2022

Modeling the Heterogeneous Temporal Dynamics of Epidemics on Networks

Andrea Joan Allen
University of Vermont

Modeling the Heterogeneous Temporal Dynamics of Epidemics on Networks

A Thesis Presented

by

Andrea Joan Allen

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Master of Science
Specializing in Complex Systems and Data Science

October, 2022

Defense Date: September 12th, 2022
Thesis Examination Committee:

Laurent Hébert-Dufresne, Ph.D., Advisor
Jeff Frolik, Ph.D., Chairperson
Chris Danforth, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

Mathematical models of infectious disease are important tools for understanding large-scale patterns of how a disease spreads through a population. Predictions of trends from disease models help guide public health prevention and mitigation measures. Most simple disease models assume that the population is randomly mixed, but real-world populations exhibit heterogeneous patterns in the way people interact. These differences in population structure can be represented by networks. Networks can then be incorporated into disease models by using various interdisciplinary concepts and tools. Yet even network disease models often overlook that populations change over time. In this thesis, two models of infectious disease are presented, for the purpose of analyzing how the spread of the disease evolves over time, particularly when the population is also changing.

To model a changing population, a sequence of different networks can each be associated with a length of time each is active for. Although, how to construct these networks from real contact data, from things like wearable sensors, is a nontrivial problem. We present a method to ascertain if temporal data can be aggregated into a single network, or not. This method underlies an algorithm for compressing real data into a time-varying sequence of networks, creating a system still tractable enough to use existing network analysis tools. We show how fine-grained temporal contact data can be compressed into just a handful of ordered, static networks while preserving the most significant temporal trends of the dynamic population.

Not only do populations change over time, but there is also inherent randomness involved in the spread of disease between individuals. To account for this, the underlying random process can be used as the basis for the disease model. Here, one particular model is presented that uses a random, or stochastic, framework to predict the temporal evolution of the spread of disease by tracking generations of infected individuals over time. We show that often the distribution of cumulative infections is heavy tailed, implying that deterministic models of spread, which present average point estimates, do not account for underlying uncertainty.

The two models presented in this thesis address the heterogeneity of the temporal dynamics of infectious disease spread through a population. These models also contribute to a body of work focused on designing models that can leverage real data about population structure and contact patterns to produce more accurate predictions and insights.

Citations

Material from this thesis has been published in the following form:

Allen, A. J., Boudreau, M. C., Roberts, N. J., Allard, A., & Hébert-Dufresne, L.. (2022). *Predicting the diversity of early epidemic spread on networks.* Phys. Rev. Research, 4(1), 013123.

Allen, A. J., Moore, C., & Hébert-Dufresne, L.. (2022). *A network compression approach for quantifying the importance of temporal contact chronology.* arXiv:2205.11566.

There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after.

-J.R.R. Tolkien, *The Hobbit*

Acknowledgements

First and foremost, I would like to thank my advisor, Laurent Hébert-Dufresne, Ph.D., who guided me through creating this thesis and its accompanying body of work. I am immensely appreciative of his patience with and trust in me throughout my time as a graduate research assitant, and thankful for his support of my career path as it winds in and out of academia.

I thank my co-authors Mariah C. Boudreau and Nicholas J. Roberts for their camaraderie throughout our research, writing, and publication process for the paper included in this thesis. I extend my gratitude to all of my collaborators and friends at the Laboratory for Structure and Dynamics, in the Complex Systems Center at the University of Vermont (UVM), for listening to my work and offering ideas and validation over the past two years.

I would also like to thank Juniper Lovato at the Complex Systems Center at UVM and Cris Moore, Ph.D., at the Santa Fe Institute (SFI), for their past and continued support of my work. Both of them, along with Dr. Hébert-Dufresne, were responsible for piquing my interest in complex systems science through the Research Experience for Undergraduates (REU) program at SFI in 2016 during my undergraduate years. Their guidance, openness, and trust allowed me to pursue the field of complexity science over the years preceeding and culminating in my graduate degree.

Lastly, I sincerely thank the members of my thesis defense committee for their time spent reviewing my work. I am deeply thankful to all who have been involved with and supportive of my work during my time at UVM.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The role of mathematical modeling in the study of infectious diseases is to gain insight about, and make quantifiable predictions for, the nature of the spread of a disease through a population. Mathematical models can help predict if, when, and how fast the disease will move through a population, which guides public health preparedness and mitigation measures. A variety of modeling tools have been developed to assess if an infectious disease will become an epidemic, and if so, how large the epidemic will be, how quickly it will spread, and what the dynamics will look like as the disease moves through a population. Modeling disease spread through a population is a highly complex task, as there are a vast amount of possible models and parameters to choose from. Often overlooked in simple models is the fact that populations change over time, affecting the way a disease can spread. In addition, disease spread is a process involving some inherent randomness. This thesis addresses two families of models best suited for dealing with these themes, in the hope of contributing to the development of mathematical disease models that can more accurately capture the complex nature of disease spread.

The challenge with any mathematical model of disease spread is to develop one that is tractable and interpretable, while also capturing enough detail about the structure and dynamics of the population and disease to obtain accurate estimates that support the questions of interest. All models are built differently, with a broad range of applications in mind: some were developed to primarily predict the end state of an epidemic, while other models are best suited for studying the course that the spread of a disease takes along the way. As the statistician George Box famously said, "All models are wrong; some models are useful" [1]. To extend this, there is no perfect model for all situations and problems.

Mathematical disease modeling goes back over three hundred years, experiencing a resurgence in the twenty-first century under the umbrella of complex systems and network science. Over the past two decades, the field of network science has been thoroughly developed and applied to the study of infectious disease spread. As a naturally interdisciplinary problem, disease modeling using network science has been developed using tools from biology and mathematics, along with physics, statistics, and engineering disciplines. Each of these individual disciplines has contributed a set of tools to the field of network disease modeling to harness the complexity of the problems of interest and work toward solutions. As a result, many of the concepts in this thesis are drawn from statistical physics and stochastic process theory along with general mathematical concepts.

This thesis focuses on two problems: (i) understanding the inherent randomness of disease spread and (ii) handling the structural and temporal variations in the population. We discuss two principled approaches for how to model the spread of the disease through heterogeneously structured populations, balancing tractability, precision, and capturing uncertainty, to understand the temporal evolution of the disease.

1.1 Mathematical Models of Disease

The earliest recognized account of mathematical disease modeling took place in the 18th century, with a model introduced by Daniel Bernoulli for smallpox, which described the increase in average life expectancy as a consequence of widespread inoculation against the disease [2]. Some two hundred years later, the foundations of modern mathematical disease models were developed by Kermack and McKendrick from 1927 to 1933, in a series of publications which introduced the concept of compartmental disease models [3, 4, 5]. In these models, the population of interest is divided into categories pertaining to individuals' disease status. In the simplest models, that is S (susceptible), I (infectious), and R (recovered or "removed"). These models are suitable for infectious disease in which the disease can be passed from one infected individual to another, and an infected individual is infectious for some period of time, after which they either recover and remain immune, die or become removed from the population – the SIR model, or become susceptible once more– the SIS model. These classical compartmental models have been used as the basis for mathematical epidemiological modeling work well into the next century [6, 7, 8, 9, 10].

The SIS and SIR models, and their variations, were built on the compartmental framework to describe how the fraction of the total population in each compartment evolves over time. The models are parameterized by rates pertaining to biological characteristics of the disease, namely the infection rate (β) and recovery rate, (γ). In the simplest models, these two rates are incorporated into sets of ordinary differential equations which govern the flow of the fraction of the total population in a given epidemiological state at a given time. Under the SI and SIR models, all nodes that eventually become infected subsequently recover (or are removed). In the SIS model, after initial growth of the infected compartment, a steady or endemic state is reached, where every new infection is matched on average by a recovered node so that a constant fraction of the population is infectious at any given time.

1.2 Modeling Epidemics on Networks

The basic compartmental models make one main assumption, which allows for deriving tractable models but is largely unrealistic. The basic models assume that the population is homogeneously mixed [7], meaning all individuals are equally likely to interact with any other. Under this assumption, the disease spreads through the population at a rate proportional to the fraction of infected individuals in the population, which corresponds to the law of mass action for chemical systems [8], when the population is large enough [11]. This representation allows for the formalization of the disease spread as a system of differential equations. Homogeneous mixing, while making for simpler models, is unrealistic, as not all members of a large population are equally likely to interact with one another. In general, populations experience some form of contact structure [12, 13, 14], and the crucial conclusions drawn from epidemiological models will yield different results, depending on the contact structure taken into account [15].

Network structures are a way to encode heterogeneity in contact patterns [16, 13, 17]. Empirical studies have found that many real-world networks have heterogeneous degree distributions, meaning that not all members of the population have the same or close to the same number of contacts. Real networks in biology, sociology, and technology exhibit heterogeneous structure [12, 18, 17, 19]. In another study, networks were derived from population mobility and census data [20] and the resulting contact network was found to exhibit properties of a small-world network [18], in which most individuals interact with a small, local neighborhood, with few long-range connections bridging the gaps between them. These results are precisely why using networks for disease modeling is a solution to relax the homogeneous mixing assumption of traditional compartmental disease models.

1.2.1 Network Definitions

Network science has its origins in the field of graph theory, dating back to the well-known mathematician Leonhard Euler in the 1700s [21]. Random graph theory underlies the fundamental concepts of the families of networks we consider [22]. Network theory as an independent field is more recent, with the bulk of modern research in the field emerging in the late 1990s and early 2000s, and is often concerned with the network being the structure on which some other process plays out [23].

A network is defined by a collection of *nodes* that represent individuals, and the *edges* between them, representing connections. A node can represent a person or animal for physical contact networks, a web page or server for the Internet, for some common examples. Edges between nodes represent a connection, contact, or interaction between two nodes.

The *adjacency matrix* A of a network is the NXN matrix in which the entry A_{ij}

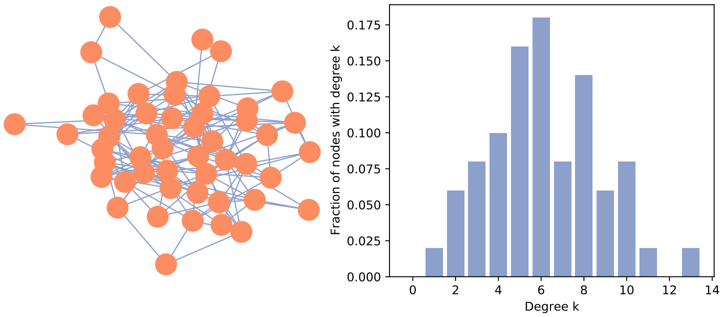

Figure 1.1: A sample network (left) and its degree distribution (right).

denotes the presence (and sometimes direction or weight) of the edge between node i and node j. The *degree* of a node is the number of edges it has, equal to the sum over A_{ij} for a given node i. An example of an adjacency matrix for a network of 4 nodes would be

$$A_{ij} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

The *degree distribution* of a network is the probability distribution over the degrees of the nodes in the network, as shown in Fig. 1.1. When selecting a node i at random, the probability that node i has degree k is proportional to the fraction of nodes in the network that have degree k. The degree distribution is a summary statistic of a network, and is a practical alternative network descriptor than the full exact adjacency matrix.

A *temporal network* refers to a network in which the connections between nodes change over time. There are many ways to represent temporal networks, but in this thesis, we define a temporal network as a sequence of *snapshots*, which are each a static or fixed network that is active for some duration of time. A temporal network's snapshots can be easily defined as a sequence of adjacency matrices.

1.2.2 Modeling Epidemics on Networks

Incorporating a network to represent more detailed contact structure can help make a disease model more accurate, albeit more complex. Modeling epidemics on networks is a broad field with many branches and approaches [24]. Two key frameworks divide the literature of epidemic modeling on networks into deterministic and stochastic models, which are deeply related. Each have their advantages and disadvantages; deterministic models capture the average behavior of the epidemic, while stochastic models account for the randomness of spread.

