Statistical Methods for Network Epidemic Models

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Mark Burch, B.A., M.Sc.

Graduate Program in Biostatistics

The Ohio State University

2016

Dissertation Committee:

Grzegorz A. Rempala, Ph.D., D.Sc., Advisor
Joseph Tien, Ph.D.
Bo Lu, Ph.D.
Abstract

The 2013-2016 Ebolavirus outbreak in West Africa has resulted in more than 34,000 suspected, probable, and confirmed cases of the disease and more than 11,000 deaths. Due to the threat to Public Health posed by large-scale epidemics of diseases such as Ebolavirus and others, it is imperative to study the propagation dynamics of these diseases in human communities. Recently, a novel class of epidemic models utilizing networks of infectious contacts has been proposed to describe these dynamics. However, historically the use of these models has largely been confined to theoretical studies of the spread of human disease. In this work we contribute several new methods to the existing body of work on such models in order to enhance their application to real-world epidemic modeling problems. First, we present novel results concerning the deterministic dynamics of these models in large populations. Next, we study novel statistical methods to harness common forms of Epidemiological data in order to model the spread of ongoing epidemics, estimate the level of intervention needed, and quantify uncertainty surrounding these insights. Lastly, we derive new analytical results concerning various interventions and the probability that the epidemic dies out early on. For illustration, these methods are applied to datasets obtained from recent Ebolavirus outbreaks in the Democratic Republic of the Congo in 2014 and the 2013-2016 Ebolavirus outbreak in West Africa.
To Emma, my family, and the many others that have supported me along the way.
Vita

May 22, 1989 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Born - Cincinnati, Ohio

2011 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B.A. Mathematics and Economics

2013 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . M.Sc. Statistics

2013-present . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Graduate Research Associate,
The Ohio State University.

Publications

Research Publications


Fields of Study

Major Field: Biostatistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iii</td>
</tr>
<tr>
<td>Vita</td>
<td>iv</td>
</tr>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Organization</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Table of symbols and notation</td>
<td>4</td>
</tr>
<tr>
<td>2. Scientific and epidemiological context of Ebolavirus</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Scientific understanding of Ebolavirus</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Ecology of Ebolavirus</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Social characteristics and the spread of Ebolavirus</td>
<td>10</td>
</tr>
<tr>
<td>2.3.1 Epidemiological studies of Ebolavirus in West Africa</td>
<td>16</td>
</tr>
<tr>
<td>2.4 Conclusion</td>
<td>20</td>
</tr>
<tr>
<td>3. Models of infectious disease</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Historical models of epidemics</td>
<td>22</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Law of large numbers for the stochastic SIR model</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Diffusion approximation</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Estimation for counting processes</td>
</tr>
<tr>
<td>3.2</td>
<td>Large population limits of network-based models</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Edge-based compartmental modeling</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Pair approximation approach</td>
</tr>
<tr>
<td>4.</td>
<td>Models of Ebolavirus</td>
</tr>
<tr>
<td>4.1</td>
<td>Models of previous Ebolavirus outbreaks</td>
</tr>
<tr>
<td>4.2</td>
<td>Models of the West Africa Ebolavirus outbreak</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Data usage</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Modeling approaches</td>
</tr>
<tr>
<td>4.3</td>
<td>Clinical trials</td>
</tr>
<tr>
<td>4.4</td>
<td>Dynamic network-based modeling approach</td>
</tr>
<tr>
<td>5.</td>
<td>Dynamic Network-Based Model of Ebolavirus</td>
</tr>
<tr>
<td>5.1</td>
<td>Layered Configuration Model random networks</td>
</tr>
<tr>
<td>5.1.1</td>
<td>The Excess Degree Distribution</td>
</tr>
<tr>
<td>5.1.2</td>
<td>SIdaR process</td>
</tr>
<tr>
<td>5.1.3</td>
<td>SIdaR Law of Large Numbers</td>
</tr>
<tr>
<td>5.1.4</td>
<td>Relationship to edge-based limiting systems</td>
</tr>
<tr>
<td>5.1.5</td>
<td>Pairwise limiting systems</td>
</tr>
<tr>
<td>5.2</td>
<td>The SEIdaR process</td>
</tr>
<tr>
<td>5.3</td>
<td>The hybrid process</td>
</tr>
<tr>
<td>5.4</td>
<td>Conclusion</td>
</tr>
<tr>
<td>5.5</td>
<td>Proof of the $SIdaR(r,k)$ law of large numbers</td>
</tr>
<tr>
<td>6.</td>
<td>Branching Process Approximation</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>6.2</td>
<td>Branching process approximation</td>
</tr>
<tr>
<td>6.3</td>
<td>DRC dataset</td>
</tr>
<tr>
<td>6.4</td>
<td>Epidemic model</td>
</tr>
<tr>
<td>6.5</td>
<td>Statistical inference</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Index case offspring distribution</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Post-index case offspring distribution</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Summary of common mathematical notation used in future chapters.</td>
</tr>
<tr>
<td>2.1</td>
<td>Table of important outbreaks of Ebolavirus considered in this document.</td>
</tr>
<tr>
<td>2.2</td>
<td>Table of epidemiological estimates obtained from select studies in the literature. The unit of measure where applicable is assumed to be $d=$days. Where applicable we use the country identifiers L=Liberia, SL=Sierra Leone, G=Guinea, and DRC=Democratic Republic of the Congo.</td>
</tr>
<tr>
<td>3.1</td>
<td>Table of transitions and rates for the stochastic (Markovian) SIR model.</td>
</tr>
<tr>
<td>5.1</td>
<td>Listing of possible transitions of the $SI_{daR}(r,k)$ process and corresponding notation.</td>
</tr>
<tr>
<td>5.2</td>
<td>Transitions for the $SI_{daR}(r,k)$ process according to the $2r+1$ possible event types with corresponding rates. Network arrangements corresponding to the transitions are also given with $\hat{j}$ and $\check{j}$ denoting, respectively, active and deactivated edges of type $j$ between nodes (denoted $u$, $v$ and $w$). Here, $N^l_{Y,u}$ denotes the set of $l$-neighbors of node $u$ with disease status $Y$.</td>
</tr>
<tr>
<td>5.3</td>
<td>Listing of possible transitions of the $SEI_{daR}(r,k)$ process and corresponding notation.</td>
</tr>
<tr>
<td>5.4</td>
<td>Transitions for the $SEI_{daR}(1,0)$ Markov process according to the 3 possible event types with corresponding rates.</td>
</tr>
</tbody>
</table>
7.1 Transitions for the approximate process formed by using the limiting equation to compute $[SI]$. ................................. 139

8.1 Table of parameters, values, and references for the simulation study. .... 157

8.2 Table of parameters, estimated value, and standard errors for the three methods in the simulation study. OLS(P) denotes the least squares estimate with partial (i.e. excluding dyads) information while OLS(C) denotes the least squares estimate with complete (i.e. including dyads) information. 164

8.3 Table of parameters, estimated value, and bootstrap confidence intervals for the OLS and hybrid estimators applied to the Guinea dataset. .... 171
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Comparison of the Negative Binomial distribution ($\hat{r} = .959$, $\hat{\mu} = 6.11$) fit against the contact tracing dataset of Fallah et al. [35].</td>
</tr>
<tr>
<td>3.1</td>
<td>A realization of a stochastic $S - I - R$ model plotted against the solution of the limiting ODE.</td>
</tr>
<tr>
<td>4.1</td>
<td>Table of compartments and transitions in the model proposed by Legrand et al.</td>
</tr>
<tr>
<td>4.2</td>
<td>Table of compartments and transition rates in the compartmental model proposed by King et al.</td>
</tr>
<tr>
<td>4.3</td>
<td>Table of compartments and transition rates in the compartmental model proposed by Scarpino et al.</td>
</tr>
<tr>
<td>4.4</td>
<td>Table of compartments (S,E,I,R) and network dynamics (D,A) for the dynamic network-based modelling approach considered in the next chapter.</td>
</tr>
<tr>
<td>6.1</td>
<td>The empirical secondary case distribution in the DRC outbreak dataset (neglecting the index case), as given by Maganga et al. [77].</td>
</tr>
<tr>
<td>6.2</td>
<td>Estimated posterior densities for the parameters of interest for the non-Markovian MCMC sampler. Green line denotes the posterior mean.</td>
</tr>
</tbody>
</table>
6.3 Final outbreak size distribution based on 20,000 simulations of the branching processes from the posterior parameter distribution. The actual outbreak size of 69 based on the DRC dataset (black line) is shown for comparison. ........................................... 133

8.1 The estimated (limiting) number of infected individuals over time for the simulated dataset with parameters in Table 8.1 ........................................... 159

8.2 The bootstrap estimates of $\theta$ from the OLS estimator with complete (including dyad) information. ......................................................... 160

8.3 The bootstrap estimates of $\theta$ from the OLS estimator with incomplete (excluding dyad) information. ......................................................... 161

8.4 The bootstrap estimates of $R_0$ from the OLS estimator with complete (including dyad) information. ......................................................... 162

8.5 The bootstrap estimates of $\theta$ from the hybrid estimator. ................................. 163

8.6 The estimated number of infected individuals over time for the Guinea dataset when $\gamma = \sigma = .1$. ......................................................... 167

8.7 Estimated density of $\hat{R}_0$ for the Guinea dataset when $\gamma = \sigma = .1$ ........ 168

8.8 Estimated bounds for the number of infected individuals over time for the Guinea dataset when $\gamma = \sigma = .1$. ......................................................... 169

8.9 Bivariate scatterplots of the bootstrap estimates from simulated trajectories of the $SEI\alpha R(1,0)$ model with parameters $\hat{\theta}_{hyb}$. ........................................... 170
Chapter 1: Introduction

The 2013-2016 West Africa Ebola virus outbreak has brought to light the relevance of epidemic modeling at a time when many acute illnesses are no longer considered a credible threat to Public Health. The increasing connectivity between regions, countries, and continents makes understanding the mechanisms that drive the spread of new acute illnesses and our ability to control them at the source more important now than ever before. Ebolavirus in particular is deadly, poorly understood, and capable of causing a large-scale threat to Public Health.

Mathematical modeling of infectious disease in human and non-human populations has a long history dating back to the Susceptible-Infectious-Recovered (SIR) model of Kermack and McKendrick in the 1920s [60]. Extensions and generalizations of this model were numerous throughout the remainder of the 20th century but the same essential elements largely did not change. More recently the field has seen the introduction of random network models intended to address some well-noted criticisms of spatially homogeneous models. Nevertheless, further generalization and analysis of the Mathematical properties of network-based models is necessary to increase their relevance in Public Health applications.
and to assess their relevance to real-world outbreaks. Our goal is to provide some of this
analysis in ways that allow us to model the outbreak dynamics of Ebolavirus in human pop-
ulations. This is accomplished with a careful review of the scientific literature pertaining
to Ebola Virus and the mathematical literature pertaining to the existing epidemic model-
ing tools in human populations. A strong understanding of both will allow us to make the
proper extensions.

The present document is organized with this in mind. Our end goal can be thought of
as threefold. First, we create a network-based modeling framework able to incorporate the
existing epidemiological knowledge concerning the spread of Ebolavirus. Second, we de-
velop the mathematical framework to derive analytical results and understand the dynamics
induced by this class of models. Lastly, we compare these dynamics to the empirical course
of the outbreak in West Africa and the Democratic Republic of the Congo. In turn we will
use this analysis to gain insights into Public Health interventions used during the previous
outbreak as well as future outbreaks.

1.1 Organization

The rest of this thesis is organized as follows. Chapter 2 will summarize the current sci-
entific and epidemiological literature relevant to the study of the spread of Ebolavirus. This
includes details from previous Ebolavirus outbreaks as well as epidemiological insights
gained from the most recent outbreak in West Africa.

Chapter 3 presents the classical approaches to epidemic modeling and more recent work
on network-based approaches. This study includes the derivation of large volume limit
results that provide a link between stochastic epidemic models and deterministic counterparts. These results form a foundation in studying the large-scale (population-level) dynamics of the models under study and motivate similar methods used in later sections.

Chapter 4 takes a closer look at some specific models of Ebolavirus that have been studied during previous outbreaks. We review the essential features of these models to be retained when applying them in a network-based framework.

Chapter 5 will give the assumptions of our network-based model and present the analytical framework to rigorously derive large volume limits for network-based models. In addition, we present an approximate stochastic system that is more amenable to statistical analysis and which possesses equivalent limiting dynamics to the network-based models under study. These results form an important starting point for the statistical methods proposed in Chapter 7.

Chapter 6 proposes statistical estimation methods for estimating epidemic parameters early in an epidemic; that is, before we may apply techniques that rely on the assumption of being in the large volume limit. We apply this method to a dataset from the 2014 outbreak in the Democratic Republic of the Congo and demonstrate how these insights may be used to model the final size of an epidemic and estimate the basic reproductive number $R_0$.

Chapter 7 proposes Least Squares and approximate likelihood methods to estimate the epidemiological parameters of the model specified in Chapter 5. The analytical properties of these estimators are studied and basic properties established for estimation of $R_0$. 
Chapter 8 applies the Least Squares and approximate likelihood methods of Chapter 7 to data from Guinea during the 2013 – 2016 Ebolavirus outbreak. We compare the performance of the estimation methods to perfect (network-wide) information in a simulation study. Lastly, we demonstrate the utility of this model by describing the level of Public Health interventions necessary to stop the propagation of the outbreak in Guinea and estimating $R_0$.

Chapter 9 provides our concluding remarks on both the outbreak in West Africa and further extensions of the model not explored here.

1.2 Table of symbols and notation

For easier referencing, we summarize the notation used throughout the remainder of the document.

For a stochastic jump process observed up to time $t$ with transitions at times $\{t_i\}_{i=1}^m$, $t_m \leq t$ of type $\{v_i\}_{i=1}^m$ the process history is defined as $\mathcal{H}_t = \{t_i, v_i\}_{i=1}^m$, i.e. it consists of the type and timing of these events.
Symbol | Description
---|---
\( \beta \) | Rate of infection
\( \sigma \) | Rate of infection onset \((E \to I)\)
\( \gamma \) | Rate of infection onset \((I \to R)\)
\( \delta_j \) | Rate of \( j \)-type contact de-activation
\( \eta_j \) | Rate of \( j \)-type contact activation
\( D_v \) | Random degree of a node \( v \)
\( \mu, K \) | Mean degree, i.e. \( E(D_v) \)
\( \sigma^2 \) | \( \text{Var}(D_v) \)
\( p_k \) | \( P(D_v) = k \)
\( r \) | Number of network layers
\( d \) | the dimension of \( \theta \)
\( \theta \) | vector of free parameters (excluding Chapter 5)
\( \theta(t) \) | \( P(v \in S|d_v = 1) \) (parts of Chapter 5 only)
\( \psi(x) \) | Probability generating function
\( N \) | Population size
\( \mathcal{G}_r(\psi,N) \) | space of \( r \)-Layered Configuration Model networks
\( [AB] \) | Total network connections s.t. \( i - j, i \in A, j \in B \)

**Stochastic Process Notation**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S, E, I, R )</td>
<td>Susceptible, Exposed, Infectious, Recovered</td>
</tr>
<tr>
<td>( X_S, X_E, X_I )</td>
<td>Number of susceptible, exposed, infectious individuals</td>
</tr>
<tr>
<td>( X_{AB} )</td>
<td>Number of ( A - B ) connections, ( A, B \in {S, E, I} )</td>
</tr>
<tr>
<td>( X_{\not{A}B} )</td>
<td>Number of inactive ( A - B ) connections, ( A, B \in {S, E, I} )</td>
</tr>
<tr>
<td>( j )-neighbor ( X_j )</td>
<td>Neighbor of type ( j ) in a multilayer network</td>
</tr>
<tr>
<td>( i \in A ) ( X_{AB,i} )</td>
<td>the number of ( j )-neighbors of ( i ) of disease status ( B )</td>
</tr>
<tr>
<td>( C_l(t) )</td>
<td>Counting process of the number of type-( l ) events</td>
</tr>
<tr>
<td>( \lambda^l(t) )</td>
<td>The hazard of type-( l ) events in a jump process</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Limiting initial condition of a stochastic process</td>
</tr>
<tr>
<td>( \mathcal{H}_t )</td>
<td>History of stochastic process</td>
</tr>
</tbody>
</table>

Table 1.1: Summary of common mathematical notation used in future chapters.
2.1 Scientific understanding of Ebolavirus

Ebolavirus was first characterized after an outbreak in the Central African nation of Zaire in 1976. At the time, it was noted that the hemorrhagic symptoms induced by the infection resulted in a high mortality rate and that the virus could be described as similar to Marburg Virus. Antigen tests performed using indirect immunofluorescence indicated that the pathogen was distinct from the previously identified Marburg Virus although similarities between the symptoms they induced were noted. The name Ebolavirus was proposed because the village of Yambuku (the location of the first identified case) resides along the Ebola River in Southern Zaire (now Democratic Republic of the Congo (DRC)) [55].

Ebolavirus and the closely related Marburg Virus are examples of a broader family of viruses referred to as Filoviridae. Ebola is a genus of Filoviridae which contains the recognized species Ebolavirus, Sudan ebolavirus, Reston ebolavirus, Tai Forest ebolavirus, and Bundibugyo ebolavirus [92]. Historically, all significant outbreaks in human communities
have been classified as either Ebolavirus or Sudan Ebolavirus and it is has not been es-
tablished if these genetic differences result in differential risk to human or other primate 
populations [92]. As of March 2014 the outbreak in West Africa was considered to be the 
species Ebolavirus, although a significant number of mutations have been observed as the 
number of cases has increased [92]. Recent estimates suggest that the overall nucleotide 
variation within the same species of Ebolavirus are in the 3 – 5% range but this figure is 
significantly higher for Marburg virus [23].

Our intent is to summarize the characteristics of Ebolavirus and social characteristics in 
affected (and non-endemic) areas that are important to understanding its spread in human 
populations. This includes the timeline of disease progression, typical symptoms, empirical 
findings of infectious contact patterns, and methods of transmission between individuals. 
While a number of distinct species of the virus have been identified to date it is for our 
purposes sufficient to consider these subtypes as essentially equivalent. For example, upon 
identification of a new subtype of the virus in Ivory Coast in 1994, the clinical presentation 
described is effectively the same as the typical clinical presentation noted elsewhere [92]. 
Reston ebolavirus (found in the Phillipines) is the exception to this rule as it has not been 
found that this induces a symptomatic response in humans [37]. That is, it is our under-
standing that the genetic variants among the other known species of Ebola do not greatly 
impact any of the basic characteristics of interest here. The same is assumed to be true for 
human populations exposed to Ebola. Heterogeneity of viral strains and human popula-
tions (for example, in available treatment and Public Health response) can possibly impact
the duration, infectiousness, and case fatality in an epidemic; however, we will expect that certain characteristics will be found in common between most outbreaks.

Infection with Ebola typically induces a severe hemorrhagic fever in humans with a case fatality rate that is estimated near 70% for the outbreak under study in West Africa [107]. There are some reports that clinical presentation of the symptoms may vary across species but these differences have not been studied extensively. In humans, Ebola may incubate for a period of 2-21 days after initial infection (estimated 10 – 11 days in West Africa outbreak) and eventually induces Ebola hemorrhagic fever with multisystem effects [107]. A short list of the early symptoms includes severe fever, nausea, diarrhea, vomiting, chest pain, dysnea, cough, ocular oedema, hypotension, conjunctivitis, headaches, and coma [37]. Severe hemorrhaging is typically observed at the peak of the infection and is observed in only about one half of infections. Human-to-human transmission is thought to occur mostly through direct contact with bodily fluids. As a result, control measures to stop the spread of the disease typically includes contact tracing, careful burial practices, and patient quarantine among others.

2.2 Ecology of Ebolavirus

It is presently believed that the existence of Ebolavirus in humans is a case of zoonosis [37]. The reservoir species for Ebolavirus was not known well into the 1990s. A reservoir species represents an ecological host that is typically asymptomatic to infection of the virus but is essential to its persistence. After the 1995 Kikwit, Democratic Republic of the Congo (DRC) outbreak significant research was conducted to identify the reservoir species
with proposed species ranging from bats to plants [92]. Fruitbats are now considered the reservoir species but are believed to be part of a more complex inter-species transmission network.

In particular, many questions remain about the inter-species dynamics between ape populations, fruit bats, and human populations. Nonhuman primate species may be relevant to our study because Ebolavirus has adversely impacted populations of several central African ape species [16, 112]. The impact of Ebolavirus on nonhuman primates is not our primary goal here but remains an important topic both because of the role of human-ape contact in the start of human epidemics and the effect of the virus on endangered ape populations. The general framework presented here may also be applicable to nonhuman primates with appropriate modifications to reflect different social behaviors and the impact of this on contact networks.

Fruitbats were proposed as a possible reservoir based on their geographic distribution prior to the studies following the Kikwit outbreak. Finally in the early 2000s several studies established fruit bats as the primary reservoir species by discovering Ebolavirus antibodies and Ebola RNA in *Hypsognathus monstrosus, Epomops franqueti,* and *Myonycteris torquata* [105, 69] In the last ten years, Zaire Ebolavirus has been discovered in a number of other similar fruit bat species [96]. To date a host species is still unknown for Tai Forest, Bundibugyo, and Sudan Ebolaviruses. The diverse geographic distribution of these fruit bat species is of particular concern because it may open an increasingly large number of human and primate species that are vulnerable to zoonotic events. Given the threat posed
to Public Health on a global scale by widespread infection of Ebolavirus and the number of remaining questions concerning the occurrence of zoonotic events, these questions merit further scientific inquiry moving forward.

### 2.3 Social characteristics and the spread of Ebolavirus

We consider the social characteristics pertinent to transmission of the virus. Given knowledge about the spread of the virus and that the handling of bodily fluids may vary greatly across cultures this forms an important aspect of any model for Ebolavirus. This is particularly important to our network-based approach as the network should reflect which people in a population could have potentially infectious contact. Therefore, for a given person in the population of interest we define the neighborhood of that person to be all other people that person could infect. We typically expect this to include those people that might have contact with bodily fluids of the infected individual. Research to date suggests that human-to-human transmission of the virus with no direct contact between bodily fluids is either rare or impossible [56]. Thus, at first pass we might expect that family members, close friends, sexual partners, and healthcare workers constitute the majority of contacts of an infected person. Other potential infectious contact is known to occur in burial of those that have been infected and passed away. It is our goal here to review the literature on previous outbreaks to see if information exists about the size and other relevant information to modeling infectious neighborhoods.
Table 2.1: Table of important outbreaks of Ebolavirus considered in this document.

<table>
<thead>
<tr>
<th>Year</th>
<th>Country</th>
<th>Outbreak Size</th>
<th>Species</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>Zaire (DRC)</td>
<td>318</td>
<td>Ebolavirus</td>
<td>[18]</td>
</tr>
<tr>
<td>1995</td>
<td>Zaire (DRC)</td>
<td>315</td>
<td>Ebolavirus</td>
<td>[61]</td>
</tr>
<tr>
<td>2000-2001</td>
<td>Uganda</td>
<td>425</td>
<td>Sudan Ebolavirus</td>
<td>[91]</td>
</tr>
<tr>
<td>2013-2016</td>
<td>Several</td>
<td>34000+</td>
<td>Ebolavirus</td>
<td>[107]</td>
</tr>
<tr>
<td>2014</td>
<td>DRC</td>
<td>69</td>
<td>Ebolavirus</td>
<td>[77]</td>
</tr>
</tbody>
</table>

First, it is known that transmission within the same household is a major driver of disease dynamics. Researchers found in the 1995 Kikwit (DRC) outbreak that 27 early infections of Ebola resulted in the secondary infection of 28 family members out of 173 [32], this indicates a probability of transmission to a given contact of approximately 16%. Similar patterns of risk for household members of primary infections have also been observed in several other detailed studies, such as the 2000 Uganda outbreak [38]. Given that most historical estimates of the basic reproductive number for Ebolavirus are between 1 and 3, this perhaps suggests that intra-family transmission may even constitute the majority of infections. It is obvious that the number of secondary infections seen within households will depend on social characteristics of the population, the strain of the disease, and other factors. However, the lesson that can be learned from this and other outbreaks is that family members of the infected are at disproportionately high risk of infection. The work by Francesconi et al. (2003) in the aftermath of the 2000 Uganda outbreak remains one of the few cases of detailed contact tracing carried out in an Ebola outbreak. This information
is useful as it provides insight into the relevant parametric assumptions that we can use to define a contact network later on.

It has been observed in the present and previous outbreaks that a large number of healthcare workers are infected while caring for ill patients. For example, in the 1995 Kikwit outbreak it was found that greater than 25% of the total cases (80/315) were healthcare personnel that had become infected before implementation of appropriate preventative measures [61]. The presence of the infection within the healthcare worker community has been well documented in the present West Africa outbreak as well[107]. Therefore, both of these factors may well contribute to the overall dynamics of disease propagation and the interaction between them is not currently understood. In particular, one fears a situation in which the disease burden among healthcare workers is so high that it prevents treatment of those that are sick which in turn allows the disease to spread uninterrupted. We discuss the mathematical details of this in a later chapter.

We first discussed the disease burden of healthcare workers and immediate family members because these are aspects of Ebola transmission that we expect are invariant to the specific population under consideration. For example, both secondary cases observed in the United States during the current outbreak were healthcare workers that had cared for an imported case. We would expect that in the event of an outbreak in most locales, healthcare workers would be at elevated risk. However, there are other aspects of Ebola transmission that are specific to certain populations. Ebolavirus is transmitted through direct contact with bodily fluids of infected individuals and so cultural practices regarding touching and
physical contact are important aspects of the possibility of transmission [61]. In this we wish to highlight the interplay between cultural practices and human behavior, which in turn can have important ramifications for disease transmission dynamics.

These descriptions of disease transmission are inherently qualitative at this level. To relate this to the end goal we will model the epidemic process as taking place on a network, where neighbors (members of the population) are connected in the graph theoretic sense if potentially infectious contact occurs between them. Therefore, cultural knowledge about burial practices and the place of traditional healers, as they are often referred to in the academic literature, in any given population may be considered in parameterizing the degree distribution that is ultimately assumed. For example, if it is common practice to touch the dead for all members of a population, then a degree distribution that would lead to a more complete graph should be employed. If a certain subset of the population is expected to have distinctly different connectivity properties from others, then a mixture distribution may be appropriate. The specifics of this will be discussed in reviewing the literature on network-based models.

As an example, Hewlett et al. study the inter-relationship between cultural practices and transmission in the context of the 2000 Uganda outbreak [49]. The important aspects of this study are the perception of disease by the local population, their cultural practices regarding treatment of the sick, and an understanding of how these practices can both contribute to or inhibit the spread of Ebola.
The research of Hewlett and Amola (2003) on the Northern Uganda outbreak is perhaps the most detailed research into the sociocultural context of Ebola, human behavior, and the epidemiology of the disease [49]. In this outbreak there were a total of 425 total cases with diagnoses occurring between August 2000 and January 2001. The epidemic affected primarily a single ethnic group called the Acholi with a population of about 470,000 in the North-Central part of Uganda. The authors describe a dichotomous cultural belief system for disease; in particular, that any given illness is classified as either a yat or a gemo (meaning epidemic) each of which has a different origin, symptoms, treatment, etc. A yat appears less serious and is marked by inflammation, aches, and can be acquired in a number of ways. The treatment for this is to see a special healer that will use his jok (spirits he has acquired) to find and remove the yat. Biomedical treatments are also often sought for these illnesses in addition to the services of the special healer. Early on in the Ebola epidemic, the sickness was initially thought to be a yat. As washing the body of the dead and touching special healers was allowed prior to consideration as a gemo, this may have contributed to the early spread of the disease in this case. As the epidemic worsened in October 2000 it became considered a gemo. There is greater uncertainty around illnesses that are described as gemo, as they are thought to come into a community much like the wind although not thought to be carried literally by the wind. For example, smallpox is considered a gemo. Additionally, the special healers do not have the same ability to cure gemo as they do yat and so there is an entirely different set of practices for treatment of those infected with an illness that is considered gemo [49].
The recommended course of action for a family member infected with an illness that is a *gemo* is the following. First, it is recommended that the ill patient is isolated in a house that is removed from the rest of the village and marked by elephant grass. He is to be cared for only by other survivors of the epidemic and once symptoms have cleared, the ill person is supposed to remain isolated for a lunar cycle until the *gemo* has cleared from him, and if he passes then his body is to be cared for by other survivors of the *gemo*. At the village-wide level there are additional recommendations for controlling the spread of the illness. Members of the village are recommended to limit their movements, no food from outside sources is to be consumed, there is no ritual dancing, sexual relations are to be suspended, and pregnant women and children are to be especially careful to avoid the ill [49].

It is seen that many of the cultural practices associated with the treatment of *gemo* are similar to the measures that Public Health officials would recommend in a similar situation. In addition, it is necessary to consider attitudes towards biomedical treatments of disease and the local attitudes towards these practices when outbreaks do occur. While this particular ethnic group often co-administers biomedical treatments along with traditional healing techniques and has practices that would limit the spread of Ebolavirus, there is great diversity of cultural practices in areas where Ebola is endemic and areas affected by the current outbreak. The preceding is intended to give a description of how these practices can affect transmission dynamics and stop the spread of an epidemic that has already begun. We move onto some of the mathematics for classical models of infectious disease in human populations as a starting point for the development of our network-based methods.
2.3.1 Epidemiological studies of Ebolavirus in West Africa

In this section we examine literature that has been published regarding the epidemiology of the disease during the West Africa outbreak and the smaller outbreak in DRC reported by Maganga et al. [77]. In particular, these studies form a starting point for quantitative study of epidemiological parameters either explicitly or to be incorporated as prior information. Lastly, the studies give insight into data availability, which informs useful data structures to consider for parameter estimation later on.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmission rate (per contact)</td>
<td>.61 d</td>
<td>10</td>
<td>2.7 d - 3.7 d</td>
<td>0.04 - 2.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asymptomatic Period (mean)</td>
<td>9.6 d</td>
<td>10 d</td>
<td>7.1 - 10 d</td>
<td>8.5 d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infectious Period (mean)</td>
<td>9 - 10 d</td>
<td>8 - 10 d</td>
<td>10 d</td>
<td>7.5 d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asymptomatic Period (median)</td>
<td>9 d</td>
<td>10 d</td>
<td>9.5 d</td>
<td>9.5 d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contacts (mean/median)</td>
<td>6.16/5.74</td>
<td>5.74 (mean)</td>
<td>5.4</td>
<td>4.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days until Hospitalization</td>
<td>2 - 5</td>
<td>2 - 5</td>
<td>5.3</td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reproductive Number $R_0$</td>
<td>1.29 (L, SL)</td>
<td>1.95 - 3.57 (L, SL, L1)</td>
<td>2.7 (L, SL, UNAFLIC)</td>
<td>1.6 (DRC)</td>
<td>1.7 - 2.0</td>
<td>1.73 (DRC)</td>
<td>1.71 (L)</td>
<td></td>
</tr>
<tr>
<td>Case Fatality Ratio</td>
<td>58 - 83%</td>
<td>53 - 81% (SL)</td>
<td>70.0% (DRC, L, SL)</td>
<td>85% (L, SL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Table of epidemiological estimates obtained from select studies in the literature. The unit of measure where applicable is assumed to be $d = \text{days}$. Where applicable we use the country identifiers L=Liberia, SL = Sierra Leone, G = Guinea, and DRC = Democratic Republic of the Congo.

For the class of network-based models that we study, the selection of a parametric model for the degree distribution (that is, the distribution of the number of connections of all individuals in the network) is inevitably an important one [95] [10] [84]. Given the relationship between the degree distribution and induced dynamics in similar models and in our own analytical results this decision should be aided where possible by sociological studies in the literature as well as contact tracing studies performed during the course of the epidemic at hand. While several extensive contact tracing studies were performed as part
of the global response to the West Africa outbreak, this data contains sensitive personal information that has prevented it from being made publicly available. Nevertheless, using published data from several contact tracing studies in Liberia and Guinea we may extract a sufficient amount of information to inform our assumptions in later sections [31] [35].

The contact tracing data provided by Fallah et al. consists of Ebola cases reported between February and December 2014 in Monteserrado County, Liberia [35]. A contact tracing study was performed between July and October 2014 for individuals infected during that time. The Liberian Ministry of Health provided data on a subset of 1,585 individuals to this data is given in Supplement 2 of [35]. For many individuals the data consists of their disease status (confirmed, probable, etc.), whether they were a healthcare worker, and in some cases the de-identified numeral assigned to that patient. In particular, this latter information may allow us to an extent to evaluate the appropriateness of Configuration Model random networks or to provide descriptive statistics about some subgraph. The provided data makes it possible to fit a parametric statistical model to the degree distribution. Below we compare the fitted distribution function (parameters estimated via Maximum Likelihood) against the empirical quantiles from the dataset and find very good agreement with the Negative Binomial distribution. The shape and right skew are consistent with information reported elsewhere, for example [3].

Dixon et al. performed a detailed retrospective study on contact tracing data from the Kindia and Faranah districts in Guinea. In both districts the data come from contact tracing studies that were performed during the height of the Guinea outbreak between September
Figure 2.1: Comparison of the Negative Binomial distribution ($\hat{\rho} = 0.959$, $\hat{\mu} = 6.11$) fit against the contact tracing dataset of Fallah et al. [35].

and December 2014. In these data, the authors report on the follow-up of contacts of approximately 90 case-patients in Kindia and 62 case-patients in Faranah. The authors report that the median time to isolation was 3 days in each district, which is consistent with values reported elsewhere [107]. Moreover, the authors report a median of 16 contacts in Kindia prefecture with an interquartile range of (11.2, 28) and a median of 9 contacts in Faranah with an interquartile range of (5.5, 15.5). Other contact tracing studies in Guinea suggest that a Negative Binomial distribution provides a very good fit to the empirical values [35] and has been shown to be an appropriate model in several other contexts [10]. Therefore,
a Negative Binomial model for the contact distribution with a mean consistent with these summary statistics will be considered an appropriate model for numerical investigations we conduct later.

The study by Dixon et al. [31] similarly provides social context regarding the type of personal relationship between those listed as contacts. In both prefectures the most common type of contact is a family member (approximately 40% of contacts) while neighbors are also a large number of observed contacts (41%) in Kindia. Healthcare workers comprised only about 2% of contacts in Kindia and none were listed for Farranah. Moreover, in Farranah the secondary attack rate (probability of infecting a given contact) was computed to be 12.3% for contacts that were members of the same household and 4.8% for other contacts. However, this difference was not statistically significant. In the data from Farranah the secondary attack rate among household members was approximately 4.8% and was much higher than the secondary attack rate of 0.4% among contacts that were not members of the same household [31]. In both cases the secondary attack rate is consistent with the values reported by Hewlett et al. in studying the outbreak in Uganda while offering some suggestion that different types of contacts may have asymmetric risk associated with transmission. A model that we introduce later is able to accommodate such heterogeneities.

Faye et al. study specific chains of transmission areas surrounding Conakry, Guinea between March and August 2014 [36]. In total this study comprises information on 152 individuals that were found to be in a chain of transmission during that period out of a total of about 193 confirmed cases in that area during the same time period [36].
authors find that 82% of recorded transmission events took place in the community (i.e.
not in a hospital or funeral) and that approximately 72% of cases involved transmission to
family members [36]. These results are consistent with the studies of Dixon and others that
implicate within household transmission as a primary driver of the epidemic. The authors
do not find evidence of asymmetric risk of infection among age groups or gender. The
authors report a hospitalization rate of approximately 70% over this time period, suggesting
that this should be accounted for. Lastly, the authors report that after safe burial practices
were implemented in April 2014 that the number of transmissions occurring in this context
were minor [36].

