenvectors. The method will give us an approximation of the eigenvector corresponding to the largest eigenvalue. Then, I will use the Rayleigh Quotient method to obtain an approximation to the corresponding eigenvalue.
The power method is a very simple algorithm. It starts with a given arbitrary vector and generates a sequence of iterations as:

\[ \hat{x}^k = A x^{k-1} \]  
\[ x^k = \frac{\hat{x}^k}{\|\hat{x}^k\|} \]  

Note that for each iteration we need to perform a matrix times vector multiplication which is \( O(n^2) \). This method converges to the eigenvector corresponding to the dominant eigenvalue because:

\[ \vec{x}^1 = A \vec{x}^0, \vec{x}^2 = A \vec{x}^1 = A^2 \vec{x}^0 \]  
Continuing in this manner:

\[ \vec{x}^k = A \vec{x}^{k-1} = A^2 \vec{x}^{k-2} = ... = A^k \vec{x}^0 \]  

Because we have assumed that \( A \) has a complete set of linearly independent eigenvectors, these vectors can be used as a basis for \( IR^n \). Consequently there are constants \( c_i \) such that:

\[ \vec{x}^0 = \sum_{i=1}^{n} c_i \vec{v}_i \]  

Using this expression in our formula for \( x_k \) and the fact that \( A^k \vec{v}_i = \lambda^k_i \vec{v}_i \) gives:

\[ \vec{x}^k = A^k \vec{x}^0 = A^k \left[ \sum_{i=1}^{n} c_i \vec{v}_i \right] = \sum_{i=1}^{n} c_i A \vec{v}_i = \sum_{i=1}^{n} c_i \lambda_i \vec{v}_i \]  

We now factor out the dominant eigenvalue term \( \lambda_1^k \) to get:

\[ \vec{x}^k = \lambda_1^k \left[ c_1 \vec{v}_1 + c_2 \left( \frac{\lambda_2}{\lambda_1} \right)^k \vec{v}_2 + ... + c_n \left( \frac{\lambda_n}{\lambda_1} \right)^k \vec{v}_n \right] \]  

As \( k \to \infty \) all the terms in the expression except the first approach zero because we have assumed that \( \lambda_1 > \lambda_i \) for all \( i \neq 1 \).

Therefore, \( \vec{x}^k \) would be the eigenvector corresponding to the dominant eigenvector estimated by the Power Method.

We can calculate the eigenvalue by using the Rayleigh quotient:

\[ \lambda = \frac{(\vec{x}^k)^T A \vec{x}^k}{(\vec{x}^k)^T \vec{x}^k} = \frac{(\vec{x}^{k+1})^T \vec{x}^{k+1}}{(\vec{x}^k)^T \vec{x}^k} \]
An alternate way to calculate the approximate eigenvalue is to take the ratio of components of successive iterates:

$$\lambda = \frac{(A\vec{x}^k)_l}{(\vec{x}^k)_l}$$  \hspace{1cm} (3.14)

Table 3.1 compares the run time for calculating the eigenvector corresponding to the dominant eigenvalue by factorization method versus the Power Method. Results clearly indicate that the Power Method converges faster.

<table>
<thead>
<tr>
<th>Matrix size/ time in seconds</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular method</td>
<td>0.3097</td>
<td>1.5271</td>
<td>11.6374</td>
<td>37.1356</td>
</tr>
<tr>
<td>Power method</td>
<td>0.0023</td>
<td>0.0436</td>
<td>0.0199</td>
<td>0.0410</td>
</tr>
</tbody>
</table>

3.2.1.2 Example. Here I have used an adjacency matrix to target the most important nodes in a network to design effective intervention methods. Figure 3.1 is an example of a simple network with nodes that has important connectivity, which influences the dynamics of other nodes in the network. I have used the equation 3.3 to calculate the eigenvalue of this network. Then I got the largest one and its corresponding eigenvector. The index of nodes corresponding to sorted eigenvector are: 13, 1, 7, 12, 2, 8, 6, 5, 9, 11, 3, 4, 10. The color map in figure match the priority of nodes in the given list.

Figure 3.1: Example of network composed of 13 nodes
3.2.1.3 Comparison with graph theory measurements. As mentioned in the review section, some concepts of the graph method such as centrality and betweenness has been used in community detection algorithms. Here I have tested some of them and have compared the result with the eigenvector method. To this end, I have calculated the betweenness and centrality of the network in Figure 3.1. The Networkx package (Hagberg et al. (2008b)) has been used for these calculations. Table 3.2 shows the result. The index column indicate the order of important nodes for each method. The results clearly imply that the eigenvector method gives the exact same node importance orders as other graph theory based methods. Therefore, instead of using community detection methods, I have used the eigenvector method for finding targeted nodes in a network to design intervention methods. Since I am using an adjacency matrix as a representation of the contact network, an eigenvector method which works with matrix features is a great choice.

<table>
<thead>
<tr>
<th>Node/Measures</th>
<th>Eigenvector Index</th>
<th>Centrality Value</th>
<th>Closeness Value</th>
<th>Betweenness Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td>0.57</td>
<td>0.272</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.25</td>
<td>0.5</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0.25</td>
<td>0.4</td>
<td>0.007</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>0.25</td>
<td>0.4</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0.25</td>
<td>0.4</td>
<td>0.007</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>0.25</td>
<td>0.5</td>
<td>0.053</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.5</td>
<td>0.57</td>
<td>0.272</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>0.25</td>
<td>0.5</td>
<td>0.053</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>0.25</td>
<td>0.4</td>
<td>0.007</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>0.25</td>
<td>0.4</td>
<td>0.007</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0.25</td>
<td>0.4</td>
<td>0.007</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>0.25</td>
<td>0.5</td>
<td>0.053</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.5</td>
<td>0.56</td>
<td>0.56</td>
</tr>
</tbody>
</table>

3.2.2 Eigenvector Centrality for Weighted Adjacency Matrices

The eigenvector of the dominant eigenvalue of the adjacency matrix gives the nodes based on their centrality importance. That value is calculated based on the node’s degree or node’s neighbors. Therefore, it works correctly when all vertices and edges in a graph have the same value, or in other words, when the network is unweighted and the matrix is symmetric.
The adjacency matrix of my model is weighted. As explained in section 2, the transition rate, or the probability that one can be infected while having a close contact with an infectious person is considered as a node weight. Each individual has a different rate or probability of being infected by contacting an infected person. This rate depends on the person’s immunity, vaccination background, etc. This node weight is mapped to the edge weight and therefore the adjacency matrix would not be symmetric anymore. Figure 3.2 is a simple example of a weighted network. Figure 3.2 (a) is a network with node weights. The outgoing edge of a node has the weight of the node. Thus, the incoming edge has its outgoing node’s weight. By this mapping strategy we get a network with weighted edges. Figure 3.2 (b) is a network that its edge’s weight is based on the node’s weight in 3.2 (a).

![Figure 3.2: Weighted network example. (a) Network with weighted nodes. (b) Network with weighted edges.](image)

The eigenvector centrality and also other centrality measurements based on graph theory use an unweighted symmetric adjacency matrix. Hence, the centrality value reported by these methods is not accurate for weighted networks. There are few studies on centrality calculation of weighted
networks. Newman (2004) simply mapped the weighted network to layers of unweighted networks. Here, I have used PageRank methods using eigenvalue and eigenvector properties of a matrix. PageRank is an algorithm used by Google Search to rank websites in their search engine results Page et al. (1999). PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. This method is very similar to my method except for having a scaling factor to handle the weights in the network. The formulation is based on the eigenvalue problem; I calculate the eigenvector of the matrix to find the important nodes in a graph.

The formulation is almost the same as equation 3.3 but using a weighted matrix plus a damping parameter. The vector \( R \) is a column vector of PageRanks for each node. PageRank is calculated as:

\[
p_i = \sum_{j \rightarrow i} \frac{p_j}{l_j}.
\] (3.15)

The \( l \) is the number of out-links or column sums. This is modified by adding a damping factor \( d \), usually set to around 0.85. Thus,

\[
p_i = \frac{1 - d}{n} + d \sum_{j \rightarrow i} \frac{p_j}{l_j} \tag{3.16}
\]

where \( n \) is the total number of nodes in the network. Then the modified adjacency matrix \( M \) is calculated by dividing column \( i \) by the number of out-links \( l_i \). Now the series of equations in equation 3.15 can be compressed to the matrix equation. Thus, it essentially reduces to an eigenvalue problem (3.3).

\[
R = \frac{1 - d}{n} 1 + dMR
\] (3.17)

where \( 1 \) is a \( n \times 1 \) vector, full of 1s. It can be rewritten as:

\[
R = (\frac{1 - d}{n} E + dM)R \tag{3.18}
\]

\[
R = \tilde{M}R \tag{3.19}
\]

\[
\tilde{M}R = 1R \tag{3.20}
\]

44
where $E$ is a $n \times n$ matrix full of 1s. Equation 3.20 can be visualized as an eigenvalue problem.

To solve this system, I can use the original method of solving for the eigenvector using a factorization method with complexity of $O(n^3)$ or using the power method explained above. Table 3.3 shows the results of the PageRank and centrality method. For the network composed in Figure 3.2, the PageRank has been hand coded and centrality method has been calculated using Networkx package (Hagberg et al. (2008b)). The centrality method report the important nodes based on their degree, but PageRank method finds the important nodes considering their weight (Table 3.3). For example with the centrality method, node 5 with weight 8 has higher order than node 3 with weight 35. But with the PageRank method, node 3 has higher order than node 5.