In almost all models, the goal is to track the size of the S (susceptible), I (infectious), and R (recovered) compartments. Deterministic models use differential equations to track the time derivatives of the sizes of the compartments that change given the flow between them, governed by infection and recovery rates β and γ that come from the biological parameters. The differential equations can be solved for the temporal evolution of the sizes of the compartments, as well as for threshold conditions for the spread of the disease, namely the basic reproduction number, R_0 [25, 26], which is the average number of infections caused by a single individual over the course of the epidemic. A deterministic model is used to treat temporal networks in Chapter 2.

Alternatively, the compartmental models can be generalized as stochastic processes [27, 28], and analyzed by borrowing techniques from statistical physics. For the SIR model, the infection and recovery of individuals are governed by a Poisson process that results in Markovian dynamics for each node's trajectory through the S, I, and R compartments. In the stochastic case, now the sizes of the compartments are represented by discrete random variables, with $P(I(t) = j) = p_j(t)$, where $I(t)$ is the size of the infected compartment at time t. Infection is transmitted at a rate proportional to β in the deterministic model, and recovery happens at rate γ for an individual. Therefore, the state of the whole system can be modeled as a Markov chain [28], where the state is defined by the sizes of the three compartments, $(S(t), I(t), R(t))$ at a given time t. The transition $(S(t), I(t), R(t)) \rightarrow (S(t) - 1, I(t) + 1, R(t))$ happens with some probability – derived from the corresponding rates and current state of all the nodes in the system [29].

One can follow the time evolution of the entire system using master (or Kolmogorov) equations that describe the probabilities for every value the state of the system can take. Alternatively, one can also treat the analysis of the stochastic model using branching processes, which are informed by summary statistics of the network like the degree distribution. This method is discussed in detail in Chapter 3.

Stochastic and deterministic models are fundamentally linked and one can be derived from the other mathematically using the master equations and underlying Markov process of the stochastic framework. At their core, both families of models describe the same system. The choice to use a stochastic framework versus a deterministic one lies with the problem under study.

With larger quantities of data supporting a model, or if the research question surrounds mainly tracking the time evolution of a spreading process, deterministic tools can be readily applied. The temporal evolution of the spread of the disease can be easily modeled with a deterministic model. However, the results are given in single point-estimates for each time point. Stochastic models address the fact that in reality, an $R_0 > 1$ does not necessarily mean an epidemic is for certain. Stochastic models produce probability distributions for the size of the epidemic, instead of a point estimate that represents the mean. Stochastic models may be preferable when less is known about the disease, population structure, or other missing data, where it may be helpful to have more than just a single prediction or point estimates explaining the expected spreading process outcome.

With deterministic network epidemiological models, the underlying stochastic processes are approximated to their average, or expected, collective behavior, resulting in a consistent solution based on the model's inputs. Within both sets of approaches there are a myriad of model techniques that balance tractability with accuracy. In this thesis, we look at a deterministic *mean-field* model to take advantage of the simplicity of deterministic models while incorporating network structure. Then we use stochastic model for understanding the uncertainty of spread during the early times of a disease spread with little data available and a non-detailed contact structure.

1.3 Deterministic Mean-Field Models of Disease Spread on Networks

The classic SIS and SIR models discussed in Sec. 1.1 are designed to approximate the expected collective behavior of the stochastic events of individual nodes and contacts by applying the law of mass action, and to deterministically model the evolution of the population compartments. As noted, a deterministic model is one in which the same inputs and initial conditions lead to the same outcome on every instance of the model.

Deterministic disease models are usually set up with a set of differential equations as a simple way to track the flow of the SIR/SIS compartments. Eq. (1.1) give the three coupled nonlinear differential equations for a basic SIR model that describe the rate of change of the sizes of the three compartments. The solution to Eq. (1.1) is shown in Fig. 1.2.

$$\frac{dS_t}{dt} = -\beta(I_t S_t), \qquad \frac{dI_t}{dt} = \beta(I_t S_t) - \gamma I_t, \qquad \frac{dR_t}{dt} = \gamma I_t \tag{1.1}$$

Nodes recover at rate γ, and infection transmits between infected and susceptible individuals at rate β. It is easy to see from Eq. (1.1) how the fully-mixed assumption works: the contact rate, β, is applied to the number of pairs of nodes where the disease can be transmitted, one from the susceptible compartment and one from the infected compartment. In this model, on average, every infected node has a disease-causing contact with every susceptible node, resulting in $I_t S_t$ pairs.

There are many ways to relax the fully-mixed assumption of the basic models and incorporate network structure while still maintaining a relatively simple model. In Chapter 2, we use the explicit *adjacency matrix* of the network in order to capture the exact relationships between nodes, and use a deterministic framework to approximate the collective stochastic behavior of individual nodes. This way, we utilize the power of deterministic modeling without assuming a fully-mixed population.

This model, known as the *quenched mean-field* model, is described by a set nonlinear differential equations defined for each node. At time t, the rate of change of

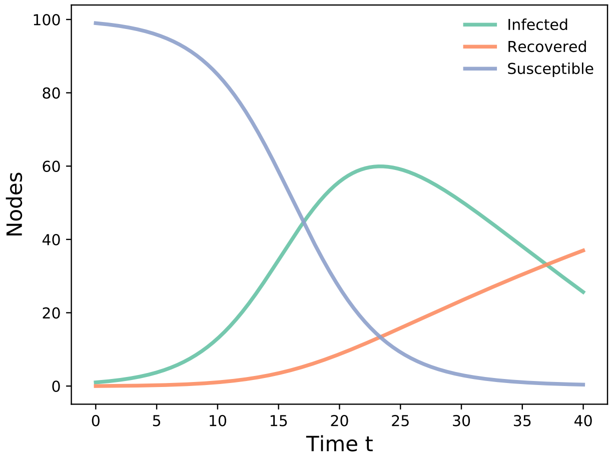

Figure 1.2: Solution of an SIR disease model.

the probability of node i being infected is

$$\frac{dp_i}{dt} = (1 - p_i(t))\beta \sum_j A_{ij} p_j(t) - \gamma p_i(t) \tag{1.2}$$

where $A_{ij} = 1$ if there is an edge between node i and j, and 0 otherwise. The solution to the node-wise probability of being infectious over time is illustrated by the top panel of Fig. 1.3. For the whole population, the set of equations can be defined as a vector

$$\frac{dP_t}{dt} = (1 - P_t)\beta A P_t - \gamma P_t \tag{1.3}$$

where $P_t = [p_0(t), p_1(t), ...p_N(t)]$. From there, the size of the infected compartment at time t is

$$I_t = \sum_N P_t, \tag{1.4}$$

depicted in the bottom panel of Fig. 1.3.

This type of model falls into the category known as a mean-field deterministic

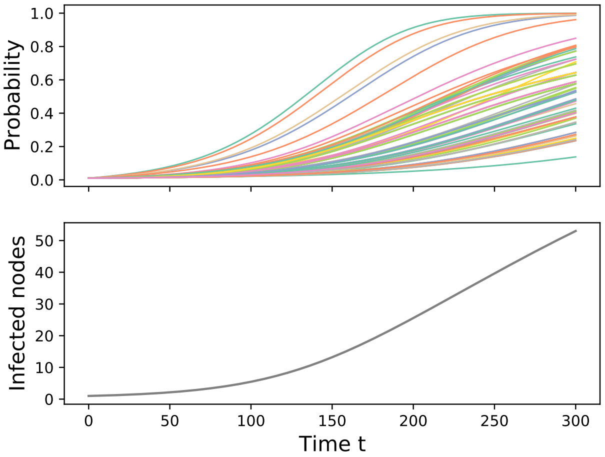

Figure 1.3: Solution of a quenched mean-field model on a network. Top panel shows the time evolution of the probabilities of each node being infectious at time t. The bottom panel shows the time evolution of $\sum P_t$, the total number of expected infectious nodes at time t.

model. Mean-field models reduce the complexity of a system by making approximations over sub-systems or identifiable components. Mean-field theory, borrowed from statistical physics and widely adapted for network science, is the theory of reducing a high-dimensional complex system to a lower dimension for tractability, where the collective behavior of a system or sub-systems is introduced in place of keeping track of a high number of individual components. Tools from statistical physics have revamped the ability to model contagion models [30, 31, 32], since the spread of a disease over a network of contacts is similar to models of non-equilibrium problems in statistical physics [33].

A *quenched* mean-field theory approach assumes that the network is static or fixed, in contrast, an *annealed* approach assumes that while the degree distribution of the network is fixed, specific contacts between nodes are not fixed and the identities of contacts are arbitrary, which is the case in a fully-mixed scenario. Mean-field

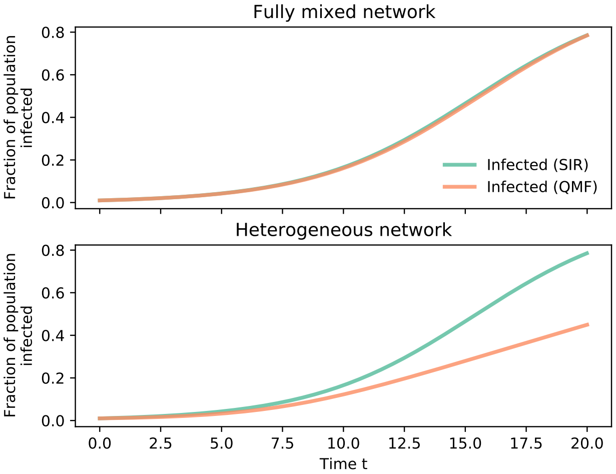

Figure 1.4: An SIR model vs. a quenched mean-field model on a network. Top panel: For a fully connected network, in which every node is connected to every other node, the two models result in the same solution. For a heterogeneous network, in which some nodes have more contacts than others, using a QMF model accounts for these patterns in the spread of the epidemic and yields a different result than that of the basic SIR model.

approaches for both regimes are widely studied [34]. Under a quenched mean-field approach, also known as individual-based mean-field theory, systems of differential equations are derived which govern the probabilities of each node in the network of being infected, susceptible, or removed/recovered at time t. The equations are derived from the underlying Markov chain, which would have q^N states where q is the number of compartments [33]. Averaging over all nodes yields mean-field equations over the network to instead reduce the problem to deterministic approximations of the trajectories of each compartment over time. The effect of using a quenched mean-field model to account for heterogeneous patterns of connections using a network is illustrated by Fig. 1.4, in which an SIR model is compared against the QMF (quenched mean-field) model for two types of networks.

To solve for long-term properties of the disease on the network under a mean-field approach, properties of the adjacency matrix can be used to solve for the system to find steady state dynamics and threshold conditions, specific to the exact contact structure of the network [35, 36].

Instead of focusing on threshold conditions and long-term behavior, we utilize the mean-field deterministic framework to approach a different kind of problem: temporal networks. In Chapter 2 we introduce an approximation of a quenched mean-field model for a disease process over a network to help quantify the importance of chronology in temporal networks. We consider a temporal network to be a sequence of static networks valid for finite, consecutive periods of time. We apply quenched mean-field theory to each of the static networks and use this approach to understand the dynamics on and of the temporal structure.

1.4 Temporal Network Models

All the disease spreading models on network models described in the previous sections have made one major assumption: all the edges between nodes exist uniformly in time. In reality, contacts between individuals are often dynamic. Even with network models where the edges are considered arbitrary, the degree distribution is assumed to be static and so the structure of the network remains static. Relaxing this assumption and accounting for the temporal dynamic changes of the network structure itself can have significant effects on the study of disease spread [37, 38].