Other important epidemiological information pertains to the course of the disease in hu-
mans. This includes but is not limited to the length of the incubation period of the disease,
the rate at which infected individuals were quarantined (if at all), burial practices regarding
the dead, and the length of the infectious period. Fortunately there has been substantial
research on these topics although it is not always known the extent to which these factors
may differ between locales. On the other hand at the country level there has been extensive
research conducted regarding rates of case-patient isolation, length of incubation period,
and length of infectious period [107]. Many of these findings are summarized in Table 2.2
and it seen that a good deal of consistency was found across various studies.

2.4 Conclusion

In this section we have reviewed the essential scientific and epidemiological history of
Ebola and its transmission needed to place the disease contextually. This study will allow
us to identify the essential drivers of disease dynamics in human populations and accommodate these features in our modeling efforts later on. We proceed to review some essential Mathematical tools for modeling the spread of infectious disease in human populations.
Chapter 3: Models of infectious disease

3.1 Historical models of epidemics

The model of Kermack and McKendrick forms a starting point for the mathematical modeling of infectious disease both historically and intuitively [59]. Suppose that each individual in a population of a fixed size is assumed to be either susceptible (S), infectious (I), or recovered (R). For \( t > 0 \) let \( S(t) \) denote the fraction of susceptible individuals in the population, \( I(t) \) denote the fraction of infectious individuals, and \( R(t) \) denote the fraction of recovered individuals with the constraint that they add to 1. Letting \( N \) denote the population size, \( \beta > 0 \) denote the rate at which an infective individual passes the infection to susceptible individuals in the population, and \( \gamma > 0 \) denote that rate at which infective individuals recover (we will remain consistent with this notation moving forward) they proposed a system composed of the following set of non-linear differential equations:
\[ \dot{S} = -\beta SI \]
\[ \dot{I} = SI - \gamma I \]
\[ \dot{R} = \gamma I \]
\[ S(0) = S_0, I(0) = I_0, R(0) = 0, \]

where the notation \( \dot{S} \) denotes the derivative with respect to time. Basic analysis of this system has formed the core of understanding for the dynamics of infectious disease modeling since its introduction [7]. Various generalizations have been proposed, for example, with an additional exposed state to account for diseases in which there is delay between when a person is infected with an illness and when they are infectious. Further generalizations deal with asymmetric levels of susceptibility and infectiousness between groups of individuals (i.e. multitype models). In general, these classes of compartmental models all make similar assumptions that particular types of individuals in a population mix homogeneously and therefore one obtains similar analytic properties for many of them. As a result, it often suffices to study analytical properties of the classical SIR formulation with the understanding that the same techniques are applicable to generalizations with minimal additional work other than accounting for added disease states (compartments).

The essential part of the analysis of this model can be done by first noting that one of the quantities can be ignored due to the linear constraint \( S(t) + I(t) + R(t) = 1 \) for every \( t > 0 \). Next, one can consider \( \frac{\dot{S}}{R} = -\frac{\beta}{\gamma} S \) to see that a partial solution to the system is given by \( S(t) = S(0)e^{-\frac{\beta}{\gamma} R(t)} \). Since the initial fraction of infective individuals, \( I_0 \), is usually
thought to be small at the start of an epidemic one may suspect that the system will tend
towards one of two equilibriums depending on if $I(0)$ is positive or not. This motivates the
definition of the basic reproductive number as $R_0 = \frac{\beta}{\gamma}$ for this model. The basic repro-
ductive number is an essential analytical property of most epidemic models and is defined
as the average number of secondary cases caused by an infectious individual in a popu-
lation of susceptible individuals. A well-known threshold theorem states that an outbreak
occurs with probability tending to 1 when $R_0 > 1$ while the outbreak remains small (i.e. the fraction of individuals infected stays below a certain point) when $R_0 \leq 1$ [7].

Much more interesting for the purposes of statistical analysis is a similar stochastic
formulation. We will refer to this model as simply the Stochastic SIR model and will
show that under a suitable scaling converges in probability (in the appropriate space of
right-continuous functions) to the equations of Kermack and McKendrick. Our purpose
in doing this is to present some results needed to derive limiting sets of differential equa-
tions for such stochastic formulations, the construction of the likelihood function in this
setting, statistical parameter estimation, and consistency of those estimates. Our treatment
in this section is particularly thin in the area of finite-population dynamics for the reason
that we are concerned primarily with large volume limits (i.e. to describe macro-population
dynamics) in our study. That is, given the magnitude of the epidemic and the spatial reso-
lution of the data for analysis (mostly aggregated at a national level), our primary interests
lie in system dynamics of large populations. For a complete treatment of finite-population
results and additional topics, Britton (2000) is generally regarded as the standard textbook [7].

3.1.1 Law of large numbers for the stochastic SIR model

To begin, consider the evolution of a bivariate stochastic process \(((S(t), I(t)) : t \geq 0)\) where we can largely ignore the \(R\) quantity because of the constraint that \(S(t) + I(t) + R(t) = N\). The evolution of this process is constructed such that each infectious individual in the population infects each susceptible individual in the population according to the points of a Poisson process with rate parameter \(N^{-1}\beta\). Moreover, each infectious individual recovers according to a Poisson process with rate parameter \(\gamma\). This implies immediately that the mean length of the infectious period is \(\gamma^{-1}\). Thus, because of the assumption that these events (jumps) occur according to independent exponential distributions we retain the Markov property and the law of the induced process can be described by the state-dependent transition rates given in the table below, for \(i\) and \(j\) in \(\{1, \ldots, N\}\) such that \(S + I + R = N\) for all \(t\). Generally speaking this is accomplished by showing that as the total population size \(N\) is sent to infinity that the \(N\)-normalized stochastic processes \((S, I, R)\) will

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>At Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>((i, j))</td>
<td>((i - 1, j + 1))</td>
<td>(N^{-1}\beta ij)</td>
</tr>
<tr>
<td>((i, j))</td>
<td>((i, j - 1))</td>
<td>(\gamma i)</td>
</tr>
</tbody>
</table>

Table 3.1: Table of transitions and rates for the stochastic (Markovian) SIR model.
converge to the system of Kermack and McKendrick. Two fundamental results that will aid in this are the Poisson Law of Large Numbers and Gronwall’s Inequality. These results are stated below.

**Proposition 1. Poisson Law of Large Numbers**

Assume that \( (Y(t) : t \geq 0) \) is a unit Poisson Process. Then almost surely

\[
\lim_{N \to \infty} \sup_{s \leq t} |N^{-1} Y(Ns) - s| = 0 \tag{3.1}
\]

The proof of this is easily obtained by decomposing the process at some time \( t \) as equal in law to the sum of independent processes over sub-intervals spanning \([0, t] \). In our case, a simple coupling argument will show that the quantities \( S \) and \( I \) in the stochastic SIR model can be written in terms of unit Poisson Processes.

**Proposition 2. Gronwall’s Inequality**

Assume \( f \) is a real function satisfying \( 0 \leq f(t) \leq a + b \int_0^t f(s) ds \) for bounded, positive constants \( a \) and \( b \) and all \( t \geq 0 \). Then \( f(t) \leq ae^{bt} \) for all \( t \geq 0 \).

The use of Gronwall’s inequality is not immediately clear. However, it states that if \( f \) is a reasonably well-behaved function that we can bound by \( a + b \int_0^t f(s) ds \), then \( f(t) \) cannot be too large. The Law of Large Numbers is derived for a class of Poisson jump processes of which the stochastic SIR model arises as a special case.

We consider a sequence of continuous time Markov processes indexed by \( N \geq 1 \) such that for every \( N \) the process \( X_N(t) = (X_N : t \geq 0) \) is defined on the state space \( \mathbb{Z}^d \) with transition rates that can be written in the form \( \lambda^N_{x,x+k}(t) = NB_k(N^{-1} x(t)) \) for each \( x,z \in \mathbb{Z}^d \).
where \( k \in K \) denotes the set of possible jumps that the process can have and the \( B_k(N^{-1}x) \) are continuous functions in \( x \) for every \( k \). The set \( K \) is required to be finite such that the system can only jump in a finite number of directions. Thus, we have completely described the possible transitions that the process can make and the rates at which those jumps will occur for any state in the state space. Equivalently, one may describe the evolution of this process as follows:

\[
P(X_N(t+h) - X_N(t) = k|X_N(t) = x) = hNB_k(N^{-1}x(t-)) + o_p(h), \quad k \in K
\]

\[
P(X_N(t+h) - X_N(t) = 0|X_N(t) = x) = 1 - hN \sum_{k \in K} B_k(N^{-1}x(t-)) + o_p(h).
\]  

A Poisson process satisfying (3.2) will be henceforth referred to as a Density Dependent Markov Jump Process (DDMJP). The term density-dependent refers to the fact that the rate functions are written explicitly in terms of the \( N \)-scaled quantities in the system. Viewing \( N \) as the size of the system (i.e. volume in a biochemical sense) then this choice of name makes sense. A convenient representation of the evolution of such a process is in terms of independent (though coupled) unit Poisson processes \((Y_k, k \in K)\) and is often referred to as the “trajectory” equation, given below:

\[
X_N(t) = X_N(0) + \sum_{k \in K} kY_k \left( \int_0^t NB_k(N^{-1}X_N(s))ds \right).
\]  

(3.3)

For a vector \( x \in \mathbb{R}^r \) we will take \( ||x|| \) to be the Euclidean norm. If \( Y(t) \) is a unit Poisson process we will use the notation \( \hat{Y}(t) = Y(t) - t \) to denote the centered analog of this process. We begin by stating the result and give the proof thereafter.
Theorem 3.1.1. Strong Law of Large Numbers for DDMJPs

Suppose that \( \lim_{N \to \infty} \bar{X}_N = \bar{x}_0 \) and that for every compact (closed and bounded) \( J \in \mathbb{R}^d \) the drift function \( F(x) = \sum_{k \in K} kB_k(x) \) is Lipschitz continuous. Thus for each \( J \) there exists a positive constant \( M_J \) such that for every \( x, y \in J \), \( ||F(x) - F(y)|| < M_J||x - y|| \). Then \( \lim_{n \to \infty} \sup_{s \leq t} ||\bar{X}_N(t) - \bar{x}(t)|| = 0 \) a.s. where \( \bar{x}(t) \) is the solution to the integral equation

\[
\bar{x}(t) = \bar{x}_0 + \int_0^t F(\bar{x}(s))ds
\] (3.4)

Thus, it is seen that the result states that a stochastic process satisfying (3.2) appropriately scaled will have a deterministic limit.

Proof. We consider the behavior in some neighborhood \( J \) of \( (x(s) : 0 \leq s \leq t) \). For each \( k \in K \) define the finite constant \( \hat{B}_k = \sup_{x \in J} B_k(x) \). Since \( J \) is a compact set and the \( B_k \) are continuous functions this constant is finite. Next, from (3.3) consider the difference

\[
||\bar{X}_N(s) - \bar{x}(s)|| = ||\bar{X}_N(0) - \bar{x}(0) + N^{-1} \sum_{k \in K} k\hat{Y}_k(N\int_0^s F(\bar{X}_N(u))du) + \int_0^s (F(\bar{X}_N(u)) - F(\bar{x}(u)))du||
\]

Which by the triangle inequality we write as

\[
\leq ||\bar{X}_N(0) - \bar{x}(0)|| + N^{-1} \sum_{k \in K} ||k|| \sup_{u \leq s} |\hat{Y}_k(N\hat{B}_k u)| + \int_0^s M_J||\bar{X}_N(u) - \bar{x}(u)||du
\] (3.5)

Gronwall’s Inequality can now be applied (take \( a \) as the first two terms and \( b = M_J \)) to show that the right hand side is bounded above by:

\[
(||\bar{X}_N(0) - \bar{x}(0)|| + N^{-1} \sum_{k \in K} ||k|| \sup_{u \leq s} |\hat{Y}_k(N\hat{B}_k u)|)e^{MJs}
\] (3.6)
Now taking the supremum over all $s$ in the interval $[0,t]$ yields:

$$\sup_{s \leq t} ||\bar{X}_n(s) - \bar{x}(s)|| \leq (||\bar{X}_n(0) - \bar{x}(0)|| + N^{-1} \sum_{k \in K} ||k|| \sup_{u \leq s} |\hat{Y}_k(N\hat{B}_ku)|)e^{Mjs}. \quad (3.7)$$

We need only to make two more observations to complete the proof. First, note that the first term in the parenthesis converges to 0 in the limit by assumption. Second, the right term in the parenthesis will converge to 0 almost surely by applying the Poisson LLN. The exponential term does not depend on $N$ and so we need to do no more with it. This completes the proof. This theorem suffices to show by verifying the assumptions that the Stochastic SIR model converges to the system of Kermack and Mckendrick as desired. This can be done easily by noting for example that the transition corresponding to an infective recovering has a “jump” of $(0, -1)$ and a density-dependent rate of $N\gamma I(t)/N$. Similarly the jump occurring due to a new infection is $(-1, 1)$ and the density-dependent rate is $N\beta SI/N^2$.

The power of this result lies in that systems with inhomogeneous behavior (such as multiple classes of susceptibility, multiple stages of infection, etc.) can be expressed in terms of a homogeneous system by explicitly accounting for this heterogeneity [7]. In a later chapter we introduce likelihood-based estimation methods for models of this type. However, for the time being we turn our attention to reviewing network-based approaches to epidemic modeling.
3.1.2 Diffusion approximation

Perhaps unsurprisingly, a form of Central Limit Theorem (CLT) also exists for the class of Poisson processes introduced in the previous section when the scaling is done by $\sqrt{N}$ rather than $N$. This CLT is useful for deriving a CLT for estimators and so the result is stated here. A similar result has been claimed for network-based models but only appears to apply to the initial progression of the epidemic [43]. Thus we state the result for Density Dependent Processes but an analogous result for network-based models is an open question.

![Figure 3.1: A realization of a stochastic $S-I-R$ model plotted against the solution of the limiting ODE.](image)

**Theorem 3.1.2. Central Limit Theorem for Density Dependent Processes**

Suppose $\partial F$ is continuous and that $\lim_{N \to \infty} \sqrt{N} \tilde{X}_N = x_0 = v_0$, and $G(x) = \sum_{k \in K} kk^T B_k(x)$ denotes the local covariance matrix. Then $\sqrt{N}(\tilde{X}_N(t) - x(t)) \to V$ where $V(t)$ is a Gaussian
Process defined by the integral equation

\[ V(t) = v_0 + \sum_{k \in K} kW_k \left( \int_0^t B_k(x(s))ds \right) + \int_0^t \partial F(x(s))V(s)ds \]  

(3.8)

Thus \( V \) is a Gaussian process with covariance matrix

\[ \text{Cov}(V(t), V(r)) = \int_0^{\min(r,t)} \Psi(t,s)G(x(s))(\Psi(r,s)^T)ds \]  

(3.9)

Such that \( \Psi \) is defined as a (matrix) function that solves the differential equation

\[ \partial \Psi_s(t,s) = -\Psi(t,s)\partial F(x(s)), \quad \Psi(s,s) = I \]

And \( \partial \Psi_s \) signifies the partial derivative with respect to \( s \)

The proof of this theorem is not necessary for results we derive but is due to Kurtz (1976) [66]. However, we point out that a similar result has not been derived for the classes of network-based models studied later on and that such a result has many applications to efficient simulation and estimation.

3.1.3 Estimation for counting processes

We review estimation methods via Maximum Likelihood for counting processes. The density-dependent process in (3.2) arises as a special case of this class of counting processes; however, we later apply this method to a process that is not density dependent (i.e., satisfies (3.11) below but not (3.2)). Consider a jump process \( X(t) \in Z^d \) defined as follows. Let the finite constant \( p > 0 \) denote the number of distinct transitions that can occur and for each \( l = 1, \ldots, p \) let \( \lambda_l(t; \theta) \) denote the hazard of transition \( l \) at time \( t \) and depends on a
vector of parameters $\theta = \{\theta_1, \ldots, \theta_d\} \in \Theta \subset \mathbb{R}^d$. Moreover, let $C_l(t)$ denote that counting process that increases by 1 each time a jump of type $l$ occurs and let $\mathcal{H}_t$ denote the process history up to time $t$ consisting of the time of each jump and the type. Define the total hazard of any event occurring as

$$\lambda_0(\theta; X(t)) = \sum_{l=1}^{p} \lambda_l(\theta; X(t)).$$

(3.10)

By this specification, the processes $\{C_1, \ldots, C_p\}$ satisfy

$$P(C_l(t + dt) - C_l(t) = 1|\mathcal{H}_t) = \lambda_l(t)dt + o_p(dt), \quad l = 1, \ldots, p,$$

$$P(X(t + dt) - X(t) = 0|\mathcal{H}_t) = 1 - dt \sum_{l=1}^{p} \lambda_l(t) + o_p(dt).$$

(3.11)

Note that the process described in (3.11) is more general than the one described in (3.2). This is done for consideration of a stochastic model later on satisfying this broader condition but does not satisfy the form of (3.2). When the hazard functions $\lambda_1, \ldots, \lambda_p$ are constant between jumps, then the probability that a jump occurs at time $t_i$ given the previous jump occurred at time $t_{i-1}$ is

$$\lambda_0(\theta; X(t_{i-1}))e^{-\lambda_0(\theta; X(t_{i-1}))(t_i - t_{i-1})},$$

(3.12)

by definition of (3.2) and standard results of independent Poisson Processes. Moreover, conditional on an event occurring at time $t_i$, the probability that it was of type $v_i$ is thus

$$P(\text{event of type } v_i|X(t_{i-1})) = \frac{\lambda_{v_i}(\theta; X(t_{i-1}))}{\lambda_0(\theta; X(t_{i-1}))}$$

(3.13)
Consider the likelihood corresponding to an event occurring at time \( t_i \) of type \( v_i \). Conditional on the state of the process at time \( t_{i-1} \) the probability that the event occurs at time \( t_i \) and is of type \( v_i \) is

\[
\lambda_{v_i}(\theta; X(t_{i-1})) e^{-\lambda_0(\theta; X(t_{i-1}))(t_i - t_{i-1})},
\]

and essentially repeating this argument for all event times yields the form of the likelihood.

As before let \( \mathcal{H}_t \) denote the history of the process up to time \( t \) consisting of event times 

\[ 0 = t_0 \leq t_1 \leq \ldots \leq t_m \leq t \]

and the event type at each \( t_i \). Then the likelihood can be written

\[
L(\theta; \mathcal{H}_t) = e^{-\lambda_0(\theta; X(t_m))(t - t_m)} \prod_{i=1}^{m} \lambda_{v_i}(\theta; X(t_{i-1})) e^{-\lambda_0(\theta; X(t_{i-1}))(t_i - t_{i-1})}
\]

\[
= \exp \left( \sum_{l=1}^{n} \left( \int_{0}^{t} \log(\lambda_l(\theta; X(s-)))dC_l(s) - \int_{0}^{t} \lambda_l(\theta; X(s-))ds \right) \right). \tag{3.15}
\]

The latter is a perhaps more useful representation. Letting \( l(\theta; \mathcal{H}_t) = \log(L(\theta; \mathcal{H}_t)) \) and \( \partial_k \) the partial derivative with respect to \( \theta_k \), the score equations are seen to be

\[
S_k(\theta; t) = \partial_k l(\theta; \mathcal{H}_t) = \sum_{l=1}^{n} \int_{0}^{t} \left( \frac{\partial_k \lambda_l(\theta; s-)}{\lambda_l(\theta; s-)} (dC_l(s) - \lambda_l(\theta; s-)ds) \right). \tag{3.16}
\]

Note that \( E(dC_l(s)|\mathcal{H}_s) = \lambda_l(\theta_0; s-) \) and therefore the score processes evaluated at the true parameter value \( \theta_0 \) are zero-mean martingales [7]. The Maximum Likelihood Estimator (MLE) is found by setting (3.16) to 0 for every \( k = 1, \ldots, d \) and this solution yields the estimator \( \hat{\theta}_{MLE} \).

**Lemma 1. Consistency of the MLE, Theorem VI.1.1**  
Andersen (1993)

Under mild regularity conditions (i.e. Condition VI.1.2 (A)-(E) in Andersen (1993), ch.6
(3.17) \[ P \xrightarrow{N \to \infty} 0, \quad N \to \infty \]

Proof. The full argument is found in [6] and is largely similar to other arguments for consistency of the MLE.

As an example, the resulting estimators of $\beta$ and $\gamma$ may be derived for the stochastic SIR model [7] to be

\[
\hat{\beta} = \frac{N_I(t)}{\int_0^t S(s)I(s)ds}, \\
\hat{\gamma} = \frac{N_R(t)}{\int_0^t I(s)ds},
\]

where $N_I(t)$ denotes the total number of infections up to time $t$ and $N_R(t)$ denotes the total number of recoveries up to time $t$. One may apply the Martingale Central Limit theorem to derive asymptotic Normality of this estimator [7]. Like the traditional MLE for independent random variables, one may also show that the solution to (3.16) is asymptotically Normal where $\mathcal{N}(\mu, \Sigma)$ will represent the multivariate Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$. This is stated below.

**Lemma 2. Normality of the MLE, Theorem VI.1.2 Andersen (1993)**

Under mild regularity conditions (i.e. Condition VI.1.1 (A)-(E) in Andersen (1993), ch.6 [6]), the estimator $\hat{\theta}_{MLE}$ as defined as the solution (3.16) is asymptotically normal that is,

\[
\sqrt{N}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{L} \mathcal{N}(0, \Sigma), \quad N \to \infty
\]
where $\Sigma = \Sigma(\theta_0)$ denotes the asymptotic covariance matrix of $\hat{\theta}_{MLE}$.

Together these results essentially imply that Maximum Likelihood Estimation for counting processes largely works similar to the case of independent, identically distributed random variables. This, combined with the fact that (3.11) is very general (i.e. the processes are not even required to be Markovian) gives good reason to use the MLE when possible, such as when the entire process history $\mathcal{H}_t$ is fully or approximately observed. However, there are times, especially in Epidemiological and Biochemical settings where this is not possible. To address this, Rempala (2012) recently has studied an alternative estimator for this problem making use of Theorem 3.1.1. This result hinges on the following argument. Suppose that $N$ is sufficiently large and that the initial condition $\bar{x}_0$ is known. Suppose that the ($N$-scaled) state of the system $y^*(t_i)$ is measured on some sequence grid of points $0 < t_1 < t_2 < \ldots < t_m$ and that the system size $N$ is known. Then by the Law of Large Numbers it should be true that

$$y^*(t_i) \approx y(t_i; \theta_0) = \bar{x}(0) + \int_0^{t_i} F(\bar{x}(s; \theta_0)) ds$$

i.e., $i = 1, \ldots, m$. \hspace{1cm} (3.20)

In other words, when the size of the system is sufficiently large the measured data should closely approximate the deterministic solution given in Theorem 3.1.1 evaluated at the true parameter value $\theta_0$. This estimator is commonly employed in the epidemic modeling literature to a variety of stochastic models [62] and its properties have been recently studied by Rempala [98] as well as Linder and Rempala [72]. A potential advantage of this estimator
estimator is that the entire trajectory (i.e., all of $H_t$) is not required for straightforward estimation. The estimator is defined as follows.

$$\hat{\theta}_{LSE} = \min_{\theta \in \Theta} \sum_{i=1}^{m} ||y(t_i; \theta) - y^*(t_i)||_2, \quad (3.21)$$

Rempala (2012) notes that the definition of $\hat{\theta}_{LSE}$ in (3.21) implies that any solution (which is not necessarily unique) to the optimization problem simultaneously satisfies

$$\sum_{i=1}^{m} J(t_i; \theta) (y(t_i; \theta) - y^*(t_i)) = 0 \quad (3.22)$$

Where $J(t; \theta) = \{\partial_j y_k(t; \theta)\}$ denotes the Jacobian of the limiting ODE. The above equation is easily shown to be equivalent to

$$\sum_{i=1}^{m} J(t_i; \theta_0) (y(t_i; \theta_0) - y^*(t_i)) = \sum_{i=1}^{m} J(t_i; \theta) (y(t_i; \theta) - y^*(t_i; \theta_0)). \quad (3.23)$$

Under the condition that $J(t; \theta)$ is bounded as $N \to \infty$ then the left hand side of the above equation converges to 0 in probability by Theorem 3.1.1. To further ensure that the solution is unique in some neighborhood of $\theta_0$ and non-trivial we will also require that

$$y(t_i; \theta_1) = y(t_i; \theta_2), \forall i = 1...m \implies \theta_1 = \theta_2, \quad (3.24)$$

as well as,

$$\sum_{i=1}^{m} \partial_k y(t_i; \theta) (y(t_i; \theta) - y^*(t_i)) \neq 0, \quad \theta \neq \theta_0 \quad (3.25)$$

for some $k = 1..r$. 

36
**Theorem 3.1.3.** Under the assumptions of Theorem 3.1.1 and additional conditions (3.24) and (3.25) above, the OLS estimator \( \hat{\theta}_{LSE} \) is consistent in the sense that

\[ ||\hat{\theta}_{OLS} - \theta_0|| \to 0, \quad N \to \infty \]  \hspace{1cm} (3.26)

The proof of this theorem is made straightforward as a result of Theorem 3.1.1 and the additional assumptions made above. As in Rempala (2012) the conditions imply that (3.23) is invertible as a function of \( \theta \) in the limit and this completes the proof. Lastly, the same argument will essentially work for any stochastic system indexed by \( N \) converging to a deterministic ODE in the limit. We apply this argument in chapter 7 to show consistency of the OLS estimator for a class of network-based models with a deterministic limit. Rempala (2012) also shows that the OLS estimator for DDMJP is asymptotically Gaussian under the assumptions of Theorem 3.1.2 [98]. Given the argument presented there, the needed condition is that the \( \sqrt{N} \)-scaled process has some diffusion approximation, i.e. such a result ought to hold in the case of a network-based model if an analogous result to Theorem 3.1.2 can be derived. This remains an open question but is in our view an important step for this class of models.

### 3.2 Large population limits of network-based models

A standard assumption of the models studied in the preceding section is that all individuals in a given population interact homogeneously. That is, that a given individual is no more likely to infect one individual in the population than another. While there are situations where this assumption is more tenable than others, a common critique of such models
is that this ignores many individual-level sources of variability [10] [117]. This immediately suggests network-based models in general as an attractive evolution of homogeneous mixing models because of their ability to accommodate individual-level heterogeneities such as superspreading [81]. In addition, mis-specification of the model can lead to very different dynamics of the epidemic process [10]. Network-based models attempt to avoid the assumptions of homogeneous mixing models as follows. We will assume that the epidemic evolves on a network $G = \langle V = \{1, \ldots, N\}, E \in \binom{N}{2} \rangle$ where individuals represent vertices and edges represent the capacity to transmit the infection. From the point of view of both realism and intuition, this model agrees with how contact-tracing studies are conducted in applied epidemiology and is considered by many to be a fundamentally more realistic model of disease spread. On the other hand, two challenges that are immediately present are specification for a model of the network itself and repercussions of not observing the network. We discuss the first of these issues now and leave the question of the latter for the time being.

A benefit of the network-based approach is that it is a generalization of the classical stochastic SIR approach by taking the contact network to be the complete graph $K_N$. However, much of the more recent literature has focused on using random graph models to try to better approximate the structure of contacts in human communities. The famous $G(N, p)$ model is perhaps the simplest random graph in which each vertex is independently with every other vertex with some probability $p > 0$. Previous studies of network-based models include $G(N, p)$ graphs [89], those that encapsulate a household structure [9], and a
particular type of dynamic random graph [5]. Two somewhat recent review papers give a comprehensive review of advancements and other open questions [50] [95].

The starting point of our model is a generalization of the $G(N, p)$ called the Configuration Model [90]. A potential shortcoming of the $G(N, p)$ model is that as the size of the graph approaches infinity, the degree distribution (i.e. the distribution of the degree of all vertices $v \in V$) converges to a Poisson distribution. The Configuration Model instead draws the degree $d_v$ of each node $v$ independently from an arbitrary degree distribution $D_v$ with some finite mean $\mu < \infty$. Next, each node $v$ is given $d_v$ (drawn from $D_v$) half-edges or “stubs” which are then paired off at random with each edge in the final graph $G$ being the connection of two half edges. Note that this model requires that $\sum_{v \in V} d_v$ to be even and does not prevent the existence of multiple edges between the same nodes or self-loops. However, under suitable regularity conditions the graph is simple with positive probability [53].

Various properties of this class of random graphs have been studied such as the presence of a giant component (a connected subset of nodes of size $O(N)$) [87] [88] and chromatic number [39]. A strong knowledge of these properties is largely not essential as long as the generating mechanism is well understood.

After the graph $G$ has been generated, the epidemic is modeled as a continuous-time Markov Chain much in the same way as in stochastic SIR model. Let $S, I, R \subset V$ denote the disjoint sets of vertices that are currently Susceptible, Infectious, and Recovered. Obviously these sets are dependent on time as the epidemic propagates. The rules for the spread
of the epidemic is that if we have two vertices say $i \in I$ and $j \in S$ and $i \sim j$ then the infection is passed from $i$ to $j$ according to an exponential distribution with rate parameter $\beta$. After his initial infection, node $i$ also recovers from the infection according to an exponential distribution with rate parameter $\gamma$. Thus, if $i$ recovers before he passes the infection along to $j$ then $j$ will remain susceptible so long as some other neighbor of $j$ does not infect him. In addition, the timings of all disjoint events are assumed to be independent from one another. With little additional effort one may include a latency period. The basic reproductive number for this model is given by

$$R_0 = \frac{\beta}{\beta + \gamma} \sum_{k \geq 1} \mu^{-1}k(k - 1),$$

(3.27)

with a similar threshold result holding such that an epidemic occurs with positive probability when $R_0 > 1$ [53].

Clearly an epidemic model defined in this way introduces complications that are not present in homogeneous mixing models. In particular, the force of the infection depends not only on the number of susceptible and infectious individuals but also how many connections they have and the way they are configured in the network. The result of this is a system of potentially much greater complexity. Therefore, a variety of approaches have been proposed to attempt to simplify the resulting model while still explaining the full range of possible dynamics. Some approaches group vertices by degree and model the “effective degree” of these groups of individuals [73, 13, 14, 94]. One approach is based on Edge-Based Compartmental Modeling– proposed in a series of works first by Erik Volz [111] and later Miller and Volz [85]. An earlier approach is known as the “pair approximation”
method and has existed for some time [58, 57, 97]. Both approaches make asymptotic arguments (i.e. as the size of the network goes to infinity) and this is the approach we will also adopt to study population-level dynamics. While both approaches are initially heuristic there is reasonable evidence that the results of Miller and Volz are correct [53] [30]. Later, we provide some previously unknown relationships between these approaches and in particular verify conditions under which the pair-approximation approach is actually asymptotically exact.

There are two additional modifications to this model that may make it more realistic at the cost of additional complexity. First, the network may contain multiple layers (i.e. different types of connections) that will allow for different mechanisms of disease transmission. For example, in the case of Ebolavirus there may be separate layers corresponding to community-type contact, funeral-type contact, and healthcare-type contact as some contact tracing datasets have classified these types of connections. Network models of this type have received attention in recent years [64]. Networks with these types of connections are sometimes referred to as multi-relational networks as well [15, 22, 29, 115]. Second, the structure of the network itself may change over time. Considerable attention has been given to situations where this occurs independently of the progression of the disease [4, 5, 11, 34, 70, 71, 111]. On the other hand, an emerging line of research is in cases where changes in the structure of the network are directly linked to progression of the disease [40]. In such cases, the presence of the infection will cause either formation/breakage of
an individual’s connections. This may provide more realistic forecasting of Public Health interventions and richer mathematical dynamics [102, 103, 46].

We begin by a brief introduction of approaches to simplifying limiting results on static, single layer networks. We will find later in Chapter 5 that many such principles can be generalized to the cases described above. We pay particular attention to the Edge-Based Compartamental Modeling approach and the “pair approximation” approach as these have the benefit of keeping the dimension of the resulting system manageable.

3.2.1 Edge-based compartamental modeling

A key analytical property in studying network-based models is whether results similar to Theorem 3.1.1 exist as the number of vertices in the graph $N$ approaches infinity. In a landmark paper, Erik Volz suggested that this was indeed the case for the Markovian SIR epidemic on a static Configuration Model network and surprisingly the resulting limit is of similar dimension to the classical SIR model [110].

The resulting low dimensional system is perhaps surprising given that the method works for any degree distribution $D_v$ with somewhat mild constraints and in doing so addresses one of the fundamental shortcomings of the classical SIR model; namely, that there is heterogeneity among people in a population with respect to the size of their infectious contacts neighborhood and that the size of this neighborhood is typically much smaller than the whole population. His system of equations was later refined in a paper by Miller [82].
We now outline the system of Volz-Miller under the refinement provided by Miller. Our treatment is largely heuristic as this is the method through which their equations were derived, although formal proofs of the correctness of the limiting equations on static single-layer networks have now been provided [30] [53].

The model is constructed as follows. We consider a Configuration Model (CM) random graph with degree distribution $D_v$ and let $\psi(x)$ denote the probability generating function of $D_v$. Letting $p_k$ denote $P(D_v = k)$, this means:

$$\psi(x) = \sum_{k \geq 0} p_k x^k. \quad (3.28)$$

As in the SIR model, upon being infected a given node will attempt to transmit the infection to his susceptible neighbors independently according to an exponential distribution with rate $\beta$. He will recover according to an independent realization of an exponential random variable with rate parameter $\gamma$. After this period he recovers and can no longer pass the infection to his neighbors. The CM network does not prevent self-loops or multiple edges but this does not pose a problem as the network is simple with positive probability [53].

The probability generating function will be drawn upon frequently. At some time $t > 0$ consider a “test” node $u$ that has not transmitted the infection. For the most part we can take this to mean that we pick a node uniformly from the set of susceptibles without any trouble but technically speaking the node $u$ does not have to be susceptible. Suppose also there is some probability $\theta(t)$ that describes the fraction of edges at time $t$ over which the infection has been transmitted. The fraction of edges over which the infection has not been transmitted is the same as the probability of picking any edge uniformly from the set of all
edges and computing the probability that the infection has been transmitted over that edge. Moreover, a vertex remains susceptible at time \( t \) if and only if the infection has not been passed over any of the edges in his neighborhood. Hence, the probability that a node of degree \( k \) would remain susceptible at time \( t \) is given by \( \theta(t)^k \). Since this can be done for every \( k \), the fraction of susceptibles at any time \( t \) can be written neatly by the law of total probability as:

\[
S(t) = \sum_{k \geq 0} p_k \theta^k = \psi(\theta(t))
\]  

(3.29)

We have written the last term as \( \theta(t) \) in order to emphasize that this quantity depends on time. To see why note that early in the epidemic the fraction of nodes of which the infection has not been transmitted is very close to 1 whilst as the infection saturates the population this quantity will decrease. Since the recovery dynamics work essentially the same as in the stochastic SIR model, the system can be expressed as

\[
S(t) = \psi(\theta(t))
\]

\[
\dot{R} = \gamma I
\]

\[
I(t) = 1 - S - R,
\]

where \( \dot{R} \) denotes the derivative with respect to time. The above system is not closed as we have not explained the dynamics of \( \theta \), however, doing so will yield a closed system of differential equations to describe the evolution of the process. As a result, it is easy to see that \( \theta \) is strictly decreasing in time. To derive the differential equation for \( \theta \), we must look
at the number of infectious individuals that are on the other "end" of the edges that have not transmitted the infection. To this end, set:

\[ \theta = \phi_S + \phi_I + \phi_R, \]  

(3.30)

where \( \phi_I \) represents the probability that the infection has not been transmitted over any randomly selected edge of our test node \( u \) and that the neighbor to which that edge connects is also infective at time \( t \). Given that each infectious node transmits the infection at rate \( \beta \) independently of all other edges and each infectious node (i.e. on the other end of a \( \phi_I \)-edge) recovers independently at exponential rate \( \gamma \), it must be the case that:

\[ \dot{\theta} = -\beta \phi_I. \]  

(3.31)

However, we are still not done because we must close the system with respect to \( \phi_I \). The outflow for this quantity will be whenever the infected neighbor \( v \) of \( u \) recovers and this occurs at rate \( \gamma \) for every infectious node independently. Hence, the infectious neighbors of \( u \) become recovered neighbors of \( u \) at rate \( \gamma \phi_I \), which yields the equation for \( \phi_R = \gamma \phi_I \).