Table 3.3: Comparing the PageRank method with the centrality measurement for a weighted network

<table>
<thead>
<tr>
<th>Method</th>
<th>Node orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>2, 3, 5, 4, 6, 7, 8, 9, 1, 10, 11</td>
</tr>
<tr>
<td>Centrality</td>
<td>2, 5, 4, 6, 7, 8, 9, 10, 11, 1, 3</td>
</tr>
</tbody>
</table>

### 3.3 Result

Pre-outbreak strategies such as vaccination and during-outbreak responses such as case isolation, prophylactic treatment, and travel restrictions are usually tested on static networks. In other words, the social contact among individuals as well as transmission probability are constant during the course of an epidemic. However, real life is a dynamic process. Even without applying any intervention methods, the contact network and transmissibility are dynamic and change over time. In case of an epidemic, people naturally reduce their communication and avoid being in crowded placed to insure their health and decrease their risk of getting sick. People also try home remedies such as using a mask, or washing hands more frequently to increase their immunity (known as self interventions). All these behaviors happen dynamically in a population and affect the course of an epidemic. Beside these natural controlling methods, the intervention methods applied by governments in a whole society have a dynamic nature as well. For example, closing schools changes the connectivity in a network and hence the contact network will have a new structure that needs to be considered in the model.
The model I have implemented in chapter 2 works with dynamic networks as a representation of a social contact network structure. Therefore, in case of an intervention, my model allows for different network structures at different times during the course of an epidemic: for example in early detection, before, and after applying intervention methods. The application of the intervention method changes the network, and remains for the rest of the simulation or until the next intervention. Taking this modified network structure into account is crucial to have an accurate result. I have generated a set of dynamic networks to test the impact of intervention strategies.

3.3.1 Dynamic Network: Degree Decreased

As explained above, the connectivity decreases naturally in the course of an epidemic since people try to reduce their risk by being more isolated. To this end, I have generated a set of dynamic networks where the degree of random nodes decreases gradually over time. For this specific simulation, I reduced the degree of 50 randomly picked nodes out of a total of 1000 nodes by 40 percent at arbitrary time steps 30, 60, 90, 120, and 200. Therefore, at each of these time steps of the total of 1000 time steps in my simulation, the new network structure which has less connectivity is used as a representation of the social contact network in my model.

3.3.2 Intervention Methods

Method designers and policy makers try to choose the best intervention strategy and to apply it on the targeted group at a proper time during an epidemic to get the best results and control the epidemic in a cost-efficient manner. The majority of methods detect the group in a population that has the largest impact on the dynamics of a disease and target them for intervention methods.

An intervention is specified as either changing the behaviors of the intervened people or changing the infectivity/vulnerability of the susceptible people. These two different strategies result in two different intervention methods:

- **Intervention strategy 1: Edge Removal**
  Changing the behavior means decreasing or eliminating the connectivity of infected or highly
at risk infectious persons in the population. This means that the transmissions are hindered at the intervened nodes, not on the person-person interactions.

- **Intervention strategy 2: Immunity Increasing**
  This second class of interventions include vaccination and antiviral administration. A vaccinated person still visits the same places, and meets the same people. However, the disease cannot propagate via her. If we give an antiviral to a person, she will become less likely to get infected and even if she is infected she will be less likely to infect other people.

### 3.3.2.1 Experiment with intervention strategy 1: Edge removal.
Edge removal intervention strategies focus on finding the important nodes in network and reduce their connectivity or even disconnect them from network. Public space closure is an example of this class of methods. To detect the importance nodes, I have used the algorithm explained in 3.2.1, using equations 3.3 and 3.7. To test the impact of intervention method, two different methods are tested.

- **First Method: reduction of connectivity**
  The first method removes different percentages of targeted nodes at a particular time in the early stages. For example, here, I have disconnected 10%, 20% and 30% of targeted nodes.

- **Second Method: timing of intervention**
  The second method tests the impact of intervention methods applied at different times in the course of an epidemic to determine the time window of highest impact. To achieve this aim, I have applied the strategy of disconnecting edges on a constant number of nodes at different times. For example, here, I have disconnected 30% of targeted nodes at time steps 30, 50, and 100.

The two intervention methods are tested on BA, ER and WS dynamic networks generated using algorithms explained at 3.3.1.

Figure 3.3 shows the results. Figure 3.3 (a) shows clearly that disconnecting the important nodes has a high impact on spread of disease. More node removals tend to less infected individuals. Total number of infected is decreased from 820 to 300 by removing 30 percent of important nodes in network. The impact of the 10% node removal compared to no intervention is higher compared to 20% and 30%, because this specific BA network has around 100 nodes of 1000 with high degree, removing those targeted nodes has higher impact than removing other nodes with less degree.
Figure 3.3 (c) shows that the number of infected is strongly reduced with edge removal. But since the ER network has bell shaped degree distribution, it is similarly responsive to 10, 20 and 30 percent edge removal.

Edge removal method has the same impact on WS network, Figure 3.3 (e). There are fewer infected individuals after applying intervention method. Since the nodes are connected to their neighbors in WS network, disconnecting nodes, make networks having some almost isolated nodes to be infected later than sooner. Therefore the epidemic happening later. Overlay we can say that the edge removal strategy has higher impact on BA network than ER and WS network.

The second column of figure 3.3 shows the result of second intervention method (timing the intervention). Results clearly implies that applying intervention methods sooner has higher impact than later. By applying methods sooner, we not only get less infected individuals, but also the epidemic is been postponed. These time stretching, give the policy makers extra time to apply more effective methods to control the epidemic. Similar to the edge removal method, this timing method has higher impact on BA than ER and WS network. Pointing out that these are simulations on single run, so there are some stochastic effects like the the line of result of time 30 at 3.3 (f).

3.3.2.2 Experiment with intervention strategy 2: Immunity increasing. This strategy mainly increases the person’s immunity or reduces the transition probability. In other words, the node’s weight in the network is decreased. This can be represented by a weighted adjacency matrix. Therefore, at the implementation level, the intervention simulation must be capable of handling a weighted network. I have used the PageRank method described in section 3.2.2. To test the impact of an intervention method, two different methods are used:

- **First Method: Increasing vaccinated individuals**
  This tests the impact of the node’s weight reduction on different percentage of population. For example, here, to compare the effect of the method, different percentages of nodes such as 10, 20, 30, 50, and 70 percent have been chosen and their weight is been reduced by 60%.

- **Second Method: Increasing immunity rate**
  This tests the impact of intervention methods by increasing the immunity rate on the same number of individuals in population. For example, here, I have used the same number of people to be immunized, 30 percent, but have increased the immunity rate to 80% and 90%.
Figure 3.4 shows the result of applying two intervention methods. The first column in Figure 3.4, clearly implies that having more individuals with reduced vulnerability results in having less infected. But the impact of this method is more on postponing the course of an epidemic rather decreasing the total number of infected. The total number of infected after applying the intervention method in Figure 3.4 (a), (c), and (e) is almost the same and around 800. But the time of epidemic emerge is been strongly affected in all tree network. The emerging point is shifted from 100 to 200 in (a). And from 100 to almost 400 in (c). WS network, 3.4 (e) is been the most responsive to this method and the emerging point is shifted to almost 600 rather than 100. This postponing effect, will give people more time to find ways to reduce their risk of being infected.

The second column of Figure 3.4 shows the result of increasing the immunity of 30 percent of targeted nodes in a network. We see some improvement on; number of the infected is reduced by increasing the immunity rate. This method has higher impact on decreasing the number of total infected individuals than method 1. There are almost 100 fewer infected individuals in Figure 3.4 (d) and around 200 fewer infected individuals in Figure 3.4 (e) and (f). The epidemic is even more delayed in the WS network, Figure 3.4. The Immunity Increasing method has the highest impact on WS network.

### 3.4 Discussion

In this chapter the impact of two different intervention methods is tested and discussed. The first method acts as an isolation method, where the connections are eliminated and edges in the network are removed. The second method represents low-impact and high-impact immunization strategies such as using masks, using anti-viral drugs, and vaccination.

The result of edge removal intervention method on specific networks and parameter sets that I have used in Figure 3.3 clearly show that edge removal has a high impact on controlling epidemics. The number of infected individuals are significantly reduced and in some cases, such as with the WS network (Figure 3.3(e) and (f)), the epidemic is delayed as well. The result of edge removal method implies that the larger the number of targeted nodes to remove, the less the number of infected people at the end. The timing of the removal is also important and sooner is better than
later (Figure 3.3 second column). For this specific set of parameters, starting at time 30 has a significant impact on disease spread.

Comparing the results of Figure 3.3 and Figure 3.4 shows that "edge removal" intervention has higher impact than "immunity increasing" strategy. For the edge removal case, we get as few as 250 infected individuals compared to around 800 without intervention. But for "immunity increasing" strategy the minimum number of infected individuals are around 600 compared to 800 infected individuals without any treatment.

The trial that modeled immunization methods suggests that for having a higher impact, we have to either apply more effective intervention methods such as better immunization, or apply the intervention methods to more people.

The result in Figure 3.4 (a), (c) and (e) for 10%, 20% and 30% shows that by reducing the vulnerability by a small percentage the number of infected individuals are not decreasing but are increasing in some cases, although the course of the epidemic has been delayed.

Why is this increase in infected individuals happening? In my model, the probability of being infected is increased by contacting infected people and having a lower immunity rate. Increasing the immunity rate during an epidemic, decreases the rate of being infected but does not eliminate the risk. In contrast, isolation methods remove a person from the network. As a result, there is no connection and thus the probability of being infected does not rise. In summary, the probability of being infected increases at a slower rate until it reaches a threshold, where the person can be labeled as infected. That is why we see the same number of infected at the end in Figure 3.4 though the epidemic happens at a slower rate.