Not only are the contacts between individuals not constant in time, but the timings between contacts are generally not uniformly or Poisson distributed. Instead, contacts are known to display “bursty” dynamics, following power-law or other heavy-tailed,

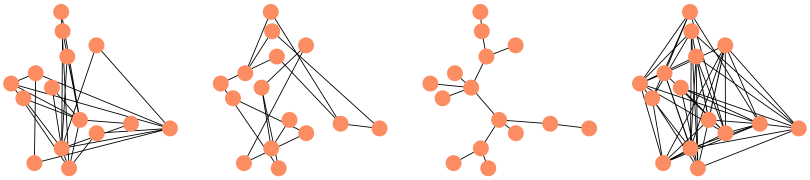

Figure 1.5: Visualization of four snapshots of a temporal network.

heterogeneous distributions [39]. High-resolution data from wearable sensors [40] or email and call logs [41] have helped studies discover temporal contact patterns between individuals and communities and discover such underlying heterogeneous contact patterns.

One way to partially capture the time-varying properties of an empirical contact network is to aggregate sequences of contacts for a short duration of time into a series of static networks, which each comprise a *snapshot* of the temporal network [42], visualized in Fig. 1.5. Often, segmenting the network into snapshots is used for extracting temporal *motifs* [43], or sub-networks that appear more often than others [44, 45] to find recurring contact patterns. Other research has been done on how the rhythms of daily human life affect contact dynamics [46, 47, 48]. Research has also been done to determine properties analogous to the static network counterparts, such as the epidemic threshold for temporal networks [49].

In Chapter 2 we introduce a quenched mean-field approach for the analysis of temporal networks. The goal of this method is to address the lack of treatments for the regime in which assuming the limits of fully quenched or annealed systems would discard critical information, or maintain a more complicated system at an avoidable cost. One goal is to be able to simplify temporal network data, but only to the point that crucial structural information about the chronology of the temporal contacts is maintained. Another goal is to also introduce a tool that can be used to ascertain temporal patterns in the data, such as periodicity when contacts are more heterogeneous vs. maintain a homogeneous and low-density contact structure, by picking up on these structures using network epidemic dynamics as the magnifying lens.

1.5 Stochastic Models of Disease Spread on Networks

As noted, deterministic mean-field models approximate over the collective average behavior of the system in order to provide solutions for the average temporal evolution of spread. However, for some problems it might be useful and necessary to capture the randomness of individual-level interactions. In Fig. 1.6, realizations of stochastic simulations of disease spread demonstrate the variability of the time evolution of a given disease spreading through a population with the same initial conditions.

Disease spread is inherently a random process. Not every interaction between an infected individual and a susceptible one will result in the susceptible individual becoming infected, for reasons ranging from variations in individual immunity, the duration of the interaction, or many other biological or situational reasons. There are other sources of noise beyond the individual level involved in modeling the course of an epidemic, such as uncertainty in detection of infections. To model a spreading process mechanistically, the spread should be modeled by the collective stochastic processes driven by the interactions of each individual in the population [50]. Writing down individual-level equations at the level of individual detail comes at the expense

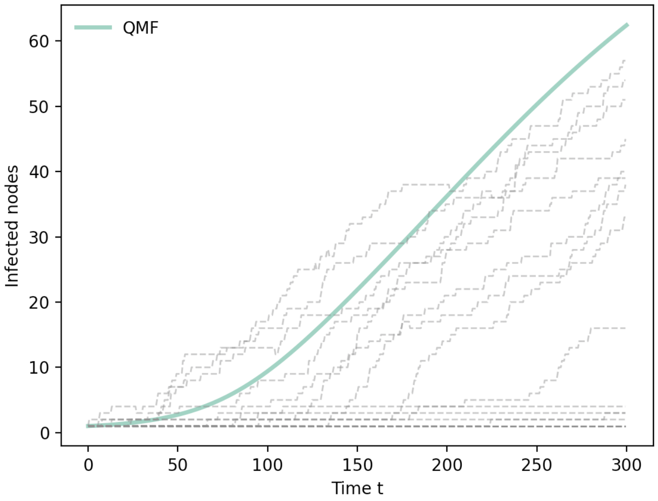

Figure 1.6: Deterministic model of disease spread versus stochastic simulations. The green curve shows the solution to a quenched mean-field model (QMF) that is deterministic. The grey dashed curves each show the result of a single, event-driven simulation on the same underlying network of 100 nodes with the same β and γ values.

of model tractability. Thus, various approximations and simplifications have been made over the course of the development of the network epidemiology literature to build classical compartmental disease models that account for the stochastic nature of disease spread.

Under a stochastic framework, individuals transmit the disease to one another according to some probability distribution, where their transmission and recovery rates from the classical disease models become analogous to “reaction rates” β and γ [27]. The infection rate of an individual, with mean β, and time to recovery, with mean γ^{-1}, are governed by Poisson processes that result in Markovian dynamics for each individual’s trajectory through the S, I, and R compartments [28].

The propagation of the disease in this way can be treated as a percolation problem, utilizing concepts from statistical physics and graph theory. A percolation problem describes how some material flows through a substrate, following the connections between individuals that make up said substrate. In this case, we can describe the spread of an epidemic through a network of contacts as a percolation process. Several works have shown how the long-term properties of an epidemic of SIR type can be treated with branching process and bond percolation tools for epidemic spreading [51, 52, 53], and the properties of branching processes on networks representing the Internet have been studied [54, 55].

Epidemics have been studied as percolation problems on small-world networks by Moore and Newman [56], and further studied on general networks [57, 58, 59]. The class of SIR epidemic models can be solved exactly on networks defined by an arbitrary degree distribution [60], by describing the epidemic process with a percolation model.

To formalize an epidemic process as a percolation problem, we derive a variable known as the *transmissibility* T, the expected probability of infection across an edge between two nodes. We derive T using β, the contact rate per time between two connected nodes, and γ, the average rate of recovery. Letting τ be a random variable representing the time a single node remains infectious, we compute the probability of the node transmitting infection across an arbitrary edge by then as

$$T(\tau) = 1 - \lim_{\delta t \to 0}(1 - \beta\delta_t)^{\tau/\delta t} = 1 - \exp^{-\tau\beta} . \tag{1.5}$$

We evaluate the probability of a particular τ by looking at the cumulative distribution over τ, given by

$$F(\tau) = 1 - \lim_{\delta t \to 0}(1 - \gamma\delta t)^{\tau/\delta t} = 1 - \exp^{-\gamma\tau}, \tag{1.6}$$

derived using γ, the average rate of recovery. Taking the derivative of Eq. (1.6) we obtain the probability mass function over τ,

$$f(\tau) = \gamma \exp^{-\gamma\tau} . \tag{1.7}$$

Finally, to find the probability of transmission T, we calculate the average probability of a node transmitting before its recovery, given that the node recovers at time τ. The average transmissibility T without loss of generality is

$$T = \int_{\tau=0}^{\infty} T(\tau) f(\tau) d\tau = \frac{\beta}{\beta + \gamma}. \tag{1.8}$$

The transmissibility T is therefore a probabilistic quantity capturing the randomness of infectious disease transmission, by representing the probability that transmission occurs between two nodes, given the known parameters β and γ that reflect the average rates for the disease.

The solution of the percolation model can be solved using probability generating functions [61]. Probability generating functions provide a way to encode the probability distributions of discrete random variables into a power series representation. As a result, power series tools can then be used to derive important quantities and transform the probability distribution in various ways. In the case of percolation on networks, the random variable in question is generally the degree of a node (chosen at random) and so the power series terms naturally represent the possible values for a node's degree from zero to infinity. Quantities pertaining to the network structure can be easily derived from the first and second moments of the generating function, and then can aid in the transformation of the generating function into a new function pertaining to new random variables, such as the number of infections given a specific transmission probability. The solutions provide analytical methods for computing the probability of an epidemic occurring, the final size of such an epidemic as a fraction of the network, and the epidemic threshold, the level for which any transmissiblity above it may result in an epidemic.

The beauty of the percolation problem framework is that by using generating functions, the static properties of the long-term spreading process can be solved explicitly for any network with a given degree distribution. However, for emerging diseases and applied problems, often the quantity of interest is the time evolution of the spreading process, and there is less concern for the final size or long-term effects. Understanding the evolution of the spread can help with decision making for mitigation and preparedness measures. As discussed, traditional disease models for time-evolution are formulated by using differential equations and a deterministic framework, where the expected behavior of the spreading process leads to a single average trajectory for the disease. By using generating functions broken down by epidemic generation, Nöel *et al.* [62] showed that the precision of the percolation framework can be applied to develop the temporal progression of a disease while still communicating the uncertainty around the average behavior. In Chapter 3, we revisit the Nöel et al. framework to address two themes. First, that continuous-time event-driven stochastic simulations validate the results generated by the theoretical framework, and find that there is a loose mapping of continuous time to epidemic generations, which can be applied to real early data on emerging disease outbreaks to make predictions for the uncertainty of continued spread. Second, we demonstrate that the probability distributions for cumulative case counts during early generations of spread are surprisingly fat-tailed and flat, suggesting that there is not a well-defined expected number of cumulative cases on heterogeneous networks, in contrast to traditional disease models with point estimates for the average trajectory of case counts.

1.5.1 Stochastic Simulation

Simulating disease spread over a network stochastically is useful for a few key reasons. First, it can be used as an exploratory tool to observe the behavior of a disease given specific parameters and a specific network. Second, it can be used to validate newly developed models to determine their accuracy. In general, mathematical models can be solved faster than performing a robust amount of simulations, but simulations are extremely useful for model validation.

In this body of work, we performed individual-based stochastic simulations for the purpose of model validation. We use a Gillespie algorithm [63] with original code available [64] also referenced in the main body of the text. In brief, the Gillespie algorithm follows the underlying Markov process of the disease spread over the network, and every step of the simulation involves two parts: Determining the continuous time duration until the next event occurs, and which event (infection between two individuals, or recovery of one infected individual) will occur. The time until this next event is chosen from an exponential distribution with mean equal to the combined rate of all possible events. The specific next event is chosen proportionally to its rate compared to the other possible events. Keeping track of every single possible individual event and its rate and history is computationally expensive, which is why the stochastic simulation is used primarily for model validation as opposed to full exploratory results.

1.6 Contributions of Thesis

In this thesis, we focus on introducing models that close the gap between existing theoretical models and the random dynamics of real-time epidemics on networks. While the following chapters are rooted in theoretical models, they can be applied to and adapted easily to real data, without the need for unavailable future data that is sometimes required to calibrate such models.

First, we present a novel framework for assessing the sensitivity of temporal contact data to aggregation into static network representations in Chapter 2. We use a deterministic disease-spreading framework as the basis of the assessment tool. This choice makes the framework best-suited for analysis of disease-related contact data, though the framework could be extended to other contexts.

Next, in Chapter 3 we analyze a stochastic disease model based on probability generating functions that enables tractability of generations of infection [62]. The formalism can be parameterized by real network data to be used for generating probabilistic forecasts of the sizes of future epidemic generations, which could help with epidemic mitigation measures.

Finally, in Chapter 4, we apply the stochastic branching process model to a time-varying network, using methods from both Chapters 2 and 3. The application of the stochastic analysis to temporal networks suggests the utility of combining deterministic and stochastic approaches to support accurate, short-term disease models that are able to capture the specifics of the population, while producing estimates of uncertainty of epidemic outcomes.

The work presented in this thesis addresses the challenges of short-term disease modeling when little is known about the contact patterns underlying the spreading process. Often times in disease modeling, the average descriptor of the system may not capture the most important characteristics of the system. Chapter 3 addresses this by providing temporal distributions of epidemic sizes of progressive generations, still with regard to a static contact network. Meanwhile, Chapter 2 addresses how to reduce a set of temporal contact data to a a new set of condensed networks corresponding to its most salient temporal characteristics. Finally, Chapter 4 applies the generational modeling framework to a temporal network, to show an idealized version of a disease model that captures heterogeneity in both the population structure and variation in time.