It will be easier to compute \( \phi_S \) and \( \phi_R \) explicitly and then use this to solve for \( \phi_I \).

To find \( \phi_S \) we must consider the probability that a randomly selected neighbor \( v \) of \( u \) is susceptible. The answer to this will be the total number of edges owned by susceptible neighbors, divided by the total number of number of edges in the graph. To this end, we introduce the notion of the neighbor degree distribution which describes the distribution of any randomly selected neighbor of \( u \). The probability that the neighbor has degree \( k \) can in
general be written as:

\[ p^n_k = \frac{kp_k}{\mu}. \]  

(3.32)

The intuition of this is the following. The probability that a neighbor has degree \( k \) is proportional to the fraction of edges “owned” by nodes of degree \( k \). The total number of edges owned by nodes of degree \( k \) will be \( Np_kk \) by the LLN in a large graph and the total number of edges anywhere in the graph is \( N\mu \) where \( \mu = E(D_v) \). Hence, we arrive at the above expression. From our prior calculations the probability that an initially susceptible node of degree \( k \) is still susceptible at time \( t \) is given by \( \theta^k \). Thus, we can find \( \phi_S \) to be:

\[ \phi_S = \sum_{k \geq 1} \theta^{k-1} p^n_k = \sum_{k \geq 1} \frac{k\theta^{k-1}p_k}{\mu} = \frac{\psi'(\theta)}{\psi'(1)} \]  

(3.33)

Where the derivative of the generating function is with respect to its argument, not time. Lastly, \( \phi_R \) can be computed because of the proportional flow with constant \( \frac{\gamma}{\beta} \) out of \( \phi_I \) into \( \phi_R \) and the flow out of \( \phi_I \) and out of \( \theta \) altogether (when the infection is transmitted over an edge). This and the fact that they have the same initial condition shows that \( \phi_R = \frac{\gamma(1-\theta)}{\beta} \).

One more simple substitution will close the system. Recalling that \( \dot{\theta} = -\beta \phi_I \),

\[ \dot{\theta} = -\beta(\theta - \phi_S - \phi_R) = -\beta(\theta - \frac{\psi'(\theta)}{\psi'(1)} - \frac{\gamma(1-\theta)}{\beta}) \]  

(3.34)
Hence we have closed the system as needed, and the final set of differential equations is given below.

\[
\dot{\theta} = -\beta (\theta - \phi_S - \phi_R) = -\beta (\theta - \frac{\psi'(\theta)}{\psi'(1)} - \frac{\gamma(1 - \theta)}{\beta}) \\
\dot{R} = \gamma I \\
S = \psi(\theta) \\
I = 1 - S - R \\
S(0) = 1 - \alpha, I(0) = \alpha, R(0) = 0, \theta(0) = 1
\] (3.35)

Although the approach of Miller-Volz lacks the intuition of the stochastic SIR model, the resulting set of differential equations is easy to solve numerically and there is great flexibility in the selection of the degree distribution \(D_v\). In particular, the only real requirement is that the degree distribution have finite 1st and 2nd moments. It is not known if similar sets of equations can be derived in the case that either the infection or waiting times do not follow exponential distributions, though our later results suggest a method of doing this.

The proof provided by Janson (2014) of the convergence of the stochastic system to the one of Miller-Volz is accessible with a familiarity of the techniques presented in the previous section [53]. The correctness of the limit in this specific case was also verified in [30], however, under more stringent assumptions on \(D_v\).

### 3.2.2 Pair approximation approach

An older approach to the problem is often referred to as “pair approximation” which may be described as follows [97]. Assume that \(\beta\) denotes the rate of infection across an edge between a susceptible individual \(i\) and an infectious individual \(j\). Let the quantity
\([SI](t)\) denote the total number of such edges in the network at some time \(t \geq 0\). Then the rate (hazard) at which susceptible individuals are infected across the entire network is \(\beta[SI](t)\). However, this approach creates the challenge of modeling the dynamics of \([SI](t)\). To do so requires studying the connections between three nodes simultaneously. For example, \([ISI]\) refers to the total number of connections in the network between an infectious individual \(i\) connected to a susceptible individual \(j\) that is also connected to an infectious individual \(k\). Other quantities involving connections between three nodes are defined analogously. With this in mind, the dynamics may be written as:

\[
\begin{align*}
\dot{S} &= -\beta[SI] \\
\dot{I} &= \beta[SI] - \gamma I \\
[\dot{SI}] &= \beta[SSI] - (\beta + \gamma)[SI] - \beta[ISI] \\
[\dot{SS}] &= -2\beta[SSI].
\end{align*}
\] (3.36)

Closure of the system will eventually require modeling the evolution of triples, quadruples, and so forth unless the process is able to be cut off at this point. The “pair approximation” is a heuristic argument that in a large graph the number of triples \([SSI], [ISI], etc.\) can be expressed in terms of dyads and the species comprising them. Let \(X, Y, Z \in \{S, I, R\}\) and consider three connected nodes \(u, v, w\). The approximation argues
\[ = [uYZ]p(u = X|v = Y, w = Z) \]
\[ \approx [uYZ]p(u = X|v = Y) \]
\[ = \mu [YZ]\frac{[XY]}{\mu Y} \]
\[ = \frac{[XY][YZ]}{Y} \]

as has been described elsewhere [7] [97]. In the second line the approximation states for a large graph with few short cycles that the influence of \( w \) on \( u \) should be negligible, given that the status of \( v \) is known. In the third line, the number of \( [uYZ] \) dyads should be approximately \( \mu [YZ] \) where \( u \) is the average degree in the graph. Lastly, the probability is equal to the total number of \( [XY] \) dyads divided by the total number of \( [\bullet Y] \) dyads, which is approximately \( Y\mu \) by the same argument. Note that if true this would immediately allow closure of the above system by writing \( [ISI] = F([SI], S) = \frac{[SI]^2}{S} \) and a similar argument could be used to replace the \( [SSI] \) term. Thus, assuming the correctness of the pair approximation it is seen that the dynamics may be captured in terms of a system with similar complexity to the EBCM approach. The pair approximation approach is perhaps more attractive as the dynamics are described in terms of intuitive quantities. However, there are some problems with the approximation above that we will elaborate on further in Chapter 5.

Applying the pair approximation (3.37) to (3.36) yields the following system:
\[
\dot{S} = -\beta [SI] \\
\dot{I} = \beta [SI] - \gamma I \\
[\dot{SI}] = \beta \frac{[SS][SI]}{S} - (\beta + \gamma)[SI] - \beta \frac{[SI]^2}{S} \\
[\dot{SS}] = -2\beta \frac{[SS][SI]}{S}. \\
\]

(3.38)

We present both approaches as different attempts at essentially the same goal—describing epidemic dynamics as an infection spreads over a contact network. We will make use of some techniques proposed by Miller and Volz and introduce others that more closely resemble the pair approximation in Chapter 5. Finally, we note that network-based models can also be easily modified to accommodate diseases that do not follow a simple $S-I-R$ mechanism, for example, by allowing multiple stages of infection, multiple types of infection, etc. Over the course of several papers Miller and Volz have used their heuristic derivation to provide the limiting set of equations under these assumptions [86] [84]. Likewise, the approach of Rand can be adapted to these situations with appropriate accounting of newly introduced species and dyads. In a later chapter we provide a rigorous proof of some of these generalizations, sufficient conditions for equivalence of these two approaches, and a method of proof that is easily adapted to other disease mechanisms.
Chapter 4: Models of Ebolavirus

In response to the declaration of a Public Health emergency and the rapid rise in the number of cases of Ebolavirus, a number of models were proposed to describe the observed dynamics and aid in forming a Public Health response to the ongoing crisis. A recent review paper has included 66 models that attempted to describe various aspects of the population-level propagation of the disease [28]. In this section we review models of previous Ebola outbreaks as well as the approaches taken in modeling the current outbreak. Chretien et al. classify each model according to whether it considered 6 aspects of the epidemic; namely, estimation of $R_0$, intervention efficacy, spread, forecasting, viral phylogenetics, and clinical trials. All papers reviewed were considered to have addressed 4 or fewer of these characteristics of the epidemic while the methods we propose can reasonably be used to address at least estimation of $R_0$, intervention efficacy, temporal spread of the virus, forecasting, and to some extent clinical trials. Hence, we focus this section on those models that address similar characteristics of the epidemic as our own.
4.1 Models of previous Ebolavirus outbreaks

We now review the existing mathematical and statistical literature for models of Ebolavirus that have been proposed in response to prior outbreaks. Flow diagrams are a common visual representation of model assumptions by listing the possible classification of an individual in the population and giving a sense for the dynamics modeled. Therefore, we make use of these diagrams to aid in understanding the differences among these models.

The first model that is most in the spirit of the models under consideration here was given by Chowell et al. (2004) [25]. The authors looked at the cumulative case data from two of the historically largest outbreaks of Ebola (prior to 2014) which were the Congo 1995 and Uganda 2000 outbreaks which both resulted in approximately 400 cases [25]. A Markovian SEIR model was fit to the cumulative incidence data using least squares estimation. The basic reproductive number was estimated to be 1.83 for the Congo outbreak and 1.34 for the Uganda Outbreak, which can be obtained from the estimated infection and recovery rates in the same manner that was introduced in the section on Classical SIR theory. To quantify the role of Public Health interventions such as quarantine and safe burial practices the authors have introduced a time-dependent infection rate that aims to capture the effects of these measures on the spread of the disease. Letting $\beta_0$ denote the infection rate with no intervention, $\tau$ denote the time of the intervention, and $\beta_1$ denote the infection rate after the intervention has been fully adopted the authors assume that this rate
varies continuously as:

\[
\beta(t) = \begin{cases} 
\beta_0 & \text{if } t < \tau \\
\beta_1 + (\beta_0 - \beta_1)e^{-(t-\tau)} & \text{if } t \geq \tau
\end{cases}
\]

All three parameters were considered free and were estimated as part of the least squares fit. Conditional upon the parameter estimates the authors have simulated stochastic epidemic trajectories to attempt to quantify the impact of the intervention on the total number of cases and have found that a delay of as little as two weeks can double the total number of reported cases. While this model demonstrates that homogeneous mixing models may reasonably describe observed data, this assumption is in general not consistent with known chains of transmission among Ebola cases [36].

The next major Ebola-specific model proposed was given by Lekone and Finkenstadt (2006) [68] and was used to model the 1995 Zaire outbreak. This model differed in a few key respects from the model of Chowell et al.; specifically, it was based on a discrete time domain, so this may simply be viewed in a sense as a discretized set-up of the model considered by Chowell et al. This is in the spirit of what are typically called “chain-binomial” models, due to the likelihood being decomposable as a product of Binomial densities. The model is constructed as follows. Let the quantities \((S(t), E(t), I(t), R(t))\) denote the number of Susceptible, Exposed, Infected, and Recovered individuals at time \(t > 0\) with the assumption that \((S(0), E(0), I(0), R(0))\) is given by the initial condition \((S_0, E_0, I_0, R_0)\). Next assume that \(B(t)\) is the number of susceptible individuals that have contracted the infection, \(C(t)\) is the number of cases by date of first symptoms, \(D(t)\) is the number of cases removed
from the infectious population during a given time interval. Lastly, let $\tau^*$ denote the extinction time and $t^*$ denote the time that control measurements were implemented. The authors remark that they have decided to simplify the function for the effect of the control measures over time versus the Chowell et al. paper for reasons of identifiability. For all $t$ we will assume that the dynamics at $t + h$ ($h = 1$ day in our case) that the transitions are explained by:

$$
S(t + h) = S(t) - B(t)
$$

$$
E(t + h) = E(t) + B(t) - C(t)
$$

$$
I(t + h) = I(t) + C(t) - D(t)
$$

$$
S(t) + E(t) + I(t) + R(t) = N
$$

Where $N$ denotes the entire population size, and the quantities $B(t), C(t), D(t)$ are all random variables with the following distributions:

$$
B(t) \sim Binomial(S(t), P(t))
$$

$$
C(t) \sim Binomial(E(t), p_C)
$$

$$
D(t) \sim Binomial(I(t), p_R),
$$
and the binomial probabilities in each case are computed from:

\[ P(t) = 1 - e^{-\frac{\beta(t)I(t)}{N}} \]

\[ p_C = 1 - e^{-\alpha h} \]

\[ p_R = 1 - e^{-\gamma h} \]

\[ \beta(t) = \begin{cases} \beta & \text{if } t < t^* \\ \beta e^{-q(t-\tau)} & \text{if } t \geq t^* \end{cases} \]

Therefore, it is seen that this model represents a form of discretization of the continuous-time model of Chowell et al. Under the homogeneous mixing assumption the hazard of a given susceptible individual being infected at time \( t \) is given as \( N^{-1} \beta(t)I(t) \). Therefore, from the general theory of Poisson processes, his probability of infection in a given interval is

\[ 1 - e^{-\int_{t^*}^{t+\tau} N^{-1} \beta(u)I(u)du} = P(t) \]

when the quantities in the integrand are constant over the course of the interval.

This information completely specifies the likelihood function which may now be written as:

\[ L(\alpha, q, \beta, \gamma) = \prod_{i=1}^{t^*} g_1(B(t))g_2(C(t))g_3(D(t)). \]

The \( g \) terms denote the transition binomial densities for each quantity. Maximizing the likelihood with respect to the free parameters on the left hand side yields parameter estimates, and the basic reproductive number can be obtained via the Continuous Mapping Theorem to be \( \hat{R}_0 = \frac{\beta}{\gamma} \). The result of this model was \( \hat{R}_0 = 1.36 \) on an aggregate level as the authors
did not consider context-specific transmission in this analysis. Largely this estimate agreed with that found by Chowell et al.

Another model of Ebolavirus was proposed by Legrand (2007) [67]. The outbreaks studied were again the Zaire 1995 and Uganda 2000 outbreaks due to availability of the data and size of the eventual outbreak. In contrast to the previous two models, this model attempted to quantify the role of transmission that infection within hospitals, the community, and burial rituals played in driving the virus dynamics. The model is similar to that of Chowell in that it assumes homogeneous mixing between compartments but creates explicit compartments for the community, hospital, and funerary aspects of transmission. The stochastic compartmental model is specified fully by the jumps and transition rates specified by the information in the table below, and conditional on available data the likelihood can be constructed.

As an example the authors consider the possibility that the infection may take several different courses that effect transmission dynamics. An infectious individual may enter a
hospital ($I \rightarrow H$), be treated in the home but eventually pass away ($I \rightarrow F$), or be treated at home and recover there ($I \rightarrow R$). Each such transition has a specified rate and when multiple transitions are possible it is assumed that each occurs with a given probability computed from the relative hazard of each transition. After obtaining Maximum Likelihood Estimates for the free parameters, the authors estimate the basic reproductive number $R_0$ to be 2.7 for both of the outbreaks that were analyzed. This is of course somewhat larger than what was found in the other two models that have been discussed here. The contribution of this model is to explicitly include Ebola-specific aspects of transmission, such as through burial procedures, but in a homogeneous mixing framework. As a result this model neglects the type of individual heterogeneity that has been reported in other places [77] [3]. A model with identical structure to that of Legrand was proposed by Rivers et al. to model the West Africa outbreak and estimated the basic reproductive number to be 1.78 in Sierra Leone and 2.22 in Liberia [99].

### 4.2 Models of the West Africa Ebolavirus outbreak

In this section we discuss a number of issues related to modeling the West Africa Ebola outbreak. Many such models retain similar features to the ones used to describe historical outbreaks. However, as a result of recent advances in other areas and novel sources of data (for example, phylogenetic information) the breadth of models expanded beyond what was reviewed previously.
4.2.1 Data usage

Most models consider the cumulative incidence of the epidemic over time as this is often the epidemiological data that is most readily available, and similar data was made public during the West Africa outbreak [1]. Other studies have used primary epidemiological data (such as a list of infected individuals compiled from hospitalization records) which may give more detailed information than what was discussed in 2. This data may include information on contact tracing [116], sociological relationships between contacts [31], and progression of the disease after infection [107]. Other studies have used phylogenetic information obtained from viral sequencing studies, for example this data is used extensively by Gire et al., Baize et al., and others [41] [8]. The practice of using such aggregate incidence data for model estimation has been called into question for epidemic modeling and the analysis of Chretien et al. suggests that such models using cumulative incidence data report systematically narrower standard errors for $R_0$ than models using disaggregated data [63] [28]. The cause of this seems likely to be systematic under-estimation of uncertainty when aggregated data is used for model fitting [62]. In short, the tendency shown in many studies is to use aggregate incidence data and to fit a deterministic model treating any discrepancy between the data and model as extrinsic, independent, Gaussian noise. However, when the data are cumulative in nature then the tendency will be that both measurement errors increase with later temporal measurements and that there is a high degree of dependency among these errors [63].
4.2.2 Modeling approaches

In chapter 3 we have already touched on some of the relevant decisions made in the process of modeling epidemic dynamics. While deterministic models are still commonly employed we will focus our attention on other stochastic models. Moreover, most models may be classified as compartmental (e.g. assume some sort of homogeneous mixing), network-based, or agent-based. Some agent-based models of the spread of Ebola in West Africa were published, for example, one that addressed the risk of global spread of the disease [42]. Nevertheless most other published works tend to be classifiable as either compartmental or network-based. We review four models in particular that represent justifiable though different approaches to the problem than the one that we present next.

Other models specifically account for Public Health interventions during the outbreak. For example, as part of the international response to the outbreak in 2014 a number of Ebola Treatment Units (ETUs) were established in an attempt to provide care for the overwhelmed infrastructure in Sierra Leone, Guinea, and Liberia [114, 24, 78]. However, there is difficulty in obtaining reliable information on the number of individuals admitted to these facilities, when they were established, and records of patient discharge. While these interventions likely played a role in the eventual slowing of the epidemic process they are ways to incorporate this more indirectly, such as by considering a more general notion of quarantine which may also happen outside of an ETU. We now review several particular models that are of interest.
King et al. argue in favor of modeling approaches that fit de-aggregated data of the number of infected individuals (or cases) rather than the popular approach of fitting cumulative incidence data [62]. Their argument hinges on several examples that show fitting cumulative incidence data tends to yield confidence intervals with coverage that is much smaller than the nominal value when the data are generated under realistic assumptions of time-dependent variation. The central issue is that Least Squares fitting of deterministic model to aggregate incidence data implicitly assumes error homoscedasticity and independence [62]. The authors call this assumption into question given that the incidence is a strictly increasing quantity and therefore subject to a time-dependent covariance structure. It stands to reason that the variance of such measurements ought to increase as the epidemic progresses. Moreover, there is little reason to assume that the discrepancies between reported data and progression of the epidemic should be temporally independent.

To model the spread of Ebolavirus in West Africa the authors propose an $S - E^m - I - R$ compartmental model in which the exposed state is broken up into several stages. This is a commonly employed technique to accommodate an incubation period that is not exponentially distributed, as assumed in the classical SEIR model. Based on data found in the WHO report the shape parameter for the incubation period ($m$) was estimated to be 3. The force of infection on a given individual is given by a homogeneous mixing mechanism as $\lambda(t) = \mathcal{R}_0 \gamma N^{-1} I$ [62]. Moreover, the model accommodates a reporting error through a Hidden Markov Model framework, although the authors report that this parameter was not identifiable given the available data and was fixed. The authors use numerical routines
developed in other work to fit the described Hidden Markov Model and ultimately report confidence intervals for $R_0$ of $(1.1, 1.3)$ in Guinea, $(1.7, 2.2)$ in Liberia, and $(1.2, 1.4)$ in Sierra Leone. These estimates are consistent with other estimates published elsewhere [28].

In short, this model accounts for several important aspects of Ebolavirus transmission such as the distribution of incubation period and potential errors in data reporting. However, it differs fundamentally from our model proposed later due to the assumption of a spatially homogeneous infection process.

Scarpino et al. propose a mechanistic SEIR network-based model that accounts for contact clustering with the addition of a clustering coefficient $\phi$, which accounts for the presence of triangles in the network [100]. However, we show in Chapter 5 the simple pair approximation used is exact only under strict conditions and its validity with clustering
present in the network is unknown [52]. The authors derive an unbiased estimator for the network clustering coefficient and from contact tracing data estimate this quantity to be \( \hat{\phi} = .21 \) and similarly estimate the mean number of contacts per individual to be 5.74. The authors use known estimates from the literature to estimate other free parameters such as the mean infection period. The resulting system of differential equations is of dimension 14 and the authors do not report any studies regarding identifiability of the model [100]. Lastly, to fit the remaining free parameters the authors use cumulative incidence data from Liberia and assume that the data is Normally distributed around the mean given as the solution to the resulting system of differential equations (i.e. what King et al. recommended against). The authors then estimate remaining free parameters by Least Squares utilizing the Gaussian form of the likelihood.

Figure 4.3: Table of compartments and transition rates in the compartmental model proposed by Scarpino et al.
Among the proposed mechanistic models, this model bears the strongest resemblance to the approach we present in the next chapter. However, as we will see there are a few key differences. In order to close the system of equations the authors apply a triangular pair-approximation (an analogue of the pair approximation in Chapter 3 when the network has clustering). Given our results in Chapter 5 it is likely that this pair approximation is only valid when the degree distribution is Poisson. Empirical data on the degree distribution as well as reports of cases of superspreading both cast doubt on this assumption [116] [77] [3]. The authors also assume that the network is static throughout the epidemic. Lastly, the authors assume no intrinsic stochasticity in the system by assuming that the measured data is due to a difference in reporting and the deterministic underlying dynamics, rather than that the evolution of the system itself is stochastic. One can certainly argue that stochastic variation from both sources is likely present.

Yamin et al. use the same primary contact tracing data as Scarpino et al. and point out the considerable right skew in the contact distribution. Moreover, they combine this information with temporal information on viral load on survivors/non-survivors observed in previous outbreaks to approximate the progression of the illness in infectious individuals in the population. It has been suggested that viral load in an infectious individual may play a strong role in the likelihood of transmitting the illness to others [108] [36]. Therefore, the authors propose a branching-process type model in which the number of secondary cases caused by a given infectious individual depends on the number of contacts he has, his viral load over the course of this period, survival status, and the length of this
infectious period. Parameters used in the model were estimated from the literature and applied to the proposed stochastic model to simulate the early longitudinal course of the epidemic. While this makes this model a viable strategy for predicting the early behavior of the outbreak it implies that it fails to account mechanistically for the entire course of the outbreak. Nonetheless this approach is similar to the branching process approximation we propose for modeling small outbreaks where the available data is the distribution of secondary cases. That model is explored in further detail in Chapter 6 where we also present parameter estimation methods for such models.

The authors estimate $R_0 = 1.73$ and give a 95% confidence interval of $(1.63, 1.83)$ for Monteserrado County, Liberia. Interestingly, the authors conclude that the average number of secondary infections caused by an individual that ultimately survives is 0.66 compared to 2.36 for an individual that ultimately dies from the infection. The authors believe that increased viral load (higher in those individuals that do not survive) greatly increases the probability of viral transmission to one’s contacts.

Merler et al. propose a spatial agent-based model accounting for household size in the affected countries. This model incorporates context specific transmission (hospital, household, and burial ceremony) with various other parameters such as the rate of admission to healthcare facilities and incubation period fixed at values reported elsewhere [79]. The unknown rates of transmission were fit to a proprietary dataset furnished by the WHO containing information on the context in which the infection was believed to have occurred. Agent-based models of this form are quite flexible for forecasting of the spatio-temporal
spread of the epidemic, which the authors have done here. For Liberia the authors estimate $R_0$ to be 1.84 when the data is assumed to have perfect reporting and a corresponding confidence interval of $(1.60, 2.13)$. As with other studies the majority of transmission events are estimated to take place in the household (> 50%) and in hospitals. In contrast, burial ceremonies are estimated to be responsible for less than 10% of the total cases reported. The estimated context-specific rates of transmission are shown to vary significantly depending on the reporting rate. As an example, the transmission rate within hospitals is estimated to be $\beta_H = .33$ when perfect reporting is assumed but this quantity falls to .21 in the case of 50% under-reporting [79]. Nonetheless, the authors explore a variety of scenarios including interventions such as increasing the number of Ebola Treatment Units (ETUs) and perform extensive simulations of the model. As an example the authors report a predicted final size interval of $(12279, 107913)$ in the case that no intervention is taken [79].

### 4.3 Clinical trials

Several candidate vaccines for Ebolavirus have been recently studied. Here we summarize some results of a phase III clinical trial conducted in Guinea and analyzed by Henao-Restrepo et al. The authors report that a virus-based vaccine which contains a glycoprotein unique to Zaire Ebolavirus is able to induce a significant immune response [48]. The authors conduct a ring vaccination cluster-randomized trial in which a cluster is defined as a confirmed case-patient, his contacts, and contacts of contacts. Randomization is done such that every enrolled individual within the same cluster either (i) immediately receives
the vaccine or (ii) receives the vaccine 21 days after enrollment in the trial. In all the au-
thors successfully enrolled 90 clusters in an approximately 1:1 randomization ratio and
enrolled approximately 4000 individuals either qualifying as primary contacts or contacts-
of-contacts.

The authors report that statistical analysis was performed via Fisher’s Exact test, com-
paring the proportion of clusters in which at least one transmission event was reported
across the randomized groups. However, we note that this method ignores the actual count
of secondary cases caused within a given chain of transmission. Thus, it may be possible to
apply methods from Chapter 6 on a Branching Process Approximation to utilize this infor-
mation. The interim analysis performed by the authors offers some evidence that the vac-
cine is effective, for example, by noting that there were 0 reported cases among individuals
that immediately received the vaccine compared to 11 confirmed cases among individuals
receiving the delayed vaccination. We do not re-analyze the data of this trial using methods
from later chapters but note that the study design appears conducive to utilizing methods
we propose.

4.4 Dynamic network-based modeling approach

Combining ideas from classical approaches to network models with knowledge about
the epidemiology of the transmission of Ebola suggests the following modeling approach.
We classify all individuals in a given population as exactly one of Susceptible (S), Exposed
(E), Infectious (I), or Recovered (R). To account for individual-level heterogeneity (such
as superspreading) we use a network-based model that will allow for an arbitrary contact
distribution. The empirical evidence reviewed in Chapter 2 suggests that a right-skewed distribution such as the Negative Binomial is a suitable model for the epidemic in West Africa. Moreover, the known progression of Ebolavirus infection suggests that individuals are often isolated during their sickness either because of cultural beliefs such as those described by Hewlett et al. or because of an organized effort to admit infected individuals to Ebola Treatment Units [49, 107]. In mathematical terms this can be done by allowing an infectious individual to activate(A) or de-activate(D) his infectious contacts after entering this state, i.e. allowing the structure of the contact network to change and linking this change to the progression of the epidemic. Finally, as transmission of Ebola is commonly observed among family members, caretakers, and funeral attendees we seek a model that can potentially account for different transmission mechanisms [31] [68, 36]. This model should also allow for asymmetric risk to be associated with different types of contact. Lastly, we note that many of these aspects of disease transmission (i.e. multiple modes of transmission, etc.) have been pointed out as important features of future network-based models in a recent review of these models by Pellis et al. [95].

No model can address every possible issue relevant to transmission. For example, as we will study the spread of the illness at a national level we necessarily assume in some way that behavior is homogeneous within the same country. Of course this cannot be entirely true– even small countries are comprised of sociologically diverse areas whose behaviors will have some influence over disease transmission. Moreover, we are not able to fully
account for other network properties that may influence transmission of the illness, for example, contact clustering within households, hospitals, and neighborhoods. This particular aspect of network science requires further study as more reliable data becomes available.

The assumption of a Markovian infection and recovery process is likewise difficult to verify but can be combated through “staging” the course of the illness as we have done here and is seen in similar models [62]. In regards to model estimation we assume reasonable accuracy regarding reporting of new cases. While reports exist about under-reporting early in the epidemic the effort by the WHO to collect accurate data was increased in response [107]. Lastly, while there is some evidence that transmission probability is associated with viral load, the absence of reliable data on this phenomenon makes it challenging to estimate. However, this information can be incorporated by adding additional states to the model if it is known.
Chapter 5: Dynamic Network-Based Model of Ebolavirus

This section presents a dynamic network-based model of Ebolavirus that is compatible with the current scientific and epidemiological characteristics known to impact disease dynamics. We define explicitly a stochastic process evolving on a dynamic random network that models the spread of the disease in a human population. Moreover, to study population-level dynamics we derive a Law of Large Numbers for the $N$-scaled process that governs its large population behavior. The model presented is quite general and allows for multiple types of transmission.

5.1 Layered Configuration Model random networks

The epidemiological evidence presented in previous chapters suggests a model that includes multiple methods of Ebolavirus transmission is desirable. Even if data is not available to estimate the contribution of each specific mode of transmission at present, this model is useful as a theoretical construct and its properties studied. We consider the construction of a Layered Configuration Model Random Network, which is an extension of the
Configuration Model random network studied by Miller and Volz in order to accommodate these more complex dynamics.

Let \( r \) denote the number of layers, each potentially corresponding to a type of infectious contact. For any vectors \( \mathbf{x} = (x_1, \ldots, x_r) \), \( \mathbf{k} = (k_1, \ldots, k_r) \) in \( \mathbb{R}^r \), denote \( \mathbf{x}^\mathbf{k} = \prod_{i=1}^{r} x_i^{k_i} \). The probability generating function (pgf) of the multivariate degree distribution is given by

\[
\psi(\mathbf{x}) = \sum_{\mathbf{k}} p_{\mathbf{k}} \mathbf{x}^\mathbf{k}, \quad \sum_{\mathbf{k}} p_{\mathbf{k}} = 1
\]

where \( p_{\mathbf{k}} = P(k_1, \ldots, k_r) \) is the probability of a node being of degree \( \mathbf{k} \), that is, having \( k_i \) neighbors in layer \( i = 1, \ldots, r \).

Conditional on the sequence of \( N \) realizations from the degree distribution, we construct the network as follows. Each node is assigned a collection of half-edges in each layer corresponding to its degree, and then half-edges within each layer are paired uniformly at random. This construction is referred to as the “layered” Configuration Model because by restricting attention to only layer \( j \), the network is a realization of a configuration model with the degree distribution given by the \( j \)-th marginal distribution of \( \psi \) [90]. We call set of such graphs as the *layered configuration model* (LCM) and denote it by \( \mathcal{G}_r(\psi, N) \).

### 5.1.1 The Excess Degree Distribution

In a single-layer graph, the *excess degree* of a node \( u \) is calculated by following an edge to \( u \) from a neighbor \( v \) and counting the number of other neighbors (excluding \( v \)) of \( u \) [90]. However for layered networks it is necessary to extend this notion. For a given node \( u \) we will define the \( i \)-degree of \( u \) as the number of connections he has in layer \( i \).
obtain the resulting distribution we define $P_{j|i}(l)$ as the probability that a randomly selected $i$-neighbor (i.e. neighbor in layer $i$) of a node $u$ has $j$-degree equal to $l$. Then, by LCM construction, $P_{j|i}(l)$ is given as

$$P_{j|i}(l) = \sum_{k : k_j = l} k_i p_k / \mu_i$$

where $\mu_i = \partial_i \psi(1) = \sum_k k_i p_k$ is the average $i$-degree, $\partial_i$ denotes the partial derivative with respect to $x_i$, and $1$ is the vector of ones in $\mathbb{R}^r$. The intuition for this is as follows. The probability that a given $i$-neighbor of $u$ has degree $p_k$ is $k_i p_k / \mu_i$ where $\mu_i = E(D_i)$. To obtain the probability that this $i$-neighbor has $j$-degree equal to $l$ we must then sum over all $p_k$ such that $k_j = l$. Next, let $\psi^{ex}_{j|i}$ denote the pgf of the excess $j$-degree distribution of a node randomly selected as a $i$-neighbor. Then,

$$\psi^{ex}_{j|i}(x_j) = \sum_l P_{j|i}(l)x_j^l = \sum_k \frac{k_i p_k}{\mu_i} x_j^{k_j} = \frac{1}{\mu_i} \partial_i \psi(\tilde{x}^j)$$

where $\tilde{x}^j$ is the vector of ones with the $j$th coordinate replaced by $x_j$. The *average excess $j$-degree of an $i$-neighbor* is then given by

$$\mu^{ex}_{j|i} = \partial_j \psi^{ex}_{j|i}(1) = \frac{1}{\mu_i} \partial^2_{ij} \psi(1).$$

Finally, define the *normalized average excess $j$-degree of an $i$-neighbor* as

$$\kappa_{ji} = \frac{\mu^{ex}_{j|i}}{\mu_j} = \frac{\partial^2_{ij} \psi(1)}{\partial_i \psi(1) \partial_j \psi(1)}.$$  \hspace{1cm} (5.2)

It is important to note that when $r = 1$, we see $P_{j|i}(k) = k p_k / \mu$, which is commonly referred to as the neighbor degree distribution. Thus the excess degree distribution becomes $q_k = (k + 1) p(k + 1)/\mu$ and $\kappa$ is the ratio of the mean excess degree to the mean degree [109]. This quantity will play an important role in further analyses.
### 5.1.2 SIdaR process

Assume that we have a realization of an LCM $\mathcal{G}_r(\psi, N)$ specifying the contact network for a population of size $N$. We retain the S-I-R framework in which individuals are classified based on disease status [60]. $S$, $I$ and $R$ correspond to susceptible, infected, and recovered (or removed) individuals. Edges within each layer represent potentially infectious contact of a given type, for example between community or family members. Moreover, we assume contacts are permitted to change in response to the infectious status of an individual. Specifically, we assume that nodes will either activate or deactivate their edges, and assume that an infectious node drops (resp. activates) edges in layer $j$ at rate $\delta_j$ (resp. $\eta_j$). We further assume that at most one of $\delta_j$ and $\eta_j$ are nonzero. Thus, all edges in deactivating layers are initially activated, and all edges in activating layers are initially deactivated. For simplicity let $j = 1, \ldots, k$ denote the deactivating layers in the network (with $\eta_j = 0$) and let $j = k + 1, \ldots, r$ denote the activating layers (with $\delta_j = 0$). Then conditional on the process history there are $2r + 1$ event types that may occur: infection ($I$) along an edge of any of the $r$ types, drop ($d$) of a deactivating edge or activation ($a$) of an activating edge, and recovery ($R$). The timings of all events are assumed to follow independent exponential clocks with the following rates:

We denote the set of nodes that are susceptible at time $t \geq 0$ as $S_t$ and define the sets of infectious and recovered nodes as $I$ and $R$ respectively. However, for notational convenience in many cases we will suppress the dependence on time. The end goal of describing
\( \beta_j \) rate of infection along \( j \)-edges (\( S \xrightarrow{j} I \)), \( j = 1, \ldots, r \)

\( \delta_j \) rate of deactivation (drop) of \( j \)-edges, \( j = 1, \ldots, k \)

\( \eta_j \) rate of activation of \( j \)-edges, \( j = k+1, \ldots, r \)

\( \gamma \) rate of recovery (\( I \xrightarrow{} R \)).