To make this method effective, we have to immunize more people in the population. In the language of public health, it means we have to have more effective immunization methods such as vaccination or apply less effective methods on more persons in population. To show that, I have tested two scenarios. First, with an immunity rate of 0.6, meaning the transition rate of targeted nodes are reduced by 0.6, I have increased the number of people in the population that have been immunized somehow. Second, I have used the same number of people to be immunized, 30 percent,
but have increased the immunity rate to 0.8 and 0.9. These higher immunity rates reflect use of medication and permanent immunization measures such as anti-viral drugs and vaccination.

Results in Figure 3.4 suggest that to reach the threshold where an intervention method becomes effective, we have to apply the method to at least 50% of the population.

3.5 Conclusion

The focus of this chapter was the implementation and testing of various intervention methods on dynamic networks.

A weighted adjacency matrix was used as a representation of the contact network in my model. As a result, the intervention model I have used has the advantage of using matrix properties such as eigenvalues and eigenvectors. To detect the important nodes in a network, and to target them for intervention methods, I have used eigenvector centrality measurements. To deal with the computational complexity issues when extremely large networks are available, I used the Power Method to calculate the eigenvector. The Power Method has a complexity of $O(n^2)$, whereas regular decomposition methods have a complexity of $O(n^3)$.

This eigenvector centrality measurements work well on unweighted networks. In other words, the centrality measurements are based on the degree of a node and its neighbors. In my model, I have taken the transition probability, or the probability that each person can be infected while connected to an infectious person, as a node weight. Therefore, the network representation in the model is a weighted adjacency matrix. To detect the important nodes in a weighted matrix, I have mapped the weighted matrix to a directed matrix with weighted edges. A reliable method that works with weighted matrices is PageRank. Even though, it has been originally implemented for Google’s search engine and is designed to rank connected webpages, I modified it to detect the targeted nodes in a network. The Power Method was also used in this later approach to calculate the eigenvectors.
Figure 3.3: Edge Removal Intervention Results. Rows: BA, ER and WS networks. First Column: First intervention method (edge reduction). Second Column: Second intervention method (timing of intervention)
Figure 3.4: Immunity Increasing Intervention Results. Rows: BA, ER and WS networks. First Column: First intervention method (Increasing vaccinated individuals). Second Column: Second intervention method (Increasing immunity rate).
CHAPTER 4

BASIC REPRODUCTION NUMBER ESTIMATION

4.1 Background

Controlling the spread of epidemic diseases is one of the key factors in public health. One needs to predict the behavior of the disease propagation in an early stage. This prediction would then enable policymakers to design and apply the most effective intervention methods to curb the epidemic.

One of the important parameters to predict the propagation of the disease is the basic reproduction number $R_0$; it is the number of infected individuals one already infected generates on average during the time of being infected. The basic reproduction number is also defined as the average number of infections caused by a host infected early in an epidemic. The basic reproduction concept was first introduced by Alfred Lotka Lotka (1912) and Ronald Ross Yorke and Macfie (1924), but its first modern application in epidemiology was by George MacDonald et al. (1968).

If $R_0 < 1$ the epidemic is not self-sustainable, but if $R_0 > 1$ an epidemic is possible. Larger $R_0$ will lead to an epidemic with certainty. Therefore $R_0 = 1$ is the epidemic threshold, the separation point where a disease dies out or where it grows exponentially. For example, the 1918 H1N1 flu ($R_0$ of 1.4–2.8) killed 50M people. The 2009 H1N1 ($R_0$ of 1.4–1.6) killed 284,000. The more contagious the disease the higher is the $R_0$. $R_0$ is around 2 for Hepatitis C and Ebola and around 18 for Measles Organization et al. (2007).

Traditional models do not consider the contact network and calculate the $R_0$ only based on epidemiological parameters such as transition and recovery rate. More recent models take the contact network into account, they formulate the $R_0$ considering network properties.
$R_0$ is routinely estimated for different infectious diseases by public health organizations worldwide. These estimates constitute an important resource for monitoring and comparing disease outbreaks. The use of $R_0$ is not entirely unproblematic. First, it is difficult to estimate $R_0$ in models (Heffernan et al., 2005; Jones, 2007; Van den Driessche and Watmough, 2008) and from outbreak data (Dietz, 1993; Heffernan and Wahl, 2006; Massad et al., 2010). Second, the result that $R_0 = 1$ defines an epidemic threshold rests on very coarse assumptions (Heffernan et al., 2005; Anderson and May, 1991b; Li et al., 2011).

One way to improve the $R_0$ estimation is taking contact network properties that shape the spread of a disease into account. For example one could use node degree distribution and average of node degree (Moreno et al., 2002; May and Lloyd, 2001), covariance coefficient of node degree, and clustering coefficients (Watts and Strogatz, 1998; Miller, 2009b; Smieszek et al., 2009), therefore the estimator of $R_0$ can have many forms. Many researches have proposed new estimators for $R_0$. This chapter is an overview of different formulation of $R_0$, discussing their limitations and weaknesses; I then contrast these with my own formulation and show results on specific simulated networks.

### 4.2 Traditional Formulation of $R_0$

Traditional epidemic models, for example the SIR model, do not take the connectivity in a network into account. Therefore, the basic reproductive ratio $R_0$ is only based on the transition and recovery rate. (H Trottier, 2000; Diekmann et al., 1990; Anderson and May, 1985). It is:

$$R_0 = \frac{N \beta}{\gamma},$$

where $N$ is the number of individuals; $\beta$ and $\gamma$ are the disease transition rate and the recovery rate, respectively.

$R_0$ can be formulated differently for different epidemiological models, such as SIS or SEIR, considering their own unique behavior and corresponding parameters. For example, $R_0$ for the SEIR model can be formulated as:
\[ R_0 = \frac{k \beta \lambda}{\mu (k + \mu)(\gamma + \mu)} \]  

(4.2)

where \( \beta \) is the effective contact rate, \( \lambda \) is the birth rate of susceptible, \( \mu \) is the mortality rate, 
\( k \) is the progression rate from exposed (latent) to infected, \( \gamma \) is the removal rate (Jones (2007)). 
Keeling and Rohani (2007) have extensively discussed and formulated the \( R_0 \) for different infection models.

### 4.3 \( R_0 \) Formulation Using Contact Network

#### 4.3.1 Degree Distribution

The basic statistic to characterize the structure of a network is the node degree distribution: \( P(i) \); It is the probability that a randomly chosen node will have degree \( i \), that means that it is linked to \( i \) other nodes. With this probability distribution, we can calculate the mean degree \( \langle k \rangle \) which is the mean of the probability distribution and its second derivative, the \( \langle k^2 \rangle \) which is mean-square degree:

\[
\langle k \rangle = \sum_k k P(X = k),
\]

(4.3)

\[
\langle k^2 \rangle = \sum_k k^2 P(X = k).
\]

(4.4)

Youssef and Scoglio (2011); Newman (2002); Karrer and Newman (2011); Volz and Meyers (2009); Keeling and Rohani (2007); Kamp et al. (2013) have formulated \( R_0 \) using the node degree distribution of the network as:

\[
R_0 = T \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle},
\]

(4.5)

where \( T \) is the transmissibility. Miller et al. (2011) have formulated \( T \) as:

\[
T = \frac{\beta}{\beta + \gamma}
\]

(4.6)
Using the probability generating function concept and its derivative, we can extract:

\[
\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = \frac{G''_0(1)}{G'_0(1)} \tag{4.7}
\]

The probability generating function and its first and second derivative formula is known for random networks that have been generated based on known degree distributions, such as the Poisson distribution or the Power Law distribution. But for random networks with node degree distribution that do not follow any specific distribution, the average node degrees need to be calculated by extracting the distribution from the network structure.

### 4.3.2 Coefficient of Variation

The coefficient of variation \(C_v\) of the degree distribution of a network is another parameter explaining network characteristics. It is defined as

\[
C_v = \frac{\sigma}{\mu} \tag{4.8}
\]

Where \(\sigma\) is the standard deviation and the \(\mu\) is the mean of the degree distribution.

May (2006) has formulated the \(R_0\) using \(C_v\) as:

\[
R_0 = \frac{\beta}{\gamma} \langle k \rangle (1 + C_v^2) \tag{4.9}
\]

For contact networks with a uniform degree distribution such as lattice networks where each node is connected exactly to \(n\) neighbors, the \(C_v\) is equal to zero and the formula only relies on epidemic parameters and mean degree of network degree distribution. Campbell and Salathé (2013) have used this formulation for their \(R_0\) calculation and epidemic prediction.

### 4.3.3 Clustering Coefficient

The clustering coefficient of a node in a network is the probability that the neighbors of this node are also connected to each other. The clustering coefficient of a node ranges between 0 and 1. The average clustering coefficient \(C\) of the whole network is the average of the clustering coefficients of all individual nodes. The clustering coefficient is a measure of the local connectivity or "cliqueness"
of a network. Having a high clustering coefficient is associated with the robustness of a network. Luce and Perry (1949) formulated the clustering coefficient as:

\[
C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of vertices}}
\]  (4.10)

or, more detailed

\[
C_i = \frac{\lambda_G(v)}{\tau_G(v)},
\]  (4.11)

where \( \lambda_G(v) \) is the number of triangles on \( v \in V(G) \) for undirected graph \( G \). \( \lambda_G(v) \) is the number of subgraphs of \( G \) with 3 edges and 3 nodes, one of which is \( v \). \( \tau_G(v) \) is the number of triples on \( v \in G \). Thus, \( \tau_G(v) \) is the number of subnetworks with 2 edges and 3 nodes, one of which is \( v \) and such that \( v \) is incident to both edges. It can simply be written as:

\[
\tau_G(v) = C(k_i, 2) = \frac{1}{2}k_i(k_i - 1)
\]  (4.12)

The effect of clustering on epidemics has been studied by several researchers, for example Britton et al. (2008); Miller (2009a); Newman (2009); Badham and Stocker (2010); Isham et al. (2011); Gleeson et al. (2010); Volz et al. (2011).