Together, the two stochastic and deterministic modeling approaches presented build upon the mathematical modeling tools for making inferences about the course of an epidemic, not just the final size or steady state. Models like these support data-driven decision making for mitigation measures, and can aid in designing interventions that rely on knowing how – and how fast– the disease will spread. When vast quantities of data are readily accessible, it is important to have the tools at hand to make as many simplifications as possible without compromising the structures that help us understand the dynamics of the system.

Bibliography

[1] George EP Box. Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier, 1979.

[2] Klaus Dietz and J. A. P. Heesterbeek. Daniel Bernoulli's epidemiological model

revisited. *Mathematical biosciences*, 180(1-2):1–21, 2002. ISBN: 0025-5564 Publisher: Elsevier.

[3] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics-I. *Proc. R. Soc. Lond., A*, 115(772):700–721, 1927. Publisher: The Royal Society London.

[4] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics—II. The problem of endemicity. *Proc. R. Soc. Lond., A*, 138(772):55–83, 1932. Publisher: The Royal Society London.

[5] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics—III. Further studies of the problem of endemicity. *Proc. R. Soc. Lond., A*, 141(772):94–122, 1933. Publisher: The Royal Society London.

[6] Norman TJ Bailey. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[7] Roy M. Anderson and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[8] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000. ISBN: 0036-1445 Publisher: SIAM.

[9] M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princet. Univ. Press, 41 Williams St, Princeton, New Jersey 08540, 2007.

[10] Odo Diekmann and Johan Andre Peter Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons, 2000.

[11] Odo Diekmann, Johan Andre Peter Heesterbeek, and Johan AJ Metz. The legacy of Kermack and McKendrick. *Publications of the Newton Institute*, 5:95–115, 1995. Publisher: Cambridge University Press.

[12] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002. ISBN: 0027-8424 Publisher: National Acad Sciences.

[13] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999. ISBN: 1476-4687 Publisher: Nature Publishing Group.

[14] Alberto Aleta, Guilherme Ferraz de Arruda, and Yamir Moreno. Data-driven contact structures: from homogeneous mixing to multilayer networks. *PLoS computational biology*, 16(7):e1008035, 2020. ISBN: 1553-734X Publisher: Public Library of Science San Francisco, CA USA.

[15] Odo Diekmann, Hans Heesterbeek, and Tom Britton. *Mathematical tools for understanding infectious disease dynamics*, volume 7. Princeton University Press, 2013.

[16] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002. Publisher: APS.

[17] Luis A. Nunes Amaral, Antonio Scala, Marc Barthelemy, and H. Eugene Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152, 2000. ISBN: 0027-8424 Publisher: National Acad Sciences.

[18] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998. ISBN: 1476-4687 Publisher: Nature Publishing Group.

[19] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187, 1999. ISBN: 0378-4371 Publisher: Elsevier.

[20] Stephen Eubank, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004. ISBN: 1476-4687 Publisher: Nature Publishing Group.

[21] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140, 1741.

[22] Béla Bollobás. Random graphs. In *Modern graph theory*, pages 215–252. Springer, 1998.

[23] Mark Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[24] M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals.* Princet. Univ. Press, 41 Williams St, Princeton, New Jersey 08540, 2007.

[25] J. A. P. Heesterbeek. A brief history of R0 and a recipe for its calculation. *Acta Biotheor*, 50(3):189–204, 2002. Place: Netherlands.

[26] Matt J. Keeling and Bryan T. Grenfell. Individual-based Perspectives on R0. *Journal of Theoretical Biology*, 203(1):51–61, March 2000.

[27] N. G. Van Kampen. *Stochastic processes in chemistry and physics.* North Holland, Amsterdam, 1981.

[28] Sheldon M. Ross. Stochastic Processes. John Wiley & Sons. *New York*, 1996.

[29] I. Z. Kiss, J. C. Miller, and P. L. Simon. *Mathematics of Epidemics on Networks: from exact to approximate models.* Springer, 2019.

[30] Joaquin Marro and Ronald Dickman. Nonequilibrium Phase Transitions in Lattice Models. *Nonequilibrium Phase Transitions in Lattice Models*, page 344, 1999. ISBN: 0521480620.

[31] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001. Publisher: APS.

[32] Malte Henkel, Haye Hinrichsen, Sven Lübeck, and Michel Pleimling. *Non-equilibrium phase transitions*, volume 1. Springer, 2008.

[33] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessan-

dro Vespignani. Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87(3):925, 2015. Publisher: APS.

[34] Guillaume St-Onge, Jean-Gabriel Young, Edward Laurence, Charles Murphy, and Louis J. Dubé. Phase transition of the susceptible-infected-susceptible dynamics on time-varying configuration model networks. *Phys. Rev. E*, 97(2):022305, February 2018. Publisher: American Physical Society.

[35] Yang Wang, D. Chakrabarti, Chenxi Wang, and C. Faloutsos. Epidemic spreading in real networks: an eigenvalue viewpoint. In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pages 25–34, 2003.

[36] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic Thresholds in Real Networks. *ACM Trans. Inf. Syst. Secur.*, 10(4), January 2008. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[37] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, October 2012.

[38] Shweta Bansal, Jonathan Read, Babak Pourbohloul, and Lauren Ancel Meyers. The dynamic nature of contact networks in infectious disease epidemiology. *Journal of biological dynamics*, 4(5):478–489, 2010. ISBN: 1751-3758 Publisher: Taylor & Francis.

[39] Alexei Vazquez, Balazs Racz, Andras Lukacs, and Albert-Laszlo Barabasi. Impact of non-Poissonian activity patterns on spreading processes. *Physical review letters*, 98(15):158702, 2007. Publisher: APS.

[40] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PloS one*, 5(7):e11596, 2010. ISBN: 1932-6203 Publisher: Public Library of Science San Francisco, USA.

[41] J.-P. Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18):7332–7336, 2007. ISBN: 0027-8424 Publisher: National Acad Sciences.

[42] Dan Braha and Yaneer Bar-Yam. Time-dependent complex networks: Dynamic centrality, dynamic motifs, and cycles of social interactions. In *Adaptive Networks*, pages 39–50. Springer, 2009.

[43] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007. ISBN: 1471-0064 Publisher: Nature Publishing Group.

[44] Paolo Bajardi, Alain Barrat, Fabrizio Natale, Lara Savini, and Vittoria Colizza. Dynamical patterns of cattle trade movements. *PloS one*, 6(5):e19869, 2011. ISBN: 1932-6203 Publisher: Public Library of Science San Francisco, USA.

[45] Qiankun Zhao, Yuan Tian, Qi He, Nuria Oliver, Ruoming Jin, and Wang-Chien

Lee. Communication motifs: a tool to characterize social communications. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1645–1648, 2010.

[46] Petter Holme. Network dynamics of ongoing social relationships. *EPL (Europhysics Letters)*, 64(3):427, 2003. ISBN: 0295-5075 Publisher: IOP Publishing.

[47] R. Dean Malmgren, Daniel B. Stouffer, Andriana SLO Campanharo, and Luis A. Nunes Amaral. On universality in human correspondence activity. *science*, 325(5948):1696–1700, 2009. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

[48] Hang-Hyun Jo, Márton Karsai, János Kertész, and Kimmo Kaski. Circadian pattern and burstiness in human communication activity. *New J Phys*, 14(1):013055, 2012.

[49] Eugenio Valdano, Luca Ferreri, Chiara Poletto, and Vittoria Colizza. Analytical Computation of the Epidemic Threshold on Temporal Networks. *Physical Review X*, 5(2):021005, April 2015. Publisher: American Physical Society.

[50] Hakan Andersson and Tom Britton. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media, 2012.

[51] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.3240060204.

[52] Donald Ludwig. Final size distribution for epidemics. *Mathematical Biosciences*, 23(1):33–46, 1975.

[53] P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63(2):157–172, 1983.

[54] Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Network Robustness and Fragility: Percolation on Random Graphs. *Phys. Rev. Lett.*, 85(25):5468–5471, December 2000. Publisher: American Physical Society.

[55] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the Internet to Random Breakdowns. *Phys. Rev. Lett.*, 85(21):4626–4628, November 2000. Publisher: American Physical Society.

[56] Cristopher Moore and Mark EJ Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678, 2000. Publisher: APS.

[57] E. Kenah and J. M. Robins. Second look at the spread of epidemics on networks. *Phys. Rev. E*, 76(3):036113, 2007. Publisher: APS.

[58] L. Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bull New Ser. Am Math Soc.*, 44(1):63–86, 2007.

[59] Eben Kenah and Joel C. Miller. Epidemic Percolation Networks, Epidemic Outcomes, and Interventions. *Interdisciplinary Perspectives on Infectious Diseases*, 2011:543520, February 2011. Publisher: Hindawi Publishing Corporation.

[60] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):016128, July 2002. Publisher: American Physical Society.

[61] Herbert S. Wilf. *generatingfunctionology*. CRC press, 2005.

[62] P.-A. Noël, B. Davoudi, R. C. Brunham, L. J. Dubé, and B. Pourbohloul. Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E*, 79:026101, 2009.

[63] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977. ISBN: 0022-3654 Publisher: ACS Publications.

[64] andrea-allen/epintervene: Initial Release v1.0.0, 2021.

Chapter 2

A network compression approach for quantifying the importance of temporal contact chronology

Abstract

Studies of dynamics on temporal networks often represent the network as a series of “snapshots”, static networks active for short durations of time. We argue that successive snapshots can be aggregated if doing so has little effect on the overlying dynamics. We propose a method to compress network chronologies by progressively combining pairs of snapshots whose matrix commutators have the smallest dynamical effect. We apply this method to epidemic modeling on real contact tracing data and find that it allows for significant compression while remaining faithful to the epidemic dynamics. This chapter discusses the method and applications. Sec. 2.5 provides

further detail into its development and validation.

2.1 Introduction

Modern data collection methods such as radio frequency identification [1] or Bluetooth signal [2] have made the collection of high resolution temporal interaction data simple and widely available. Temporal interactions have rich dynamics in continuous time, yet we often want to combine intervals of temporal data into simpler, static structures—typically a series of "annealed" networks consisting of the aggregate interactions over some time interval—in order to compress the data, reduce analytical complexity, or even to streamline data collection efforts. For example, digital contact tracing protocols ping devices at fixed intervals to save energy and lighten data requirements. However, it is nontrivial to determine when and how to aggregate temporal data without losing critical information about the dynamics of the interactions.

Many methods currently exist to represent and analyze temporal networks [3]. Much recent work focuses on simplifying temporal networks through patterns in the network structure and dynamics; providing algorithms for detecting temporal system states [4], dynamical approach for generating simplified models of temporal network data [5], tools to identify community structure in time-varying networks [6], data-driven approaches to model dynamics on temporal networks by determining change points [7], and methods to represent key temporal features as static networks [8, 9]. Purely from a dynamical perspective, epidemic spread on temporal networks is well-studied [10, 11, 12, 13], as are synchronization [14, 15, 16] and control dynamics [17].

Taken together, these previous studies outline the limits in which one can ignore chronology in network structure. When the dynamics *on* the network are much faster than the dynamics *of* the network, the static limit of the network is appropriate, in which many dynamical steps of the network structure can be combined to yield the same behavior of the dynamics on the network. When the dynamics on the network are slower than the changes of the network, then it is safe to average over many timesteps, referred to as the annealed limit. In both limits, the dynamical history of the network can be compressed in a clear way. In between these limits, it is less clear to what extent the temporal history of the network structure can be compressed while remaining faithful to the dynamics on the network, which are explored in [18].