**Table 5.1**: Listing of possible transitions of the \( SIdaR(r,k) \) process and corresponding notation.

the evolution of this system in terms of species \((S,I,R)\) and dyads between individuals suggests introduction of the following quantities.

For any \( u \in S \) let \( X^j_{SI,u} \) and \( X^j_{SS,u} \) denote the number of infectious and susceptible active \( j \)-neighbors of \( u \). Also, let \( X^j_{SI,u} \) and \( X^j_{SS,u} \) denote the number of inactive infectious and susceptible \( j \)-neighbors of \( u \). Similarly for any \( u \in I \) let \( X^j_{IS,u} \) and \( X^j_{IS,u} \) denote the number of susceptible active and deactivated, respectively, \( j \)-neighbors of \( u \). The pair approximation model that was presented in Chapter 3 suggests that tracking the dynamics of such dyad quantities will be sufficient to describe the dynamics of the epidemic.

Therefore, we consider aggregate variables that represent the total number of nodes or pairs of neighboring nodes (i.e. dyads) with a given disease status. The aggregate variables may be written in terms of the number of connections of individuals as follows. The total number of \( j \)-edges between susceptible and infectious individuals is denoted \( X^j_{SI} \) and is given by \( X^j_{SI} = \sum_{u \in S} X^j_{SI,u} \). We denote the aggregated dyad counts as vectors in \( \mathbb{R}^r \), e.g. \( X_{SI} = (X^1_{SI}, \ldots, X^r_{SI}) \) and likewise for \( X_{SI}, X_{SS}, \) and \( X_{SS} \). It is important to note that \( X_{SS} \) and \( X_{SS} \) count the edges twice. We let \( X(t) = (X_S,X_I,X_{SI},X_{SI},X_{SS},X_{SS})(t) \) denote the state of the aggregate stochastic process (with dimension \( 4r + 2 \)) at time \( t \geq 0 \) where \( X_S \) and \( X_I \)
denote the number of susceptible and infectious nodes, respectively. Note that the number of recovered individuals is given by $X_R = N - X_S - X_I$ and so, for the sake of simplicity, we ignore the equation for $X_R$ throughout. The transitions for the aggregate process are listed in Table 5.1.2.

We refer to the aggregate process as the $S1daR(r,k)$ model in order to emphasize the activation (a) and deactivation (d) dynamics. While we believe the network-based model to generally present a more realistic description of population-level dynamics than homogeneous mixing models, this assumption comes at a price. The evolution of the aggregated variables under study becomes complicated, largely due to the aggregation of the nodes that unravels the Markov property, as studied in other places [5].

Note that a dyad subscript (e.g. $X_{SI}$) is understood throughout to denote a (row) vector in $\mathbb{R}^r$. For simplicity we take multiplication, division, integration and ordering of such vectors to be coordinatewise. All state variables depend on $N$ but this is ignored notationally as the dependence is clear.

Consider the $S1daR(r,k)$ process $X(t)$ on the LCM $\mathcal{G}_r(\psi, N)$ with transitions as outlined in Table 5.2. The Doob-Meyer decomposition theorem [80] guarantees the existence of a zero-mean martingale $M(t) = (M_S, M_I, M_{SI}, M_{SI}, M_{SS}, M_{SI})(t)$ such that

$$X(t) = X(0) + \int_0^t \mathcal{F}_X(X(s)) ds + M(t) \quad (5.3)$$
Table 5.2: Transitions for the SlidaR(r,k) process according to the 2r + 1 possible event types with corresponding rates. Network arrangements corresponding to the transitions are also given with ^ and ~ denoting, respectively, active and deactivated edges of type j between nodes (denoted u, v and w). Here, \( N_{Y,u}^j \) denotes the set of l-neighbors of node u with disease status Y.

<table>
<thead>
<tr>
<th>Event</th>
<th>Rate</th>
<th>Transition</th>
<th>Arrangement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infection of ( u ) by ( v, j = 1, \ldots, r )</td>
<td>( \beta_j X_{SI}^j )</td>
<td>((X_S, X_l) \rightarrow (X_S - 1, X_l + 1))</td>
<td>( v - u )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \Delta X_{SI}^j = -1 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>((X_{SS}^j, X_{SI}^j) \rightarrow (X_{SS}^j - X_{SS,u}^j, X_{SI}^j + X_{SS,u}^j))</td>
<td>( v - u - w ) for ( w \in N_{S,u}^j )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>((X_{SS}^j, X_{SI}^j) \rightarrow (X_{SS}^j - X_{SS,u}^j, X_{SI}^j + X_{SS,u}^j))</td>
<td>( v - u - w ) for ( w \in N_{S,u}^j )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( X_{SI}^j \rightarrow X_{SI}^j - X_{IS,u}^j )</td>
<td>( v - u ) for ( w \in N_{I,u}^j )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( X_{SI}^j \rightarrow X_{SI}^j - X_{IS,u}^j )</td>
<td>( v - u ) for ( w \in N_{I,u}^j )</td>
</tr>
</tbody>
</table>

Deactivation of j-edge, \( j = 1, \ldots, k \) \( \delta_j X_{SI}^j \) \( X_{SI}^j \rightarrow X_{SI}^j - 1 \)

Activation of j-edge, \( j = k + 1, \ldots, r \) \( \eta_j X_{SI}^j \) \( (X_{SI}^j, X_{SI}^j) \rightarrow (X_{SI}^j - 1, X_{SI}^j + 1) \)

Recovery of infected \( u \) \( \gamma X_l \) \( X_l \rightarrow X_l - 1 \)

where the integrable function \( \mathcal{F}_X(X) = (\mathcal{F}_S, \mathcal{F}_I, \mathcal{F}_{SI}, \mathcal{F}_{SS}, \mathcal{F}_{SS}) (X) \) is given by

\[
\mathcal{F}_S(X_{SI}) = - \sum_{i=1}^{r} \beta_i X_{SI}^i, \\
\mathcal{F}_I(X_l, X_{SI}) = \sum_{i=1}^{r} \left( \beta_i X_{SI}^i \right) - \gamma X_l, \\
\mathcal{F}_{SI}^j(X_{SI}, X_{SI}, X_{SS}) = \sum_{i \in S} \left( \sum_{i=1}^{r} \beta_i X_{SI,i}^j (X_{SS,i}^j - X_{SI,i}^j) \right) - (\gamma + \delta_j)X_{SI}^j + \eta_j X_{SI}^j, \quad j = 1, \ldots, r, \\
\mathcal{F}_{SI}^j(X_{SI}, X_{SI}, X_{SS}) = \sum_{i \in S} \left( \sum_{i=1}^{r} \beta_i X_{SI,i}^j (X_{SS,i}^j - X_{SI,i}^j) \right) - (\gamma + \eta_j)X_{SI}^j + \delta_j X_{SI}^j, \quad j = 1, \ldots, r, \\
\mathcal{F}_{SS}^j(X_{SI}, X_{SS}) = - 2 \sum_{i \in S} \beta_i X_{SI,i} X_{SS,i}^j, \quad j = 1, \ldots, r, \\
\mathcal{F}_{SS}^j(X_{SI}, X_{SS}) = - 2 \sum_{i \in S} \beta_i X_{SI,i} X_{SS,i}^j, \quad j = 1, \ldots, r, \\
\mathcal{F}_{SS}^j(X_{SI}, X_{SS}) = - 2 \sum_{i \in S} \beta_i X_{SI,i} X_{SS,i}^j, \quad j = 1, \ldots, r.
\]
To elucidate the intuition of this decomposition we explain the evolution of the $S$ and $SI$ equations. The rate at which susceptibles become infected depends only on the state variable $X_{SI}$ since the infection is transmitted independently across such dyads. The hazard at which susceptibles are lost due to infection in a layer $l$ is equal to the number of these connections across the entire network ($X_{SI}^l$) times the independent rate at which the infection is transmitted across these connections ($\beta_l$). Summing across all layers in the network yields the total infectious pressure appearing on the right hand of the $S$ equation in (5.4).

To understand the dynamics of the $SI$ quantity, consider the infection of a susceptible node $i$. When $i$ is infected by the disease, a new $SI$ dyad of type $j$ is created between $i$ and all of his susceptible $j$-type neighbors. By the notation we have defined this quantity is given by $X_{SS,i}^j$ (since $i$ was susceptible prior to the infection occurring). On the other hand, all $SI$ connections of which $i$ was the susceptible node now become $I-I$ dyads. Therefore, the net change in the number of $SI$ dyads of type $j$ is given by $X_{SS,i}^j - X_{SI,i}^j$. This change in the dyads occurs regardless of through which layer $i$ was infected. Lastly, the hazard of infection that $i$ experiences in layer $l$ is given by $\beta_l X_{SI,i}^l$. Summing over all of the independent layers gives the total hazard of infection for node $i$ to be $\sum_{l=1}^r \beta_l X_{SI,i}^l$.

We now define two more variables necessary to describe the evolution of the process in the large graph limit. Let $X_{S\bullet}(t) = (X_{S\bullet}^1(t), \ldots, X_{S\bullet}^r(t))$ where $X_{S\bullet}^j(t)$ is the number of $j$-edges belonging to susceptible nodes at time $t$. We partition the collection of susceptible nodes $S$ by their degree $k \geq 0$ so that $S = \cup S_k$ and $X_S = \sum_k X_{S_k}$. Note

$$X_{S\bullet}(t) = \sum_k k X_{S_k}(t).$$  \hfill (5.5)
We also define $\theta = (\theta_1, \ldots, \theta_r)$ by
\[
\theta(t) = \exp\left(-\beta \int_0^t \frac{X_{SI}(s)}{X_{S\bullet}(s)} \, ds\right),
\]
where $\beta = (\beta_1, \ldots, \beta_r)$. We may interpret $\theta_j(t)$ as the probability that no infection has occurred along a $j$-edge by time $t$ in a susceptible node of $j$-degree one, given that the node was susceptible at $t = 0$. Said differently, $\theta^1$ is the probability that a susceptible node of multi-degree 1 has not been infected through any layer by time $t$, given that the node was susceptible at $t = 0$. Note that we may equivalently define $\theta$ as the solution to the integral equation
\[
\theta(t) = \theta(0) + \int_0^t \mathcal{F}_\Theta(X_{SI}(s), X_{S\bullet}(s), \theta(s)) \, ds
\]
where $\theta(0) = 1$ and
\[
\mathcal{F}_\Theta(X_{SI}, X_{S\bullet}, \theta) = -\beta \frac{X_{SI}}{X_{S\bullet}}.
\]

As is seen in the Theorem 5.1.1 below, the quantity $\theta$ plays a key role in describing the evolution of $X(t)$ in the large graph limit. The use of such a variable was introduced by Volz and Miller in their edge-based approach that we reviewed in Chapter 3. In some ways this quantity may be less attractive to study, lacking the intuition of other quantities in the system [110] [82]. In the case of a static graph with $r = 1$ layer, the large graph limit of $\theta$ corresponds to the same variable in the standard SIR edge-based model [85]. Shortly, we will consider conditions on the degree distribution under which this quantity may be eliminated from the resulting final system.

The stochastic process defined in Section 5.1.2 is complex and difficult to analyze. In particular, the martingale decomposition given in (5.4) requires us to be able to describe the
“neighborhood” of a susceptible node \( i \) in order to go any further. In the following section we present a limit theorem (Theorem 5.1.1) that shows this stochastic process converges to a relatively simple system of ODEs as the number of nodes \( (N) \) tends to infinity. The limiting ODEs retain a high level of intuition and are simple enough to be amenable to analysis. In the case of a finite but large population, analysis of this deterministic system provides a good approximation to disease dynamics, an approach that we explore in Chapter 7. In the general case, the evolution of the quantities of interest, \( X(t) \), will involve a function of the \( \theta \) variable defined in the preceding section. In Section 5.1.4, we state corollaries that relate our result to edge-based models in the special case of static graphs [86]. Finally, in Section 5.1.5 we show that, for a particular class of degree distributions the evolution of \( X(t) \) decouples from \( \theta \), which reveals a perhaps surprising connection between our limiting system and the low-dimensional models obtained via pair approximation in Chapter 3.

5.1.3 SIdaR Law of Large Numbers

All limits considered below, unless otherwise noted, are for \( N \to \infty \). We use \( \xrightarrow{P} \) to denote convergence in probability in the space of right-continuous with finite left limits (càdlàg) stochastic processes on a random network constructed as an LCM \( \mathcal{G}_r(\psi,N) \). We say that a sequence of random variables \( Y_N \to \infty \) with high probability (w.h.p.) if \( P(Y_N > k) \to 1 \) for any \( k > 0 \). Let \( 0 < T < \infty \). We assume the following conditions:

(I) For \( 0 < t \leq T \), \( X_{S^\bullet}(t) \to \infty \) w.h.p.
(II) The fractions of initially susceptible, infected, and recovered nodes converge, respectively, to some $\alpha_S, \alpha_I, \alpha_R \in [0,1]$, i.e.

$$X_S(0)/N \xrightarrow{P} \alpha_S, \quad X_I(0)/N \xrightarrow{P} \alpha_I, \quad X_R(0)/N \xrightarrow{P} \alpha_R.$$ 

Furthermore, $\alpha_S > 0$, $\alpha_I > 0$, and the initially infected and recovered nodes are chosen randomly.

(III) $\sum_k ||k||^2 p_k < \infty$.

Condition (I) implies that, for large graphs, the infection does not deplete the population of susceptible individuals completely. That is, some proportion of individuals remain susceptible on $[0,T]$ and, hence, $\theta$ is well-defined in (5.6). Furthermore this condition also implies the average $j$-degree of a randomly chosen node is finite. This follows as $\partial_j \psi(1)$ is positive since $0 < \liminf N^{-1}X_s \leq \partial \psi(1)$.

Condition II implies that the initial conditions for the dyads, scaled by $N$, also converge in probability, i.e.

$$X_{SI}^j(0)/N \xrightarrow{P} \alpha_{SI}^j, \quad X_{SI}^j(0)/N \xrightarrow{P} \alpha_{SI}^j, \quad X_{SS}^j(0)/N \xrightarrow{P} \alpha_{SS}^j, \quad X_{SS}^j(0)/N \xrightarrow{P} \alpha_{SS}^j, \quad j = 1, \ldots, r$$

(5.9)

where, for the deactivating layers $j = 1, \ldots, k$,

$$\alpha_{SI}^j = \alpha_S \alpha_I \mu_j, \quad \alpha_{SI}^j = 0, \quad \alpha_{SS}^j = \alpha_S^2 \mu_j, \quad \alpha_{SS}^j = 0, \quad j = 1, \ldots, k$$

and, for the activating layers $j = k+1, \ldots, r$,

$$\alpha_{SI}^j = 0, \quad \alpha_{SI}^j = \alpha_S \alpha_I \mu_j, \quad \alpha_{SS}^j = 0, \quad \alpha_{SS}^j = \alpha_S^2 \mu_j, \quad j = k+1, \ldots, r.$$

79
we obtain the initial condition $\alpha_{SI}^j$ for $j = 1, \ldots, k$ as follows. By assumption, all deactivating layer edges are initially activated and all activating layer edges are initially deactivated. Then by II, the limiting probability of selecting a random node that is susceptible is $\alpha_S$. The average number of $j$-edges this node has is $\mu_j$, and the limiting probability that any given edge connects to an infected node is $\alpha_I$. Therefore, $\alpha_{SI}^j = \alpha_S \alpha_I \mu_j$. The initial conditions for other dyad quantities are obtained similarly. We denote $\alpha = (\alpha_S, \alpha_I, \alpha_{SI}, \alpha_{\tilde{SI}}, \alpha_{SS}, \alpha_{\tilde{SS}})$ in further analysis.

Condition (III) implies that $\sum_k k^2 p_k < \infty$ for $j = 1, \ldots, r$ such that the degree distribution has finite second moments. This further implies that the multigraph considered when matching half-edges uniformly at random is a simple graph with positive probability [53].

We introduce one further quantity that plays a key role in describing how network structure affects the large graph limit. For $1 \leq j, l \leq r$, let $\tilde{\kappa}_{jl}$ be defined by

$$\tilde{\kappa}_{jl}(x) = \frac{\psi(x) \partial^2 \psi(x)}{\partial_j \psi(x) \partial_l \psi(x)}.$$  

(5.10)

The quantity $\kappa_{jl}(\theta)$ can be interpreted as the ratio of the average excess $j$-degree of a susceptible node chosen randomly as an $l$-neighbor of an infectious node to the average $j$-degree of a susceptible node.

Lastly we give the function that by Theorem 5.1.1, describes the evolution of $(X(t), \theta(t))$ in the large graph limit. Let $(x, x_\Theta) = (x_S, x_I, x_{SI}, x_{\tilde{SI}}, x_{SS}, x_{\tilde{SS}}, x_\Theta)$ and define $\mathcal{H}(x, x_\Theta) =
(\mathcal{H}_X, \mathcal{H}_\Theta)(x, x_\Theta)$ where $\mathcal{H}_X = (\mathcal{H}_S, \mathcal{H}_I, \mathcal{H}_{SI}, \mathcal{H}_{SS}, \mathcal{H}_\Theta)$ and $\mathcal{H}_\Theta$ are given by

$$\mathcal{H}_S(x_{SI}) = -\sum_{l=1}^r \beta_l x^l_{SI},$$

$$\mathcal{H}_I(x_I, x_{SI}) = \sum_{l=1}^r \left( \beta_l x^l_{SI} \right) - \gamma x_I,$$

$$\mathcal{H}_{SI}^j(x_{SI}, x_{SI}, x_{SS}, x_\Theta) = \sum_{l=1}^r \left[ \beta_l \bar{\kappa}_{jl}(x_\Theta) \frac{x^l_{SI} x^j_{SI} - x^j_{SI}}{x_S} \right] - \left( \beta_j + \gamma + \delta_j \right) x^j_{SI} + \eta_j x^j_{SI}, \quad j = 1, \ldots, r,$$

$$\mathcal{H}_{SS}^j(x_{SI}, x_{SS}, x_\Theta) = -2 \sum_{l=1}^r \beta_l \bar{\kappa}_{jl}(x_\Theta) \frac{x^l_{SI} x^j_{SS} - x^j_{SI}}{x_S}, \quad j = 1, \ldots, r,$$

$$\mathcal{H}_{\Theta}^j(x_{SI}, x_\Theta) = -\beta_j \frac{x^j_{SI}}{\alpha S \partial_j \psi(x_\Theta)}.$$

**Theorem 5.1.1** (law of large numbers). Assume conditions (I–III) for the LCM $\mathcal{G}_{r}(\psi, N)$. Then, for any $0 < T < \infty$,

$$\sup_{0 < t \leq T} \| (X(t)/n, \theta(t)) - D(t) \| \to 0$$

where $D(t) = (D_X, D_\Theta)(t)$ is the solution of

$$D(t) = D(0) + \int_0^t \mathcal{H}(D(s))ds \quad (5.12)$$

with initial conditions $D_X(0) = \alpha$ and $D_\Theta(0) = 1$. 

81
Thus, Theorem 5.1.1 specifies the large graph limit of the aggregated $SIdaR(r,k)$ process on $\mathcal{G}_r(\psi,N)$ under conditions (I–III). Or similarly, $(X(t)/N, \theta(t))$ converges uniformly in probability on any finite interval $[0,T]$ to the solution $D(t)$ of the deterministic set of equations given above by (5.12).

5.1.4 Relationship to edge-based limiting systems

We consider two special cases of the large graph limit theorem for multilayer networks. First, we consider a static network which results from $\delta_j = \eta_j = 0$ for $j = 1, \ldots, r$. Corollary 5.1.0.1 shows that in this case our system (5.12) is equivalent to an edge-based model with multiple modes of transmission. The model is one proposed by Miller and Volz [86] but with a modification to allow for a large number of initially infected nodes (following [83] where Miller modifies the standard SIR edge-based model for such a scenario). In the case that the initially infected nodes are randomly chosen (which in our case follows from II), the resulting model is derived from the methods in Chapter 3 to be:

$$\dot{\theta}_j = -\beta_j \theta_j + \beta_j \alpha_S \frac{\partial_j \psi(\theta)}{\partial \psi(\theta)} + \gamma(1 - \theta_j) + \beta_j \alpha_R, \quad j = 1, \ldots, r$$

$$\dot{R} = \gamma I$$

$$S = \alpha_S \psi(\theta)$$

$$I = 1 - S - R$$

$$\theta_j(0) = 1, \quad R(0) = \alpha_R, \quad S(0) = \alpha_S, \quad I(0) = 1 - \alpha_S - \alpha_R.$$  

Details of the equivalency, i.e. proof of Corollary 5.1.0.1, are given in [52].
**Corollary 5.1.0.1.** Assume $\delta_j = \eta_j = 0$ for $j = 1, \ldots, r$ and the conditions of Theorem 5.1.1 hold. Then, the conclusions of Theorem 5.1.1 hold where $D(t)$ is equivalent to the solution of the edge-based model with multiple modes of transmission (5.13).

We further consider the special case of a static, single layer graph (i.e. $r = 1$ and $\delta_1 = \eta_1 = 0$). In the case $r = 1$, (5.13) reduces to the well-known edge-based SIR model of Volz and Miller et al. [85, 110], which has been proven to be the large graph limit of the SIR stochastic process on a static configuration model graph [53, 30]. Thus our proof of Theorem 5.1.1 can be viewed as a generalization of these results by allowing for heterogeneity in contact type and changes in the network in response to propagation of the disease. By taking $r = 1$ in Corollary 5.1.0.1 we have provided an alternative proof of this fact.

**Corollary 5.1.0.2.** Assume $r = 1$, $\delta_1 = \eta_1 = 0$ and the assumptions of Theorem 5.1.1 hold. Then, the conclusions of Theorem 5.1.1 hold where $D(t)$ is equivalent to the solution of the edge-based SIR model (5.13) with $r = 1$.

We describe further details of the equivalence between the $Sl\mathit{d}aR(r,k)$ model and the results of Miller and Volz in [52]. In conclusion, we see that 5.1.1 provides an alternative proof of the correctness of the system considered by Miller and Volz and significantly expands the class of network-based models of heterogeneous degree for which a Law of Large Numbers has been rigorously derived. Perhaps equally important in our view is a novel relationship between these systems and pair approximation models which we now discuss further.
5.1.5 Pairwise limiting systems

In this section we consider a certain class of degree distributions where the left hand side $\bar{\kappa}_{ij}$ as defined by equation (5.10) is constant. Upon inspection of the definition of $\bar{\kappa}_{ij}$ it may seem if a constant relationship exists between $\psi$ and its derivatives then this may be possible. Given that the condition holds, the $\bar{\kappa}_{ij}$ in (5.11), which in general depend on $\theta$ may be replaced by the appropriate constant instead. As this is the only place $\theta$ appears on the right hand side for dyad-related quantities it would appear that this might make the $\theta$ quantity superfluous, i.e. it can be eliminated. In the end, this offers a simplification of the limiting system (5.11) and, in fact, the system of differential equations defining the large graph limit will coincide with the model derived via the pairwise approach. Therefore this may be considered a sufficient condition for the large graph correctness of the pairwise approach to hold.

Poisson-type distributions

We define a multivariate Poisson-type (PT) distribution to be a distribution with a pgf $\psi$ that satisfies

$$\bar{\kappa}_{ij}(x) \equiv \bar{\kappa}_{ij}(1) = \kappa_{ij}, \quad i, j = 1, \ldots, r$$

(5.14)

where we recall the definition of the normalized average excess degree, $\kappa_{ij}$, in equation (5.2). At first glance, (5.14) may lack apparent intuition. However, several useful discrete distributions can be shown to satisfy the condition. For example, in the single layer case this condition is equivalent to

84
\[ \partial \psi(x) = \partial \psi(1) \psi^\kappa, \]  

(5.15)

which is satisfied by the univariate Poisson (\(\kappa = 1\)), Binomial(\(n, p\)) (\(\kappa = \frac{n-1}{n}\)), and Negative Binomial(\(r, p\)) (\(\kappa = \frac{1+r}{r}\)) distributions. As a simple example, consider the probability generating function of the univariate Poisson(\(\lambda\)) distribution given by

\[ \psi(x) = \exp\{\lambda(x - 1)\}, \]  

(5.16)

whose derivative is easily shown to be

\[ \partial \psi(x) = -\lambda \exp\{\lambda(x - 1)\}. \]  

(5.17)

Since \(\partial \psi(1) = \lambda\) and \(\kappa = 1\) it is then easily checked that (5.15) is satisfied. The other distributions that we have listed may be checked in similar fashion.

Some special cases of interest that fall within this class are given. A \(k\)-regular graph (where all nodes have degree \(k\)) is a special case of the binomial distribution. The geometric distribution is a special case of the negative binomial distribution and recently Bansal et al. have shown that the geometric distribution (i.e. the discrete analog of the exponential distribution) gives the best fit for several empirical contact networks [10]. Moreover, the published Ebolavirus contact tracing dataset published by Fallah et al. is well approximated by the Negative Binomial distribution as shown in Chapter 2 [35].

In the multilayer case with \(r > 1\), it is straightforward to show that independent layers that satisfy the univariate PT-condition also satisfy (5.14). The estimation results presented
in the following chapter remain valid without this condition but it will be seen that some simplification can be found when it is satisfied.

**Pairwise model**

We seek to formalize some of the simplifications alluded to in the previous section. Suppose the that $\psi$ satisfies condition (5.14), then the limiting system (5.11) defining $H = (H_X, H_\Theta)$ has a particularly simple form. Substituting the constant $\kappa_{jl}$ for $\tilde{\kappa}_{jl}(x_\Theta)$ decouples $H_X$ from $H_\Theta$. We consider the resulting model in this section and introduce some new notation to do so. Let $[XY]_j$ and $[\tilde{XY}]_j$ denote, respectively, the number of activated and deactivated edges of type $j$ between a node of status $X$ and a node of status $Y$. Let $[XYZ]_{ij}$ denote the number of triples with an active $i$-edge between nodes of status $X$ and $Y$ and an active $j$-edge between node the node of status $Y$ and one of status $Z$. Similarly, $[\tilde{XYZ}]_{ij}$ will denote such triples where the $i$ edge is dormant.

Under the correlation equations approach of Rand [97], triples are needed to describe the evolution of pairs, quadruples (e.g. $[XYZW]_{ijk}$) are needed to describe the evolution of triples, and so forth. As discussed in Chapter 3, a pair approximation for triples is used in order to close the system at the level of pairs [97]. For consideration of triples in the multilayer setting, we must take into account the edge types and the appropriate excess degrees as defined in Section 5.1. Let $p(u = X | A)$ denote the probability that a node $u$ has disease status $X \in \{S, I, R\}$ given an edge (or triple) arrangement $A$ for $u$. We derive the
pair approximation of $[XYZ]_{ij}$ as follows:

$$[XYZ]_{ij} = [uYZ]_{ij}P(u = X|[uYZ]_{ij})$$  \hspace{1cm} (5.18)

$$\approx [uYZ]_{ij}P(u = X|[uY]_i)$$  \hspace{1cm} (5.19)

$$= \mu_{i|j}^{ex}[YZ]_j \frac{[XY]_i}{\mu_{iY}}$$  \hspace{1cm} (5.20)

$$= \kappa_{ij} \frac{[XY]_i[YZ]_j}{Y}$$  \hspace{1cm} (5.21)

with $\kappa_{ij}$ as defined in equation (5.2). The intuition is as follows. The first equality states that the number of such triples may be found as the number of $[uYZ]_{ij}$ triples (where $u \in \{S, I, R\}$) times the probability that $u = X$ conditional on the $[YZ]_{ij}$ connection. The second line is the pair approximation, as in [7], in which for a network with few short cycles the probability that $u = X$ should be approximately independent of the state of third node conditional on the status of the second. In the third line, the total number of triples of the form $[uYZ]_{ij}$ may be computed as $[YZ]_j \mu_{i|j}^{ex}$ by following the connection along the j-edge. That is, there are $[YZ]_j$ of these dyads and the average number of i-type neighbors of a given j-neighbor was defined earlier as $\mu_{i|j}^{ex}$. Finally, the total number of i-neighbors of all nodes $v \in Y$ will be given by $\mu_{iY}$. Hence, the probability that $u \in X$ conditional on $v \in Y$ is $\frac{[XY]_i}{\mu_{iY}}$. The last equation applies the definition given in (5.2). It is also straightforward to calculate $[XYZ]_{ij} = [XYu]_{ij}P(u = Z|[XYu]_j)$ in order to show that $\kappa_{ji} = \kappa_{ij}$.

Applying the correlation equations approach using the pair approximation (5.21) to the $SIdaR(r,k)$ dynamics described in Section 5.1.2 results exactly in the same equations as
the limiting system (5.12) in the case of a PT distribution:

\[
\frac{dS}{dt} = -\sum_{j=1}^{r} \beta_j [SI]_j, \\
\frac{dI}{dt} = \sum_{j=1}^{r} \beta_j [SI]_j - \gamma I, \\
\frac{d[SI]_j}{dt} = \sum_{l=1}^{r} \beta_l \kappa_{jl} \frac{[SS]_j [SI]_l}{S} - \sum_{l=1}^{r} \beta_l \kappa_{jl} \frac{[SI]_j [SI]_l}{S} \\
- (\beta_j + \gamma + \delta_j) [SI]_j + \eta_j [\tilde{SI}]_j, \quad j = 1, \ldots, r, \\
\frac{d[\tilde{SI}]_j}{dt} = \sum_{l=1}^{r} \beta_l \kappa_{jl} \frac{[SS]_j [SI]_l}{S} - \sum_{l=1}^{r} \beta_l \kappa_{jl} \frac{[\tilde{SI}]_j [SI]_l}{S} \\
- (\eta_j + \gamma + \delta_j) [\tilde{SI}]_j + \delta_j [SI]_j, \quad j = 1, \ldots, r, \\
\frac{d[SS]_j}{dt} = -2 \sum_{l=1}^{r} \beta_l \kappa_{jl} \frac{[SS]_j [SI]_l}{S}, \quad j = 1, \ldots, r, \\
\frac{d[\tilde{SS}]_j}{dt} = -2 \sum_{l=1}^{r} \beta_l \kappa_{jl} \frac{[SS]_j [SI]_l}{S}, \quad j = 1, \ldots, r.
\]

Here, we have derived the system (5.22) using absolute pair and triple counts. However, if we scale all variables by the graph size \( N \), the nondimensional quantities satisfy the same system of equations (this holds for any \( N \) and, hence by continuity of the solution, also in the limit). From here on we will consider only the scaled variables, which will be convenient when we state the law of large numbers in Corollary 5.1.0.3. Accordingly, we set the initial conditions to be

\[
(S, I, [SI]_1, \ldots, [SI]_r, [\tilde{SI}]_1, \ldots, [\tilde{SI}]_r, [SS]_1, \ldots, [SS]_r, [\tilde{SS}]_1, \ldots, [\tilde{SS}]_r)(0) = \alpha
\]

so that they agree with the initial conditions in Theorem 5.1.1 as \( N \to \infty \).

Observe that we can reduce the dimension of the system (5.22) further. We only need to keep track of the deactivated edges for the activating layers, i.e. \( j = k + 1, \ldots, r \) and
\([SS]_j \equiv 0\) for an activating layer since its initial condition is zero. Hence, we only must track \([SS]_j\) for \(j = 1, \ldots, k\). We refer to system (5.22) with its initial condition (5.23) as the pairwise model.

The discussion above is summarized in the following corollary.

**Corollary 5.1.0.3.** Assume the conditions of Theorem 5.1.1 hold for LCM \(\mathcal{G}_r(\psi, N)\). Then, the conclusions of Theorem 5.1.1 hold where \(D(t)\) is the solution of the pairwise model (5.22) if and only if \(\mathcal{G}_r(\psi, N)\) has a multivariate Poisson-type degree distribution.

Furthermore, we can consider the implications of Corollary 5.1.0.3 in the static, single layer case. If \(r = 1\) and \(\delta_1 = \eta_1 = 0\), then the pairwise model (5.22) reduces to the well-known correlation equations model of Rand [97]:

\[
\begin{align*}
\frac{dS}{dt} &= -\beta [SI], \\
\frac{dI}{dt} &= \beta [SI] - \gamma I, \\
\frac{d[SI]}{dt} &= \beta \kappa \frac{[SS][SI]}{S} - \beta \kappa \frac{[SI]^2}{S} - (\beta + \gamma) [SI], \\
\frac{d[SS]}{dt} &= -2\beta \kappa \frac{[SS][SI]}{S}.
\end{align*}
\]

**Corollary 5.1.0.4.** Assume the conditions of Theorem 5.1.1 hold for \(\mathcal{G}_1(\psi, N)\), a static graph (i.e. \(\delta_1 = \eta_1 = 0\)). Then, the conclusions of Theorem 5.1.1 hold where \(D(t)\) is the pairwise SIR model of Rand (5.24) if and only if \(\mathcal{G}_1(\psi, N)\) has a univariate Poisson-type degree distribution.

Together, Corollaries 5.1.0.1 and 5.1.0.3 imply that, in the case of a multivariate PT degree distribution on a static graph, the pairwise model (5.22) is equivalent to the edge-based model with multiple modes of transmission (5.13). Likewise, Corollaries 5.1.0.2
and 5.1.0.4 indicate that the pairwise SIR model (5.24) is equivalent to the edge-based SIR model, (5.13) with \( r = 1 \), when the distribution is PT. Also, the edge-based SIR model has previously been shown to be equivalent [51, 84] to a higher dimensional pairwise model of Eames and Keeling [33]. The latter model stratifies the susceptible nodes by degree and, hence, has dimension \( K + 3 \) where \( K \) represents the number of distinct degrees [84]. The model of dimension \( K + 3 \) was derived as an approximation to an earlier well-known model of Eames and Keeling, which is of dimension \( 3K^2/2 + 3K/2 + 1 \) [33]. On the other hand, (5.24) can be reduced to two differential equations. Separation of variables on \( d[SS]/dS \) (see, e.g., [7] for \( \kappa = 1 \)) gives \( [SS] = \alpha_S^{2(1-\kappa)} \mu S^{2\kappa} \). Subsequent inspection of \( d[SI]/dS \) yields a linear differential equation that can be solved. That is, one may express \([SI]\) explicitly as a function of \( S \):

\[
[SI] = \begin{cases} 
\frac{\beta + \gamma}{\beta(1 - \kappa)} S - \mu S^{2\kappa} - \left( \frac{\beta + \gamma}{\beta(1 - \kappa)} \alpha_S^{1-\kappa} - \mu \alpha_S\kappa - \alpha_S^{1-\kappa} \alpha_I \mu \right) S^\kappa, & \kappa \neq 1, \\
\frac{\beta + \gamma}{\beta} S \log(S) - \mu S^2 - \left( \frac{\beta + \gamma}{\beta} \log(\alpha_S) - \mu \alpha_S - \alpha_I \mu \right) S, & \kappa = 1.
\end{cases}
\]

(5.25)

Thus Corollary 5.1.0.4 characterizes a condition on the degree distribution under which the dimension of a pairwise model that is equivalent to the edge-based SIR model has been reduced from \( K + 3 \) to two.