Despite determined attempts, it is still unclear how, and to what extent, clustering affects epidemic spread. Several highly influential studies report contradictory results: The study by Newman (2003b) shows increasing clustering increases \( R_0 \). But Kiss and Green (2008) examined the same network model and show that the degree distribution also changed by changing clustering, thus in fact the variation of the degree distribution lowered the epidemic threshold and not the clustering. In 2009, new studies by Newman (2009) using generating function methods, showed that in a network with isolated triangles, clustering decreases the epidemic threshold, thus, increasing \( R_0 \). However, Miller (2009a) reported that in the same network model, assortativity, and not clustering, is responsible for the lowering of the epidemic threshold. Britton et al. (2008) have used the same randomly clustered graph as Newman (2003a) and showed that in a graph with constant mean degree, increasing clustering would decrease the \( R_0 \). But it is difficult to credit their observation since in their model, some other factors such as the variance of degree distribution
Table 4.1: Different studies finding on clustering effect on $R_0$

<table>
<thead>
<tr>
<th>Author</th>
<th>Clustering Coefficient</th>
<th>$R_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miller et al</td>
<td>Increased</td>
<td>Decreased</td>
</tr>
<tr>
<td>Keeling et al</td>
<td>Increased</td>
<td>Decreased</td>
</tr>
<tr>
<td>Newman 2003</td>
<td>Increased</td>
<td>Increased</td>
</tr>
<tr>
<td>Newman 2009</td>
<td>Increased</td>
<td>Increased</td>
</tr>
<tr>
<td>Britton et al</td>
<td>Increased</td>
<td>Increased</td>
</tr>
<tr>
<td>Ours in WEs</td>
<td>Increased</td>
<td>Increased</td>
</tr>
</tbody>
</table>

is affected while clustering is increased. Using a correlation equation formulation Keeling (2005), shows that clustering increase $R_0$. Table 4.1 summarizing these inconsistencies.

Finally, the study by Miller (2009a) is arguably the most successful treatment of epidemic dynamics on clustered networks to date. It describes a general formalism which takes both triangles and larger sub-structures into account, and finds:

$$ R_0 = \frac{\langle k^2 - k \rangle}{\langle k \rangle} T - 2 \frac{\langle n_\Delta \rangle}{\langle k \rangle} T^2 - \frac{\langle n_{\square} \rangle}{\langle k \rangle} T^3 + O(\frac{T^4}{\langle k \rangle}) $$

(4.13)

where $n_\Delta$ and $n_{\square}$ denote the average number of triangles and squares each node may be found in, respectively.

Molina and Stone (2012) have formulated $R_0$ using both clustering coefficient and mean degree, the simplest format of their formulation is:

$$ R_0 = (1 - CT) \frac{\langle k^2 - k \rangle}{\langle k \rangle} $$

(4.14)

They discuss their formulation and the clustering effect and indicate that their result clearly implies that clustering fails to have a major impact on the epidemic threshold. Because in general the term $CT \leq 1$. Nevertheless, increasing clustering will lead to an increase in this threshold, however small. Because of this small effect of clustering on $R_0$ and also the correlation between clustering and other topological properties such as assortativity, its effects might be difficult to measure both in simulation and experiment.
Despite of a large numbers of studies, the effect of clustering on epidemic spread is still not fully understood. In the next section, I will discuss my observations on the effect of clustering on epidemic spread.

4.4 Comparison of $R_0$ Estimators for Erdős-Rényi and Watts-Strogatz Networks

The degree distribution, clustering coefficient, and coefficient of variation give important information, but they do not define the structure of a network uniquely. Networks can have the same degree distribution $\langle k \rangle$ while other network properties are different resulting in different behavior of the disease spread. $R_0$ formulations that depend only on $\langle k \rangle$ are insufficient.

For example, Figure 4.1 shows two network structures with the same degree distribution. The value for mean degree, $\langle k \rangle$, square mean degree $\langle k^2 \rangle$ as well as the coefficient of variation is the same for both configurations. Calculating $R_0$ using either equation 4.5 or 4.9 generates the same value for both network structures. Table 4.2 shows the different network property values for this example. Despite that the network structures looks different their properties such as mean degree and coefficient of variation are the same. However, the clustering coefficient is different.

![Figure 4.1: Two Networks with same degree distribution](image)

4.4.1 Experiment

To test the effect of degree distribution, Coefficient of Variation, and Clustering Coefficient in $R_0$ value and formulating my own $R_0$, I have studied the Erdős-Rényi (ER; Erdős et al., 1959) and

The ER network is one of the well known random graphs where the node degree has a Poisson distribution in the limit of a large number of nodes. The WS model is a random graph generation model that produces graphs with small-world properties, including short average path lengths and high clustering. The degree distribution in the case of the ring lattice is just a Dirac delta function centered at \( \langle k \rangle \). By rewiring the network, disconnecting an edge and rewiring it again with a randomly chosen node, one can generate a new random network. By increasing the rewiring probability the degree distribution will approach a Poisson distribution. In other words, by increasing the rewiring probability, the randomness is increased and the WS approaches the ER network properties.

I used the ER network and three different types of WS network topologies, defined by different rewiring probability \( p \): highly structured \( p = 0.01 \), intermediately structured, i.e. small world \( p = 0.1 \), and highly unstructured \( p = 0.5 \). These networks will be the benchmark networks in this study to find an accurate \( R_0 \) estimator. The mean degree \( \langle k \rangle \) is the same for all three networks but their degree distribution is different. Figure 4.2 shows the degree distribution of these networks. Figure 4.3 Anonymous (2010) shows the changes on a WS network by changing rewiring probability.

![Figure 4.2](image)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Network (a)</th>
<th>Network (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \langle k \rangle )</td>
<td>2.25</td>
<td>2.25</td>
</tr>
<tr>
<td>( \langle k^2 \rangle )</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Coefficient of Variation</td>
<td>2.538</td>
<td>2.538</td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>0.21</td>
<td>0.458</td>
</tr>
</tbody>
</table>
To examine the accuracy of the $R_0$ estimation by equations 4.5, 4.9, and 4.14 for this specific network types, I calculated $R_0$ and I then run my method described in chapter 2 to estimate the total number of infected individuals and epidemic behavior. The results of the simulations are averages over 500 runs. For the model parameters in this simulation, I used a transition rate of 0.005 and a recovery rate of 0.05. Each of these equations and corresponding network property is discussed.

\subsection*{4.4.1.1 Mean Degree $\langle k \rangle$.} The mean degree and mean square degree are the same for the four benchmark networks, with the same transition and recovery rate, this leads to the same $R_0$ estimate calculated by formula 4.5. Therefore, the epidemic would spread in the same manner in all four cases. Despite of the same $R_0$ value, the course of an epidemic in fact is different; the maximum number of infected and the initial time of outbreak are different (Figure 4.4). There are more infected individuals in a WS network with lower rewiring probability than with higher probability. The ER network has a smaller number of infected individuals compared to the WS networks. Thus, equation 4.5 that relies on mean degree of the network is not a precise $R_0$ estimator for this networks.

\subsection*{4.4.1.2 Coefficient of Variation.} To evaluate the accuracy of $R0$ estimated by the equation 4.9 for the benchmark networks, first the coefficient of variation has been calculated using equation 4.8. As shown in the third row of Table 4.3, the values are the same for all network. By having the same coefficient of variation, mean degree, and epidemic parameters, the equation 4.9 would give the same $R0$ value for all studied networks. Thus, the $R0$ estimation suggest that equation 4.9 does not estimate $R0$ correctly for these networks. I can conclude that the mean
degree and coefficient of variation properties of a graph do not cover all graph properties that affect the spread of disease.

Table 4.3: Epidemic characteristic and network summary statistics. All parameter values are average of 500 simulations in network with 1000 nodes, $\beta = 0.005$ and $\gamma = 0.05$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ER</th>
<th>WS p=0.5</th>
<th>WS p=0.1</th>
<th>WS p=0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle k \rangle$</td>
<td>10.18</td>
<td>12.02</td>
<td>10.95</td>
<td>10.10</td>
</tr>
<tr>
<td>$\langle k^2 \rangle$</td>
<td>103.18</td>
<td>132.62</td>
<td>109.93</td>
<td>92.49</td>
</tr>
<tr>
<td>Coefficient of Variation</td>
<td>0.0099</td>
<td>0.0099</td>
<td>0.0099</td>
<td>0.0099</td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>0.0109</td>
<td>0.29</td>
<td>0.55</td>
<td>0.656</td>
</tr>
<tr>
<td>Spectral gap</td>
<td>4.5</td>
<td>3</td>
<td>0.867</td>
<td>0.09</td>
</tr>
<tr>
<td>Max number of infected</td>
<td>793.6</td>
<td>900</td>
<td>963.3</td>
<td>995.60</td>
</tr>
<tr>
<td>Time of early stage</td>
<td>109.4</td>
<td>102.30</td>
<td>125.3</td>
<td>173.20</td>
</tr>
<tr>
<td>$R_0^*$</td>
<td>1.0049</td>
<td>1.2040</td>
<td>1.75</td>
<td>8.9805</td>
</tr>
</tbody>
</table>

4.4.1.3 Clustering Coefficient. Many studies have reported different $R_0$ estimation for networks with clusters (Table 4.1). To examine the effect of clustering on the estimation of $R_0$ of the benchmark networks in this study, I calculate the clustering coefficient for each network and $R_0$ using equation 4.14. The values are reported in Table 4.4.