Here, we address the problem of quantifying if a sequence of networks lies in between these two limits; thereby allowing us to ignore and compress unimportant structural changes while preserving changes that affect the dynamical process. We propose a method to do so by assessing the sensitivity of pairwise network snapshots to aggregation. We consider an epidemic spreading process but abstract the dynamics to a simple diffusion process which could represent other dynamical processes on networks (e.g., synchronization of coupled oscillators or cascading failures in power grids) and formulate a pair-wise error measure using the matrix commutator of the network adjacency matrices that captures the effects on the diffusion process of aggregating snapshots. We use the error measure to successively compress a sequence of chronological snapshots by compressing the adjacent pair of snapshots with the lowest relative error. Using synthetic networks and real data, we find that this approach is successful at producing a compressed snapshot sequence that still mimics the dynamic behavior of the original sequence.

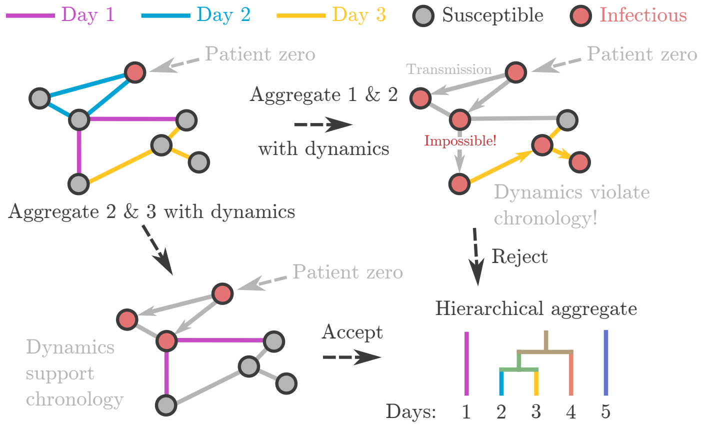

Figure 2.1: Schema of our hierarchical aggregation.

2.2 Analytical Framework

We assume that we have some temporal network data consisting of a large number of snapshots of the network structure. This comes without loss of generality as even continuous time data could be represented as a large series of mostly empty graphs. We seek to quantify the importance of temporal chronology by measuring the error we would introduce by combining any two consecutive networks into a static snapshot. Here, we identify error if the compression of networks over time introduces paths that effectively allow the contagion to progress backwards in time, as depicted in Fig. 2.1. From there, we can successively compress the pairs that minimize this error.

We use a spreading process over the temporal network as a way to observe the significance of aggregating a pair of successive network snapshots. We assume the contagion spreads at a rate β along edges connecting infectious nodes to their susceptible neighbors. We measure the effect of pairwise aggregation by computing a measure of error that scales with the difference in sizes of the infected compartments under the temporal and aggregate regimes over the duration of the snapshot pair.

2.2.1 Snapshot Definitions

We introduce the following notation required for the derivation of our method. Let $X = [X_{ij}]$ be the adjacency matrix for a static network. Assume X is valid for a duration $\delta t_X = t_1^X - t_0^X$, the t_1^X and t_0^X are the ending and starting times of X, and transmission of a spreading process occurs between contacts with rate β_X. We define a snapshot $(X, \delta t_X, \beta_X)$ by its static adjacency matrix, duration, and spreading rate.

Assuming $(X, \delta t_X, \beta)$ and $(Y, \delta t_Y, \beta)$ are snapshots such that $\delta t_Y = t_1^Y - t_0^Y$ and $t_0^Y = t_1^X$, with a uniform β, then we define X and Y as a pair of consecutive snapshots. Given consecutive snapshots $(X, \delta t_X, \beta)$ and $(Y, \delta t_Y, \beta)$, let the aggregate $\overline{(X,Y)}$ with respect to $(\delta t_X, \delta t_Y)$ be

$$\overline{(X,Y)} = \frac{\delta t_X X + \delta t_Y Y}{\delta t_X + \delta t_Y}. \tag{2.1}$$

2.2.2 Solution of Diffusion on a Network

Let A be the adjacency matrix of a network with N nodes. Let $P(t)$ be a vector of length N encoding the probability that each of the nodes from node 0 to node $N-1$ is infected at time t. Then the rate of change of $P(t)$ is given by the following differential equation:

$$\frac{dP}{dt} = \{1 - P(t)\}\, \beta A P(t) - \alpha P(t) \tag{2.2}$$

where from here forward we will drop the term αP_i since we set $\alpha = 0$ in this SI model, either because there is no recovery in the system or because it happens on a much slower timescale than contagion and/or network dynamics. This differential equation can then be solved for to obtain an approximation of $P(t)$, and the size of the infected compartment at time t would then be $I(t) = |P(t)|$.

2.2.3 Approximation of Diffusion

If we assume large N and early time t such that infections are rare enough to approximate $(1 - P(t)) \approx 1$, the solution for the entire system with all nodes can be written by the following differential equation that utilizes the matrix exponential

$$P(t) \approx \exp \beta A t \cdot P(0) \tag{2.3}$$

where $P(0)$ is vector of initial probabilities of infection for each node. Say we have two consecutive snapshots, $(A, \delta t_A, \beta)$ and $(B, \delta t_B, \beta)$. We define the operator $T(A)$ as the transmission dynamics on the snapshot, written as

$$T(A) = \beta \delta t_A A. \tag{2.4}$$

We consider two regimes of dynamics on the pair of snapshots: temporal, where we switch from A to B at the switch-point time $t_1^A = t_0^B$, and aggregate, where we consider $\overline{(A, B)}$ for the full duration $t : [t_0^A, t_1^B]$.

Under the matrix approximation respecting the chronology of the two consecutive snapshots, the solution at the end of the duration of the two snapshots is given by

$$P(t_1^B) \approx \exp\{T(B)\} \cdot \exp\{T(A)\} \cdot P(0), \tag{2.5}$$

and in the case where we aggregate the pair of snapshots,

$$P(t_1^B) \approx \exp\left\{T(\overline{A+B})\right\} \cdot P(0), \tag{2.6}$$

where $T(\overline{A+B}) = \beta(\overline{A+B})(\delta t_A + \delta t_B) = \beta(\delta t_A A + \delta t_B B) = T(A) + T(B)$, following from Eq. (2.1).

Notice that the product of the two exponential terms in Eq. (2.5) cannot be simplified trivially, without fully detailing the product of their Taylor series expansions. Consequentially, $T(A)$ and $T(B)$ cannot be summed in a single exponent, unless they are commutative matrices. We will leverage this property to help assess if the chronology of a snapshot pair is important to the epidemic diffusion dynamics over the network.

2.2.4 Aggregation Error Approximation

In Fig. 2.2 we demonstrate the effect on the SI process of aggregating a consecutive pair of snapshots compared to the temporal solution. Without computationally solving for the full dynamic solution, our aim is to capture this effect by approximating the magnitude of the difference between the temporal and aggregate solutions for a pair of snapshots. We measure and quantify this difference at two points in the process; δ_{t_A} and $\delta_{t_A} + \delta_{t_B}$, and ultimately multiply each quantity by $\delta_{t_A} + \delta_{t_B}$ to culminate in an error measure that will correspond with the integrated area between the temporal and aggregate solutions were we to solve for them computationally.

To do so, we first address the difference between the solutions after δ_{t_A}, the duration of the first snapshot. We define a matrix D_{MID} which captures the difference

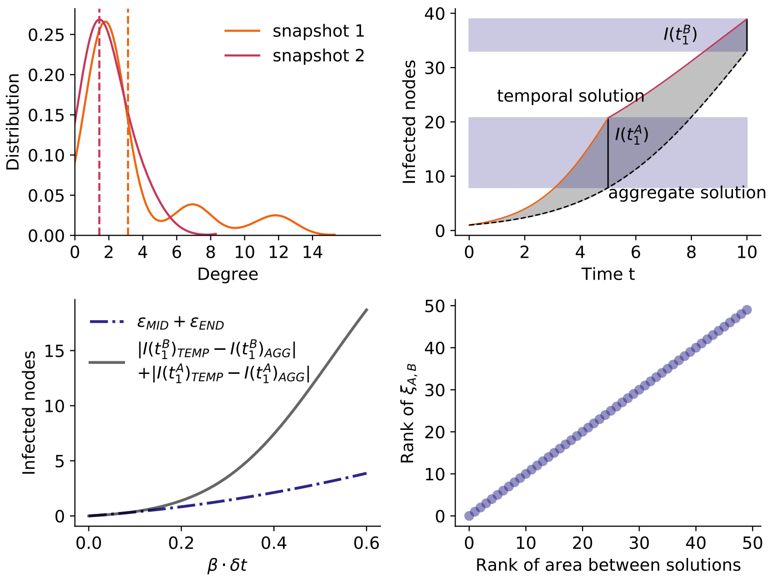

Figure 2.2: Top left: Degree distributions for two snapshots. Top right: Deterministic solution for number of infected nodes, $I(t)$, of an SI process with $\beta = 0.12$, $\delta t = 5$ on the temporal and aggregate versions of the snapshots. Bottom left: True solution difference in number of infected nodes under the temporal and aggregate regimes for varying values of βt by varying $t = [0, 5]$. Bottom right: Ranking of $\xi_{1,2}$ for snapshots 1 and 2 for increasing values of $\beta\delta_t$ compared against the integrated area between solutions.

between each matrix exponential in Eq. (2.5) and Eq. (2.6) to quantify the difference in dynamics in the temporal vs. aggregate solutions. We set

$$D_{MID} = T(A) - \beta\delta_{t_A}\overline{A+B} = \beta\delta_{t_A}(A - \overline{A+B}) \tag{2.7}$$

keeping only the leading linear terms of the difference of the two matrix exponentials.

Next, we want to capture the terminal difference after the duration of the second snapshot, $\delta_{t_A} + \delta_{t_B}$. But the matrix exponential solution from Eq. (2.5) cannot be solved in the same way. Recall that only when A and B are commutative matrices, then their effect on the dynamics will be the same irrespective of their chronology, and can therefore be aggregated without any effect. Consequentially, we utilize the extent to which A and B do *not* commute as the basis for quantifying the importance of their chronology.

The Baker-Campbell-Hausdorff formula can express the product $e^{T(B)}e^{T(A)}$ as a single exponential where we define the matrix C as

$$\begin{aligned} C = \log(\exp T(B) \exp T(A)) = T(B) + T(A) + \frac{1}{2}\left[T(B), T(A)\right] \\ + \frac{1}{12}\left\{[T(B), [T(B), T(A)]] + [T(A), [T(A), T(B)]]\right\} + ... \end{aligned} \tag{2.8}$$

in which $[T(B), T(A)] = T(B)T(A) - T(A)T(B)$, which is the *commutator* of $T(B)$ and $T(A)$. The matrix product BA records paths of length two where the first transition occurs in A and the second in B. The commutator captures paths that follow the chronology (BA), and subtract from each entry the paths that violate the chronology of the two snapshots (AB). For example, if a contact between nodes 1 and 2 occurs in snapshot A before a contact between nodes 2 and 3 in snapshot B, then a disease could spread from node 1 to 3; if these contacts occur in the opposite order, however, then this two-step transmission path cannot occur. Thus, we use the commutator to capture the sensitivity of the temporal ordering of snapshot A and B.

Similarly to Eq. (2.7), we define D_{END} as the difference matrix between the temporal matrix exponential solution using the BCH and the aggregate solution, evaluated at the end of both snapshot durations, obtaining

$$D_{END} = \exp(C) - \exp(T(\overline{A+B})) = \frac{1}{2}[T(B), T(A)] \tag{2.9}$$

to first order. Finally, we want to utilize these matrices that capture the error between the temporal and aggregate solutions, without having to define arbitrary initial conditions like in the solutions defined in Eqs. (2.3), (2.5) and (2.6). We instead take the largest singular value norms of D_{MID} and D_{END} to capture the magnitude of each matrix. We define

$$\epsilon_{MID} = \|\beta\delta_{t_A}(A - \overline{A+B})\|_{LSV} \tag{2.10}$$

and

$$\epsilon_{END} = \|\frac{1}{2}[T(B), T(A)]\|_{LSV} \tag{2.11}$$

where the LSV norm captures the peak response of the dynamics matrices to a unit vector, representing the largest possible reaction to any set of initial conditions on the network.