This relationship, we believe, establishes an important (and previously unknown) connection between two “competing” approaches to simplifying the limits of network-based epidemic models on CM random graphs. In particular, it gives conditions under which the pair approximation approach employed frequently (e.g. the model of Scarpino et al [100])
is indeed “correct.” An important follow-up question that we do not consider here is the
discrepancy introduced by application of the pair approximation when this condition does
not hold exactly. One assumes that the impact of this discrepancy is positively related with
the extent to which (5.14) is not constant.

5.2 The SEIdaR process

We extend the SIR-type model of the previous section to accommodate an exposed (but
not infectious) state. As in the case of the stochastic SIR model presented in Chapter 3
the additional exposed state may largely be dealt with through appropriate accounting of
the added compartment and additional notation. Thus we give the proof of Theorem 5.1.1
in the case of the $SIdaR(r,k)$ model for notational simplicity and to more easily relate our
results to the existing literature in which SIR-type models are considered. The proof of
Theorem 5.1.1 is given in Section 5.5. We describe the needed modifications to prove the
analogous $SEIdaR(r,k)$ Law of Large Numbers. In this section, we give the statement of
the analogous limit theorem for the $SEIdaR(r,k)$ model and some insight into the notational
and analytical modifications to make.

Throughout this section of this section, we will continue to use the same notation intro-
duced earlier in this chapter. The epidemic is assumed to take place on a network that is a
realization of the layered Configuration Model $\mathcal{G}_r(\psi,N)$ which defines the contact network.
We define analogous quantities needed for the SEIR-type model. Thus at time $t > 0$ every
node $v \in V$ is assumed to be Susceptible (S), Exposed(E), Infectious(I), or recovered(R).
We continue to assume that there are a total of $r$ layers, labeled such that $j = 1, \ldots, k$ denote
the de-activating layers (i.e. \( \delta_j > 0 \) and \( \eta_j = 0 \)) and the layers \( j = k + 1, \ldots, r \) denote the activating layers (i.e. \( \delta_j = 0 \) and \( \eta_j > 0 \)). Conditional on the history of the process there are now \( 2r + 2 \) events types that can occur: exposure (E) along an edge of any of the \( r \) types, infectiousness (symptom) onset (I), drop (d) of a deactivating edge or activation (a) of an activating edge, symptom onset, and recovery (R). As before all events are assumed to occur according to independent exponential clocks with rates as given in the table below.

<table>
<thead>
<tr>
<th>Rate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_j )</td>
<td>rate of infection along ( j )-edges (( S \xrightarrow{j} E )), ( j = 1, \ldots, r )</td>
</tr>
<tr>
<td>( \delta_j )</td>
<td>rate of deactivation (drop) of ( j )-edges, ( j = 1, \ldots, k )</td>
</tr>
<tr>
<td>( \eta_j )</td>
<td>rate of activation of ( j )-edges, ( j = k + 1, \ldots, r )</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>rate of infectiousness onset (( E \xrightarrow{} I ))</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>rate of recovery (( I \xrightarrow{} R ))</td>
</tr>
</tbody>
</table>

Table 5.3: Listing of possible transitions of the \( SEIdaR(r,k) \) process and corresponding notation.

We denote the set of susceptible nodes at time \( t \geq 0 \) as \( S \) and analogously define the sets of exposed, infectious, and recovered nodes as \( E, I, \) and \( R \) respectively. From the notation in the table we see that the exposed state modification causes a newly exposed node to wait a random period of time which is on average \( \sigma^{-1} \) before becoming infectious. This is introduced to account for the epidemiological evidence that the progression of Ebolavirus follows such a physiological presentation. For any \( u \in S \) let \( X_{SE,u}^j, X_{SI,u}^j \), and \( X_{SS,u}^j \) denote the number of exposed, infectious, and susceptible active \( j \)-neighbors of \( u \). Also, let
$X^j_{SE,u}$, $X^j_{SI,u}$, and $X^j_{SS,u}$ denote the number of inactive exposed, infectious, and susceptible $j$-neighbors of $u$. Similarly for any $u \in E$ let $X^j_{ES,u}$ and $X^j_{ES,u}$ denote the number of susceptible active and deactivated, respectively, $j$-neighbors of $u$. In addition, it is assumed that the activation (or de-activation) of edges occurs only while a node is infectious (i.e. not while the node is susceptible, exposed, or recovered). We will again consider the aggregated dyad quantities as well as the total number of susceptible, exposed, and infectious individuals. We retain the notation that a dyad subscript (e.g. $X_{SE}$) is understood throughout to denote a (row) vector in $\mathbb{R}^r$.

The aggregate variables may be written in terms of the number of connections of individuals the same as before. The number of $j$-edges between susceptible and infectious nodes is $X^j_{SI}$ and given by $X^j_{SI} = \sum_{u \in S} X^j_{SI,u}$. Thus, we let $X(t) = (X_S, X_E, X_I, X_{SE}, X_{SI}, X_{SS}, X_{\bar{SE}})(t)$ denote the state of the aggregate $SEIdaR(r,k)$ stochastic process (with dimension $6r + 3$) at time $t \geq 0$ where $X_S, X_E, X_I$ denote the number of susceptible, exposed, and infectious nodes, respectively. Note that the number of recovered individuals is given by $X_R = N - X_S - X_E - X_I$ and so we ignore the equation for $X_R$ throughout. Similar to before, the first step is in constructing the appropriate Doob-Meyer decomposition. This is achieved in similar fashion to (5.4).

Consider the $SEIdaR(r,k)$ process $X(t)$ on the LCM $\mathcal{G}_r(\psi, N)$ with transitions as outlined in Table 5.3. By the Doob-Meyer decomposition theorem, for this process there is a zero-mean martingale $M(t) = (M_S, M_E, M_I, M_{SE}, M_{SI}, M_{SS}, M_{\bar{SE}}, M_{\bar{SI}}, M_{\bar{SS}})(t)$ such that

$$X(t) = X(0) + \int_0^t \mathcal{F}_X(X(s)) ds + M(t)$$

(5.26)
and the integrable function $F_X(X) = (F_S, F_E, F_{SE}, F_{SI}, F_{SI}, F_{SS}, F_{SS})(X)$ is given by

$$F_S(X_{SI}) = -\sum_{l=1}^{r} \beta_l X_{SI}^l,$$

$$F_E(X_{SI}, X_E) = \sum_{l=1}^{r} \beta_l X_{SI}^l - \sigma X_E,$$

$$F_I(X_E, X_I) = \sigma X_E - \gamma X_I,$$

$$F_{jSE}(X_{SE}, X_{SI}, X_{SS}) = \sum_{i \in S} \sum_{l=1}^{r} \beta_l X_{SI, i}^l (X_{SS, i}^j - X_{SE, i}^j) - \sigma X_{jSE}, \quad j = 1, \ldots, r,$$

$$F_{jSE}(X_{SE}, X_{SI}, X_{SS}) = \sum_{i \in S} \sum_{l=1}^{r} \beta_l X_{SI, i}^l (X_{jSS, i}^j - X_{jSE, i}^j) - \sigma X_{jSE}, \quad j = 1, \ldots, r,$$

$$F_{jSI}(X_{SE}, X_{SI}, X_{SS}) = -\sum_{i \in S} \sum_{l=1}^{r} \beta_l X_{SI, i}^l (X_{jSS, i}^j - X_{jSI, i}^j) + \sigma X_{jSE} - (\beta_j + \gamma + \delta_j) X_{SI}^j + \eta_j X_{SI}^j, \quad j = 1, \ldots, r,$$

$$F_{jSI}(X_{SE}, X_{SI}, X_{SS}) = -\sum_{i \in S} \sum_{l=1}^{r} \beta_l X_{SI, i}^l (X_{jSS, i}^j - X_{jSI, i}^j) + \sigma X_{jSE} - (\gamma + \eta_j) X_{SI}^j + \delta_j X_{SI}^j, \quad j = 1, \ldots, r,$$

$$F_{jSS}(X_{SI}, X_{SS}) = -2 \sum_{i \in S} \sum_{l=1}^{r} \beta_l X_{SI, i}^l X_{jSS, i}^l, \quad j = 1, \ldots, r,$$

$$F_{jSS}(X_{SI}, X_{SS}) = -2 \sum_{i \in S} \sum_{l=1}^{r} \beta_l X_{SI, i}^l X_{jSS, i}^l, \quad j = 1, \ldots, r.$$  

(5.27)

We see that the addition of the exposed state results in slightly more complicated notation and more possible transitions, however, the Doob-Meyer decomposition largely resembles (5.4). The intuition for the above is essentially the same as our prior explanation with the changes introduced by the $E$ compartment. We will require the quantities $X_{SI}(t)$ and $\theta(t)$ to be defined the same as in (5.5) and (5.6). Note that even with the additional $E$ compartment that $\theta(t)$ still retains the same physical interpretation. Hence, $\theta$ can still equivalently be defined in terms of (5.7) and (5.8). The analogous conditions for the system
to behave deterministically are as follows. Since assumptions (I) and (III) are simply conditions on the number of susceptible nodes and the characteristics of the degree distribution these conditions remain unchanged. Condition (II) must be modified slightly such that the initial fraction of exposed individuals, i.e. $\alpha_E$ is defined properly. The modified condition is stated below.

(IIa) The fractions of initially susceptible, infected, and recovered nodes converge, respectively, to some $\alpha_S, \alpha_E, \alpha_I, \alpha_R \in [0, 1]$, i.e.

$$X_S(0)/N \xrightarrow{P} \alpha_S, \quad X_E(0)/N \xrightarrow{P} \alpha_E, \quad X_I(0)/N \xrightarrow{P} \alpha_I, \quad X_R(0)/N \xrightarrow{P} \alpha_R.$$  

Furthermore, $\alpha_S > 0$, $\alpha_E + \alpha_I > 0$, and the initially infected and recovered nodes are chosen randomly.

Condition (IIa) again implies that the initial conditions for the dyads scaled by $N$ converge in probability, i.e.

$$X_{jSE}(0)/N \xrightarrow{P} \alpha_{jSE}, \quad X_{jSE}^{-}(0)/N \xrightarrow{P} \alpha_{jSE}^{-},$$

$$X_{jSI}(0)/N \xrightarrow{P} \alpha_{jSI}, \quad X_{jSI}^{-}(0)/N \xrightarrow{P} \alpha_{jSI}^{-}, \quad j = 1, \ldots, r.$$  

(5.28)

For the deactivating layers $j = 1, \ldots, k$,

$$\alpha_{jSE} = \alpha_S \alpha_E \mu_j, \quad \alpha_{jSE}^{-} = 0, \quad \alpha_{jSI} = \alpha_S \alpha_I \mu_j, \quad \alpha_{jSI}^{-} = \alpha_S \alpha_I \mu_j,$$

$$\alpha_{jSS} = \alpha_S^2 \mu_j, \quad \alpha_{jSS}^{-} = \alpha_S^2 \mu_j,$$

and for the activating layers $j = k + 1, \ldots, r$,

$$\alpha_{jSE} = 0, \quad \alpha_{jSE}^{-} = \alpha_S \alpha_E \mu_j, \quad \alpha_{jSI} = 0, \quad \alpha_{jSI}^{-} = \alpha_S \alpha_I \mu_j,$$

$$\alpha_{jSS} = \alpha_S^2 \mu_j, \quad \alpha_{jSS}^{-} = \alpha_S^2 \mu_j.$$
We will denote the initial condition for the limit of the SEI\(daR(r,k)\) process \(\alpha = (\alpha_S, \alpha_E, \alpha_I, \alpha_{SE}, \alpha_{SI}, \alpha_{SI}, \alpha_{SS}, \alpha_{SS})\).

With this in mind the limiting set of differential equations for the SEI\(daR(r,k)\) process is as follows. Let \((x, x_\Theta) = (x_S, x_E, x_I, x_{SE}, x_{SI}, x_{SI}, x_{SS}, x_\Theta)\) and define \(\mathcal{L}(x, x_\Theta) = (\mathcal{L}_X, \mathcal{L}_\Theta)(x, x_\Theta)\) where \(\mathcal{L}_X = (\mathcal{L}_S, \mathcal{L}_E, \mathcal{L}_I, \mathcal{L}_{SE}, \mathcal{L}_{SI}, \mathcal{L}_{SI}, \mathcal{L}_{SS}, \mathcal{L}_{SS})\) and \(\mathcal{L}_\Theta\) are:

\[
\mathcal{L}_S(x_{SI}) = - \sum_{j=1}^{r} \beta_j x_{SI}^j,
\]

\[
\mathcal{L}_E(x_{SI}, x_E) = \sum_{j=1}^{r} \beta_j x_{SI}^j - \sigma x_E,
\]

\[
\mathcal{L}_I(x_E, x_I) = \sigma x_E - \gamma x_I,
\]

\[
\mathcal{L}_{SE}^j(x_S, x_{SE}, x_{SI}, x_{SS}, x_\Theta) = \sum_{l=1}^{r} \left[ \beta_l \tilde{k}_{jl}(x_\Theta) \frac{x_{SI}^j}{x_S} (x_{SS}^j - x_{SE}^j) \right] - \sigma x_{SE}^j, \quad j = 1, \ldots, r,
\]

\[
\mathcal{L}_{SI}^j(x_S, x_{SE}, x_{SI}, x_{SS}, x_\Theta) = \sum_{l=1}^{r} \left[ \beta_l \tilde{k}_{jl}(x_\Theta) \frac{x_{SI}^j}{x_S} (x_{SS}^j - x_{SE}^j) \right] - \sigma x_{SE}^j, \quad j = 1, \ldots, r,
\]

\[
\mathcal{L}_{SI}^j(x_S, x_{SE}, x_{SI}, x_{SS}, x_\Theta) = - \sum_{l=1}^{r} \beta_l \tilde{k}_{jl}(x_\Theta) \frac{x_{SI}^j x_{SI}^j}{x_S}
\]

\[
\mathcal{L}_{SI}^j(x_S, x_{SE}, x_{SI}, x_{SS}, x_\Theta) = - \sum_{l=1}^{r} \beta_l \tilde{k}_{jl}(x_\Theta) \frac{x_{SI}^j x_{SI}^j}{x_S}
\]

\[
\mathcal{L}_{SI}^j(x_S, x_{SE}, x_{SI}, x_{SS}, x_\Theta) = - \sum_{l=1}^{r} \beta_l \tilde{k}_{jl}(x_\Theta) \frac{x_{SI}^j x_{SI}^j}{x_S}
\]

\[
\mathcal{L}_{SS}^j(x_S, x_{SI}, x_{SS}, x_\Theta) = - 2 \sum_{l=1}^{r} \beta_l \tilde{k}_{jl}(x_\Theta) \frac{x_{SI}^j x_{SI}^j}{x_S}, \quad j = 1, \ldots, r,
\]

\[
\mathcal{L}_{SS}^j(x_S, x_{SI}, x_{SS}, x_\Theta) = - 2 \sum_{l=1}^{r} \beta_l \tilde{k}_{jl}(x_\Theta) \frac{x_{SI}^j x_{SI}^j}{x_S}, \quad j = 1, \ldots, r,
\]

\[
\mathcal{L}_{\Theta}^j(x_{SI}, x_\Theta) = \frac{\beta_j x_{SI}^j}{\alpha_S \delta_j \psi(x_\Theta)}.
\]
Lastly we state the SEIR-modified version of Theorem 5.1.1, which is also similar after including the modifications in this section.

**Theorem 5.2.1 (SEIdaR law of large numbers).** Assume conditions (I),(IIa),(III) for the LCM $\mathcal{G}_r(\psi,N)$. Then, for any $0 < T < \infty$,

$$
\sup_{0 < t \leq T} ||(X(t)/N, \theta(t)) - D(t)|| \overset{P}{\rightarrow} 0
$$

where $D(t) = (D_X, D_\theta)(t)$ is the solution of

$$
D(t) = D(0) + \int_0^t \mathcal{L}(D(s))ds \tag{5.30}
$$

with initial conditions $D_X(0) = \alpha$ and $D_\theta(0) = 1$.

**Proof.** The proof follows closely the argument for Theorem 5.1.1 which is given in Section 5.5. We remark on the slight adaptations needed at the end of Section 5.5.\hfill $\square$

Because of the structure of the resulting system, the ways in which the final limiting system was previously simplified largely still apply. For example, if the degree distribution satisfies (5.14) then the quantity $\theta$ still decouples in the limiting set of equations and the result is equivalent with the appropriate pairwise model.

To illustrate this point, we consider the example of the $SEIdaR(1,0)$ model under the assumption that the PT-condition is satisfied with $r = 1$ de-activating layer in which connections are dropped at rate $\delta$. By application of the pair approximation the final system is equivalent to:
\[
\frac{dS}{dt} = -\beta [SI] \\
\frac{dE}{dt} = \beta [SI] - \sigma E \\
\frac{dI}{dt} = \sigma E - \gamma I \\
\frac{d[SS]}{dt} = -2\beta \kappa \frac{[SS][SI]}{S} \\
\frac{d[SE]}{dt} = \beta \kappa \frac{[SS][SI]}{S} - \sigma [SE] - \beta \kappa \frac{[SE][SI]}{S} \\
\frac{d[SI]}{dt} = \sigma [SE] - (\beta + \gamma + \delta)[SI] - \beta \kappa \frac{[SI]^2}{S}.
\] (5.31)

The above system may be simplified further by eliminating the [SS] equation through separation of variables on [SS] and S. However, its inclusion does no harm. The limiting system of equations may be further derived for cases where \( r > 1 \) and such a model may be of interest when more specific data is available. As an example, an “ideal” model of Ebola may contain \( r = 3 \) layers corresponding to the sociological contexts (community, healthcare, funeral) in which transmission is primarily driven. For practical purposes, obtaining reliable data from which to estimate layer-specific model parameters (\( \mu_j, \beta_j \), etc.) remains a significant challenge. Our numerical examples presented next will thus focus on the \( r = 1 \) layer case with the assurance that there is significant epidemiological evidence that community-type contacts have been observed as the primary driver of transmission in Guinea and elsewhere. In addition, the evidence of whether the type of contact confers asymmetric risk of infection is inconclusive [31].
5.3 The hybrid process

The analysis presented so far also motivates an alternative stochastic model to the network-based ones to this point. That is, upon deriving the limiting set of equations that describe the evolution of the epidemic it is generally speaking possible to derive an analogous jump process with equivalent limiting dynamics. The attractiveness of this is twofold. First, the limiting system of equations governing population-level dynamics will match the network based model from which we derive the “approximate” process. Second, counting processes (i.e. including but not limited to DDMJP) are a common model in describing biochemical systems and therefore one may harness existing estimation results. This benefit of this is described in greater detail in chapter 7 but for now we derive an approximate process to the $SEIdaR(1,0)$ model.

The target limiting set of equations for this process is given in (5.31). We then construct an approximate Markov Jump process, $X^*(t) = (S^*, E^*, I^*)(t)$ on $\mathbb{Z}^3$, with jumps and transition rates defined as in the table below.

<table>
<thead>
<tr>
<th>Event</th>
<th>Rate</th>
<th>Transition</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Exposure</td>
<td>$\lambda_1^N(t; \theta) = N \beta \tilde{D}_{SI}(t; \theta)$</td>
<td>$\Delta X^* = (-1, 1, 0)'$</td>
</tr>
<tr>
<td>Symptom Onset</td>
<td>$\lambda_2^N(t; \theta) = \sigma E(t)$</td>
<td>$\Delta X^* = (0, -1, 1)'$</td>
</tr>
<tr>
<td>Recovery</td>
<td>$\lambda_3^N(t; \theta) = \gamma I(t)$</td>
<td>$\Delta X^* = (0, 0, -1)'$</td>
</tr>
</tbody>
</table>
And the quantity $D_{SI}(t; \theta)$ is computed as a forward-Euler approximation to the limiting $N$-normalized $S - I$ quantity as follows. Let $0 = t_0 < t_1 < \ldots < t_m \leq t$ denote the successive history of event times up to time $t$ and $S(t_i)$ the empirical fraction of susceptible individuals at time $t_i$. Let $0 = t_0 < t_1 < \ldots < t_m \leq t$ denote the successive history of event times up to time $t$ and $S(t_i)$ the empirical fraction of susceptible individuals at time $t_i$. Let $(\bar{D}_{SE}, \bar{D}_{SI}, \bar{D}_{SS})(0) = (\alpha_S \alpha_E \mu, \alpha_S \alpha_I \mu, \alpha_S^2 \mu)$ in agreement with the initial condition of $(D_{SE}, D_{SI}, D_{SS})(0)$ in Theorem 5.2.1 for the $SEIdaR(1,0)$ model with contact drop rate $\delta$. Then starting at $t_0$, the quantities $(\bar{D}_{SE}, \bar{D}_{SI}, \bar{D}_{SS})(t_i)$ for $i = 0, \ldots, m - 1$ are computed as the forward-Euler approximation to the limiting system as:

\[
\begin{align*}
\bar{D}_{SE}(t_{i+1}; \theta) &= \bar{D}_{SE}(t_i) + (t_{i+1} - t_i)\left(\beta \kappa \frac{D_{SS}(t_i)\bar{D}_{SI}(t_i)}{S(t_i)} - \sigma \bar{D}_{SE}(t_i) - \beta \kappa \frac{D_{SE}(t_i)\bar{D}_{SI}(t_i)}{S(t_i)} \right) \\
\bar{D}_{SI}(t_{i+1}; \theta) &= \bar{D}_{SI}(t_i) + (t_{i+1} - t_i)\left(\sigma \bar{D}_{SE}(t_i) - (\beta + \gamma + \delta)\bar{D}_{SI}(t_i) - \beta \kappa \frac{D_{SS}(t_i)\bar{D}_{SI}(t_i)}{S(t_i)} \right) \\
\bar{D}_{SS}(t_{i+1}; \theta) &= \bar{D}_{SS}(t_i) + (t_{i+1} - t_i)\left(-2\beta \kappa \frac{D_{SS}(t_i)\bar{D}_{SI}(t_i)}{S(t_i)} \right)
\end{align*}
\]

(5.32)

This process evolves as a stochastic process on the state space $Z^3$ and whose limit satisfies an equivalent set of differential equations to the $SEIdaR(1,0)$ process under the PT condition given in Theorem 5.2.1. The process is Markovian because the transition rates listed in Table 5.4 are constant between jumps and the update of these rates depends only on the current value of the process through $S(t)$. The equivalence with the system of differential equations in (5.31) is stated in the following corollary.

**Corollary 5.3.0.5.** Define the hybrid process process $X^* = (S^*, E^*, I^*)(t)$ with initial condition $N^{-1}X^*(0) \to (\alpha_S, \alpha_E, \alpha_I)$ as $N \to \infty$ and $(\bar{D}_{SE}, \bar{D}_{SI}, \bar{D}_{SS})(0) = (\alpha_S \alpha_E \mu, \alpha_S \alpha_I \mu, \alpha_S^2 \mu)$. Also assume the conditions for Theorem 5.2.1 on $\mathcal{G}_1(\psi, N)$ where $\psi$ satisfies the univariate
PT-condition. Then the N-normalized process $N^{-1}X^*(t)$ converges uniformly in probability to $(D_S, D_E, D_I)(t)$ where $D(t)$ denotes the solution given in Theorem 5.2.1 on any finite interval $[0,T]$ as $N \to \infty$.

The proof of the above follows as a consequence of an argument similar to the one presented in Chapter 3 for Theorem 3.1.1 and by our construction of the process. We may also construct an approximate process when the degree distribution does not satisfy the PT-condition, however, given that the Negative Binomial distribution is appropriate we do not consider it for the time being. The motivation for this approximate stochastic process revolves around computation of the likelihood function which we describe more completely in Chapter 7.

5.4 Conclusion

In this chapter we present a law of large numbers for a complex epidemic process evolving on a Layered Configuration Model random graph. The model is able to address the dynamics of Ebolavirus described in several places [107] [31] [49] and the limiting system of equations rigorously studied [52]. This result generalizes previous work by Janson et al. and Decreusefond et al. [53] [30] by allowing multiple sources of transmission and the structure of the network to change over time. We believe the approach is sufficiently flexible to derive results not considered here (for example, non-Markov infectious period). We reveal a new relationship between edge-based systems and the limiting equations obtained by applying the pair approximation of Rand [97]. In particular, our results yield conditions
under which this approximation is “correct”, i.e. the system obtained is equivalent. Lastly, we consider an \( S - E - I - R \)-type modification of this model and derive the corresponding limiting system. A corresponding Markov Jump Process is further introduced which satisfies the same limiting set of equations and will play a role in our subsequent analysis.

### 5.5 Proof of the \( SIdaR(r, k) \) law of large numbers

We provide the proof of two useful lemmas and then the proof of Theorem 5.1.1. We then describe the modifications needed to prove Theorem 5.2.1 in the same manner.

As alluded previously in describing (5.4), to derive the limit for 5.1.1 we are required to describe the neighborhood (in terms of infectious status) of a given susceptible node. This distribution is summarized in Remark 5.5.0.1 and states that in a finite graph, the neighborhood of a susceptible node may be described by a multivariate hypergeometric distribution. Asymptotically, this distribution approaches a particular multinomial distribution that may be described in terms of \( \theta \), essentially verifying a conjecture of Miller and Volz [85] about the asymptotic distribution of this neighborhood. In all, our final method of proof via Gronwall’s Inequality somewhat follows the style of Theorem 3.1.1, however, it is seen that several aspects of this require more technical arguments.

Lemma 3 shows that \( X_{S^k} \) and \( X_{S^*} \) can be expressed in the limit as functions of \( \theta \) given by (5.6). Lemma 4 shows that the dynamics of the \( N \)-scaled process on the finite graph converges (in probability) to the dynamics described by the ODE system (5.11) involving \( \theta \). Theorem 5.1.1 then follows from Lemma 4 as an application of Doob’s martingale and Gronwall’s inequalities, much the same as in Theorem 3.1.1.
For convenience we will take operations on vectors such as multiplication, division, integration and ordering to be coordinatewise. We first provide an important remark about the neighborhood of a susceptible node of degree \( k \) in layer \( j \), i.e.
\[(X_{S_{k},i}^{j}(t), X_{S_{k},i}^{j}(t), X_{S_{k},i}^{j}(t), X_{S_{k},i}^{j}(t)) \text{ for } i \in S_{k}.
\]
The distribution of the neighborhood is critical when we consider the expectations of
\[Q^{j,(k)}_{i} = X_{S_{k},i}^{j}(X_{S_{k},i}^{j} - X_{S_{k},i}^{j}), \quad \tilde{Q}^{j,(k)}_{i} = X_{S_{k},i}^{j}(X_{S_{k},i}^{j} - X_{S_{k},i}^{j}),
\]
which are mixed moments with respect to the neighborhood distribution, as is clearly required by (5.4).

Recall that we consider the evolution of the \( SI_{d}AR(r,k) \) process on a realization of the LCM random graph that has been generated by time \( t = 0 \). One may also consider a construction where the network is revealed dynamically as infections occur, as was used in [30, 53]. Under this equivalent construction, all half-edges of a susceptible node \( i \in S_{k} \) remain unpaired until infection occurs. However, one could force the pairing off all unpaired edges at time \( t > 0 \), uniformly at random, in order to derive the neighborhood of some susceptible node \( i \). The following remark is perhaps most easily understood by keeping this equivalent construction in mind and recalling the definition of the probability space in Section 5.1.

**Remark 5.5.0.1.** For \( k_{j} \geq 1 \) and \( i \in S_{k} \), and conditionally on the process history up to time \( t \), the vector
\[(X_{S_{k},i}^{j}(t), X_{S_{k},i}^{j}(t), X_{S_{k},i}^{j}(t), X_{S_{k},i}^{j}(t)) \text{ follows the multivariate hypergeometric distribution with}
\]
probability mass function below:
\[
P(\mathbf{X}_{SI,i}^j(t) = n_{SI}^j, \mathbf{X}_{SI,i}^j(t) = n_{SI}^j, \mathbf{X}_{SS,i}^j(t) = n_{SS}^j, \mathbf{X}_{SS,i}^j(t) = n_{SS}^j) = \frac{\binom{n_{SI}^j(t)}{n_{SI}^j} \binom{n_{SS}^j(t)}{n_{SS}^j} \binom{n_{SI}^j(t)}{n_{SI}^j} \binom{n_{SS}^j(t)}{n_{SS}^j} \binom{n_{SI}^j(t) - X_{SI,i}^j(t)}{n_{SI}^j - n_{SI}^j} \binom{n_{SS}^j(t) - X_{SS,i}^j(t)}{n_{SS}^j - n_{SS}^j} \binom{n_{SI}^j(t) - X_{SI,i}^j(t)}{n_{SI}^j - n_{SI}^j} \binom{n_{SS}^j(t) - X_{SS,i}^j(t)}{n_{SS}^j - n_{SS}^j}}{\binom{k_j}{k_j}},
\]
\[\text{(5.33)}\]
supported on the four-simplex \(0 \leq n_{SI}^j + n_{SI}^j + n_{SS}^j + n_{SS}^j \leq k_j\). Essentially, this states that the \(j\)-type neighborhood of \(i\) is a multivariate hypergeometric distribution that depends on the number of “free” \(j\)-type edges connecting to susceptible, infectious, and recovered nodes.

This implies
\[
E[\mathbf{X}_{SI,i}^j] = k_j \frac{n_{SI}^j}{X_{SI}^j}
\]
where the expectation is taken with respect to (5.33). Also, based on the LCM definition the neighborhoods of \(i\) (conditional on the number of stubs he has) are distinct layers are independent, i.e.
\[
P(\mathbf{X}_{SI,i}^l(t) = n_{SI}, \mathbf{X}_{SS,i}^l(t) = n_{SS}) = \prod_{l=1}^{r} P(\mathbf{X}_{SI,i}^l(t) = n_{SI}^l, \mathbf{X}_{SS,i}^l(t) = n_{SS}^l).
\]

Hence, the mixed moments are given as
\[
E \left[ Q_{i}^{l,(k)} \right] = k_j k_l \frac{X_{SI}^l(t)}{X_{SI}^l(t)} \left( \frac{X_{SS}^l(t)}{X_{SS}^l(t)} - \frac{X_{SI}^l(t)}{X_{SI}^l(t)} \right),
\]
\[\text{(5.34)}\]
for \(l \neq j\) and
\[
E \left[ Q_{i}^{jj,(k)} \right] = k_j (k_j - 1) \frac{X_{SI}^j(t)}{X_{SI}^j(t)} \left( \frac{X_{SS}^j(t)}{X_{SS}^j(t)} - \frac{X_{SI}^j(t) - 1}{X_{SI}^j(t) - 1} - k_j \frac{X_{SI}^j(t)}{X_{SI}^j(t)} \right).
\]
\[\text{(5.35)}\]
Likewise, for any \(1 \leq j, l \leq r\),
\[
E \left[ \tilde{Q}_{i}^{l,(k)} \right] = k_j k_l \frac{X_{SI}^l(t)}{X_{SI}^l(t)} \left( \frac{X_{SS}^l(t)}{X_{SS}^l(t)} - \frac{X_{SI}^l(t)}{X_{SI}^l(t)} \right).
\]
We now prove the first lemma which makes use of the above. This lemma shows that $X_{S_k}$ and $X_{S^*}$ can be expressed in the limit as functions of $\theta$ as given by (5.6).

**Lemma 3.** Recall the definition of $X_{S_k}$ and $X_{S^*}$ given in (5.5). In addition, assume condition (I) and $\sum_k ||k||p_k < \infty$. Then

(a) $\sup_{0 < t \leq T} |N^{-1}X_{S_k} - \alpha_S p_k \theta^k| \xrightarrow{P} 0$ for any $k \geq 0$, and

(b) $\sup_{0 < t \leq T} ||N^{-1}X_{S^*}(t) - \alpha_S \theta(\psi(\theta(t)))|| \xrightarrow{P} 0$.

**Proof.** (a). Note $X_{S_k}(t) = \sum_{i=1}^N Z_i(t)$ where $Z_i(t) \in \{0, 1\}$ denotes the indicator process of whether node $i$ is of degree $k$ and susceptible at time $t > 0$. Recall from Remark 5.5.0.1 that $EX_{S_k} = k_j X_{S_k} / X_{S^*}$ for $i \in S_k$. We claim that

$$EZ_i(t) = P(i \in S_k(t)) = N^{-1}S(0)p_k \theta^k.$$  

(5.36)

Indeed, $EZ_i(t) = P(Z_i(t) = 1) = P(i \in S_k(t)) = P(i \in S_k(t) | i \in S_k(0))P(i \in S_k(0))$ where $P(i \in S_k(0)) = p_k N^{-1}S(0)$ by Assumption II and

$$P(i \in S_k(t) | i \in S_k(0)) = \exp \left( -\sum_{j=1}^r \beta_j \int_0^t E \left[ X_{S_{j,i}}(s) \right] ds \right)$$

$$= \exp \left( -\sum_{j=1}^r \beta_j \int_0^t k_j \frac{X_{S_{j,i}}(s)}{X_{S^*}(s)} ds \right) = \theta^k.$$
Equation (5.36) then implies that \( \{X_{Sk}(t) - S(0)p_k \theta^k(t)\}_{t \geq 0} \) is a càdlàg martingale process with mean zero and finite variation. By the triangle inequality,

\[
P \left( \sup_{0 < t \leq T} \left| N^{-1}X_{Sk}(t) - \alpha_S p_k \theta^k(t) \right| > \varepsilon \right) \leq P \left( \sup_{0 < t \leq T} \left| N^{-1}X_{Sk}(t) - N^{-1}S(0)p_k \theta^k(t) \right| > \frac{\varepsilon}{2} \right) + P \left( \sup_{0 < t \leq T} \left| N^{-1}S(0)p_k \theta^k(t) - \alpha_S p_k \theta^k(t) \right| > \frac{\varepsilon}{2} \right).
\]

The second term tends to zero by assumption Assumption II and for the first term we have, by Doob’s martingale inequality,

\[
P \left( \sup_{0 < t \leq T} \left| N^{-1}X_{Sk}(t) - N^{-1}S(0)p_k \theta^k(t) \right| > \frac{\varepsilon}{2} \right) \leq \left( \frac{\varepsilon}{2} \right)^{-2} N^{-2} \text{Var} X_{Sk}(T).
\]

Since there are at most \( N \) jumps for \( X_{Sk} \) and each is of size one, it follows that the quadratic variation of \( X_{Sk} = O(N) \), which gives the needed result.

\( (b) \) By equation (5.5), we may write

\[
N^{-1}X_{S*}(t) - \alpha_S \theta \partial \psi(\theta(t)) = \sum_k kN^{-1}X_{Sk}(t) - \alpha_S \sum_k k p_k \theta^k(t).
\]

Consider arbitrarily large \( L \) and \( \varepsilon > 0 \). By Markov’s inequality, we have

\[
P \left( \sup_{N \geq 1} \sup_{0 < t \leq T} \left\| \sum_k kN^{-1}X_{Sk}(t) - \alpha_S k p_k \theta^k(t) \right\| > \varepsilon \right) \leq \varepsilon^{-1} \sum_{|k| > L} |k| \sup_{0 < t \leq T} \left| N^{-1}X_{Sk}(t) - \alpha_S p_k \theta^k(t) \right| \leq 3 \varepsilon^{-1} \sum_{|k| > L} |k| |p_k|
\]

106
since EXS_k/n ≤ 2p_k for N sufficiently large and α_S ≤ 1, therefore we may apply the Monotone Convergence Theorem. So the tail of the sum is negligible since L is arbitrary and, by assumption, \( \sum_k ||k||p_k < \infty \). The result then follows since in (a) we showed convergence for fixed \( k \).