$R_0$ is decreasing while the clustering coefficient is increasing, except for ER and WS $p = 0.5$. Table 4.3 shows that the clustering coefficient is positively correlated with the number of infected
individuals, thus $R_0$ should also be positively correlated with the cluster coefficient, but is not. We can conclude that equation 4.14 does not calculate an accurate $R_0$ for the benchmark networks studied here.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ER</th>
<th>WS p=0.5</th>
<th>WS p=0.1</th>
<th>WS p=0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering Coefficient</td>
<td>0.0109</td>
<td>0.29</td>
<td>0.55</td>
<td>0.656</td>
</tr>
<tr>
<td>$R_0$</td>
<td>9.126</td>
<td>9.768</td>
<td>8.587</td>
<td>7.670</td>
</tr>
</tbody>
</table>

### 4.5 My Formulation

The topology of a contact network strongly affects the probability of a disease outbreak, because each social contact network is a complex random network with unique structure and properties. Therefore, $R_0$ should be formulated in a way to estimate the best value regarding the corresponding network properties. Equations 4.5 and 4.9 are not the best $R_0$ estimator for the benchmark networks studies here. These equations only rely on mean degree, mean square degree and the coefficient variation of degree distribution. Additionally, many studies take the clustering coefficient into account, but Table 4.1 and Table 4.4 show that the effect of clustering is not been fully understood.

Taken together, the results suggest that the network properties such as mean degree, clustering coefficient, and coefficient of variation are not sufficient to project the network structure shaping the epidemic. My own improvements of the estimation of $R_0$ is based on the use of the matrix representation of the contact network. This matrix representation allows me to use eigenvalue and corresponding eigenvector to reveal important network information.

#### 4.5.1 $R_0$ Based on Eigenvalue Properties

I incorporated eigenvalues properties as a new scalar for the $R0$ formulation to improve the estimation.
Using Singular Value Decomposition methods, the adjacency matrix $A$ of the contact network can be written as:

$$A = U \Sigma V^T,$$

where $U$ and $V$ are orthogonal matrices and $\Sigma$ is an $m \times n$ rectangular diagonal matrix with decreasingly ordered non-negative real numbers known as the singular values. I will treat the eigenvalues and singular values equivalently.

The matrix $A$ can also be expressed as the sum of rank 1 matrices:

$$A = \tilde{u}_1 \sigma_1 \tilde{v}_1^T + \tilde{u}_2 \sigma_2 \tilde{v}_2^T + \ldots + \tilde{u}_p \sigma_p \tilde{v}_p^T = \sum_{i=1}^{r} \tilde{u}_i \sigma_i \tilde{v}_i^T$$

(4.16)

where $r$ denotes the rank of $A$. If we truncate this series after $p$ terms, then we have an approximation to $A$ that captures the most important features of the network. The best way to get this $p$ is to eliminate the smaller singular values. To do so, I calculated the rate of changes among singular values and eliminated those with less than an arbitrary cutoff value of 0.005. To calculate this rate for any given network type, the average of all singular values for all simulations with a particular network was calculated. Figure 4.5 shows the average of the first 200 eigenvalues for each network type. The rate of change of the singular values differs for different network topologies, even though different networks can have the same mean $\langle k \rangle$ and $\langle k^2 \rangle$.

As shown in Figure 4.5, the rate of changes among singular values are different for different networks. This figure is the outcome of one simulation. The variation of the results of different simulations using the same parameters is high. Singular values in the matrix $\Sigma$ are ordered decreasingly but the isolation between the singular values is high, therefore the arbitrary cutoff value does not indicate the well-defined smallest rank of the matrix accurately. Hence, the rate of change would be different for different simulation on a particular network type and that would not be a good metric to improve $R_0$ estimation.

### 4.5.2 $R_0$ Based on Spectral Gap Measurement in Eigenvalue Properties

Networks with similar $\langle k \rangle$ and $\langle k^2 \rangle$ can still lead to difference of the spread of infectious diseases, because their own unique structure is not considered correctly in the $R_0$ estimation. As shown in
Figure 4.5: First 200 eigenvalues of different networks.

Figure 4.5 by increasing the rewiring probability in WS network, we increase the randomness and get networks with structure more similar to ER network. Thus, the feature we are changing here is network randomness and connectivity, but still keep $\langle k \rangle$ and $\langle k^2 \rangle$ the same.

Spectral gap is a matrix property that projects the randomness feature in a network. The spectral gap is the difference between the largest and second largest eigenvalue. If the difference is large, then the network has good connectivity, expansion, and randomness properties.

I have used the spectral graph properties of the adjacency matrix to project the randomness in the $R_0$ estimator. I have added this spectral gap metric as a scalar to the equation 4.5 to allow better estimation of the epidemic behavior. There is an inverse relationship between spectral gap and maximum number of infected (Table 4.3). So I used the $\frac{1}{\text{Spectral Gap}}$ to take this inverse relationship into account. The maximum number of infected reported are high (Table 4.3) implying that $R_0$ is higher than 1 and an epidemic is evident. So I have $1 + \frac{1}{\text{Spectral Gap}}$ as a multiplier in
my formula. My $R_0^\ast$ then is:

$$R_0^\ast = (1 + \frac{1}{SG}) \frac{\beta}{\beta + \gamma} \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}$$

(4.17)

where $SG$ is spectral gap value.

The last row of Table 4.3 also shows our $R_0^\ast$ calculation. The $\langle k \rangle$, $\langle k^2 \rangle$, maximum number of infected and $SG$ values are averages over our 500 simulation runs with transmission rate $\beta = 0.005$ and recovery rate $\gamma = 0.05$.

Additionally, I tested my $R_0^\ast$ formulation on another set of networks with $\langle k \rangle = 4$. The WS networks are generated by increasing rewiring probability as in the earlier networks. Table 4.5 shows that decreasing the spectral gap leads to an increase of the total number of infected individuals. The result conducts the same conclusion as for the cases where $\langle k \rangle = 10$, reported in Table 4.3.

The $R_0^\ast$ estimation for ER and WS $p = 0.5$ is less than 1. Meaning no epidemic is happening. But there is a chance of an epidemic in WS $p = 0.1$ and WS $p = 0.01$. The $R_0^\ast < 1$ may suggest that there is no epidemic outbreak, but whether that is accurate is not clear, except that the networks with the values below 1 also peak with many fewer infected individuals than the networks with $R_0^\ast > 1$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ER</th>
<th>WS p=0.5</th>
<th>WS p=0.1</th>
<th>WS p=0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle k \rangle$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$\langle k^2 \rangle$</td>
<td>15.03</td>
<td>13.5</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Spectral gap</td>
<td>0.859</td>
<td>0.429</td>
<td>0.083</td>
<td>0.006</td>
</tr>
<tr>
<td>Max number of infected</td>
<td>604</td>
<td>668</td>
<td>746</td>
<td>923</td>
</tr>
<tr>
<td>$R_0^\ast$</td>
<td>0.537</td>
<td>0.712</td>
<td>2.348</td>
<td>30.18</td>
</tr>
</tbody>
</table>

4.5.2.1 Clustering Coefficient and $R_0^\ast$. As shown in Table 4.3, the clustering coefficient values are increasing for these benchmark network. The total number of infected individuals and
also my $R_0^*$ are increasing, too. My observations and my $R_0^*$ corroborate the results of Newman (2003b, 2009); Britton et al. (2008).

4.6 Discussion

The ER and WS networks with different rewiring probabilities have been used as benchmark networks in my study. If we increase the rewiring probability of WS network, we increase the randomness and the network structure moves toward ER network structure. These benchmark networks with different rewiring probabilities, all have same mean and square mean degree, yet the network topology is different. Hence, the behavior of epidemic and total number of infected is also different.

I have shown that previously formulated $R_0$s are poor estimators for an specific set of random networks. I used the mean degree, clustering coefficient and coefficient of variation of nodes in the formulas designed by other studies to calculate the $R_0$. Since the mean degree, the square of mean degree, and covariance of coefficient are the same for the benchmark networks, the $R_0$ value is the same for all of those. But the total number of infected and time for epidemic emerge is different among these networks. Thus the earlier $R_0$ estimators based on these network properties are poor estimators for these benchmark networks.

To formulate the $R_0$ than can capture the randomness in network, I have tried two different methods. First, I used the SVD method to get the rate of changes of eigenvalues of adjacency matrix. This metric is not an accurate metric because there is a high variance among different simulated networks. Therefore, the arbitrary cutoff value is not able to detect the first few important nodes accurately.

For the second method, I used the spectral gap property of the eigenvalues of the corresponding adjacency matrix. I have used this metric as a multiplier in my formulation. I have shown that this new formulation estimate the $R_0$ that corroborates the disease behavior. The $R_0$ values calculated for these benchmark networks with different mean degree implies that spectral gap is an accurate metric and overcomes these problems and shows promise.
CHAPTER 5

ESTIMATION OF DYNAMIC SOCIAL NETWORKS USING GENEALOGIES OF THE PATHOGENS

5.1 Background

There are typically four main factors that influence the propagation of infectious diseases. The social contact network among the host individuals, the transition rate, the recovery rate, and the genealogy of the pathogen. Current gene sequencing technologies have opened a new gateway in epidemiology research. Sequencing the genome of pathogens has become more affordable and faster. I will review the studies that have modeled the propagation of epidemic disease spread taking into account genealogy data and present a method to choose among social networks using the genetic data of the pathogen.