We then scale these values by the duration of time these particular snapshots will cover in the overall spreading process, approximating the effect visualized in Fig. 2.2, in our final error measure $\xi_{A,B}$, defined as

$$\xi_{A,B} = (\epsilon_{END} + \epsilon_{MID})(\delta t_A + \delta t_B). \tag{2.12}$$

We want to highlight that $\xi_{A,B}$ vanishes for the base case in which $T(B) = T(A)$. All terms in Eq. (2.9) and Eq. (2.7) cancel out, since for any step in one network exists an equivalent step in the other network, and therefore the positive and negative terms sum to zero. When $T(B)T(A) = 0$, all terms in Eq. (2.9) are directly zero since there exist no path that can go from one network to the other, however the switchpoint error Eq. (2.7) may be nonzero.

Eq. (2.12), while not estimating a particular mechanistically defined quantity, results in a scalar value that preserves the ranking of snapshot pairs from least to most induced error when computed over all possible pairs A, B and solved computationally. As such, $\xi_{A,B}$ is useful as a tool for selecting the pair with the least induced error from a sequence of pairs.

2.2.5 Compression Algorithm

Given a set of temporal data as a sequence of M snapshots, we can use the framework to compress the snapshots into $M - j$ snapshots via a greedy algorithm. First, the number of desired iterations j is set. For steps from 1 to j,

1. The error $\xi_{A,B}$ from Eq. (2.12) is computed for each pair A, B of ordered, consecutive snapshots.
2. Identify the pair $A^*, B^* = \text{argmin}_{A,B}(\xi_{A,B})$ to be compressed.
3. Replace snapshots $(A, \delta t_A, \beta)$ and $(B, \delta t_B, \beta)$ with their aggregate, $(\overline{A+B}, \delta t_A + \delta t_B, \beta)$.

2.3 Results

The proposed aggregation algorithm produces a set of temporal snapshots that are able to better support the spreading dynamics of the fully temporal network than the set of evenly divided and compressed snapshots. To assess the performance of the compression algorithm against an evenly distributed compression, we integrate the dynamics of Eq. (2.2) over the full temporal set of snapshots, $x(t)_{TEMP}$, as well as over the system defined by the new sets of even snapshots, $x(t)_{EVEN}$, and the snapshots produced by the algorithm, $x(t)_{ALG}$. We define a validation error measure d_{EVEN} (and d_{ALG}, respectively) as

$$d_{EVEN} = \int \frac{|x(t)_{EVEN} - x(t)_{TEMP}|}{x(t)_{TEMP}} dt \tag{2.13}$$

with d_{ALG} defined analogously.

We apply the algorithm to synthetic networks in Fig. 2.3 and real data in Fig. 2.4. The error metric defined in Eq. (2.12) picks up on the sensitivity of pairs of snapshots to aggregation, and can be assessed at any level of snapshot resolution. We show in the top panel of Fig. 2.4 how the sensitivity of certain temporal ranges is maintained over a large range of resolution, which allows for pre-aggregation of data to improve the speed of the algorithm. The error metric allows us to identify the daily patterns of the contact data at a glance. Once integrated in the compression algorithm, the middle panel shows how we can aggregate over nights and capture the daily activity in one or two snapshots. As seen in the bottom panel, our algorithm compresses more than twice as much than evenly distributed compression while retaining a given level of error on the resulting dynamics.

2.4 Discussion

The error term ξ obtained in Eq. (2.12) provides a fast approach to estimating aggregation error using the matrix exponential. There are four interesting applications for the ξ error term. First, it can directly provide bounds of accuracy when studying dynamics on temporal networks with tools developed for epidemics on static networks. Our analytical error estimate starts with a description of epidemic dynamics but was

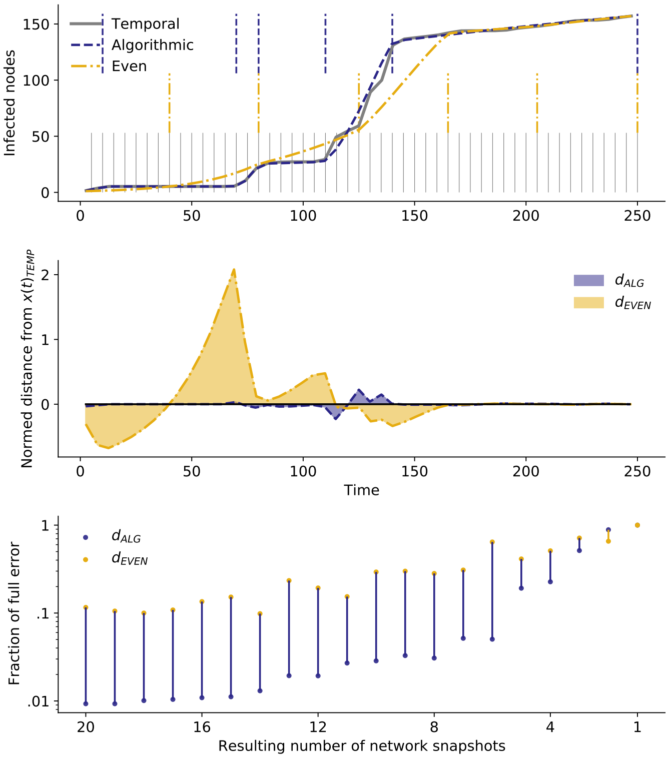

Figure 2.3: Compression algorithm run on a series of 50 synthetic network snapshots, compressed into 6 aggregate snapshots, with a $\beta = 0.0017$ and intervals of $t = 5$ such that $\beta\delta_t \in [0.0085, 0.42]$. Blue dashed lines represent the boundaries for the resulting snapshots from our algorithm, yellow represent the boundaries for the evenly distributed aggregated snapshots. Middle panel shows the normalized distance from the temporal curve over time for each solution. Bottom panel shows the shaded area from the middle panel as a function of number of aggregated snapshots, as a fraction of the error induced by full aggregation.

boiled down to a simple diffusion rate. While other mechanisms (saturation and recovery) were ignored, Fig. 2.2 showed how our approximation preserved the ordering of spreading process solutions relative to one another. Consequently, our tool could very likely be used to estimate error around other type of diffusion dynamics.

Second, it can help compress large sequences of temporal networks by combining any consecutive pair of network $T(A)$ and $T(B)$ into an aggregate if the expected error on the aggregate is smaller than some threshold. We can apply this process

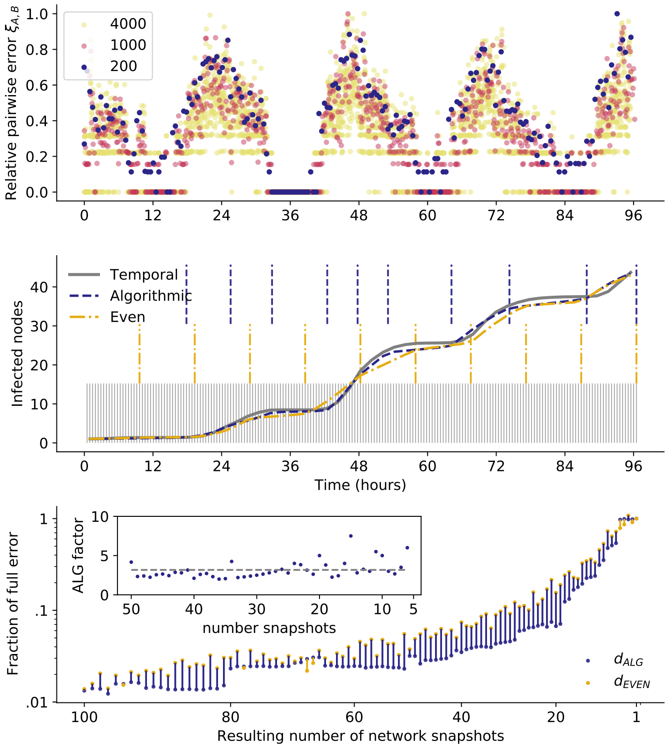

Figure 2.4: Application to a hospital contact network [19]. Top panel: the error measure $\xi_{S(t),S(t+1)}$ computed for consecutive snapshot (S) pairs at 3 different levels of pre-aggregation. The hospital contact dataset contains contacts for approximately $9,000$ unique timestamps. We pre-aggregate by evenly coarse-graining the data to 4,000, 1,000 and 200 snapshots. Mid panel: the compression algorithm run on a the 200 snapshots to generate 10 snapshots, compared to the fully temporal dynamics and the dynamics under even compression. The pre-aggregated 200 snapshots each have duration 1737 seconds with mean 27.34 contacts per snapshot. We used a $\beta = 0.000015$ such that $\tau \in [0.025, 0.5]$. Vertical lines show boundaries for the resulting aggregated snapshots. Bottom panel shows the sum of the shaded area in middle panel function of resulting number of aggregated snapshots, relative to the error induced by the full aggregation. The inset shows by what factor our algorithm can further compress the snapshot sequence while producing an error less than the even aggregation at the number of snapshots shown.

recursively and hierarchically to compress the data to a much smaller sequence of networks while keeping track of the duration of each network snapshot. As shown in Fig. 2.4 using real temporal interaction data, this approach allowed us to consistently meet a certain level error while decreasing the number of required network snapshots by a factor of almost 2.

Third, the error can be used to to estimate the accuracy of data collection in the first place by testing how compressible it could be. This might help focus data collection efforts by identifying places and times with fast temporal variations, as in the top panel of Fig. 2.4. Fourth, the error can be used on non-temporal data to compare the structure of any two networks $T(A)$ and $T(B)$ that share some of the same nodes. At its core, our approach is a network comparison tool: How different are networks when compared to their average?

Limitations of the method include its sensitivity to the spreading process parameters, specifically keeping $\beta\delta t$ within the appropriate range for the matrix exponential approximation, which requires the dynamics to be slow compared to the timescale of network snapshots. Another limitation is the greediness of the algorithm, which means it can get stuck in sub-optimal compression sequence when compressing a long sequence to a handful of snapshots. Future work might explore how to better predict the optimal stopping point of temporal compression.

Altogether, we hope that our work will inspire more tools to compress temporal network data which is an area rich in possible applications.

2.5 Supplemental Information

In this supplement, we describe the development of the error measure $\xi_{A,B}$ and justify its design against our other hypotheses for relevant error measures. We also provide an analysis of the data used in the applications in the main text. This section provides insight into why specific approximation choices were made for the error measure, how the framework was applied to real and synthetic data, and justification of the choice of parameters in the spreading process model.

2.5.1 Developing the Error Measure $\xi_{A,B}$

The theoretical model of a spreading process used in the main body the text relies on an approximation of the matrix exponential and Baker-Campbell-Hausdorff formula. Both formulas are defined by Taylor series expansions of their closed forms, requiring a cutoff for computation as well as simplicity. In the paper, we used a first-order approximation, showing that even that level of crude approximation captured the appropriate quantities to power the network chronology assessment tool. However, similar results are achievable using a higher order approximation, which also underlies the more mechanistic motivation for the model. The simpler approximation (first order) produced comparable results to the higher-order approximations, validating the choice of the simpler approximation. In this supplement, we show the third-order approximations that we first tested for use in the error measure compared against the approximation to first-order. We also justify the use of the largest singular value (LSV) norm for quantifying the size of the matrix term computed in the error measure, by comparing the compression results against the use of a more mechanistically motivated error measure.

Error Measure with Higher Order Approximation

The error measure $\xi_{A,B}$ was at first devised from a desire to best approximate the mechanistic model of the epidemic process over the network. This means that we first designed $\xi_{A,B}$ to be the approximate difference in number of infected nodes between the temporal and aggregate solutions of a spreading process, using the diffusion approximation and BCH formula as a proxy for the deterministic differential equation solution. In this section, we provide the detailed derivation of the original error measure following this framework. Later, we show that the results obtained using this more detailed, mechanistic framework were comparable to the simplified framework presented in the main text. This supplement therefore is useful for justifying the motivation of our framework, and validating the use of the first-order approximations used in the main text.