Before the next lemma we remark on the boundedness of a few variables under consideration.

**Remark 5.5.0.2.** We note that \( \alpha_S \theta \partial \psi(\theta) \leq \partial \psi(1) \) and for sufficiently large \( N, N^{-1}X_{SI} \leq N^{-1}X_{\bullet} \leq 2\partial \psi(1) \) (with the same holding for \( N^{-1}X_{SS} \)). By Assumption (I), \( N^{-1}X_{\bullet} \) is bounded away from 0 on \([0,T]\) for finite \( T \). By definition of \( \theta \) this implies \( \theta \) is also bounded away from 0. Furthermore, by Lemma 3b, we can take the same lower bound for \( \alpha_S \theta \partial \psi(\theta) \). Let \( \xi > 0 \) be a uniform lower bound for \( N^{-1}X_{\bullet}, \theta \) and \( \alpha_S \theta \partial \psi(\theta) \). We will use the notation \( [\xi, 2\partial \psi(1)]^r := [\xi_1, 2\partial_1 \psi(1)] \times \ldots \times [\xi_r, 2\partial_r \psi(1)] \) and hence may write \( X_{\bullet}/N \in [\xi, 2\partial \psi(1)]^r \).

Let \( F_X \) and \( F_\Theta \) be defined as in (5.4) and (5.8) and \( H \) as in (5.11). Define \( \Delta(t) = (\Delta_X, \Delta_\Theta)(t) \) where \( \Delta_X = (\Delta_S, \Delta_I, \Delta_{SI}, \Delta_{SS}, \Delta_{SS}) \) and \( \Delta_\Theta \) are given by

\[
\Delta_X(t) = N^{-1}F_X(X(t)) - H_X(N^{-1}X(t), \theta(t))
\]

(5.37)

and

\[
\Delta_\Theta(t) = F_\Theta(X_{SI}(t), X_{\bullet}(t), \theta(t)) - H_\Theta(N^{-1}X_{SI}(t), \theta(t)).
\]

(5.38)

Lemma 4 shows that \( \Delta_X \) and \( \Delta_\Theta \) converge to zero uniformly in probability. The convergence of \( \Delta_\Theta \) to zero follows almost directly from Lemma 3. The convergence of \( \Delta_X \) to
zero is more subtle and involves comparing the empirical (network-wide) hypergeometric mixed moments, i.e. moments of the form $\sum_{i \in S_k} Q_{i}^{l,k}$ and $\sum_{i \in S_k} \tilde{Q}_{i}^{l,k}$ to the limiting moments of a multinomial distribution whose probabilities are expressed in terms of $\theta$.

This phenomenon is first conjectured by Volz in [110] and generalized in [85] and so a major consequence of our proof is in verifying this (in a somewhat general case than a static SIR model). We will use the fact that as $N \to \infty$, the hypergeometric mixed moments are approximately multinomial and we can replace $X_{S*}$ with a function of $\theta$ by Lemma 3.

Therefore, it will be convenient to define the following compensators.

Let $C_{h}^{j,k}(t) : [0, T] \to \mathbb{R}$ be given by

$$C_{h}^{j,k}(t) = k_l k_j X_{Sj}^l(t) \left( \frac{X_{SJ}^j(t)}{X_{S*}^j(t)} - \frac{X_{SJ}^l(t)}{X_{S*}^l(t)} \right), \quad l \neq j,$$

and

$$C_{h}^{j,j}(t) = k_j (k_j - 1) \frac{X_{SJ}^j(t)}{X_{S*}^j(t)} \left( \frac{X_{SS}^j(t)}{X_{S*}^j(t)} - 1 \right) \frac{X_{SJ}^j(t) - 1}{X_{S*}^j(t) - 1} - k_j \frac{X_{SJ}^j(t)}{X_{S*}^j(t)},$$

so that the hypergeometric mixed moment in equation (5.34) is given by

$$E \left[ Q_{i}^{j,k} \right] = C_{h}^{j,k}(t).$$

The function and notation are explained as follows. The $C_{h}$ function represents the mean (as a function of time) hypergeometric mixed moment term in layers $j,l$ of a node of multi-degree $k$. It is useful to define the related function for the multinomial distribution, $C_{m}^{j,k} : [0, T] \times [\xi, 2\partial \psi(1)^{r}] \to \mathbb{R}$, which is given by

$$C_{m}^{j,k}(t, z(t)) = k_l k_j \frac{N - 2 X_{Sj}^l(t)}{z_l(t)} \left( \frac{X_{SS}^j(t)}{z_j(t)} - \frac{X_{SJ}^l(t)}{z_j(t)} \right), \quad l \neq j.$$
and

\[ C_{m}^{j,l,(k)}(t,z(t)) = k_j(k_j - 1) N^{-2}X_{Sl}^{j} \left( \frac{X_{Sl}^{j}}{z_j(t)} - \frac{X_{Sl}^{j}}{z_j(t)} \right) - k_j N^{-1}X_{Sl}^{j}. \] (5.43)

That is, if \( N^{-1}X_{Sl}^{j}/z_j(t) \) and \( N^{-1}X_{SS}^{j}/z_j(t) \) are considered as probabilities, then \( C_{m}^{j,l,(k)}(t,z(t)) \) represents a mixed moment with respect to the multinomial distribution with these corresponding probabilities. Observe that there exists \( L > 0 \) such that

\[ C_{m}^{j,l,(k)} \leq L ||k||^2 \] (5.44)

for any \( j,l \) (and uniformly in \( N \)) since the domain of \( z \) is bounded away from 0 and \( N^{-1}X_{Sl}, N^{-1}X_{SS} \) are uniformly bounded above by Remark 5.5.0.2. It also follows that \( C_{m}^{j,l,(k)} \) is Lipschitz continuous in \( z \).

**Lemma 4.** Under conditions (I) – (III),

1. \( \sup_{0 < t \leq T} ||\Delta_\Theta(t)|| \xrightarrow{P} 0 \), and

2. \( \sup_{0 < t \leq T} ||\Delta_X(t)|| \xrightarrow{P} 0 \).

**Proof.** (a). Define \( \mathcal{J} : [0,T] \times [\xi, 2\partial\psi(1)]^r \rightarrow \mathbb{R}^r \) given by

\[ \mathcal{J}(t,z(t)) = -\beta \theta(t) N^{-1}X_{Sl}(t)/z(t) \].

By Remark 5.5.0.2 and (I), \( (t,N^{-1}X_{Sl}) \) and \( (t,\alpha_S\theta \partial\psi(\theta)) \) are in the domain of \( \mathcal{J} \) for \( t \in [0, T] \). By definition of \( \Delta_\Theta \) in (5.38),

\[ \Delta_\Theta(t) = -\beta \frac{X_{Sl}(t)\theta(t)}{X_{Sl}(t)} + \beta \frac{N^{-1}X_{St}(t)}{\alpha_S \partial\psi(\theta(t))} = \mathcal{J}(t,N^{-1}X_{Sl}(t)) - \mathcal{J}(t,\alpha_S \theta(t) \partial\psi(\theta(t))). \]

Since \( \mathcal{J} \) is Lipschitz continuous in \( z \),

\[ \sup_{0 < t \leq T} ||\Delta_\Theta(t)|| = \sup_{0 < t \leq T} ||\mathcal{J}(t,N^{-1}X_{Sl}(t)) - \mathcal{J}(t,\alpha_S \theta(t) \partial\psi(\theta(t)))|| \]

\[ \leq L_1 \sup_{0 < t \leq T} ||N^{-1}X_{Sl}(t) - \alpha_S \theta(t) \partial\psi(\theta(t))|| \].
for some $L_1 > 0$. The result then follows from Lemma 3b since (III) implies $\sum_k ||k||p_k < \infty$.

(b) We note that, by definition in (5.37), $\Delta_S = \Delta_I = 0$. We will show that $\sup_{0 < t \leq T} ||\Delta_{SI}(t)|| \overset{P}{\to} 0$ and one may argue similarly for $\Delta_{SL}$, $\Delta_{SS}$ and $\Delta_{SS}$ which together imply $\sup_{0 < t \leq T} ||\Delta_X(t)|| \overset{P}{\to} 0$. First, observe $\Delta_{SI} = (\Delta_{SI}^1, \ldots, \Delta_{SI}^r)$ and so it suffices to show

$$
\sup_{0 < t \leq T} |\Delta_{SI}^j(t)| \overset{P}{\to} 0
$$

for some $j = 1, \ldots, r$.

Let $1 \leq j \leq r$. We can rewrite $\Delta_{SI}^j$ as

$$
\Delta_{SI}^j = N^{-1} \mathcal{H}^j_{SI}(X_{SI}, X_{SS}) - \mathcal{H}^j_{SI}(N^{-1}X_{SI}, N^{-1}X_{SS}, \theta)
$$

$$
= N^{-1} \sum_{i \in S} \left( \sum_{l=1}^r \beta_l X_{SL,i}^l (X_{SS,i}^j - X_{SI,i}^j) \right) - \sum_{l=1}^r \left[ \beta_l N^{-2} X_{SI}^j (X_{SS}^j - X_{SI}^j) \frac{\partial^2 \psi(\theta)}{\alpha_S \theta_j \theta \partial_j \psi(\theta) \partial_l \psi(\theta)} \right]
$$

$$
+ \beta_j N^{-1} X_{SI}
$$

$$
= \sum_k N^{-1} \sum_{i \in S_k} \left( \sum_{l=1}^r \beta_l X_{SL,i}^l (X_{SS,i}^j - X_{SI,i}^j) \right)
$$

$$
- \sum_{k \neq j} \beta_l N^{-2} X_{SI}^j \left( X_{SS}^j - X_{SI}^j \right) \frac{k_j k_l p_k \theta^k}{\alpha_S \theta_j \theta \partial_j \psi(\theta) \partial_l \psi(\theta)}
$$

$$
+ \sum_k \beta_j N^{-1} X_{SI}^j \left( X_{SS}^j - X_{SI}^j \right) \frac{k_j (k_j - 1) p_k \theta^k}{\alpha_S \theta_j \theta \partial_j \psi(\theta)^2} + \beta_j N^{-1} X_{SI}^j \frac{k_j p_k \theta^k}{\theta_j \theta \partial_j \psi(\theta)}
$$

$$
= \sum_k \left[ N^{-1} \sum_{i \in S_k} \left( \sum_{l=1}^r \beta_l X_{SL,i}^l (X_{SS,i}^j - X_{SI,i}^j) \right) - \sum_{l=1}^r \beta_l \alpha_S p_k \theta^k C_{m}^{j,(k)}(t, \alpha_S \theta \partial \psi(\theta)) \right].
$$

Thus, define

$$
\Delta_{SI}^{j,(k)}(t) = N^{-1} \sum_{i \in S_k} \left( \beta_l X_{SL,i}^l (X_{SS,i}^j - X_{SI,i}^j) \right) - \beta_l \alpha_S p_k \theta^k C_{m}^{j,(k)}(t, \alpha_S \theta \partial \psi(\theta)).
$$
so that $\Delta_{SI}^j = \sum_{l=1}^r \sum_k \Delta_{SI}^{jl,(k)}$. Ignoring the $\beta_l$ term, note that the left hand term above is, by design, the network-wide empirical moment for layers $j, l$ and the term on the right is the desired multinomial moment. Hence, it suffices to show

$$\sup_{0 < t \leq T} \left| \sum_k \Delta_{SI}^{jl,(k)}(t) \right| \xrightarrow{P} 0 \quad \text{for } j, l = 1, \ldots, r. \quad (5.46)$$

This is done in what follows in two separate steps. Consider an arbitrary pair $(j, l)$ $1 \leq j, l \leq r$. We first derive a bound for the contribution to this term from the tail of the degree distribution. That is, as $M \to \infty$

$$\sup_{N} \sup_{0 < t \leq T} \left| \sum_{||k|| > M} \Delta_{SI}^{jl,(k)}(t) \right| \xrightarrow{P} 0. \quad (5.47)$$

To this end, observe that

$$\left| N^{-1} \sum_{||k|| > M} \sum_{i \in S_k} \beta_l X_{SI,i}^j(t)(X_{S_S,i}^j(t) - X_{SI,i}^j(t)) \right| \leq N^{-1} \sum_{||k|| > M} \sum_{i \in S_k} |\beta_l k_l| \leq N^{-1} \beta_l C' \sum_{||k|| > M} ||k||^2 X_{S_k}(t) \leq 2 \beta_l C' \sum_{||k|| > M} ||k||^2 p_k \quad (5.48)$$

for some $C' > 0$ since $X_{S_k}/N \leq 2 p_k$ for $N$ sufficiently large. From the bound on $C_m^{jl,(k)}$ in (5.44), we also have

$$\left| \sum_{||k|| > M} \beta_l \alpha_S p_k \theta^k C_m^{jl,(k)}(t, \alpha_S \theta \partial \psi(\theta)) \right| \leq \beta_l L \sum_{||k|| > M} ||k||^2 p_k. \quad (5.49)$$

Then (5.47) follows from (5.48) and (5.49) together with (III).
Next we show that \( \sup_{0 < t \leq T} \left| \Delta_{SI}^{jl,(k)}(t) \right| \xrightarrow{P} 0 \) for any fixed \( k \). We write

\[
|\Delta_{SI}^{jl,(k)}(t)| = \left| N^{-1} \sum_{i \in S_k} \left( \beta_l X_{SI,i}^j (X_{SS,i}^j - X_{SI,i}^j) \right) - \beta_l \alpha_S p_k \theta^k C_m^{jl,(k)} (t, \alpha_S \theta \partial \psi(\theta)) \right|
\]

\[
\leq N^{-1} \beta_l \left| \sum_{i \in S_k} Q_i^{jl,(k)} - X_{SI}^{jl,(k)} (t) \right| + N^{-1} \beta_l X_S \left| C_h^{jl,(k)} (t) - C_h^{jl,(k)} (t, N^{-1} X_{S*}) \right| + \beta_l N^{-1} X_S C_m^{jl,(k)} (t, N^{-1} X_{S*}) - \alpha_S p_k \theta^k C_m^{jl,(k)} (t, N^{-1} X_{S*}) + \beta_l \alpha_S p_k \theta^k \left| C_m^{jl,(k)} (t, N^{-1} X_{S*}) - C_m^{jl,(k)} (t, \alpha_S \theta \partial \psi(\theta)) \right|
\]

(5.50)

where the inequality is obtained by the triangle inequality and addition/subtraction of the terms \( C_h^{jl,(k)} (t) \) and \( C_m^{jl,(k)} (t, N^{-1} X_{S*}) \). We now show that each of these terms tends to zero uniformly in probability.

By Remark 5.5.0.1 and equation (5.41), the process \( M^{jl,(k)}_h(t) = \sum_{i \in S_k} Q_i^{jl,(k)} - X_{SI}^{jl,(k)} (t) \) is a zero-mean, piecewise-constant càdlàg martingale that jumps only if infection/recovery of a node of degree \( k \) (or a neighbor of a node of degree \( k \)) occurs or activation/drop of a \( j \)-edge or \( l \)-edge belonging to a node of degree \( k \) occurs. Recall that, for each layer, either activation or drops are possible, not both. Consider events impacting a node \( u \in S_k \). For infection or recovery events, of which there are at most \( 2(1 + k_j + k_l) \) corresponding to infection and recovery of \( u \) itself or one of its \( j \)-or \( l \)-neighbors), the jump size is at most \( k_j k_l \). For deactivation and activation events of an \( l \) or \( j \)-edge, of which there are at most \( k_l + k_j \) affecting \( u \), the jump size is also at most \( k_j k_l \). Recall that the number of nodes of degree \( k \) is approximately \( N p_k \) for large \( N \). The quadratic variation of \( M^{jl,(k)}_h(t) \) is the sum of its squared jumps (see e.g. Chapter 9 in [7])
and thus satisfies

\[ [M_{hl}^{j,l}(k)](t) = \sum_{s \leq t} (\delta M_{hl}^{j,l}(s))^2 \leq 2Np_k(1+k_j+k_t)(k_jk_t)^2 + Np_k(k_t+k_j)(k_jk_t)^2 \leq L_4||k||^2N \]

for \(0 < t \leq T < \infty\) and some \(L_4 > 0\). Since \(E[M_{hl}^{j,l}(k)](t) = E(M_{hl}^{j,l}(k)(t))^2 = O(N)\), Doob's martingale inequality implies

\[
\sup_{0 < t \leq T} \left| N^{-1}M_{hl}^{j,l}(k)(t) \right| \xrightarrow{P} 0,
\]

i.e. the term in (5.50) tends to zero uniformly in probability.

In consideration of the term in (5.51), note that \(N^{-1}X_{S_k} \leq 1\) and

\[ C_{hl}^{j,l}(k) = C_m^{j,l}(k, N^{-1}X_{S_k}) \]

for \(j \neq l\). For the case \(l = j\), we have

\[
\begin{align*}
&\sup_{0 < t \leq T} \left| C_{hl}^{j,l}(k)(t) - C_m^{j,l}(k, N^{-1}X_{S_k})(t) \right| \\
&= \sup_{0 < t \leq T} \left| k_j(k_j - 1) \frac{X_{S_l}^j(t)}{X_{S_k}^j(t)} \left( \frac{X_{S_k}^j(t) + X_{S_S}^j(t) - X_{S_l}^j(t)}{X_{S_k}^j(t)(X_{S_k}^j(t) - 1)} \right) \right| \\
&\leq \frac{2L'||k||^2}{X_{S_k}^j(T) - 1}
\end{align*}
\]

for some \(L' > 0\) and since \(X_{S_k}^j(t)\) is non-increasing on \([0, T]\). Thus, the term in (5.51) tends to zero uniformly in probability by (I).

For the term in (5.52), we observe

\[
\begin{align*}
&\sup_{0 < t \leq T} \left| N^{-1}X_{S_k} C_{m}^{j,l}(k, N^{-1}X_{S_k}) - \alpha_S p_k \theta^k C_m^{j,l}(k, N^{-1}X_{S_k}) \right| \\
&\leq L||k||^2 \sup_{0 < t \leq T} \left| N^{-1}X_{S_k} - \alpha_S p_k \theta^k \right| \xrightarrow{P} 0
\end{align*}
\]

by the bound on \(C_m^{j,l}(k)\) in (5.44) and Lemma 3a.
Finally, since $C^{j.l.}_{m}(t,z(t))$ is Lipschitz continuous in $z$, we have

$$\sup_{0 < t \leq T} \left| C^{j.l.}_{m}(t,N^{-1}X_{S^\bullet}) - C^{j.l.}_{m}(t,\alpha_S \theta \partial \psi(\theta)) \right| \leq L' \sup_{0 < t \leq T} \left| N^{-1}X_{S^\bullet} - \alpha_S \theta \partial \psi(\theta) \right|$$

for some $L' > 0$ and so the term in (5.53) tends to zero uniformly in probability by Lemma 3b. Therefore, recalling also (5.47) we conclude that (5.46) holds and hence (5.45) follows.

We may now complete the derivation of Theorem 5.1.1 via Gronwall’s inequality (see, e.g., [7]).

**Proof of Theorem 5.1.1.** Recall the definition of $\Delta = (\Delta_X, \Delta_\Theta)$ from equations (5.37) and (5.38). Note that, by equations (5.3) and (5.7),

$$(X(t)/n, \theta(t)) = (X(0)/n, \theta(0)) + \int_0^t H(X(s)/n, \theta(s))ds + \mathcal{E}(t)$$

where

$$\mathcal{E}(t) = (N^{-1}M(t), 0) + \int_0^t \Delta(s)ds.$$}

We note each coordinate of the vector process $M(t) = (M_S, M_I, M_{SI}, M_{\tilde{S}I}, M_{SS}, M_{\tilde{S}S})(t)$ is a pure jump, càdlàg, zero mean, martingale process. Consider the process $M_{SI}^j$ which, by equation (5.3), jumps only if infection of a node, recovery of a node or a $j$-edge drop/activation occurs at time $s$. Recall that, for each $j$, either activations or drops are possible, not both. Consider events corresponding to a node of degree $k$. For infection and recovery events the jump size, $\delta M_{SI}^j(s)$, is not greater than that node’s $j$-degree, and for activation and deactivation events, of which there are at most $2k_j$ affecting that node, the
jump size is one. Since the number of nodes of degree $k$ is approximately $np_k$ for large $n$, the corresponding quadratic variation process satisfies

$$E[M_{SI}^j](t) = E \sum_{s \leq t} ( \delta M_{SI}(s))^2 \leq 2n \sum_k k^2 p_k + 2n \sum_k k_j p_k \leq 4n \sum_k k^2 p_k = O(n)$$

by (III). Consequently, Doob’s martingale inequality implies $\sup_{0 \leq t \leq T} |N^{-1}M_{SI}(t)| \xrightarrow{P} 0$. A similar argument applies also to $M_{\tilde{S}I}$, $M_{S\tilde{S}}$, and $M_{\tilde{S}S}$ as well as $M_S$ and $M_I$, both of which make only unit jumps. Since by Lemma 4 we have $\sup_{0 \leq t \leq T} ||\Delta(t)|| \xrightarrow{P} 0$, this implies $\sup_{0 \leq t \leq T} ||\mathcal{E}(t)|| \xrightarrow{P} 0$.

Note that $\mathcal{H}$ is a (vector valued) Lipschitz continuous function on its domain, which we can take to be $[0, 1]^2 \times ([\xi, 2 \partial \psi(1)]^R) \times [\xi, 1]^R$ by Remark 5.5.0.2. Together with Gronwall’s inequality, this implies

$$\sup_{0 < s \leq t} ||(X(t)/N, \theta(t)) - D(t)|| \leq ||(X(0)/N, \theta(0)) - D(0)|| + \sup_{0 < s \leq t} \left|\int_0^t [\mathcal{H}(X(s)/N, \theta(s)) - \mathcal{H}(D(s))]ds\right|$$

$$+ \sup_{0 < s \leq t} ||\mathcal{E}(s)||$$

$$\leq ||(X(0)/N, \theta(0)) - D(0)|| + L \int_0^t \sup_{0 < s \leq t} ||(X(s)/N, \theta(s)) - D(s)||ds + \sup_{0 < s \leq t} ||\mathcal{E}(s)||$$

$$\leq \left(||(X(0)/N, \theta(0)) - D(0)|| + \sup_{0 < s \leq t} ||\mathcal{E}(s)||\right) e^{Lt} \xrightarrow{P} 0,$$

for some $L > 0$, since the first term in the parenthesis tends to zero by Condition II and (5.9). The assertion follows when we take $t = T < \infty$.

**Modifications for SEIR** The essential change induced by the introduction of the $E$ state is that the neighborhood of a node which we need in (5.4) is now a different hypergeometric
distribution. The other aspects the proof given in the last section remain unchanged. Thus, we give the hypergeometric distribution which describes the neighborhood of a node in the SEIdaR \((r, k)\) model with the understanding that the rest of the proof follows the same reasoning as before.

In Remark 5.5.0.1 we noted that the disease status configuration of a susceptible node \(u\)'s \(j\)-neighbors is described by a hypergeometric distribution. With the addition of the \(E\) compartment the only substantive change is that these neighbors may also be exposed rather than only susceptible or infectious. Thus, the hypergeometric distribution for the \(j\)-layer of a susceptible node of degree \(k\) at time \(t\) becomes:

\[
P(X_{SE,i}(t) = n_{SE}, X_{SI,i}(t) = n_{SI}, X_{SS,i}(t) = n_{SS}, X_{SE,i}(t) = n_{SE}) = \frac{\binom{n_{SE}}{X_{SE}(t)} \binom{n_{SI}}{X_{SI}(t)} \binom{n_{SS}}{X_{SS}(t)} \binom{n_{SI}}{X_{SI}(t) - X_{SE}(t) - X_{SS}(t) - X_{SI}(t) - X_{SS}(t)}}{\binom{k}{j}},
\]

supported on the six-simplex \(0 \leq n_{SE} + n_{SI} + n_{SS} + n_{SE} + n_{SI} + n_{SS} \leq k\). Each layer remains independent of every other layer (conditional on the number of neighbors), and so it still holds that

\[
P(X_{SE,i} = n_{SE}, X_{SI,i}(t) = n_{SI}, X_{SS,i}(t) = n_{SS}) = \prod_{l=1}^{r} P(X_{SE,i}(t) = n_{SE}, X_{SI,i}(t) = n_{SI}, X_{SS,i}(t) = n_{SS}).
\]

The analogous mixed-moment terms in (5.27) now become:

\[
Q^{j,(k)}_i = X_{SI,i}(X_{SS,i} - X_{SE,i}) \quad \text{and} \quad \tilde{Q}^{j,(k)}_i = X_{SI,i}(X_{SS,i} - X_{SE,i}),
\]

116
and are now computed with respect to (5.55). The proof of Lemma 3 is essentially un-
changed. The proof of Lemma 4 follows similarly with the same argument but making use
of (5.55). The final proof of Theorem 5.2.1 can then proceed via Gronwall’s Inequality in
the same manner as Theorem 5.1.1.
Chapter 6: Branching Process Approximation

6.1 Introduction

The CMSEIdR model with $r = 1$ proposed in the previous chapter can be approximated by a suitable branching process in the early phases of the epidemic. This occurs because when number of cases is small, the probability that a given neighbor has already been infected tends to 0 as $N \to \infty$. In this chapter we derive the offspring distribution of this branching process and derive parameter estimation methods for analyzing outbreak data when the number of infected individuals is small. Thus, these methods may be applicable in the early phase of a large outbreak or to an outbreak such as the one that occurred in the Democratic Republic of the Congo in 2014 as reported by Maganga et al [77].

6.2 Branching process approximation

As in the case of a classical SIR [2] the early behavior of an epidemic on a CM graph can be approximated by a suitable branching process [12, 53, 17]. Such approximation allows one in turn to use the early epidemic data to statistically ascertain the probability of
a major outbreak (large number of infections among the network nodes) as well as to estimate the rate of infection spread and changes in the contact network as described below. By and large, there have been limited studies of statistical estimation for network-based models [113]. In one of the early papers on the topic, O’Neill and Britton study Bayesian estimation procedures for the $G(N, p)$ model when the network is assumed to be of reasonable enough size that the network structure may be included as missing data and imputed via a Monte Carlo scheme [20]. More recently, Groendyke has extended this approach to non-Markovian dynamics by allowing both the infection time and the recovery time to follow an arbitrary gamma distribution [44, 45]. However, these methods are generally tailored to networks of small size, where posterior sampling is feasible, and may encounter various difficulties in large networks when the imputation becomes computationally expensive and often impractical [113]. The framework presented here, on the other hand, assumes the proportion of infectives is small relative to the population size. This allows us to avoid explicit imputation of the network and makes the numerical complexity of the analysis comparable to that of a small homogenous SIR epidemic [101].

The main contribution of the current chapter is to present a statistical inference method for analyzing the early stages of an epidemic, or a small outbreak, evolving according to SIR type dynamics on a random graph. While it is widely known that patients infected with Ebola undergo an asymptomatic period (E) this does not affect the final offspring distribution that we derive. Hence, the our analytical results apply equally well to SEIR dynamics. Moreover, since our model in this section does not use temporal information its
inclusion or exclusion does not change the derived results. A novel aspect of our method is that we assume the random graph structure evolves in response to the epidemic progression. This allows us to account for changing contact patterns in response to infection, for example due to population behavioral changes or health interventions (e.g. quarantine).

We apply this method to a dataset from the smaller 2014 outbreak in the Democratic Republic of the Congo described by Maganga et al. [77]. While the West Africa outbreak received considerable attention in fall 2014 due to the dramatic rise in the number of cases, there was an independent Ebola outbreak in the Democratic Republic of the Congo (DRC) occurring at the same time. This much smaller outbreak began July 26, 2014 in Équateur Province and lasted until November 2014. The index case was a woman living in Inkanamongo village that presumably became infected by consuming bushmeat of an infected animal. Several healthcare workers involved in a postmortem cesarean section on the women subsequently became ill and generated further chains of transmission [77].

6.3 DRC dataset

There were a total of 69 confirmed infections in the 2014 DRC outbreak. The time series of cumulative case counts was reported by Maganga et al. [77]. However, the analysis method presented here does not require temporal data and instead utilizes the distribution of secondary cases, also given in [77]. The index case is believed to have caused 21 secondary cases, presumably due to her funeral acting as a super-spreading event [77]. Hence, we assume this data point to be an outlier and exclude it from our final analysis, as Maganga et al. also did in their estimation of $R_0$. The secondary case distribution for other
named contacts is shown in Figure 6.1. One patient caused three subsequent cases, two patients caused two additional cases, 30 caused one additional case and 11 patients caused zero additional cases. These may be viewed as observations from the post-index offspring distribution of the branching process approximation, i.e. the number of infections caused by an individual who is himself not the index case.

Although, as already mentioned, the temporal data measurements are not explicitly required for parameter estimation in our approach, the available temporal information can be directly incorporated into the likelihood function, or used to inform the parameterization of the prior distributions, which is what was done for the current DRC dataset. Further details are given below.
6.4 Epidemic model

Again we consider a graph $G = \langle V, E \rangle$ where the vertex set, $V$, corresponds to individuals in a population of size $N$ and the edge set, $E$, corresponds to potentially infectious contacts occurring between such individuals. The graph structure is given as a realization of a configuration model random graph with a prescribed degree distribution, $D$, where the probability that a node has degree $k$ is denoted by $q_k \equiv P(D = k)$. Each node $i \in V$ is assigned $d_i$ half-edges that are drawn at random from $D$. Then, the pool of half-edges is paired off uniformly to form the final network. This link formation model does not exclude the possibility of self-loops or multiple edges but this has a negligible effect in the large graph limit (see, e.g., discussion in [53]).

The general disease framework adopted is the standard compartmental SIR model where, for every time $t > 0$, each node $i \in V$ is classified as susceptible (S), infectious (I), or recovered (R) from the infection. Given some number of initially infectious nodes, a node $i$ becomes infective via transmission along an edge from one of his infectious neighbors. In the Markovian case, we assume that $i$ remains infectious for an exponentially distributed amount of time with rate parameter $\gamma$, which we refer to as the recovery rate. More generally, a non-exponential (e.g., gamma) recovery rate leads to the so-called semi-Markov SIR model [54]. While infectious, the infective $i$ attempts to transmit the infection to all of his susceptible neighbors according to an exponential distribution with rate $\beta$. In the case that the realization of the infection “clock” for a particular neighbor “rings” before the recovery
“clock”, then this neighbor becomes infected. The epidemic ends when there are no more infectives.

We assume the same network dynamics as the CMSEIdR(1) model of the previous chapter. Specifically, we assume that infectious individuals drop each of their contacts according to an exponential distribution with rate $\delta$. The dropped contacts, which could account for behavioral changes due to disease such as isolation or decreased mobility, cannot be reformed.

6.5 Statistical inference

6.5.1 Index case offspring distribution

For an index case $i$ with degree $d_i$, at most $d_i$ secondary infections can be produced. Conditional on the recovery time of that index case, $t_i$, the probability that infection has passed to any particular neighbor is given by

$$p_{t_i} \equiv p(t_i; \beta, \delta) = \frac{\beta}{\beta + \delta}(1 - e^{-(\beta + \delta)t_i}) \quad (6.1)$$

where the first term in the product on the right-hand side is the probability that an edge transmitted infection prior to being dropped, while the second term represents the probability that either an infection or a drop occurred before recovery. Therefore, the total number of secondary infections caused by node $i$, which we denote by $X_i$, conditional on the time of recovery $t_i$, is given by

$$P(X_i = x_i|t_i, d_i) = \binom{d_i}{x_i} p_{t_i}^{x_i} (1 - p_{t_i})^{d_i-x_i}.$$
If the recovery time is not known but assumed to follow a distribution \( f(t_i) \equiv P(T_i = t_i) \), then we may analogously define \( \pi_{d_i}(x_i) \), the conditional probability of \( x_i \) offspring given degree \( d_i \). The law of total probability implies

\[
\pi_{d_i}(x_i) = \int_0^\infty \left( \frac{d_i}{x_i} \right) p(t_i)^x(1 - p(t_i))^{d_i-x} f(t_i) dt_i.
\] (6.2)

We will assume identical recovery distributions for all individuals and, thus, the offspring distribution will be the same for all index cases.

The final form for the offspring distribution of an index case may be found by supposing that the degree, \( d_i \), of the index case is unknown. Let \( q_{d_i} \equiv P(D_i = d_i) \) denote the probability that the index case has degree \( d_i \). Therefore, the law of total probability gives

\[
P(X_i = x_i) = \sum_{d_i \geq x_i} q_{d_i} \pi_{d_i}(x_i).
\] (6.3)

That is, we sum over all possible degrees that could yield at least \( x_i \) secondary infections and weight them according to the degree distribution.

### 6.5.2 Post-index case offspring distribution

We now consider the offspring distribution for a post-index case. By definition, such an individual acquired infection from another individual in the network and, thus, has at least one neighbor. Therefore, some adjustments are needed to account for the fact that post-index cases have a degree distribution which differs from \( D \). Let \( q'_k \) denote the probability that a given neighbor in the CM network has degree \( k \). Then from the results of the previous
chapter,

\[ q'_k = \frac{kq_k}{\mu}. \]

Since at least one of the neighbors of a post-index case has already been infected, he may pass the infection to at most \( k - 1 \) of his neighbors. Let \( X'_i \) denote the offspring distribution for a post-index infection \( i \). Similarly to Eq. (6.3) we derive

\[ P(X'_i = x'_i) = \sum_{k > x'_i} q'_k \pi_{k-1}(x'_i). \]  

(6.4)

For a fixed set of parameters the basic reproductive number, \( R_0 \), can be calculated as \( E(X'_i) \), i.e. the average number of secondary infections caused by a post-index case. That is, \( R_0 \) is given by

\[ R_0 = \sum_{x'_i=0}^{\infty} \sum_{k > x'_i} q'_k \pi_{k-1}(x'_i). \]  

(6.5)

To illustrate, we assume that the degree distribution is Poisson with mean parameter \( \lambda \) and the recovery distribution is exponential with rate parameter \( \gamma \). This implies

\[ f(t_i) = \gamma e^{-\gamma t_i} \]

and

\[ q_{d_i} = \frac{\lambda^{d_i} e^{-\lambda}}{d_i!}. \]

Therefore, by Eq. (6.3), the offspring distribution for an index case is given by

\[ P(X_i = x_i|\lambda, \beta, \gamma, \delta) = \sum_{k \geq x_i} \frac{\lambda^k e^{-\lambda}}{k!} \int_0^\infty \binom{k}{x_i} p(t; \beta, \delta)^x_i (1 - p(t; \beta, \delta))^{k-x_i} \gamma e^{-\gamma t} dt, \]

125
and, by Eq. (6.4), the post-index case offspring distribution is given by

\[ P(X'_i = x'_i | \lambda, \beta, \gamma, \delta) = \sum_{k > x'_i} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} \int_0^\infty (k-1) p(t; \beta, \delta)^{x'_i} (1 - p(t; \beta, \delta))^{k-1-x'_i} \gamma e^{-\gamma} dt. \]

This expression has no simple analytical form but it is not hard to approximate the integral numerically since it can be written as an expectation against the recovery distribution. Therefore, a simple Monte Carlo sample from the desired recovery distribution allows for efficient computation of this term.