Evolutionary biologist studying epidemiology typically get the genetic data, reconstruct the evolutionary tree based on the data and then use coalescence theory to estimate the origin of disease and epidemic parameters using Bayesian inference (Bedford et al., 2010). Several researchers (Pybus et al., 2012; Drummond et al., 2002; Lemey et al., 2010; Stadler and Bonhoeffer, 2013; Leventhal et al., 2014; Kühnert et al., 2014) have implemented a diffusion model to account for the movement of the pathogen over geographical locations. Extension of this approach, known as phylodynamic analysis make inferences of epidemiological processes from viral phylogenies (Grenfell et al., 2004). Volz et al. (2009); Leventhal et al. (2012); Kühnert et al. (2014); Leventhal et al. (2012); Vaughan and Drummond (2013) mapped the coalescent theory to the SIR epidemic model. The relationship of susceptible and infected individuals is mapped to the branching process in a coalescent tree.

The accurate identification of the route of transmission taken by an infectious agent is critical to understanding its epidemiology and to the design of effective prevention methods. The transmission tree shows the path of the disease propagation with epidemiological modifications, and describes
the transmission events between infected hosts, which differs from the phylogenetic tree. The phylogenetic tree, however, describes the ancestral relationships between pathogens sampled from the hosts. The leaves of the phylogenetic trees are the sampled pathogens, and its internal nodes are the most recent common ancestors of the sampled and the transmitted pathogens. A key problem is that these trees differ both in timing of the internal nodes and in topology. These differences also become more pronounced when a higher fraction of infected hosts is sampled.

Cottam et al. (2008); Jombart et al. (2011); Lieberman et al. (2011); Teunis et al. (2013); Mollentze et al. (2014) have used a statistical framework, Bayesian inference, to estimate the transmission tree of an outbreak of an infectious disease and show how the phylogenetic tree of sampled pathogens is related to that. These models try to use genetic data and Bayesian framework to get the transmission tree. These methods typically identify transmission events as branching events in a phylogeny, and they do not consider within-host genetic diversity. This simplification may be appropriate for pathogens with a fairly long generation time.

Worby et al. (2014); Didelot et al. (2014); Ypma et al. (2013, 2011); Morelli et al. (2012) have tried to overcome limitation of previous works to construct a more accurate transition tree.

Didelot et al. (2014) first reconstruct a timed phylogenetic tree and then infer the underlying transmission network, given the observed phylogeny. This two-step approach represents an approach to reconstruct transmission events on top of a timed phylogeny.

Ypma et al. (2011) moved this approach forward by constructing an inference scheme that uses spatial, temporal, and genetic data simultaneously, but assuming these data are independent of each other. Genetic and epidemiological data are evidently correlated, and a rigorous inference scheme should estimate the likelihood of a transmission tree accounting for these correlations.

To improve the assumption, Morelli et al. (2012) uses epidemiological data (times of reporting and removal from the infected population, spatially confined hosts, their locations, and estimates of the age of an infection based on clinical signs) together with pathogen sequences obtained from infected hosts to estimate transmission trees and infection dates during outbreaks.
Ypma et al. (2013) improved their previous work and presented a consistent way to use pathogen genetic sequence data in transmission tree reconstruction by simultaneously estimating the phylogenetic tree of the pathogens and the transmission tree.

All these models use genealogical data at some level to have more accurate epidemiological parameter estimation and epidemic prediction. The constructed transmission trees show the route of propagation of a disease in a population. Even though this transmission tree is constructed based on some temporal and spatial information of sampled data, it is still not a good representation of the underlying social contact network. In order to fill in the gap, I use the genealogy information to directly reconstruct the accurate underlying social contact network that influences the propagation of disease in a population.

5.2 Hidden Markov Model

A Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved or hidden states. In a HMM, one does not know anything about what generates the observation sequence. The number of states, the transition probabilities, and from which state an observation is generated are all unknown. (Ghahramani and Hinton, 2000; Ghahramani, 2001; Li et al., 2000; Zhang, 2004; Durbin, 1998). Figure 5.1 is an example of HMM states and probabilities.

Hidden Markov Models have been well studied and used in bioinformatics and voice recognition. HMM for time series data has also been used in financing and stock market modeling. In the following pages I will present an overview of the HMM for time series data and its general application. I will then discuss its effectiveness in my proposed modeling of constructing dynamic contact network considering genealogy information of sampled pathogens from hosts.

An HMM is composed of a five-tuple: \((S, K, \Pi, A, B)\):

1. \(S = 1, ..., N\) : set of hidden states. \(S_t\) denotes state at time \(t\).
2. \(K = k_1, ..., k_M\) : output alphabet. \(M\) is the number of observations.
3. $\Pi = \pi_i$ : Initial state distribution. The probability of the hidden state that the process starts with, where $\pi_i = P(s_1 = i)$

4. $A = a_{ij}$ : State transition probability distribution. A matrix with elements representing the transition probability between hidden states, where $a_{ij} = P(s_{t+1} | s_t) \ 1 \leq i, j \leq N$

5. $B = b_j(o_t) : Emission$ probability distribution. The probability of an observation given hidden states. The emission probability matrix is an $M \times N$ matrix containing probability of each observation given all hidden states. Th probabilistic function for each state $j$ is:

$$b_j(o_t) = P(o_t | s_t = j)$$

5.2.1 Three Fundamental Problems HMMs Can Solve

There are three fundamental problems that HMM solve:

1. Given a model $\mu = (A, B, \Pi)$, and an observation sequence $O = (o_1, ..., o_T)$, it efficiently computes the probability of the observation sequence given the model : $P(O | \mu)$
2. Given a model $\mu$ and the observation sequence $O$, calculates the state sequence $(1, ..., N)$ that best "explains" the observations.

3. Given an observation sequence $O$, and a space of possible models, it adjusts the parameters so as to find the model $\mu$ that maximizes $P(O|\mu)$.

The first problem can be used to determine which of the trained models is most likely when the training observation sequence is given. The second problem reveals the hidden path that generates the observation sequence. The third problem is of most significant interest for my work because it allows estimating parameters of the model. I have explored the second and third problem to find a proper solution to combine genealogy of the pathogen with the social contact network.

### 5.2.2 Baum-Welch Algorithm

To solve the third problem, to find the model parameter that best describes the observation sequence, the Baum-Welch algorithm is commonly used. This algorithm was introduced by Leonard E. Baum and Lloyd R. Welch in the late 1960s. The Baum-Welch algorithm uses the well-known Expectation Maximization algorithm to find the maximum likelihood estimate of the parameters of a hidden Markov model given a set of observed feature vectors. This algorithm estimates the parameters that maximize the:

$$\arg\max_\mu P(O|\mu)$$  \hspace{1cm} (5.1)

It has been implemented as a forward-backward algorithm. Because many different state paths can give rise to the same observation sequence, the probabilities for all possible paths should be added to obtain the full probability. This number of possible paths increase exponentially with the length of observed sequence, therefore the brute force evaluation is not practical. The forward-backward algorithm is a dynamic programming approach to calculate this probability. In the forward algorithm the probability $\alpha_i(t)$ of observed sequences up to and including $o_t$ is been calculated. In the backward algorithm $\beta_i(t)$ is been obtained by a backward recursion starting at the end of the sequence.

Definition of intermediate variables $\alpha$ and $\beta$ for final formulation:
\[ \alpha_i(t) = P(o_1, \ldots, o_{t-1}, s_t = i|\mu) \]  \hspace{1cm} (5.2)

\[ \alpha_i(1) = \pi_i \]  \hspace{1cm} (5.3)

\[ \alpha_j(t + 1) = \sum_{i=1}^{N} \alpha_i(t) a_{ij} b_j(o_t) \]  \hspace{1cm} (5.4)

\[ \beta_i(t) = P(o_{t+1}, \ldots, o_T|s_t = i, \mu) \]  \hspace{1cm} (5.5)

\[ \beta_i(T) = 1 \]  \hspace{1cm} (5.6)

\[ \beta_i(t) = \sum_{j=1}^{N} a_{ij} b_j(o_t) \beta_j(t + 1) \]  \hspace{1cm} (5.7)

\[ p_t(ij) = P(s_t = i, s_{t+1} = j|O, \mu) \]  \hspace{1cm} (5.8)

\[ = \frac{P(s_t = i, s_{t+1} = j, O|\mu)}{P(O|\mu)} \]  \hspace{1cm} (5.9)

\[ = \frac{\alpha_i(t) a_{ij} b_j(o_t) \beta_j(t + 1)}{\sum_{m=1}^{N} \alpha_m(t) \beta_m(t)} \]  \hspace{1cm} (5.10)

\[ = \frac{\alpha_i(t) a_{ij} b_j(o_t) \beta_j(t + 1)}{\sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_m(t) a_{mn} b_n(o_t) \beta_n(t + 1)} \]  \hspace{1cm} (5.11)

\[ \gamma_i(t) = P(s_t = i|O, \mu) \]  \hspace{1cm} (5.12)

\[ = \sum_{j=1}^{N} P(s_t = i, s_{t+1} = j|O, \mu) \]  \hspace{1cm} (5.13)

\[ = \sum_{j=1}^{N} p_t(ij) \]  \hspace{1cm} (5.14)

These variables are used to re-estimate the model parameters as:
\[ \pi_i' = \gamma_i(1) \] (5.15)

\[ a_{ij} = \frac{\sum_{t=1}^{T} p_t(ij)}{\sum_{t=1}^{T} \gamma_t} \] (5.16)

\[ b_{ij} = \frac{\sum_{t=1}^{k} p_t(ij)}{\sum_{t=1}^{T} \gamma_t} \] (5.17)

### 5.2.3 Viterbi Algorithm

Finding the best state sequence given a model and the observation sequence is also known as “decoding” (Durbin (1998)). This is a problem often encountered in parsing speech and in pattern recognition. An efficient algorithm, the Viterbi algorithm, based on dynamic programming, is often used to find the best state sequence. This algorithm was described in 1967 by Andrew Viterbi.