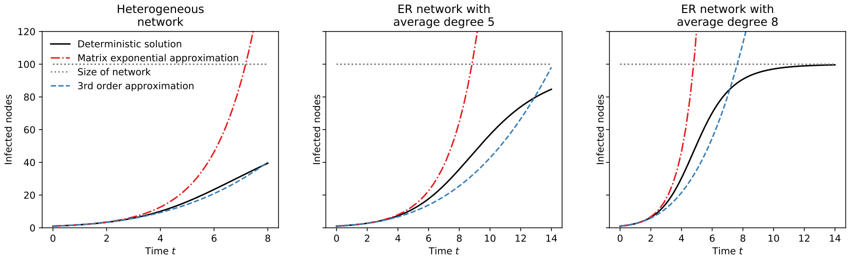

Figure 2.S1 : Matrix exponential approximations of epidemic spread on a network for increasing values of $\beta\delta_t$. A different network is shown in each panel, which vary in structure and edge density. Each panel compares the full matrix exponential solution to third-order approximation, compared against the true deterministic solution for a single static network. We use this as the basis for seeing that approximating the solution works for early time t but starts to falter once more than half the network has been infected.

Let $(A, \delta t_A, \beta)$ and $(B, \delta t_B, \beta)$ be *consecutive* snapshots with constant β. As discussed in the main text, when two matrices X and Y do not commute, then the Baker-Campbell-Hausdorff formula can express $\exp(Y)\exp(X)$ in a single exponent, where we define the matrix Z as

$$Z = \log(\exp Y \exp X) = Y + X + \frac{1}{2}[Y, X] + \frac{1}{12}\left([Y, [Y, X]] + [X, [X, Y]]\right) + \ldots \quad (2.14)$$

in which $[Y, X] = YX - XY$, the *commutator* of X and Y. Letting the matrix $C = \log(\exp T(B) \exp T(A))$, and using the Baker-Campbell-Hausdorff equation, C can be expressed as

$$C = T(B) + T(A) + \frac{1}{2}(T(BA)) - T(AB))) + \frac{1}{12}(T(BBA) + T(ABB) \\ + T(AAB) + T(BAA)) - \frac{1}{6}(T(BAB) + T(ABA)), \quad (2.15)$$

where we approximate to third order matrix products, as the higher order terms scale with $\tau^n |(A + B)^n|, n > 4$, which should fall to zero assuming $\beta\delta_t < 1$. We solve for $\exp(C)$ by using the Taylor series expansion definition, and again approximate by truncating to powers less than 4. We want to then quantify the expected difference between the sizes of infected compartments under the temporal versus aggregate regimes. We follow Equation (2.15) for the temporal regime, and use a truncated power series to third order for the aggregate regime to solve for

$$\epsilon_{END} = \sum_{i=0}^{N-1} \left[\exp(C) - \exp(T(\overline{A + B})) \right] P(0), \quad (2.16)$$

where ϵ_{END} here is meant to approximate the difference in number of nodes infected after δ_t of temporal solution and the aggregate solution. Letting D_{END} be the difference matrix shorthand between the temporal and aggregate matrices above, we compute $D_{END} = \exp(C) - \exp(T(\overline{A+B}))$ as

$$\begin{aligned} D_{END} \approx \frac{1}{2}(T(BA) - T(AB)) - \frac{1}{6}(T(BAB) + T(ABA)) \\ + \frac{1}{12}(T(BBA) + T(ABB) + T(AAB) + T(BAA)) \\ + \frac{1}{4}(T(BBA) - T(AAB) + T(BAA) - T(ABB)) \end{aligned} \tag{2.17}$$

We can further simplify D above in Eq. (2.17) as

$$\begin{aligned} D_{END} \approx \frac{1}{2}(T(BA) - T(AB)) - \frac{1}{6}(T(BAB) + T(ABA)) \\ + \frac{1}{12}T(BBA) + \frac{1}{12}T(ABB) + \frac{1}{12}T(AAB) + \frac{1}{12}T(BAA) \\ + \frac{1}{4}T(BBA) - \frac{1}{4}T(AAB) + \frac{1}{4}T(BAA) - \frac{1}{4}T(ABB) \end{aligned} \tag{2.18}$$

Simplifying, we have

$$\begin{aligned} D_{END} \approx \frac{1}{2}(T(BA) - T(AB)) - \frac{1}{6}(T(BAB) + T(ABA)) \\ + \frac{1}{3}T(BBA) - \frac{1}{6}T(ABB) - \frac{1}{6}T(AAB) + \frac{1}{3}T(BAA), \end{aligned} \tag{2.19}$$

and combining terms we obtain

$$\begin{aligned} D_{END} \approx \frac{1}{2}(T(BA) - T(AB)) - \frac{1}{6}(T(BAB) + T(ABA) \\ + T(ABB) + T(AAB)) + \frac{1}{3}(T(BBA) + T(BAA)). \end{aligned} \tag{2.20}$$

Now using Equation (2.17) to solve for Equation (2.16) we define

$$\epsilon_{END} = \sum_{i=0}^{N-1} [|D_{ij}|] \cdot P(0), \tag{2.21}$$

an approximation for the total difference between the sizes of the infected compartments after t time under the temporal and aggregate regimes, where we take the element-wise absolute value of each entry to account for time-violating transmissions that either under- or over-estimate the number of infections. We compute ϵ_{MID} the same way, by finding the difference matrix of matrix exponential solutions for the temporal and aggregate regimes after t_A time.

$$D_{MID} \approx \sum_{n=1}^{3} \frac{(\beta \delta t_A A)^n - (\beta \delta t_A \overline{A+B})^n}{n!} \tag{2.22}$$

We solve for ϵ_{MID} the same way by taking $\epsilon_{MID} = \sum_{i=0}^{N-1} [|D_{ij}|] \cdot P(0)$, which estimates the difference in infected nodes at time t_A between the temporal and aggregate regimes. Finally, we define the error $\xi_{A,B}$ as the approximated contributed effect of aggregating snapshots A and B on a spreading process with spreading rate β over the durations 0 to δt_A to δt_B, defined as

$$\xi_{A,B} = (\epsilon_{END} + \epsilon_{MID})(\delta t_A + \delta t_B). \tag{2.23}$$

The third-order approximation for $\xi_{A,B}$ is shown compared to the full solution in Fig. 2.S2 .

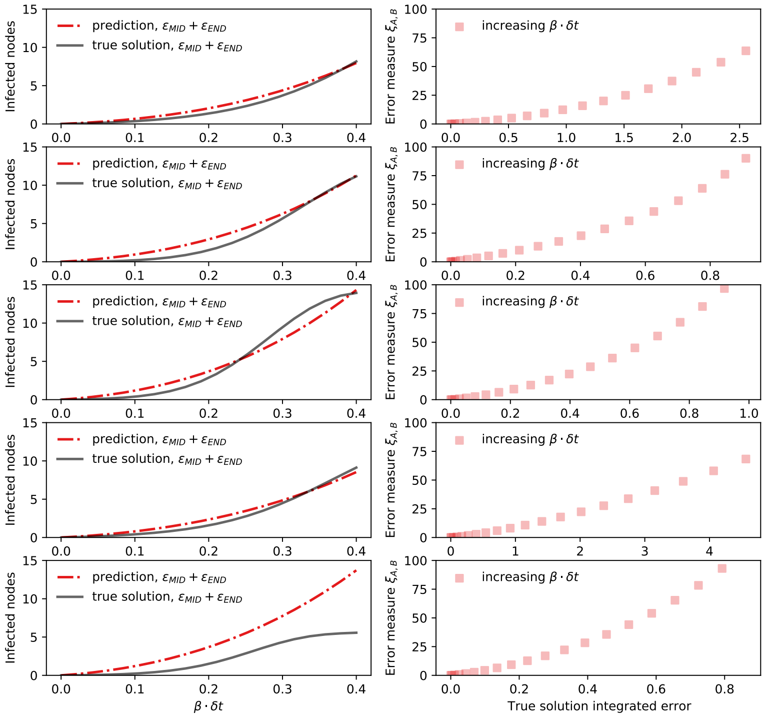

Figure 2.S2 : Left: predicted terminal and midpoint error, $\epsilon_{MID} + \epsilon_{END}$, over a range of increasing $\beta\delta_t$ using 3rd order approximation. The ϵ values are compared with the difference, from the deterministic solutions, between the temporal and aggregate networks for two snapshots. Right: Predicted $\xi_{A,B}$ values over a range of $\beta\delta_t$ values, compared against the mechanistic analogue: the integral of the area between the temporal and aggregate deterministic solutions. From the right hand set of panels it is clear that the error measure $\xi_{A,B}$ preserves monotonicity of rankings when using the integrated error.

Error Measure with First-Order Approximation and LSV Norm

To obtain the point estimates for the difference between number of nodes infected, we solved for the estimate by essentially solving an initial value problem by taking the dot product of the difference of the two matrix exponential solutions with the initial vector of infection probabilities, and marginalizing over the resulting vector. In the main text we used a simpler first-order approximation, along with the largest singular value (LSV) norm for the difference matrix that captured the difference between the temporal and aggregate solutions. This solution is simpler, but less mechanistically motivated, so in this chapter we validate that approach against the approach described in the previous section. Namely, we check that using a first-order approximation and matrix norm instead of third-order and initial-value problem (IVP) solution gives the same results when used in the compression algorithm on the same data.

Once again, for this version of the model, we let D be the difference matrix shorthand between the temporal and aggregate matrices for a pair of snapshots, and want to compute $D = \exp(C) - \exp(T(\overline{A+B}))$. We define D_{END} as the difference matrix between the temporal matrix exponential solution using the BCH and the aggregate solution, evaluated at the end of both snapshot durations, obtaining

$$D_{END} \approx \exp(C) - \exp(T(\overline{A+B})) = \frac{1}{2}[T(B), T(A)] \tag{2.24}$$

where we kept only linear terms from the Taylor series expansion of the matrix exponential and the BCH solution for the temporal regime. Similarly, ϵ_{MID} becomes

$$D_{MID} \approx T(A) - \beta\delta_{t_A}\overline{A+B} = \beta\delta_{t_A}(A - \overline{A+B}) \tag{2.25}$$

keeping only linear terms. Now we have defined the difference matrix D using a first-order approximation, for D_{MID} and D_{END}. Now we develop a way to quantify the relative scale of each D. Instead of taking the sum of the result of the dot product with $P(0)$, an initial state vector, we instead take the largest singular value norms of

D_{MID} and D_{END} to capture the magnitude of each matrix. We define

$$\epsilon_{MID} = \|\beta\delta_{t_A}(A - \overline{A+B})\|_{LSV} \tag{2.26}$$

$$\epsilon_{END} = \|\frac{1}{2}[T(B), T(A)]\|_{LSV} \tag{2.27}$$

and then define the final error measure between snapshots A and B as

$$\xi_{A,B} = (\|D_{END}\|_{LSV} + \|D_{MID}\|_{LSV})(\delta_{t_A} + \delta_{t_B}) \tag{2.28}$$

This way, we don't have to define arbitrary initial conditions to solve for the effect of the difference between solution regimes. Eq. (2.28) is the version used in the main text.

Comparing Versions of the Error Measure

The approximation defined by Eq. (2.28) is different from the third-order approximation because it moves away from a mechanistic estimate. The following figures show how the first-order approximation with the LSV norm still preserves the ranking of error between snapshot pairs, and results in the same compression algorithm choices as the third-order approximation. This analysis justifies why we chose to use the simpler, first-order approximation in the main text.