Using Eq. (6.5), we can calculate the basic reproductive number in this Markovian case. For an arbitrary degree distribution, \( R_0 \) is given by

\[ R_0 = \beta \beta + \gamma + \delta \sum_{k=0}^\infty \frac{(k-1)kq_k}{\mu}, \]  

which only differs in the inclusion of \( \delta \) from the corresponding formula on a static CM graph [85, 53]. Here \( \mu = E(D) < \infty \) by assumption. The summation in Eq. (6.6) represents the expected excess degree, i.e. the degree of a node which is necessarily a neighbor of a node, not counting the known edge. This coincides with the calculations of the previous chapter as \( \kappa = 1 \) when \( r = 1 \) and the degree distribution is Poisson. Thus \( R_0 \) is found to be (cf., e.g., [7] chapter 6)

\[ R_0 = \frac{\beta}{\beta + \gamma + \delta} \sum_{k=0}^\infty \frac{\lambda^{k-1}e^{-\lambda}}{k!} k(k-1) = \frac{\beta \lambda}{\beta + \gamma + \delta}. \]
6.5.3 Likelihood and estimation

In practice, outbreak data may not arise solely from a single index case and often separate chains of infection are tracked. Suppose the data \( \{x_1, \ldots, x_m\} \) corresponds to the number of secondary infections for each of \( m \) independent index cases and the data \( \{x'_1, \ldots, x'_m'\} \) corresponds to secondary infections caused by each of the \( m' \) post-index cases.

Let \( \Theta = (\beta, \gamma, \delta, \lambda) \) denote the vector of parameters where \( \gamma \) and \( \lambda \) represent the parameters of the recovery time and degree distributions, respectively. The offspring distributions given in Eqs. (6.3) and (6.4) allow for explicit formulation of the likelihood for \( \Theta \) which is given by

\[
L(\Theta|x_1, \ldots, x_m, x'_1, \ldots, x'_m') = \prod_{j=1}^{m} \sum_{k \geq x_j} q_k \pi_k(x_j) \times \prod_{c=1}^{m'} \sum_{l > x'_c} q'_l \pi'_l(x'_c). \tag{6.8}
\]

With the specification of the likelihood, maximum likelihood estimators (MLEs) for the rate parameters can be found by numerical optimization. Given the MLE \( \hat{\Theta} = (\hat{\beta}, \hat{\gamma}, \hat{\delta}, \hat{\lambda}) \), the corresponding estimator for the basic reproductive number can be calculated by application of the continuous mapping theorem to the expression for \( \mathcal{R}_0 \) given in Eq. (6.5). For example, the estimator following from Eq. (6.7) would be

\[
\hat{\mathcal{R}}_0 = \frac{\hat{\beta} \hat{\lambda}}{\hat{\beta} + \hat{\gamma} + \hat{\delta}}. \tag{6.9}
\]

Denote the vector of current parameters as \( \Theta = (\beta, \gamma, \delta, \lambda) \). If we denote the parameter prior distribution \( \phi(\Theta) \) then the likelihood function above may be also used to compute the Metropolis-Hastings acceptance probability in the Markov chain Monte Carlo (MCMC)
sampler. Let \( x \) denote the vector of data counts and \( \tau \) be the transition kernel. The MCMC algorithm for obtaining the posterior distribution of \( \Theta \) is then as follows

1. Initiate \( \Theta_{\text{curr}} = \Theta_0 \).

2. Obtain proposal \( \Theta_{\text{prop}} \) from \( \tau(\Theta|\Theta_{\text{cur}}) \).

3. Accept (or not) this proposal with Metropolis-Hastings probability given by

\[
\rho(x, \Theta_{\text{cur}}, \Theta_{\text{prop}}) = \min\left(1, \frac{L(\Theta_{\text{prop}}|x)\phi(\Theta_{\text{prop}})\tau(\Theta_{\text{prop}}|\Theta_{\text{cur}})}{L(\Theta_{\text{cur}}|x)\phi(\Theta_{\text{cur}})\tau(\Theta_{\text{cur}}|\Theta_{\text{prop}})}\right). \tag{6.10}
\]

4. Return to 2.

5. Repeat until convergence.

In this way, after sampler convergence, we obtain an approximate sample from the posterior distribution of \( \Theta \) and may subsequently compute its approximate \( 1 - \alpha \) credibility region, given the observed data \( x \). In particular, the credibility interval of the parameter \( \mathcal{R}_0 \) given by Eq. (6.5) may be determined.

### 6.5.4 Generalizations

Note that the likelihood formula (6.8) is valid in a non-Markovian setting such as an arbitrary recovery time distribution and could further be extended to the scenario where the transmission rate varies with time since infection. A model using this time-varying transmission mechanism was proposed by [116], however, no such data has been made publicly available for either the West Africa or DRC outbreak. Note that in this case the
formula for $R_0$ would differ from Eq. (6.5) due to a form for $p_t$ that differs from Eq. (6.1) but would remain calculable as $E(X'_t)$.

If additional data were available, such as the recovery time or number of contacts of each individual, it could be explicitly incorporated into the likelihood function (6.8) through the joint distribution of recovery times and degrees.

6.6 Analysis of the DRC dataset

To illustrate our method we perform the Bayesian posterior estimation of the model parameters for the 2014 Ebola outbreak in the DRC based on the data described in Section 6.3. The specific model considered is as given in Section 6.4, where transmission occurs according to an exponential distribution with rate parameter $\beta$, and the degree distribution is Poisson with parameter $\lambda$. Recovery time is assumed to follow a gamma distribution $\Gamma(\alpha, \beta)$. We note that, while incubation periods for Ebola range from two to 21 days [27], our method does not require consideration of latent exposure since it does not depend on infection timing.

We perform estimation via the MCMC scheme given in Section 6.5.3. Prior distributions were set to be minimally informative Gaussian (parameterized as $N(\mu, \sigma^2)$) distributions and hyper-parameters were selected based on previous estimates [77, 106]. The prior distribution for $\lambda$ was taken to be $N(\mu_\lambda = 16, \sigma^2_\lambda = 14^2)$ and for $\delta$ was taken to be $N(\mu_\delta = .01, \sigma^2_\delta = .05^2)$. To improve mixing of the MCMC scheme, $\beta$ was estimated on log-scale under an assumed $N(\mu_\beta = \log(.02), \sigma^2_\beta = 4^2)$ prior distribution. Lastly, the gamma distribution for the infectious period was chosen to have prior mode (i.e.,
Figure 6.2: Estimated posterior densities for the parameters of interest for the non-Markovian MCMC sampler. Green line denotes the posterior mean.

\[(\alpha - 1)/\beta\) distributed as \(N(\mu_m = 11, \sigma_m^2 = 6^2)\) and prior standard deviation \((\sqrt{\alpha}/\beta)\) as \(N(\mu_{sd} = 6, \sigma_{sd}^2 = 4^2)\). A Gaussian transition kernel \(\tau\) was used. Central 95\% credibility intervals were calculated for the parameters of interest as well as for \(R_0\). Results are summarized in Figure 6.2 as the histograms of posterior samples.

The 95\% credibility interval for the basic reproductive number \(R_0\) is found to be (.589, 1.15) with the \(R_0\) posterior mean of .842. The latter value numerically agrees closely with the moment-based estimate given in Maganga et al. [77] and is also consistent with the relatively small total number of confirmed infections observed during the epidemic. We further note that our interval estimate compares favorably to the 95\% \(R_0\) confidence interval of (−.38, 2.06) reported by Maganga et al. [77] indicating that for the DRC data...
the fully parametric model produces a more precise (shorter) interval. The infection rate is found to have a posterior mean of .0069 and a corresponding 95% credibility interval of (.00232,.0167). The contact drop rate is found to have a posterior mean of .0573 and corresponding credible interval of (.00340,.128). These quantities were not estimated directly by the authors in [77] so comparison here is not possible. The estimated mean for the infectious period is found to be 16.18 days with corresponding credibility interval of (6.65 days, 27.9 days). Maganga et al. did not explicitly estimate infectious period but instead gave an estimate for time from symptom onset to death with mean 11.3 days [77]. Based on a historic 1995 outbreak of Ebola in the DRC, Legrand et al. estimated the time from onset to end of infectiousness for survivors to be 10 days and time from onset to death to be 9.6 days [67]. While our estimate of infectious period is somewhat longer than these, we note that our interpretation of the quantity is the total length of time that an individual could cause infection. This could include several days after death in which the body is being prepared for and undergoing funeral rituals. Lastly, the mean number of contacts is estimated to be 16.2 with corresponding credibility interval of (8.59,22.8), which is again in close pointwise agreement with the estimate inferred from the number of contacts traced by Maganga et al. [77]. Overall, the point estimate for $R_0$ and infection and recovery rates are seen to be consistent with a small outbreak behavior observed in the DRC dataset and to agree well with the numerical values reported earlier. However, based on the same data, our parametric model is also seen to yield more precise interval estimates.
The final outbreak size of simulated branching processes from the posterior parameter distribution was used as a model diagnostic for fit to the empirical data. Conditioning on the number of infections caused by the index case as the number of independent branches, the posterior parameter samples and corresponding post-index case offspring distributions were used to simulate branching process realizations. The final outbreak sizes were calculated and the distribution is presented in Figure 6.3. Reasonable agreement is observed between this distribution and the empirical outbreak size of 69 cases [77].

To assess the sensitivity of the credibility interval for \( R_0 \) and the final size distribution to the prior variance, the additional MCMC analysis was performed. Specifically, for the unchanged values of prior means the prior variance was increased fourfold to \((\sigma^2_\beta, \sigma^2_\delta, \sigma^2_m, \sigma^2_{sd}, \sigma^2_\lambda) = (8^2, 1^2, 12^2, 8^2, 28^2)\). As expected, this change was found to increase the credible bounds for individual parameters. However, with the increased prior variance the 95% credible interval for \( R_0 \) was found to be (.582, 1.13) and no substantial change in the resulting final outbreak size distribution was noted. This indicates that the posterior distributions of these particular quantities are largely robust to the prior distribution specifications.

6.7 Conclusion

We presented likelihood-based and Bayesian parameter estimation methods for a class of stochastic epidemic models on configuration model graphs. The method is based on applying the branching process approximation, and is applicable when the number of infections is small in relation to the size of the population under study. In particular, this
Figure 6.3: Final outbreak size distribution based on 20,000 simulations of the branching processes from the posterior parameter distribution. The actual outbreak size of 69 based on the DRC dataset (black line) is shown for comparison.

includes the case when the total outbreak size is small or when we are at the onset of a large outbreak. The method are flexible, for example allowing for arbitrary degree and recovery distributions, and in principle requires only a knowledge of the distribution of secondary cases, although additional data can be incorporated into the inference procedure, as the likelihood function under branching approximation remains straightforward to evaluate under a wide range of data collection schemes.

We illustrated our approach with the analysis of data from the 2014 DRC Ebola outbreak which was originally described and analyzed in Maganga et al. [77]. Our method, under only weakly informative prior distributions, is seen to produce a considerably tighter
credibility interval for $R_0$ than the moment-based confidence estimate reported in [77]. This demonstrates the utility of the branching process approximation for small epidemics in obtaining more precise estimates of $R_0$, which is essential in assessing potential risk of a large outbreak and in determining the level of control efforts (e.g. vaccination or quarantine) needed to mitigate an outbreak. The final size comparison indicates that the observed data is within the range of model predictions, although the direct comparison of observed and model predicted offspring distributions indicates some disagreement in the observed frequency of zeros and ones. The small sample size prohibits definitive conclusions but this may indicate the need to incorporate more complex network dynamics (e.g. distinguishing between multiple types of infectious contacts).

As the statistical estimation methods appear essential to inform public health interventions, we hope that our work here will help in establishing a broader inference framework for epidemic parameters, based on the type of data usually collected in the course of an outbreak. The Bayesian approach is attractive in this context, as it naturally incorporates any prior or historical information. However, the method of this chapter only addresses the inference problem at the epidemic onset and, in particular, is not appropriate when the number of infected individuals comprises a significant portion of the population. We address estimation for such large outbreaks, possibly also incorporating more complex network dynamics, in the following chapter.
Chapter 7: Parameter estimation methods for large outbreaks

7.1 Introduction

In this section we explore methods for estimating epidemiological parameters from data that are commonly available during outbreaks. As with most of our analysis, we are most interested in asymptotic considerations with an eye towards applying them to data from population-level studies.

7.2 Estimation for network-based SIR models

In spite of a great deal of study in the quantitative epidemiology literature, there are few rigorous results available concerning parameter estimation for any class of network-based models [95]. Pellis et al. point out that such methods become increasingly relevant as more precise network data is collected in the future [95]. We have reviewed some of the existing approaches to parameter estimation for network-based models in Chapter 6. At present we wish to study parameter estimation in the case that the methods of Chapter 6 do not apply, that is, when the size of the outbreak is large and the branching process approximation is no longer appropriate.
Parameter estimation for network-based models presents challenges because the network itself is rarely observed. Therefore, separating properties of the network from characteristics of the epidemic becomes difficult. As in spatially homogeneous models, the observed data can most often only be assumed to be \( \mathcal{H}_t \), i.e. the timings of exposure, infection, and recovery events up to time \( t \). It is in theory possible to condition on the structure of the network and write the likelihood as:

\[
L(\theta | \mathcal{H}_t) = \sum_{g \in \mathcal{G}(N, \psi)} P(\mathcal{H}_t | g, \theta) P(g | \theta),
\]

and then sum over every possible network configuration. This is similar in flavor to the approaches discussed early in Chapter 6 but the network is instead imputed via Markov Chain Monte Carlo and treated as an additional parameter. Unfortunately, this approach appears to become very difficult when (i) the network becomes large (i.e. there are \( O(N^2) \) edges to impute in the structure of the network) or (ii) when the network is no longer a realization of the \( G(N, p) \) model. The methods reviewed in beginning of Chapter 6 ought to be suitable when the graph is of this form but do not apply outside of it. In addition, there are some challenges of identifiability that have been noted \cite{113}. As most networks will not have the Poisson degree distribution specified by this model, it becomes desirable to explore other methods.

We have already alluded to some ways of avoiding the use of (7.1). First, we examine how the problem can be handled when we have complete observation of some parts of the
network. Second, we apply the results from Chapter 5 to suggest that Least Squares estimation remains feasible as in the case of DDMJP. Lastly, we propose a novel approximate likelihood approach motivated by the “hybrid” model presented at the end of Chapter 5.

7.3 Score process estimator

Let \( \theta = (\beta, \sigma, \gamma, \delta, \kappa, K = \mu) \) denote the full vector of free parameters and suppose that the network parameters (fully specified by \( \kappa \) and \( K \)) are known to be \( \kappa_0 \) and \( K_0 \) and the PT-condition is satisfied. Further, we assume that the graph is observed sufficiently to know the number of \( S-I \) connections (denoted as \([SI](t)\)) in the network for \( t > 0 \). We define the counting process \( C_E(t) \) as the process that increases by 1 each time an exposure occurs. Similarly, define the counting processes \( C_I(t) \) and \( C_R(t) \) that count the cumulative number of individuals with symptom onset and recovered individuals up to time \( t > 0 \). Then because the hazard of each event type is constant between times and known throughout, the likelihood of \( \mathcal{H}_t \) can be written in the form of (3.15). Taking derivatives with respect to the free parameters \((\beta, \sigma, \gamma)\) yields the following score process equations:

\[
S_\beta(t) = C_E(t) - \beta \int_0^t [SI](s)ds \\
S_\sigma(t) = C_I(t) - \sigma \int_0^t E(s)ds \\
S_\gamma(t) = C_R(t) - \gamma \int_0^t I(s)ds 
\]

(7.2)

By definition, each counting process is 0 at \( t = 0 \) and therefore each process evaluated at the true parameter values \((\beta_0, \sigma_0, \gamma_0)\) is a zero-mean martingale as in Chapter 3. This follows because the hazard of each event \((E, I, R)\) is given as the respective rate times the
integrand. Thus, as before the score-type estimator given the trajectory up to time $t$ is the parameter value that sets the left hand side of (7.2) to 0.

Setting the left hand side of each equation to 0, the score-type estimators become:

$$\hat{\beta} = \frac{C_E(t)}{\int_0^t [SI](s)ds}$$

$$\hat{\sigma} = \frac{C_I(t)}{\int_0^t E(s)ds}$$

$$\hat{\gamma} = \frac{C_R(t)}{\int_0^t I(s)ds}$$

(7.3)

Note the similarity between (7.3) and (3.18). Consistency of the estimator as $N \to \infty$ and finite $t$ now follows from the form of the counting processes and Chapter 3, Lemma 1. The assumptions in this section are quite strong but yield an important insight; that is, the missing $[SI]$ quantity is perhaps the biggest hurdle to parameter estimation. On the other hand, the Law of Large Numbers in Theorem 5.2.1 of Chapter 5 motivates a potential solution to this problem when the population is large and the initial condition is known (or can be approximated). In particular, the result states that when the network is of sufficient size, that the missing $[SI]$ quantity (when scaled) behaves approximately deterministically. This suggests consideration of a process where the deterministic limit of $[SI]$ is used to approximate the stochastic (and missing) number of $S-I$ connections in the network and then computing estimates analogous to (7.3). For analysis at the population level this ought to yield an approximately equivalent result.

A more formal description of this process is now given. Consider a jump process $Y^*_N(t) = (S, E, I)(t)$, indexed by $N$, with rates and transitions as follows: where given an initial condition and set of parameters $\theta$ the quantity $D_{SI}(t; \theta)$ is computed as the solution
Table 7.1: Transitions for the approximate process formed by using the limiting equation to compute $\delta SI$.

<table>
<thead>
<tr>
<th>Event</th>
<th>Rate</th>
<th>Transition</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Exposure</td>
<td>$\lambda_1^N(t; \theta) = N\beta D_{SI}(t; \theta)$</td>
<td>$\Delta Y^* = (-1, 1, 0)'$</td>
</tr>
<tr>
<td>Symptom Onset</td>
<td>$\lambda_2^N(t; \theta) = \sigma E$</td>
<td>$\Delta Y^* = (0, -1, 1)'$</td>
</tr>
<tr>
<td>Recovery</td>
<td>$\lambda_3^N(t; \theta) = \gamma I$</td>
<td>$\Delta Y^* = (0, 0, -1)'$</td>
</tr>
</tbody>
</table>

We will need two further conditions on the rate functions that essentially specify their uniqueness and non-degeneracy. The needed conditions are as follows:

$$
\forall l, s \leq t \implies \theta_1 = \theta_2
$$

$$
\exists l, s \quad \delta_k \lambda^l_N(s; \theta) \neq 0, \quad 0 \leq s \leq t, k = 1,...,d.
$$

Proposition 3. Let the closure of the parameter space $\text{cl}(\Theta)$ be a compact set in $\mathbb{R}^d$ and assume conditions (7.4) above. Suppose that $\mathcal{H}_t$ is generated by the SEI$\text{da}R(1,0)$ process in Chapter 5 and the conditions of Theorem 5.2.1 are satisfied with $\alpha = D(0)$ and $\mu_0$ known. Let $\hat{\theta} = (\hat{\beta}, \hat{\sigma}, \hat{\gamma}, \hat{\delta}, \hat{k})$ denote the solution to (3.16) where for each $N$ the transition
rates are as given in Table 7.1. Then \( \hat{\theta} \) is consistent in the sense that

\[ ||\hat{\theta} - \theta|| \xrightarrow{P} 0, \quad N \to \infty, \quad (7.5) \]

**Proof.** Let \( C_1 \) denote the counting process that increases by 1 each time a new exposure occurs and let \( \lambda^1_N(s; \theta) = N\beta D_{SI}(s; \theta) \) denote the hazard of this event. Similarly, let \( C_2(s) \) and \( C_3(s) \) denote the counting processes that increase by 1 when a new symptom onset and recovery occurs. The corresponding rates for these processes are \( \lambda^2_N(s; \theta) = \sigma E(s) \) and \( \lambda^3_N(s; \theta) = \gamma I(s) \) as in Table 7.1. Let \( \theta_0 \) denote the true parameter value and \( \text{dim}(\theta) = d \).

Since the conditions of Theorem 5.2.1 are satisfied, the \( N \)-scaled quantities \( ([SI], E, I)(s) \) involved in the rates of \( (C_1, C_2, C_3)(s) \) converge to their deterministic limit \( (D_{SI}, D_E, D_I)(s; \theta) \) directly from Theorem 5.2.1. Identically, we have

\[ N^{-1}\lambda^1_N(s; \theta) \xrightarrow{P} \beta D_{SI}(s; \theta), \quad (7.6) \]

and for any \( s > 0 \),

\[ N^{-1}\lambda^2_N(s; \theta) \xrightarrow{P} \sigma D_E(s; \theta) \quad (7.7) \]

\[ N^{-1}\lambda^3_N(s; \theta) \xrightarrow{P} \gamma D_I(s; \theta). \]

Moreover, from (7.7) it follows that

\[ N^{-1}\lambda^l_N(s; \theta) \xrightarrow{P} \lambda^l(s; \theta), \quad l = 1, 2, 3 \quad (7.8) \]

uniformly on \( \theta \)-compact sets. By definition the estimator \( \hat{\theta} \) is the solution (3.16), i.e. the solution to

140
\[ 0 = S_k(t; \hat{\theta}) = \sum_{l=1}^{3} \int_{0}^{t} \frac{\partial_k \lambda^l_N(s^-; \hat{\theta})}{\lambda^l_N(s^-; \hat{\theta})} (dC^l_N(s) - \lambda^l_N(s^-; \hat{\theta}) ds), \quad k = 1, \ldots, d \] (7.9)

when the epidemic is assumed to have been observed for time \( 0 \leq s \leq t < \infty \), thus generating \( \mathcal{H}_t \). This is equivalent to:

\[
\begin{aligned}
N^{-1} \sum_{l=1}^{3} &\left( \int_{0}^{t} \frac{\partial_k \lambda^l_N(s^-; \hat{\theta})}{\lambda^l_N(s^-; \hat{\theta})} (dC^l_N(s) - \lambda^l_N(s^-; \hat{\theta}) ds) \right) \\
&= N^{-1} \sum_{l=1}^{3} \left( \int_{0}^{t} \frac{\partial_k \lambda^l_N(s; \hat{\theta})}{\lambda^l_N(s; \hat{\theta})} (\lambda^l_N(s; \hat{\theta}) - \lambda^l_N(s; \theta) ds) \right). \\
\end{aligned}
\] (7.10)

(7.11)

Without loss of generality consider the left hand side \( \theta_1 \in \Theta \) and fixed \( l = 1 \) (i.e. the first term in the sum). Call this random variable \( Z^l_1 \). We seek to show that \( N^{-1} Z^l_1 \overset{P}{\to} 0 \). Since the closure of the parameter space \( \Theta \) is a compact set and \( t < \infty \), the integrand on the left hand side is bounded for all \( 0 \leq s \leq t \) by continuity of \( \lambda^l(s; \theta) \) and its derivative for \( l = 1, \ldots, 3 \). The continuity of \( \lambda^l(s; \theta) \) in \( \theta \) follows from definition in Table 7.1. It follows from the general theory (e.g. Chapter 9 in Britton (2000) [7]),

\[ \text{Var}(Z^l_1) = E \left( \int_{0}^{t} \left( \frac{\partial_{\theta_1} \lambda^l_N(s^-; \hat{\theta})}{\lambda^l_N(s^-; \hat{\theta})} \right)^2 dC^l_1(s) \right) = O(N), \] (7.12)

as there are most \( N \) exposures between time 0 and time \( t \). Therefore, \( N^{-1} Z^l_1 \overset{P}{\to} 0 \) as desired and the argument may be repeated for \( l = 2, 3 \). Thus, the left hand side of (7.10) converges to 0 in probability. By assumption of non-degeneracy the integrand on the right hand side of (7.10) is non-zero for some \( 0 \leq s \leq t \). It follows that

\[ N^{-1} (\lambda^l_N(s; \hat{\theta}) - \lambda^l_N(s; \theta)) \overset{P}{\to} 0, \quad l = 1, \ldots, 3 \] (7.13)

141
and thus,
\[ N^{-1} \lambda^l_N(s; \hat{\theta}) \xrightarrow{P} \lambda^l(s; \theta), \quad l = 1, \ldots, 3. \] (7.14)

Lastly, as \( N \to \infty \),
\[
|\lambda^l(s; \hat{\theta}) - \lambda^l(s; \theta)| \\
\leq |\lambda^l(s; \hat{\theta}) - N^{-1} \lambda^l_N(s; \hat{\theta})| + |N^{-1} \lambda^l_N(s; \hat{\theta}) - \lambda^l(s; \theta)| \\
\leq \sup_{\theta \in cl(\Theta)} |\lambda^l(s; \theta) - N^{-1} \lambda^l_N(s; \theta)| + |N^{-1} \lambda^l_N(s; \hat{\theta}) - \lambda^l(s; \theta)| \xrightarrow{P} 0
\] (7.15)

from (7.14) and (7.8). By the Continuous Mapping Theorem it follows for any convergent subsequence of \( \{\hat{\theta}_N\} \), say \( \{\hat{\theta}_{N_j}\} \) we have \( \hat{\theta}_{N_j} \xrightarrow{P} \theta \) for \( \theta \in cl(\Theta) \) and this completes the proof.

\[ \square \]

The utility is that one obtains a consistent estimator of \( \theta \) in the SEIdaR(1,0) model (and generalizable to the SEIdaR(r,k) model) without having to directly observe the dyad-related quantities. Also, the result implies consistency of the resulting estimator when the Hybrid Process in Chapter 5 is used to construct the likelihood since \( \bar{D}_{SI}(t; \theta) \to D_{SI}(t; \theta) \) by Theorem 5.2.1.

### 7.4 Least squares estimator

Under certain circumstances where the data is reported only at discrete times (rather than continuously), least squares estimation may present an attractive alternative. As an example, WHO situation reports that were released during the West Africa were of this
form. These reports often reported the number of suspected, probable, and confirmed infections over several day intervals [107].

King et al. note that fitting a system of ordinary differential equations to cumulative case data at some set of time points via Least Squares has been a popular method of parameter estimation in the epidemiology literature [62]. On the other hand, this method is often employed without a rigorous study of the underlying analytical properties of this class of estimators.

Recently, Rempala has studied asymptotic consistency and efficiency of Ordinary Least Squares (OLS) estimators for DDMJP [98]. Therefore, for epidemic models that assume homogeneous mixing this choice of estimator appears worthwhile. As we reviewed in Chapter 3 the consistency of this estimator follows from the Law of Large Numbers for the limiting process. We apply a similar argument here using the Law of Large Numbers for the SEIIdaR\((r,k)\) process. From chapter 5 the limiting equations for the SEIIdaR\((1,0)\) model (when the PT-condition holds) is

\[
\begin{align*}
\frac{dS}{dt} &= -\beta [SI] \\
\frac{dE}{dt} &= \beta [SI] - \sigma E \\
\frac{dI}{dt} &= \sigma E - \gamma I \\
\frac{d[SS]}{dt} &= -2\beta \kappa \frac{[SS][SI]}{S} \\
\frac{d[SE]}{dt} &= \beta \kappa \frac{[SS][SI]}{S} - \sigma [SE] - \beta \kappa \frac{[SE][SI]}{S} \\
\frac{d[SI]}{dt} &= \sigma [SE] - (\beta + \gamma + \delta)[SI] - \beta \frac{[SI]^2}{S}. 
\end{align*}
\]

(7.16)
We consider data \( y(t) \subset (S, E, I, SE, SI)(t) \) measured over some finite set of observation times \( 0 = t_0 \leq t_1 \leq \ldots \leq t_m \leq t \) where \( t \) as usual denotes the end of the outbreak. The \([SS]\) equation may actually be eliminated from the above and so it is not considered. In the case of most epidemics, this data will usually consist of some combination of the \( S, E, \) and \( I \) quantities. The Law of Large Numbers presented in Chapter 5 essentially states that for the study of population-level dynamics that these counts should be close to the dimensionalized quantity in the limiting ODE model evaluated at the true parameter value \( \theta_0 \). That is, for every \( t_i \) it holds that

\[
y(t_i) \approx ND(t_i; \theta_0) \quad t_i \leq t,
\]

(7.17)

where \( D(t_i; \theta_0) \) is the solution to (7.16) for the species contained in \( y(t_i) \) evaluated at the true parameter value \( \theta_0 \) with known initial condition. In practice the left hand side of (7.17) is known for every \( t_i \) but the right hand side is not known and can be viewed as a function of \( \theta \). In the case that the population size is unknown, it is has been suggested that this may be treated as a nuisance parameter [101]. Therefore, the OLS estimator as defined in Rempala (2012) is the value (not necessarily unique) of the estimating equation

\[
\hat{\theta}_{OLS} = \arg\min_{\theta \in \Theta} (N^{-1}y(t_i) - D(t_i; \theta))^2.
\]

(7.18)

Under the same assumptions from Chapter 3 of non-degeneracy and assuming convergence of \( y(t_i) \) to its deterministic limit in Chapter 5, the above can be shown to be a consistent estimator using the same argument as in Chapter 3. An attractive feature of the
OLS estimator $\hat{\theta}_{OLS}$ is that it is consistent whether the data $y(t_i)$ are assumed to arise from the full network-based model or an approximate process with the same limit such as the Hybrid Process.

Proposition 4. Let $y(t) \subset (S,E,I,SE,SI)(t)$ denote the data measured over some finite set of observation times $0 = t_0 \leq t_1 \leq \ldots \leq t_m \leq t$ generated by the SEIdaR(1,0) model. Assume the conditions of Theorem 5.2.1 model holds and additional conditions (3.24) and (3.25) of Chapter 3. Then the OLS estimator $\hat{\theta}_{LSE}$ defined as the solution to (7.18) is consistent in the sense that

$$||\hat{\theta}_{OLS} - \theta_0|| \to 0, \quad N \to \infty \quad (7.19)$$

However, the choice of stochastic system is likely to impact resulting covariance estimation. Rempala and Linder note that while the covariance structure of the resulting OLS estimator for DDMJP takes an analytical form, it is very often numerically unstable [72]. As a result, estimating the covariance of this estimator may be handled via a parametric bootstrap procedure such as the one studied recently in Linder and Rempala [72] (note that for DDMJP this is suggested even though the covariance can be written analytically). Let $t^* = \{t_1,t_2,\ldots,t_m\}$ and $D_{t^*} = \{y^*(t_1),\ldots,y^*(t_m)\}$ denote the observed data. Given the similar analytical properties of the hybrid process to DDMJP this is the approach to variance estimation that we propose as well. This is accomplished as follows

As the samples $\{\hat{\theta}_1,\ldots,\hat{\theta}_B\}$ form an independent sample from the distribution of $f(\hat{\theta}|\theta_0)$, quantities of interest may be computed with the appropriate Monte Carlo estimate from this
Algorithm 1 The OLS bootstrap procedure for interval estimation.

1. Estimate free parameters of $\theta$ as the solution of (7.18).
2. Generate bootstrap trajectories $\{D_1, \ldots, D_B\}$ using $\theta = \hat{\theta}_{OLS}$.
3. Re-estimate parameters $\{\hat{\theta}_1, \ldots, \hat{\theta}_B\}$.
4. $\{\hat{\theta}_1, \ldots, \hat{\theta}_B\}$ then forms an independent sample from $f(\hat{\theta}|\theta_0)$.

sample. For example, since $R_0 = f(\hat{\theta}) = \frac{\hat{\beta} \hat{k} \hat{K}}{\hat{\beta} + \hat{\gamma} + \delta}$, this procedure can be used to construct a confidence interval for $R_0$ and related epidemiological quantities. Numerically, simulation is not ideal because Monte Carlo error decreases on the order of $B^{-1}$ [65], i.e. the number of bootstrap replicates that we compute. However we note that Algorithm 1 parallelizes nicely because each trajectory is simulated independently.

7.5 Hybrid estimator

The “hybrid” $SEI\text{da}R(1,0)$ process proposed in Chapter 5 motivates another method of parameter estimation for when the data is assumed to be generated by the full network-based model. That is, assume temporarily that the data $\mathcal{H}_i$ is generated by the hybrid $SEI\text{da}R(1,0)$ process instead of the $SEI\text{da}R(1,0)$ model. As in Chapter 5 the Hybrid Process is a jump process that is Markovian in that the transition rates are constant between jumps. Since the dyad species are in a sense “missing” data, the standard approach to deal with this is to impute or approximate $[SI]$ in some way. The “hybrid” process essentially
attempts to do this by approximating $E([SI]|\mathcal{H}_t)$ by using the limiting system of differential equations and the observed $S(t)$. Then from the general theory for jump processes in Chapter 3 the MLE given $\mathcal{H}_t$ for the Hybrid Process is the solution to:

$$0 = \partial l(\theta|\mathcal{H}_t) \sum_{l=1}^{3} \int_0^t \left( \frac{\partial_k \lambda_l(\theta; s-)}{\lambda_l(\theta; s-)} (dC_l(s) - \lambda_l(\theta; s-)ds) \right),$$  \hspace{1cm} (7.20)

where the population size $N$ and initial condition needed for computation of the forward Euler approximation are known. In the case that these nuisance parameters ($N$ and the initial conditions) are unknown they may be dealt with via likelihood profiling as discussed in Schwartz et al. [101]. We continue to let $l(\theta|\mathcal{H}_t)$ denote the log-likelihood of the observed process up to time $t$ with complete data. This equation is not the “correct” MLE for the full network-based model in the sense that it not equivalent to (7.1). However, one may view this as an approximate equation approach when the underlying process is assumed to be the network-based SEI$\text{daR}(1,0)$ model. When the model that generates $\mathcal{H}_t$ is assumed to be the hybrid process then we may apply the consistency results for counting processes from chapter 3.

As in the OLS case we may design a parametric bootstrap procedure as follows:

The density of some function $g(\hat{\theta})$ can then be approximated by the appropriate Monte Carlo estimate from the sample $\{\hat{\theta}^{1}_{hyb},...,\hat{\theta}^{B}_{hyb}\}$. We apply this technique in the next section to derive confidence intervals for $R_0$ and related epidemiological quantities from the Ebolavirus outbreak in Guinea.
Algorithm 2 The hybrid bootstrap procedure for standard error estimation.

1. Estimate free parameters of $\theta$ as the solution of (7.20)
2. Generate bootstrap trajectories $\{D_1, \ldots, D_B\}$ using $\theta = \hat{\theta}_{hyb}$
3. Re-estimate parameters $\{\hat{\theta}_1, \ldots, \hat{\theta}_B\}$
4. $\{\hat{\theta}^1_{hyb}, \ldots, \hat{\theta}^B_{hyb}\}$ then forms an independent sample from $f(\hat{\theta}_{hyb}|\theta_0)$.

7.6 Intervention strategies

We derive analytical results concerning vaccination in the $SIdaR(1,0)$ (and therefore $SEIdaR(1,0)$) model with one layer in which contacts de-activate at rate $\delta$. In general, our questions center upon how the availability of a vaccine may be incorporated into such a model and quantifying the impact of deploying that vaccine. Our assumption moving forward is that a vaccinated individual may still become infected, however, he is unable to transmit the illness to any of his neighbors. One may derive analytical results for a vaccine functioning in a different way through arguments similar to what we present here. We further assume that the vaccine functions perfectly, i.e. it is effective with probability 1.

For compartmental models there has been a great deal of research corresponding to the identification of subgroups that should be vaccinated in order to prevent spread of disease [76, 75, 104]. As it might be expected, the inherent individual heterogeneity in network-based models may necessitate a different course of action.