Formulation of this algorithm for fixed observation sequence is:

\[
\arg\max_{s'} P(S', O | \mu) \] (5.18)

where \( S' = (s_1, s_2, \ldots, s_T) \) is the state sequence and \( O = (o_1, o_2, \ldots, o_T) \) is the observation sequence.

\[ \delta_j(t) = \max_{s_1, \ldots, s_{t-1}} P(s_1, \ldots, s_{t-1}, o_1, \ldots, o_{t-1}, s_t = j | \mu) \] (5.19)

The \( \delta_j \) store the probability of observation sequence using the most likely path that ends at state \( i \); \( \psi_i(t) \) stores the node of the incoming edge that leads to this most probable path. The calculation is done by induction, similar to the forward-backward algorithm, but it uses maximization instead of summation.

Initialization:

\[ \delta_i(1) = \pi_i b_i(o_1) \] (5.20)

\[ \psi_i(1) = 0 \] (5.21)

Induction:

\[ \delta_i(t) = b_{ij} \max_{o} \delta_i(t-1) a_{ij} \] (5.22)

\[ \psi_i(1) = \arg\max [\delta_i(t-1) a_{ij}] \] (5.23)
Best path:

\[ s_t^* = \psi_{t+1}(s_{t+1}^*) \] (5.24)

where,

\[ s_T^* = \operatorname{argmax}\{\delta_i(T)\} \] (5.25)

### 5.3 My Model

There are few recent studies that combine the genealogy data and the epidemiological transmission network as well as epidemic parameters to predict the epidemic behavior, or estimate the epidemic parameters such as \( R_0 \). They do not take into account the underlying social contact factors. The transition tree estimated by those models can be a representation of a contact network but as a static and not a dynamic representation.

I combine both the genealogy information data and social contact network data within a unified model. The tools necessary to execute such a model, namely rapid genetic sequencing, higher computational speed and larger digital storage capacities, are becoming more and more efficient and affordable. Hence, I have devised an algorithm and constructed a model that incorporates the temporal data collection and the dynamic contact network in order to connected the underlying observed genomic patterns with social networks.

I have therefore generated a set of static networks amongst samples of the pathogen and the representation of how they switch over time.

This set of dynamic networks captures the dynamics of the social contact network of infected people in real life. This will be the first demonstration of a relationship between the observed genomic sequences of a pathogen and the social contact network that changes over time. This approach shall thus enhance our understanding of epidemiological and genomic patterns which we can then use to offer solutions and ways to arrest the progression of an epidemic.
5.3.1 HMM Formulation and Hidden and Observed State and Their Probabilities

The joint distribution of a sequence of hidden states and observations can be formulated as:

\[ P(S_1:T, O_1:T) = P(S_1)P(O_1|S_1) \prod_{t=2}^{T} P(S_t|S_{t-1})P(O_t|S_t) \] (5.26)

The Baum-Welch algorithm needs the hidden states: \( S \) and sequence observation states: \( O \) as well as transition probabilities: \( P(S_t|S_{t-1}) \) and emission probabilities: \( P(O_t|S_t) \) as input.

**Observed sequence**
The observation sequences are initially the sequenced genome data of pathogens sampled from hosts. The genome data of pathogens of infected host is sampled over time and therefore reflect a time series of genetic data. This genetic data represents the evolutionary relationship among the sampled pathogens and can be represented as a phylogenetic tree. There are different methods to construct this phylogenetic tree. There are well known packages to construct the most likely tree based on genome data such as Migrate (Beerli and Felsenstein, 1999), BEAST (Drummond et al., 2012), and Paup* (Swofford, 2003).

**Hidden states**
The hidden states are a set of networks that explains the dynamics among the contacts among individuals in population. I have generated an initial arbitrary network based on the given information such as location and time of sampled data. The information such as age, mobility, public spaces, and their traffic can be used to generate the network if those are available. The other set of networks are generated randomly by manipulating contacts among individuals.

Figure 5.2 shows the model formulation for epidemic sequential data.

**Emission Probabilities** To formulate the HMM, we must also define the transition probabilities between hidden states and emission probabilities. Here, each hidden state represents one network structure.
To calculate the emission probabilities, $P(O_t|S_t)$, the likelihood value of a tree, based on the given network, must first be calculated: $P($genealogy tree$|$network structure). Tree can be a relaxed version of a network where each node is only connected to two other nodes directly, (i.e. the daughter nodes).

To compare these two structures (tree and network) and calculate the likelihood, I generated a weighted matrix for both hidden and observed states. For the hidden states, the contact networks, I calculated the shortest path length among each two nodes and report these in a distance matrix. The Networkx package (Hagberg et al. (2008b)) is used to calculate this values. To normalize the value, each distance is divided by the largest path length.

$$N_{ij} = \text{normalized shortest path between } i \text{ and } j \quad (5.27)$$

For the observation states, the genealogy tree, I calculated the Patristic Distance matrix. The patristic distance is defined as the number of mutational differences between two tips on a tree. These pairwise distances are also reported in a distance matrix. The Python module, Dendropy Sukumaran and Holder (2010) is used for this calculation. The values are divided by the maximum divergence value.

$$G_{ij} = \text{normalized amount of divergence between } i \text{ and } j \quad (5.28)$$
Therefore, both network path lengths and divergence values are proportional and normalized. Subsequently, both the tree and the network structures have been transformed into a weighted adjacency matrix. To get the likelihood value I have calculated the Euclidean distance between the two matrices as:

\[
A = (a_{ij}); \\
A_{ij} = ||N_i - G_j||^2
\]  

(5.29)  
(5.30)

The sum of each row in the matrix or in other words, the sum of probability of each state given all observation should sum to one: \( \sum P(O_{1:t}|S_t) \). To normalize the value, each matrix element is divided by the norm one of the corresponding matrix row. Therefore, the emission probability matrix contain the probability value of the observed phylogenetic tree given the network structure.

After setting up the initiation of model parameters, the Baum-Welch algorithm is used to estimate the model parameters that best fit the observation. After having found the best transition and emission probability matrix, I have used the Viterbi algorithm to find the best state sequence path that fits the observation. Finding the hidden state sequence path plays an important role in my method, because that path indicates the dynamics of the networks. This is a key factor in my model in chapter 2 to predict the disease transmission behavior in the population.

\section*{5.4 Result}

The new model was tested with simulated data and one real data set from a study by Morelli et al. (2012).

\subsection*{5.4.1 Experiment on Simulated Data}

To test the model in simulated data, four simulated data sets are needed: Observed sequence, Hidden states, Emission probability matrix, and transition probability matrix.
To simulate the observed sequence data, I generated a simple example of seven samples of a pathogen. Early in an outbreak, there is little information available about infected individuals. Few cases are infected and not all have been tested. Over time, when the disease propagates, more samples of the pathogen are available. Thus, there are fewer sampled pathogens at early outbreak than later.

Sampled data are the DNA of pathogens which have been sequenced. In my model, I need the phylogenetic tree of sampled data as my observation sequence. A genealogy tree has been constructed at each time point when the data was sampled. Hence, the number of samples increase over time, the genealogy tree also get more tips over time. For this simulation, I have generated four arbitrary genealogy trees with increasing number of sampled pathogens. Figure 5.3 shows four trees of sampled pathogens over time.

For the hidden state simulation, I generated six arbitrary random networks of seven nodes. The network topology is different among these networks. All have the same number of nodes but different connection between nodes. Since the size of population is fixed and only the number of samples pathogens increase over time, all networks have the same number of nodes as the largest number of samples. Since these networks represent a small population, the variation on their connectivity is low, therefore the network topologies are very similar. Figure 5.4 shows five networks representing the connectivity among host individuals.

To calculate the emission probability matrix for these five networks and four trees, I have used the method in 5.3.1. The probability of each observation sequence data, the tree, given the hidden state, the network, is reported in 5.1. The bold value in each column represent the network topology.
that fit best to that particular tree. The highest value indicates that the best emission probability belongs to network #4 which is the best match for tree #4.

The transition probability matrix is set in a way that all transitions are equally likely between all hidden states. There are five hidden states here, thus each transition probability is 0.2.

I tested the model for several different observation sequences, keeping the original 5 hidden states. First, I tested a case with only one observed tree for nine time steps. With only one data point we will be only able to find the most probable underlying network. Second, I picked two
Table 5.1: Emission Probability Matrix

<table>
<thead>
<tr>
<th>Networks/Trees</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7122</td>
<td>0.6752</td>
<td>0.7409</td>
<td>0.8716</td>
</tr>
<tr>
<td>2</td>
<td>0.7180</td>
<td>0.6910</td>
<td>0.7392</td>
<td>0.8518</td>
</tr>
<tr>
<td>3</td>
<td>0.6892</td>
<td>0.6859</td>
<td>0.7735</td>
<td>0.8514</td>
</tr>
<tr>
<td>4</td>
<td>0.6679</td>
<td>0.6645</td>
<td>0.7666</td>
<td>0.9010</td>
</tr>
<tr>
<td>5</td>
<td>0.6892</td>
<td>0.6859</td>
<td>0.7735</td>
<td>0.8514</td>
</tr>
</tbody>
</table>

observed data, the first and the last tree, each tree was repeated for 4 times for a total of 8 time steps; that means we have the same collected data for some consecutive time points. Third, the observed sequence contains all four trees, but each tree is repeated a few times and the observed sequence has a length of 20. Table 5.2 shows the observed sequence and the most probable hidden state sequence calculated by the Viterbi algorithm.