The concept of heterogeneous reproductive numbers for individuals appears to have most thoroughly studied by Lloyd Smith et al. in a landmark paper [74].
set of assumptions to the classical stochastic SIR model the authors assume that for each individual $i$, the individual reproductive number $R_0^i$ is a realization of some non-negative random variable with distribution $f_{R_0}(r)$. Our early results largely extend some concepts introduced by Lloyd Smith et al. to the $SIdaR(1,0)$ (and $SEIdaR(1,0)$ model as the latent period does not affect the branching process) model and show that similar principles still apply. There have been other studies of intervention strategies for network-based models under more restrictive assumptions such as those presented by Britton et al. (2007), where the case of a constant recovery period is considered [19]. Janson et al. (2014) study the branching process approximation to a model equivalent to $SIdaR(1,0)$ with $\delta = 0$ and give conditions for the calculation of the basic reproductive number as $R_0 = E(Z)$ where $Z$ denotes the offspring distribution of a post-index case [53]. In Chapter 6 we have explicitly computed this distribution for arbitrary recovery distribution in terms of the model parameters.

Let $\mu$ and $\sigma^2$ denote the mean and variance of the degree distribution with $\pi_k = P(D_v = k)$, i.e. the distribution from which each node $v$ draws its degree independently in forming the network. We define a viable neighbor of the post-index case as a neighbor that this individual can potentially infect. Conditional on a post-index case having $d'$ viable neighbors (i.e. he has excess degree $d'$ and was not vaccinated) and length of infectious period $t_r$ we have as in Chapter 6

$$Z|d',t_r \sim Binom(d',p_t = \frac{\beta}{\beta + \delta}(1 - exp(-(\beta + \delta)t_r))).$$

(7.21)
Let $D'$ denote the random excess degree of the post-index case and assume the infectious period $T_r$ is exponential with rate parameter $\gamma$, and define $p(T_r) = p(T_r; \beta, \delta)$ as in (6.1). Then we have:

$$E(Z) = E(E(Z|d', t_r)) = E(D'p(T_r)) = E(D')E(p(T_r)) = \left(\frac{\sigma^2 + \mu}{\mu} + \mu - 1\right) \frac{\beta}{\beta + \gamma + \delta},$$

(7.22)

where $E(D') = (\frac{\sigma^2 + \mu}{\mu} + \mu - 1)$ follows from computation of the excess degree distribution, see e.g. [19]. The form of $E(p(T_r)) = \frac{\beta}{\beta + \gamma + \delta}$ under the exponentially distributed infectious period is a standard property of independent exponential random variables. We consider two possible interventions. In the first, the perfect vaccine which prevents transmission of the infection to one’s neighbors (but does not prevent one from being infected) is distributed at random to a fraction $c \in (0, 1)$ of the population. Let $Z_{vacc}^c$ denote the offspring distribution of the post-index case under this intervention. In the second intervention, no individuals are vaccinated but we are able to reduce the edge-wise rate of transmission from $\beta$ to $\beta^* = (1 - c)\beta$ for the same $c \in (0, 1)$. Let $Z_{edge}^c$ denote the offspring distribution of the post-index case under this intervention. We compare the intervention-adjusted $R_0$ for each intervention, denoted as $R_{0\text{edge}} = E(Z_{edge}^c)$ and $R_{0\text{vacc}} = E(Z_{vacc}^c)$.

**Lemma 5.** Fix $c \in (0, 1)$ and assume the Branching Process Approximation in Chapter 6 holds. Then we have $R_{0\text{edge}} > R_{0\text{vacc}}$.

**Proof.** Under the edgewise intervention, this is the same as an epidemic in which the rate of infection is simply $\beta^* = \beta(1 - c)$ and $p_{t_r}^* = \frac{\beta(1-c)}{\beta(1-c)+\delta}(1 - e^{-(1-c)\beta+\delta)t_r})$. Therefore,
proceeding as in (7.22),
\[
R_0^{\text{edge}} = E(Z_{\text{edge}}^c|c, \beta, \gamma) = E(E(Z_{\text{edge}}^c|D' = d', t_r)) = \left(\frac{\sigma^2 + \mu}{\mu} + \mu - 1\right) \frac{(1 - c)\beta}{(1 - c)\beta + \gamma + \delta},
\]
(7.23)
where the last equality follows because of the assumed independence between the recovery process and the excess degree. On the other hand, under the vaccine intervention the post-index case has 0 viable neighbors with probability \(c\). Since the vaccine is administered at random then conditional on not being vaccinated his expected number of viable neighbors is \(\left(\frac{\sigma^2 + \mu}{\mu} + \mu - 1\right)\) as before. Hence the expected number of viable neighbors is \((1 - c)\left(\frac{\sigma^2 + \mu}{\mu} + \mu - 1\right)\) such that
\[
R_0^{\text{vacc}} = E(Z_{\text{vacc}}^c|c, \beta, \gamma) = E(E(Z_{\text{vacc}}^c|D' = d', t_r)) = \left(\frac{\sigma^2 + \mu}{\mu} + \mu - 1\right) \frac{(1 - c)\beta}{\beta + \gamma + \delta},
\]
(7.24)
and comparison of the two expressions with \(c \in (0, 1)\) shows that \(R_0^{\text{vacc}} < R_0^{\text{edge}}\) as needed.

These insights are most important when \(R_0 \approx 1\) as this is the portion of \(\Theta\) in which one intervention may make an outbreak impossible while the possibility remains under a competing intervention. A different metric that becomes relevant when both intervention-adjusted reproductive numbers exceed 1 (i.e. \(R_0^{\text{edge}}\) and \(R_0^{\text{vacc}} > 1\)) is the extinction probability under each intervention. That is, it is the probability that a chain of transmission from a post-index case dies out, i.e. the outbreak is small. From the general theory of Branching
processes, the extinction probability is the solution

\[ \rho = \psi_Z'(\rho), \] (7.25)

where \( \psi_Z' \) denotes the probability generating function of the post-index case offspring distribution. This motivates comparing the extinction probability under each intervention from a single post-index case. We continue under the assumption that the infection is passed according to an exponential distribution with rate \( \beta \) and consider an arbitrary non-negative distribution for the infectious period denoted by \( f_T(t) \).

The following result will aid in the proof of this.

**Lemma 6.** For any \( s \in (0, 1) \) let the function

\[ H(X; s) = \int_{R^+} \sum_{d' \geq 0} (q_i^X + (1 - sp_i^X)) d' \frac{(d'^{d'+1}) \pi_{d'+1}}{\mu} f_T(t) dt \]

and assume the excess degree distribution satisfies \( \sum_{d' \geq 0} 2d' \frac{(d'^{d'+1}) \pi_{d'+1}}{\mu} < \infty \) such that interchangeability holds as follows:

\[ \partial^2 H(X; s) = \int_{R^+} \sum_{d' \geq 0} \partial^2 (q_i^X + (1 - sp_i^X)) d' \frac{(d'^{d'+1}) \pi_{d'+1}}{\mu} f_T(t) dt. \] (7.26)

Then \( H(X; s) \) is convex for \( X \in [0, 1] \)

**Proof.** Consider \( d' \geq 0 \) and \( t > 0 \) as fixed. Then taking the second derivative directly yields

\[ \partial^2 H(X; s) = \frac{\beta^2 d'(s+1)t^2(e^{\beta t X}(1-s) + d'(1+s))(e^{-\beta t X}(1-s) + (1-s))d'}{(se^{-\beta t X} - e^{-\beta t X} - s - 1)^2} \] (7.27)

Clearly the denominator is positive. As well, given the constraints \( s \in (0, 1), t > 0, \) and \( \beta > 0 \) one easily checks that the numerator is greater than 0. Lastly, by the assumed interchangeability this shows that \( H \) is convex. \( \square \)
Lemma 7. Assume the Branching Process approximation to the SEIdaR(1,0) model with \( \delta = 0 \) and any non-negative infectious period distribution \( f_T(t) \), \( c \in (0,1) \) such that 
\[ E(Z_{\text{edge}}^c) > 1 \text{ and } E(Z_{\text{vacc}}^c) > 1. \]
In addition, assume for \( s \in (0,1) \) and \( X \in [0,1] \) the function 
\[ H(X; s) = \int_{\mathbb{R}^+} \sum_{d' \geq 0} (q_t^X + (1 - sp_t^X))^{d'} (d' + 1) \frac{\pi_{d'+1}}{\mu} f_T(t) dt \] 
(7.28)
is convex in \( X \), where \( p_t^X = (1 - e^{-(X\beta)t}) \) and \( q_t^X = 1 - p_t^X \). Then the solutions \( \rho_{\text{edge}} \) and \( \rho_{\text{vacc}} \) to (7.25) satisfy \( \rho_{\text{vacc}} \geq \rho_{\text{edge}} \).

Proof. We begin by computing the value of the pgf for fixed \( s \). By definition, \( \psi_Z(s) = E(s^Z) \) and so our conditioning argument presented in the previous lemma remains valuable. We proceed as follows
\[ \psi_{\text{edge}}(s) = E(s^{Z_{\text{edge}}^c}) = \int_{\mathbb{R}^+} \sum_{d' \geq 0} E(s^{Z_{\text{edge}}^c}|d', p_t^*) P(D' = d', P_t^* = p_t^*) dt, \] 
(7.29)where \( p_t^* = (1 - e^{-(1-c)\beta t)}) \). We use the following facts. Given \( d' \) excess neighbors and the recovery time \( t \), we know the random variable \( Z_{\text{edge}}^c|d', t \sim \text{Bin}(d', p_t^*) \). Therefore, 
\[ E(s^{Z_{\text{edge}}^c}|p_t^*, d') = (q_t^* + (1 - sp_t^*))^{d'} \]where \( q_t^* = 1 - p_t^* \). Also, the distribution of \( D' \) remains independent of the recovery time. Write the above as
\[ \psi_{\text{edge}}(s) = \int_{\mathbb{R}^+} \sum_{d' \geq 0} (q_t^* + (1 - sp_t^*))^{d'} P(D'_{\text{edge}} = d') P(P_t = p_t^*) dt \]
(7.30)
An identical argument shows that
\[ \psi_{\text{vacc}}(s) = \int_{\mathbb{R}^+} \sum_{d' \geq 0} (q_t + (1 - sp_t))^{d'} (1 - c)(d' + 1) \frac{\pi_{d'+1}}{\mu} f_T(t) dt \] 
(7.31)
Consider a fixed $s \in (0, 1)$. Let $X$ be a Bernoulli random variable with success probability $(1 - c)$ and define $H(X; s)$ as

$$H(X; s) = \int_{R^+} \sum_{d' \geq 0} (q_t^X + (1 - sp_t^X)) \frac{(d' + 1) \pi_{d'+1}}{\mu} f_T(t) dt$$  \hspace{1cm} (7.32)

Where $p_t^X = (1 - e^{-X\beta t})$ and $q_t^X = 1 - p_t^X$. Since we assume $H$ is convex in $X \in [0, 1]$ (shown in Lemma 6), then for any $s \in (0, 1)$, we apply Jensen’s Inequality to find

$$\psi_{\text{vacc}}(s) = E(H(X); s) \geq H(E(X); s) = \psi_{\text{edge}}(s).$$ \hspace{1cm} (7.33)

Hence, the solutions to $\psi_{\text{edge}}(\rho_{\text{edge}}) = \rho_{\text{edge}}$ and $\psi_{\text{vacc}}(\rho_{\text{vacc}}) = \rho_{\text{vacc}}$ satisfy $\rho_{\text{vacc}} \geq \rho_{\text{edge}}$ as desired.

While these studies are preliminary, we do gain some useful insights from them. We see some evidence that vaccination, when possible, may be a more effective intervention than reducing infectivity-per-contact. The success of a naive vaccination strategy also suggests it is sensible when possible to identify potential “superspreaders” (i.e. highly connected) individuals in a population beforehand and to emphasize vaccination of these individuals. Doing so should further reduce $R_0$ and the probability that a small outbreak will not turn into a large one. Identifying superspreaders in a population may be accomplished by studying identifying demographic characteristics shared by these individuals.

### 7.7 Conclusion

In this chapter we have studied statistical estimation methods for the $SEIdaR(1, 0)$ model with the understanding that analogues may be derived for the full $SEIdaR(r, k)$ model.
model with appropriate data. Basic statistical properties such as consistency and bootstrap methods for interval estimation were established. Least Squares estimators remain straightforward to use and we propose a novel approximate likelihood approach as well. We use these insights to derive estimators for key epidemiological parameters such as $R_0$ and various intervention thresholds. We conclude this section with some results concerning deployment of a fixed amount of vaccine and its impact on $R_0$. These considerations suggest that targeted intervention strategies are in general more effective than non-targeted interventions.
Chapter 8: Modeling the 2013-2016 Ebola outbreak in Guinea

We begin this section by performing a simulation study to apply the methods of the previous section with realistic parameter values for an Ebolavirus outbreak to assess parameter identifiability from incidence data and approximate rate of asymptotic convergence. We pay particular attention to standard errors of the hybrid and OLS estimator in this scenario.

Next, we apply the methods of the previous section to publicly available count data that has been digitized from WHO Ebola Situation reports. In particular, our goal will be to fit the model parameters $\theta = (\beta, \sigma, \gamma, \delta, \kappa, K)$ from this data and in certain cases other parameter estimates in the literature. From this we may estimate the basic reproductive number $R_0$ and critical thresholds for the spread of the disease. We analyze the data under the assumption that it is generated by the SEIdaR(1,0) model with $r = 1$ layer that de-activates at rate $\delta$ and satisfies the PT-condition. While similar analysis can be performed in the absence of the PT-condition there is considerable simplification in certain formulas (i.e. $R_0$) when it is satisfied. Our justification for its use lies in the previously seen datasets regarding the degree distribution in similar contact networks and the data made available
by Fallah et al. [35]. In addition, we analyze the data using both the OLS and Hybrid Estimators that were proposed in Chapter 7.

### 8.1 Simulation study

We first explore a simulation study based on realistic parameter choices from the literature. The parameters are set as \( \theta_0 = (\beta_0, \sigma_0, \gamma_0, \delta_0, \kappa_0, K_0) = (0.0458, 1, 1, 1.158, 1.6, 6.04) \). Further details of the selection of these parameter values may be found in the table below and they are seen to be consistent with the range in Chapter 2. We perform the simulation with an effective population size of \( N = 15000 \) in order to assess bias and variability of the hybrid and OLS estimator for systems of this size. Miller and others have demonstrated numerically that numerical convergence to the limiting ODE works reasonably well for systems of this approximate size [85]. The agreement between the stochastic system and limiting ODE in general will be expected to depend on the underlying parameters. Hence the choice of reasonable parameter values based on the literature.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean length of Asymptomatic Period (( \sigma^{-1} ))</td>
<td>10 days</td>
<td>[107] [26]</td>
</tr>
<tr>
<td>Mean length of Infectious Period (( \gamma^{-1} ))</td>
<td>10 days</td>
<td>[107] [62] [26]</td>
</tr>
<tr>
<td>Mean Number of Contacts (( K ))</td>
<td>6.04</td>
<td>[100] [35] [31]</td>
</tr>
<tr>
<td>Normalized Neighbor Excess Degree (( \kappa ))</td>
<td>1.6</td>
<td>[35]</td>
</tr>
<tr>
<td>Contact Duration Length (( \delta^{-1} ))</td>
<td>6.3 days</td>
<td>[31] [107]</td>
</tr>
<tr>
<td>Basic Reproductive Number (( R_0 = \frac{\beta \kappa K}{\beta + \gamma + \delta} ))</td>
<td>1.47</td>
<td>[28]</td>
</tr>
</tbody>
</table>
The value of $\beta$ is selected such that the basic reproductive number for the simulation study is $R_0 = 1.46$ and produces an attack rate similar to the one described in [31]. This compares reasonably with many estimated values for the epidemic in West Africa reported from other models in the literature [28]. We simulate $B = 700$ trajectories of the network-based $SEI{d}aR(1,0)$ model with the single layer taken to be a de-activating layer with rate $\delta$. We estimate the parameters via Least Squares using two datasets. The first method includes knowledge of the full trajectory $y^*(t_i) = \{S, E, I, SE, SI\}(t_i)$ in order to assess identifiability when all possible information is known. In the second, we withhold dyad-related information and fit the trajectory via least squares against $\{S, E, I\}(t_i)$. For each simulated epidemic this information is assumed to be known at $t_1 = 1, t_2 = 2, ..., t_m = t_{end}$ where $t_{end}$ is the last measurement before the end of the outbreak. This quantity is stochastic because the outbreak will last for a random period of time. In each case the effective population size $N$ and the initial conditions $\{[SE], [SI]\}(0)$ are also considered free parameters and fit via the least squares procedure. This is done because the initial condition itself is random (initially infective individuals chosen uniformly at random) as is the “explosion” time of the outbreak. For the Hybrid method the initial condition and effective population size are likewise fit from the data via profiling. The data in this case is assumed to be the entire process history $\mathcal{H}_t$ up to the end of the outbreak consisting of the event time and event type (exposure, symptom onset, recovery) and estimation is done by solving (7.9).
Below, the deterministic trajectory is plotted against 700 realizations of the stochastic \textit{SEIdaR}(1,0) model with the given parameters. The trajectories are aligned to have peak incidence occurring at approximately the same time.

![Number of Infected Individuals vs Time](image)

Figure 8.1: The estimated (limiting) number of infected individuals over time for the simulated dataset with parameters in Table 8.1

**Estimation with Least Squares**

The bootstrap simulation procedure is performed and the data are fit using OLS with both sets of information described above. First we consider the case of complete data. 2-D projections of the bootstrap estimates for each trajectory are given below in order to assess
parameter identifiability. The scatterplots below suggest reasonable identifiability of the model parameters and initial condition for the values selected in the simulation study.

Figure 8.2: The bootstrap estimates of \( \theta \) from the OLS estimator with complete (including dyad) information.

Moreover, we may look at the estimates of \( R_0 \) to assess the variability in estimating this quantity from the simulated trajectories. By the bootstrap algorithm in chapter 7 a 95% confidence interval for \( R_0 \) may be constructed by looking at the appropriate quantiles of the parameter estimates from the bootstrap trajectories. For the OLS estimator with complete
information the mean bootstrap estimate of $R_0 = 1.43$ is and the estimated standard error is .053. This results in a bootstrap confidence interval of $(1.33, 1.53)$.

For the OLS estimator with partial ($S, E, I$) information the mean bootstrap estimate of $R_0 = 1.42$ is and the estimated standard error is .060. This results in a bootstrap confidence interval of $(1.32, 1.56)$ which agrees approximately with the above. Below, we give the estimated values of $R_0$ using the OLS estimator with complete (dyad) information.

A summary table comparing the standard error of the estimate for each parameter and a comparison with the hybrid method is given in the next section. In general, we find that
both OLS methods have low bias and acceptable variability for populations of this size in this region of the parameter space.

8.1.1 Estimation with the Hybrid Method

We also apply the hybrid method of the previous section with using the observed history $H_t$. A priori, we perhaps expect that this method produces similar performance in terms of bias to the OLS estimators but with a smaller standard error than OLS with only the \{S, E, I\} information. The reason for this being that the hybrid method uses the form of the limiting equation to approximate $[SI](t)$ directly. On the other hand the OLS estimator with complete information should perform better due to more information available. While not a formal E-M argument the hybrid performs a similar function when dyad information
is missing—approximating the missing \([SI]\) quantity with its expectation conditional on the process history contained in \(S(t)\). We assess identifiability of the hybrid estimator in the figure below.

![Hybrid Estimates from Event Data](image)

**Figure 8.5:** The bootstrap estimates of \(\theta\) from the hybrid estimator.

The estimated value of \(R_0\) is 1.43 and the resulting approximate 95% confidence interval is \((1.26, 1.55)\). It is seen that the standard error for this quantity is slightly larger than the OLS estimators.

The results in the table yield several interesting observations. First, that the resulting estimators have low bias even when \(N \approx 15,000\) and therefore convergence happens quickly for real-world applications. Second, that our hypothesis regarding standard errors
Table 8.2: Table of parameters, estimated value, and standard errors for the three methods in the simulation study. OLS(P) denotes the least squares estimate with partial (i.e. excluding dyads) information while OLS(C) denotes the least squares estimate with complete (i.e. including dyads) information.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>OLS(P) Mean</th>
<th>OLS(C) Mean</th>
<th>Hybrid Mean</th>
<th>OLS(P) SE</th>
<th>OLS(C) SE</th>
<th>Hybrid SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>.0458</td>
<td>.0450</td>
<td>.0458</td>
<td>.0457</td>
<td>.0113</td>
<td>.0011</td>
<td>.0024</td>
</tr>
<tr>
<td>$\delta$</td>
<td>.158</td>
<td>.155</td>
<td>.158</td>
<td>.1611</td>
<td>.066</td>
<td>.0072</td>
<td>.0252</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>1.60</td>
<td>1.58</td>
<td>1.58</td>
<td>1.58</td>
<td>.122</td>
<td>.113</td>
<td>.124</td>
</tr>
<tr>
<td>$K$</td>
<td>6.04</td>
<td>6.12</td>
<td>6.00</td>
<td>6.03</td>
<td>1.01</td>
<td>.314</td>
<td>.624</td>
</tr>
<tr>
<td>$R_0$</td>
<td>1.47</td>
<td>1.43</td>
<td>1.43</td>
<td>1.42</td>
<td>.060</td>
<td>.053</td>
<td>.069</td>
</tr>
</tbody>
</table>

of individual parameters appears to be correct, that is, the OLS estimator with complete data produces smaller standard errors than both other methods. Likewise, the hybrid methods seems to produce smaller standard errors than the OLS estimator with partial data. However, the estimate of $R_0$ is interesting in that all three estimators produce somewhat comparable standard errors. This seems due to $R_0$ being a highly non-linear function of the individual parameters. Further study of this phenomenon is suggested for future work as well as analytical results concerning the variance of these estimators. As suggested by the results of Rempala (2012) this should be achievable with a diffusion approximation for the $SEIdaR(1,0)$ process. All three methods appear reasonably stable in terms of parameter identifiability. We now apply a similar study to the dataset from Guinea seen earlier in this section.
8.2 Data description

We fit data from the 2013 – 2016 West Africa outbreak obtained from the Humanitarian Data Exchange [1]. In both cases the data are in the form of the number of cumulative confirmed, suspected, and probable cases reported over the course of the epidemic and have been digitized from the WHO situation reports. From this, we may obtain the approximate “complete” data \( \mathcal{H}_t \) described in the previous chapter as follows. For each time interval \([t_i, t_{i+1}]\) we assume that the time of symptom appeared uniformly in the interval for each new case, as in [101]. For an individual \( k \) let \( t_k^I \) denote the time at which symptoms first appeared and assume that the recovery parameter \( \gamma \) and the incubation period parameter \( \sigma \) are known. The complete data corresponding to individual \( k \) is the triple \((t_k^E, t_k^I, t_k^R)\) denoting the time of exposure, infection (symptom onset), and recovery respectively. When \( t_k^I \) and the rates \( \sigma \) and \( \gamma \) known, then this triple may be well approximated by the triple \( D_k = (D_k^E, D_k^I, D_k^R) = (t_k^I - I_1, t_k^I, t_k^I + I_2) \) where \( I_1 \) is drawn from an exponential(\( \sigma \)) distribution and \( I_2 \) is drawn from an independent exponential(\( \gamma \)) distribution. Performing this imputation for each individual \( k = 1, \ldots, F \) that is infected up to time \( T \) approximates the complete data \( \mathcal{H}_t \approx \{D_k\}_{k=1}^F \).

The data required for the least squares fitting (the S,E,I species) may be obtained from \( \mathcal{H}_t \) by noting

\[
I(t_i) = \sum_{k=1}^F \mathbf{1}_{d_k^I \leq t_i \leq d_k^R}.
\]

(8.1)

This is the approach taken later in this section to computing \( \hat{\theta}_{OLS} \).
8.3 Ebolavirus in Guinea

Using estimates published by the WHO and several other sources listed earlier we approximate the rate of symptom onset as $\sigma_0 = .1$ and the rate of recovery by $\gamma_0 = .1$. This implies the mean duration of each period is 10 days and that the median is 6.93 days. Using the data in the previous section we apply both the OLS and Hybrid estimation methods of Chapter 7 to this dataset. Our primary interest lies in parameter estimates of the free parameters $\theta_{free} = (\beta, \delta, \kappa, K) \in \theta$. An appealing benefit of parameterizing the degree distribution in this way is that it commits us only to the family of PT-distributions rather than a single one (i.e. Poisson, etc.). In addition, the initial conditions $((SE_0, SI_0))$ and effective population size ($N$) are both taken as free parameters and fit via likelihood profiling, as suggested in Schwartz et al. [101].

8.3.1 Estimates with hybrid method

The hybrid method (i.e. maximization of the Hybrid Likelihood of Chapter 7 ) yields a point estimate $\hat{\theta}_{hyb} = (\hat{\beta}, \hat{\delta}, \hat{\kappa}, \hat{K}) = (.0439,.2987,1.19,10.27)$ and therefore an estimated reproductive number of $\hat{R}_0 = 1.22$ using the plug-in estimator of the previous chapter. The effective population size is estimated to be $N_{eff} = 15001$ and the initial condition is estimated as $\{SE_0, SI_0\} = (1,50)$. To estimate standard errors for relevant quantities the bootstrap algorithm of the previous chapter was employed with $B = 2500$ replications in order to obtain a small Monte Carlo error ($\approx 2\%$). For each bootstrap replicate $b = 1, \ldots, B$ the data was refit via the hybrid method with the parameters $\sigma = .1$ and $\gamma = .1$ held fixed and
Figure 8.6: The estimated number of infected individuals over time for the Guinea dataset when $\gamma = \sigma = .1$.

the replicate $\hat{\theta}_b$ obtained. To obtain point estimates for the basic reproductive number ($R_0$) and critical thresholds regarding control ($\beta_{\text{crit}}, \delta_{\text{crit}}, K_{\text{crit}}$) of the epidemic, the consistent point estimates of each were applied. Thus for each bootstrap replicate we obtain

$$R_{0b} = \frac{\hat{\beta}_b \hat{\kappa}_b \hat{K}_b}{\hat{\beta}_b + \gamma_0 + \delta_b},$$

and this yields a 95% confidence interval of $(1.09, 1.33)$ for $R_0$. The approximate density of $R_0$ is given below.

This interval is consistent with other estimates of the reproductive number for Guinea which has generally been estimated to be lower than for Liberia and Sierra Leone [63] [28]
Figure 8.7: Estimated density of $\hat{R}_0$ for the Guinea dataset when $\gamma = \sigma = .1$.

[107]. Given this information we may also estimate the critical intervention thresholds required to bring $\hat{R}_0$ to unity. As a reminder from the previous chapter the critical thresholds of infection, contacts, and quarantine are given by

\[ K_{\text{crit}} = \frac{\beta + \gamma + \delta}{\kappa \beta} \]  
\[ \beta_{\text{crit}} = \frac{\gamma + \delta}{\kappa K - 1} \]  
\[ \delta_{\text{crit}} = K \kappa \beta - \gamma - \beta. \]

For the Guinea dataset we estimate these critical thresholds to be $\hat{\delta}_{\text{crit}} = .3921$, $\hat{\beta}_{\text{crit}} = .0356$, and $\hat{K}_{\text{crit}} = 8.47$. The quantity $\delta_{\text{crit}}$ represents the rate of contact de-activation (i.e. isolation) that needs to be achieved to bring $R_0$ to unity. Therefore we may interpret this
as saying that individuals should be quarantined in approximately 2.5 days in order to stop further spread of the epidemic. The quantity $K_{crit}$ similarly represents the number of contacts (with $\kappa$ constant) that would achieve the same effect. That is, the mean number of contacts needs to be reduced to 8.47 with its skew held roughly constant in order to stop further spread of the epidemic. The quantity $\beta_{crit}$ suggests that reducing the per-contact rate of infection by about 19% may be sufficient to stop future outbreaks. We may also compare the bootstrap trajectories against the observed data in order to assess model fit. In order to account for stochastic effects in the time to explosion the bootstrap trajectories are aligned so that peak incidence occurs at approximately the same time.

![Simulated Trajectories vs Guinea](image)

Figure 8.8: Estimated bounds for the number of infected individuals over time for the Guinea dataset when $\gamma = \sigma = .1$. 

169
We see that the simulated trajectories from the $SEIdaR(1, 0)$ model appear to fit the data well. Lastly we may assess the identifiability in this region of the parameter space by looking at bivariate projections of the bootstrap estimates. In particular if certain parameters are not identifiable in this region this should be visible.

Figure 8.9: Bivariate scatterplots of the bootstrap estimates from simulated trajectories of the $SEIdaR(1, 0)$ model with parameters $\hat{\theta}_{hyb}$.

The plots indicate that the 4 free parameters appear identifiable. In particular, the lack of strong correlations between point estimates in the various bootstrap trajectories gives no cause for concern regarding identifiability of these parameters. This gives confidence in the validity of the analysis of the Guinea dataset using the hybrid method. We next perform a similar analysis making use of the OLS estimator with E and I information.

170
8.3.2 Estimates with least squares method

We consider the same dataset for Guinea as in the analysis using the Hybrid Estimator. The same set of simulated trajectories is used for the fitting and to accommodate the stochastic time to deterministic behavior, the initial condition is considered free.

As before we compare the results in terms of standard errors for individual parameters and for $R_0$. These results are summarized in the table below.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Point Estimate (OLS)</th>
<th>Point Estimate (Hybrid)</th>
<th>OLS 95% CI</th>
<th>Hybrid 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>.0435</td>
<td>.0439</td>
<td>(.031, .058)</td>
<td>(.038, .055)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>.309</td>
<td>.299</td>
<td>(.165, .440)</td>
<td>(.225, .381)</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>1.27</td>
<td>1.19</td>
<td>(1.03, 1.51)</td>
<td>(.954, 1.58)</td>
</tr>
<tr>
<td>$K$</td>
<td>10.2</td>
<td>10.3</td>
<td>(8.06, 12.16)</td>
<td>(6.61, 14.51)</td>
</tr>
<tr>
<td>$R_0$</td>
<td>1.25</td>
<td>1.22</td>
<td>(1.14, 1.34)</td>
<td>(1.09, 1.32)</td>
</tr>
</tbody>
</table>

Table 8.3: Table of parameters, estimated value, and bootstrap confidence intervals for the OLS and hybrid estimators applied to the Guinea dataset.

It is seen that once again the hybrid estimator and OLS estimator produce similar (though not identical) analyses of the Guinea dataset. Our simulation study suggests that both methods do a reasonable job in comparison against the “best case” estimator (when dyad information is observed) in terms of both bias and standard errors. Our conclusion is that both methods offer reasonable analyses of the Ebolavirus outbreak in Guinea and that the interpretations provided under each analysis remains valid.
8.4 Conclusion

Our results from both studies indicate that the estimation methods presented in Chapter 7 are suitable for analyzing real-world epidemics at the population level. In contrast to the work of Scarpino et al., we have studied conditions under which the pair approximation is valid and have rigorously studied the properties and variability of our estimators [100]. Using both estimators we are able to closely approximate the observed dynamics witnessed in Guinea over the course of the outbreak. Our numerical results are consistent with estimates of similar quantities seen elsewhere, however, our model provides a more mechanistic description of epidemic dynamics and epidemic stochasticity than many others. Future work on diffusion approximation to the SEIΔaR model will provide further analytical insight into this.

Moreover, we introduce the study of critical thresholds to control epidemic spread and apply these the West Africa epidemic in order to estimate the level of Public Health intervention needed to limit the spread of the virus. We accompany these quantities with consistent estimators that follow from the results of Chapter 7. Our results suggest that efficient isolation of sick individuals remains perhaps the simplest route to containing the spread of the disease. Finally, given recent studies on ring vaccination clinical trials we believe that our methods are likely suitable for future analyses in this context. Our methods allow estimation of the per-contact rate of infection that seems to be of interest in these studies [47].
Chapter 9: Summary and Conclusion

We conclude our study of Ebolavirus transmission dynamics and leave some final thoughts relevant to modeling the spread of other illnesses.

9.1 Network-Based Epidemic Models

In this work we present several contributions to the study of network-based epidemic models and their applications in understanding the spread of Ebolavirus. These contributions include (i) a general Law of Large Numbers for epidemics on networks with multiple types of dynamic connections, (ii) new insights concerning the validity of pair approximation methods, (iii) a computationally efficient stochastic approximation via a coupled Markovian system (the “hybrid” process), (iv) statistical estimation methods for data collected in the early phases of an outbreak, (v) statistical estimation methods for large-scale outbreaks in large populations, and (vi) new analytical results concerning public health interventions and the effect on $R_0$ as well as the extinction probability of the outbreak early
on. We have applied these methods to independent outbreaks of Ebolavirus in the Demo-
cratic Republic of the Congo and Guinea to demonstrate that they provide useful models of
Ebola transmission dynamics, spread, and forecasting.

Nevertheless, there are many topics concerning network-based epidemic models that
have not been addressed in this work. To begin, empirical networks often contain a variety
of motifs and clustering that is not fully captured by Configuration Model random networks
[84]. Specification of probabilistic models of empirical networks remains an open and very
challenging question in network science that has yet to be resolved [95]. It is inevitable
that these higher-order properties of network structure will impact transmission dynamics
as well and this leaves further open problems concerning limiting dynamics and estimation
for such network models. The recent goal of Miller and Kiss to understand dynamics on
networks of arbitrary structure is, in our view, goal-worthy [84]. Thus, it is of interest to
apply the methods in this document to other situations—those with subpopulations, multiple
stages of infectiousness, time-varying risk of transmission, other network structures, etc.

We still lack an analytical description of the variability of network-based epidemic
models in the way that a Central Limit Theorem may be derived for spatially homoge-
neous models. An attempt has been made at such a result in [43] but this applies only
to the very beginning of the epidemic. Such a result remains similarly useful in deriving
analytical results for the variability of estimators, as in [98]. Similarly, there are diseases
with more complex mechanisms of transmission that are not adequately described by the
framework in this document. The most pressing example is Zika Virus which appears to
be transmissible both through the environment via the Aedes Egypti mosquito and through human-to-human sexual transmission. Thus it is seen that further generalization to this model that will allow both network-based transmission of an illness between humans and some other environmental mechanism is desirable.

9.2 Future Ebolavirus Outbreaks

While zoonotic events remain rare it is now clear the Ebolavirus is capable of causing mass infections in human populations. The epidemic has also brought to light the need for coordination across levels of a Public Health response. That is, quantitative models can be used to inform public policy and assess risk in useful ways, such as in quantifying the role of Ebola Treatment Units in stopping the spread of the illness. Moreover, contact tracing and detailed record collection are generally standard responses to ongoing epidemics; however, unfortunately much of this data has still not been made available on a public level. The international response of the mathematical modeling community making use of publicly available data was unprecedented. Thus, we suggest that data collection, curation, and dissemination be considered a priority during future outbreaks of similar infectious diseases while still ensuring that the personal information of those affected first hand is protected.

There is still work to be done in producing more descriptive models of Ebolavirus itself. The $SEIdra(r,k)$ framework is flexible enough to explain the individual contributions of funerary, community, and hospital transmission to the overall dynamics. However, suitable data from which to estimate these individual components is currently lacking for the West
Africa outbreak. A more detailed description of each of these modes of transmission in a network-based framework may yield novel insights that the one-layer model does not.

In conclusion, network-based models open up new possibilities for gaining insight into infectious disease dynamics. Nevertheless, many mathematical challenges remain as new diseases are encountered and new models of human contact networks arise. Future research in this area holds potentially great rewards in terms of public health policy and disease prevention.
Bibliography


185