Since the transition probability values are equal, changing between hidden states based on the observation are mostly relying on emission probabilities for this experiment. For the first experiment, where we only have one tree as observed state, the best sequence path that matches the observation is network #4. This outcome makes sense according to the emission probability matrix, where the best underlying network for tree #4 is network #4. The best sequence path for the second experiment is also convincing because the best network estimated for tree number #1 is network #2. The third experiment is a more realistic case, where we have more observed sequence for a longer time period. The outcome is also promising; it shows that the dynamic of the contacts among individuals have a pattern similar to their emission probabilities. The best network estimated for tree number #1 and #2 is network #2. The best network estimated for tree number #3 and #4 are networks #3 and #4 respectively.

Table 5.2: Best state sequence path

<table>
<thead>
<tr>
<th>Genealogies</th>
<th>Observed Sequence</th>
<th>Best Hidden State Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(4,4,4,4,4,4,4,4,4)</td>
<td>(4,4,4,4,4,4,4,4,4)</td>
</tr>
<tr>
<td>2</td>
<td>(1,1,1,4,4,4,4)</td>
<td>(2,2,2,4,4,4,4)</td>
</tr>
<tr>
<td>4</td>
<td>(1,1,1,2,2,2,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4)</td>
<td>(2,2,2,2,2,2,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4)</td>
</tr>
</tbody>
</table>
5.4.2 Experiment With Real Data

There are a few studies that use genealogy data to study the epidemic of diseases such as Tuberculosis, SARS, and Food-and-Mouth Disease Virus (Didelot et al., 2014; Jombart et al., 2011; Didelot et al., 2014; Morelli et al., 2012). They mainly estimate the best transmission tree that indicates the best connection path between hosts. This transmission tree indicates some contact structure between hosts based on the genealogy information, not the physical contact between hosts, but still, could reveal some information about the underlying contact network.

To test the accuracy of my method, I compared my results with the study the Food-and-Mouth Disease Virus outbreak in UK in 2001 and 2007 by Morelli et al. (2012). In addition to the sequenced genome data, the location of farms has been reported as a map and the distance between them was recorded. I did not have access to all detailed data, therefore some information had been extracted from their plots and figures.

I extracted the pathogen tree from their genealogy tree and the number of mutations between nodes. I used the number of mutations as branch length. Since the data are samples over time, I did generate four different trees over time. Since the number of mutations are increasing over time, the weights on my matrix are dynamic and change over time. These trees had been used as observation sequences. Figure 5.5 show the trees.

![Figure 5.5: Phylogenetic trees of Foot-and-Mouth virus sampled from infected hosts at four different times.](image)

For the hidden states, the set of dynamic networks, I generated five different networks. The first network is generated based on the transmission tree reported by Morelli et al. (2012) and has
been manipulated to generate the other four networks. All networks have the same number of nodes with different connectivity and edge weight. These are weighted directed networks with edge weights indicating the distance between farms (Figure 5.6).

![Networks](image)

Figure 5.6: Underlying contact networks

Table 5.3 shows the emission probabilities of four trees and five networks. The best network representation for each observation tree based on the probability value has been marked in bold. The best network that describe the genealogy tree #4, is network #5.

<table>
<thead>
<tr>
<th>Hidden States / Observed States</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6557</td>
<td>0.7444</td>
<td>0.7839</td>
<td>0.8161</td>
</tr>
<tr>
<td>2</td>
<td><strong>0.6680</strong></td>
<td><strong>0.7476</strong></td>
<td>0.7595</td>
<td>0.8249</td>
</tr>
<tr>
<td>3</td>
<td>0.6542</td>
<td>0.7415</td>
<td>0.7818</td>
<td>0.8226</td>
</tr>
<tr>
<td>4</td>
<td>0.6531</td>
<td>0.7473</td>
<td>0.7763</td>
<td>0.8233</td>
</tr>
<tr>
<td>5</td>
<td>0.6628</td>
<td>0.7409</td>
<td>0.7602</td>
<td><strong>0.8361</strong></td>
</tr>
</tbody>
</table>

I ran two experiments. First, there is one tree as observed state. Second, there are more observed sequence for a longer time period. The best underlying network that represents the tree that has all sampled pathogens is network #5. In network #5, farms are connected with their closest neighboring farms. This connectivity pattern fits with the path of propagation of disease between farms. The second experiment which is more realistic, shows the dynamics of the contact networks. There is no data available for the contact network among farms, but based on the distance between farms, the network #5 looks reasonable to have the highest probability to match with the
tree containing all sampled pathogens. The pattern of switching between contact networks (Table 5.4 second experiment) is promising regarding the number of connections and distance between infected farms.

Table 5.4: Best state sequence path

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Observed Sequence</th>
<th>Best Hidden State Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>(4,4,4,4,4,4,4,4,4)</td>
<td>(5,5,5,5,5,5,5,5,5)</td>
</tr>
<tr>
<td>Second</td>
<td>(1,1,1,2,2,2,3,3,3,4,4,4,4,4)</td>
<td>(2,2,2,2,2,2,1,1,1,5,5,5,5,5)</td>
</tr>
</tbody>
</table>

The Morelli et al. (2012) paper only estimates the transmission tree that shows how the disease propagate between farms. The location of farms and distance between them is available but no data regarding their connectivity has been reported. I used the transmission tree as a starting point for the network construction. Then, I added weighted edges between farms considering the distance between them. By using the HMM, I have estimated the set of dynamic networks that describes the dynamics of spread of a disease on contact network between host. These dynamic network reveal the connectivity among farms during the course of the epidemic. Having the dynamic contact network among host may lead us to a better epidemic prediction and intervention.

5.5 Discussion

In this chapter I reviewed the few recent studies that combine the genealogy information with epidemiological parameters to estimate the model parameters and epidemic behavior. The new generation of genome sequencing machines that generate genome data quickly and cheaply have opened a new gate for genomic epidemiology. Additionally, speed and storage capacity of computers has increased massively in recent years, and enables us to do massive simulations and analysis more affordably and quickly. Thus, we can have more precise modeling of epidemic diseases by taking more factors into account. Accurate estimation leads to more effective intervention methods and controlling of an epidemic.

In my model, I combine the genealogy information with contact network structure as well as epidemiological parameters to estimate the dynamic network representing the social behavior of the host. The transmission tree constructed by other studies using genome data only reveals the
path of propagation of a disease, and nothing related to the host contact structure. Hence, the underlying network structure is missed. Even the simple network that can be detected from a transmission tree is static and not dynamic. My HMM, in contrast to other models, produces the set of dynamic networks representing the contact network among hosts that shape the behavior of the epidemic as well as the probability of switching between them. The transition probability estimated by HMM indicate how the model switches between the underlying networks over time to emit the genealogy tree of sampled pathogens. The result of both simulated data and real data suggest that my approach is promising.

To have more precise results, I need to improve my model in two ways. First, I can get more precise phylogenetic trees from the sampled pathogens. Or, better, instead of one tree per time point I could integrate over all possible trees conditioned in the sequence data using a Bayesian phylogenetic estimator. A further improvement would be to take population genetic modeling, such as incorporation of migration among populations of pathogens into account (Beerli, 2006; Beerli and Palczewski, 2010; Palczewski and Beerli, 2014). Migration and mobility are very common in contact networks. If we can project this factor into the phylogeny tree, the calculation would be even more promising.

Second the probability calculation can be improved. The emission probability is calculated based on the distance matrix. This can be improved by treating the branch length on a tree and edge weight in networks in more sophisticated way.

5.6 Future Work

The main limitation of the models reviewed here, as well as mine, is that all are based only on the sampled data. But in reality, we have genome data of few infected hosts that have been sampled randomly. Usually, those who have visited hospitals or health care centers. Especially in my modeling considering the social contact network, the sampled nodes are much fewer than the total number of nodes in a contact network representing the whole population. Yet, to have accurate epidemic prediction we need a dynamic contact network structure as close to the real network structure as possible. To this aim, as a future work to continue my research in this area, I
will implement a method to extrapolate the contact network of a population as a whole from only a few nodes.

In order to take the identified dynamic networks developed in chapter 2 and extrapolate to a network that is representative of the entire population in districts relevant to the outbreak, I will use a network algorithm technique and epidemiological features of the underlying population. There are multiple methods that have been used to reconstruct sparse random graphs (since the extrapolated population will have less infections) with few samples or participants. These models reconstruct the hidden larger matrix based on the fewer nodes and some feature information available such as their neighbors. Methods such as randomization and Singular Value Decomposition (SVD) are simple models to reconstruct the network based on some nodes. However, since a real social contact network is a complex structure, containing information about the targeted society, such as age range, average household size, public spaces such as schools, hospitals, our algorithm will contain more complexity and parameters to consider all these aspects. SVD, which works by approximating a matrix with a smaller set (Erdős et al., 2014), is relevant to our scenario because it captures the most important features of the data which can be unique in a real-world situation. Since SVD works with matrices, it would further be suitable for me because my networks derived in chapter 2 will be represented by weighted adjacency matrices. Erdős et al. (2014); Vuokko and Terzi (2010); Anandkumar et al. (2011) describe other methods which I will consider as a base line to design my model.
BIBLIOGRAPHY


Dr. Ashki was born in 1980 in the US and was raised in Tehran, Iran. She returned to the US in 2006 to continue her graduate education. She holds a bachelor’s degree in software engineering from Azad University of Tehran, a master’s degree in bioinformatics from Indiana University, and a doctorate degree in scientific computing from Florida State University (2015). Her PhD research focused on the stochastic modeling of epidemic diseases. Aside from her academic background, she also has an extensive professional work experience in database administration, web development, and software development. Her current interests involve modeling the spread of infectious diseases and improving the already existing models. She is also working on a project to increase awareness and control the spread of sexually transmitted diseases in the middle eastern countries. Dr. Ashki is now the mother of her 9-month old daughter, Hannan. Her husband, Hannan, and she will be moving to California where Dr. Ashki will continue her endeavors in research.