Fig. 2.S3 shows the result of a comparison experiment for two separate networks. The top panel of each experiment shows the true solution from the solution of the differential equations of the sum of the difference in infected nodes at t_A (switch time) and $t_A + t_B$ (end time) for increasing values of $\beta\delta_t$ in black. The red and purple dashed lines compare the expressions $\epsilon_{MID} + \epsilon_{END}$ using the first-order and third-

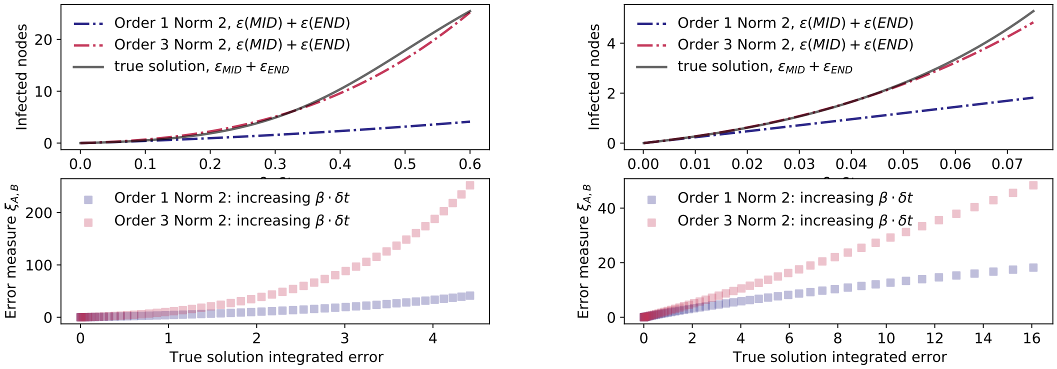

Figure 2.S3 : Comparison of first and third-order approximations of snapshot pair error: Each pair of panels shows results on a different temporal network. The top panel shows the true value of the difference in infected nodes at the mid- and end-point for the temporal and aggregate solutions. Dotted lines show the respective approximations. The bottom panel of each pair of figures shows the true solution integrated error compared to $\xi_{A,B}$ using either first or third-order approximation. As an approximation for the underlying system, the first-order approximation is worse than third-order. However, as shown, as $\beta\delta_t$ increases, the error measure $\xi_{A,B}$ under the first-order approximation increases monotonically with the true solution, validating its use as an appropriate approximation since the goal is to rank relative error.

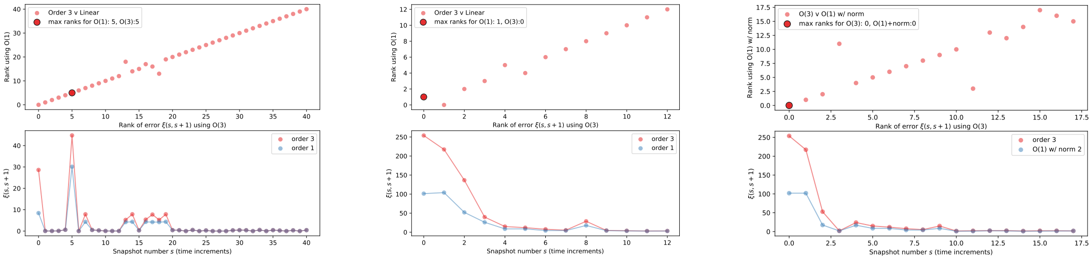

Figure 2.S4 : Comparing the ranking of pairwise snapshot error using third-order and first-order approximation, and using LSV norm vs. initial conditions solution for the synthetic network data: Left panels show comparison of the ranking of pairwise snapshot error using third-order and first-order approximation. The right panel shows the same comparison of order approximation, using the LSV norm. Overall, the ranking of relative snapshot pair error is preserved between the approximations, making either approximation a valid choice.

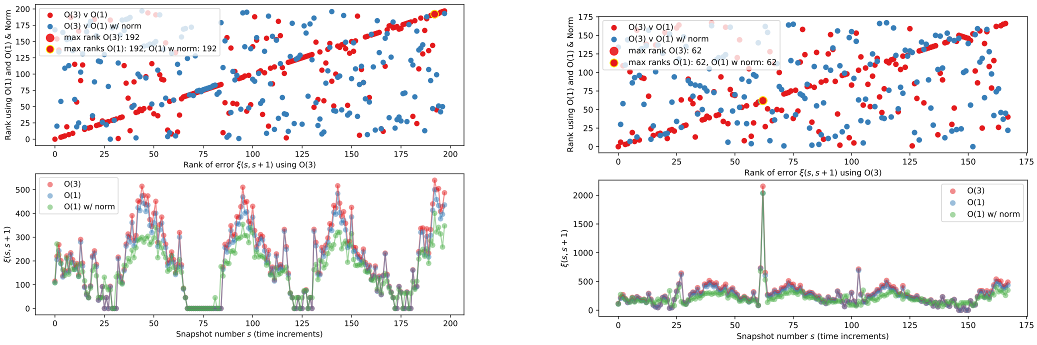

Figure 2.S5 : Rankings of pairwise snapshot error using first vs. third order approximations with and without using the LSV norm for the empirical temporal network with 200 snapshots (left) and again after 25 compressions (right).

order approximations and LSV norm described in the preceding section. The figure shows that, although the third-order approximation is a better literal approximation for the number of infected nodes, the first-order approximation preserves the ranking of snapshot pair error. Since collectively, these approximations are used to rank the importance of snapshot pair chronology, the ranking is more important than the approximation accuracy.

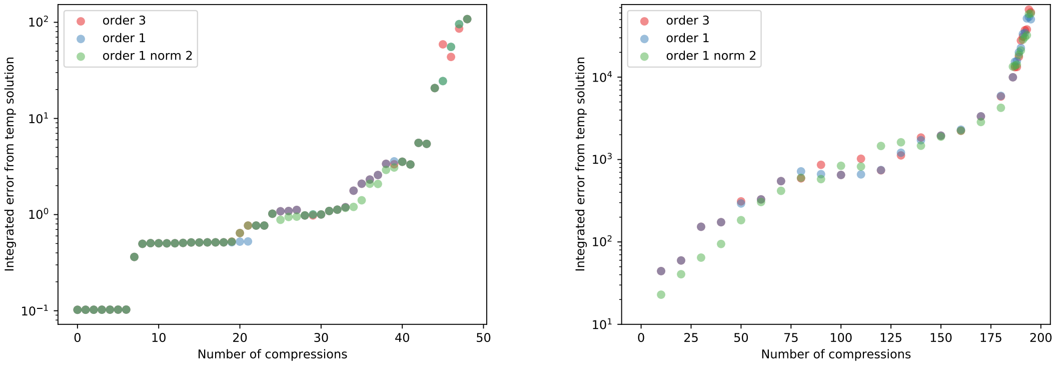

Figure 2.S6 : Error from fully temporal solution compared between approximations for the synthetic temporal network (left) and empirical temporal network (right): Each point represents the integrated area between the curve of the compressed solution to the curve of the original temporal solution, for the third-order, first-order, and first-order with LSV norm approximation of $\xi_{A,B}$. All generally produce the same relative error.

This result is further substantiated by Fig. 2.S4 , which shows the ranking of snapshot pair error ($\xi_{A,B}$) using first-order vs. third-order approximations for progressive sets of compressed snapshots from the synthetic data used in the main text. The key takeaway here is that while the value of $\xi_{A,B}$ changes slightly depending on the approximation used, the general ranking of which pair induces the lowest error is the same.

Lastly, Fig. 2.S5 shows the same ranking results for the empirical temporal data. Starting with 200 snapshots, the left panel of Fig. 2.S5 shows the value and rank of each $\xi_{A,B}$ for the $O(1)$ and $O(3)$ approximations. After 25 compressions, the right side panel of Fig. 2.S5 shows the value and rank of $\xi_{A,B}$ for the resulting compressed snapshots. It can be seen that again, the exact values of $\xi_{A,B}$ depends on the approximation, but the relative lowest ranked snapshot pair is generally consistent between the two approximations.

Finally, we test how the different approximations compare when $\xi_{A,B}$ is used to compress snapshots and the integrated error is measured from the new compressed solution vs the temporal solution. Fig. 2.S6 shows that third-order, first-order, and first-order with the LSV norm all produce comparable levels of integrated error when used in the compression algorithm. Based on these results, we chose to select the simpler, first-order and clean LSV norm to use in the computation of $\xi_{A,B}$ presented in the main text.

2.5.2 Data Analysis and Application Details

We used a data set of physical proximity contacts in a hospital [19] to apply our compression approach. In this section we provide some further detail about the data set. Also, we show the effect of pre-compressing the data on the results of the compression algorithm. Pre-compression is defined by taking the raw temporal data and aggregating time windows together to create the base layer of network data, since many of the contacts are so fine-grained there is one contact per 20-second interval. In the main text, we distributed the contacts evenly into 200 snapshots, each spanning approximately 28 minutes. This choice was somewhat arbitrary, however, we found that the level of pre-compression does not have a significant impact on the future use of the algorithm as long as the pre-compression level is not too high (otherwise one would forego the purpose of the compression algorithm). This section provides the raw data analysis and the pre-compression analysis. Data analysis of the hospital contact data used in the main text is shown in Table 2.S1 .

	Variable	Result
0	num timestamps	9453
1	max timestamp	347640
2	min timestamp	140
3	mean number of contacts per timestep	3.43
4	variance of contacts per timestep	5.89
5	mean value of how often contacts repeat	28.467
6	variance of how often contacts repeat	4671.79
7	mean number of unique contacts	17.52
8	variance of number of unique contacts	238.40
9	mean duration between time steps	10.71

Table 2.S1 : Observational statistics for the hospital contact dataset

Pre-Compression Analysis

The hospital contact temporal data was pre-compressed into 200 snapshots before applying the compression algorithm. We tested whether the results obtained by pre-compressing the data was sensitive to the level of pre-compression. In this section, we show that regardless of pre-compression level (up to a reasonable point), produces the same motifs and generally the same temporal boundaries once the algorithm was applied.

N	Pre-compression level	Average k	Average q	Average C
75.0	100	30.69	29.81	0.10
75.0	150	28.62	27.77	0.08
75.0	200	27.53	26.71	0.07
75.0	300	25.58	24.79	0.05
75.0	400	24.65	23.88	0.05
75.0	600	23.53	22.78	0.04
75.0	800	22.56	21.84	0.04
75.0	1000	21.51	20.80	0.03

Table 2.S2 : Network statistics for the pre-compressed snapshots at various levels

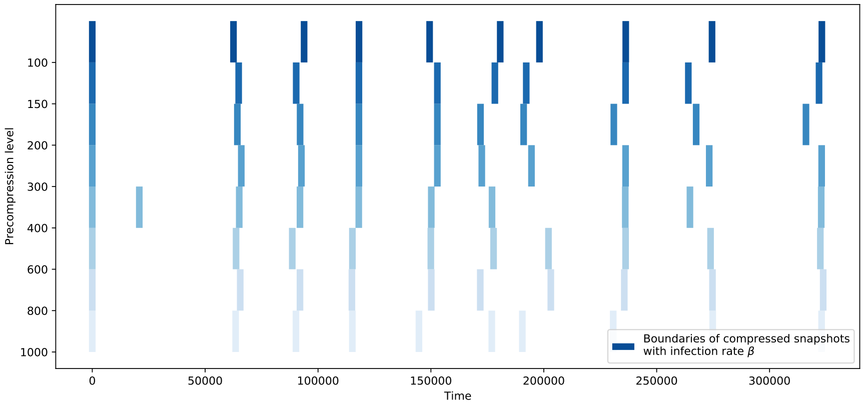

Figure 2.S7 : Locations of resulting compressed snapshot boundaries using the algorithm starting with different levels of pre-compression. Notice how more or less, the boundaries positioning is maintained regardless of the pre-compression level.

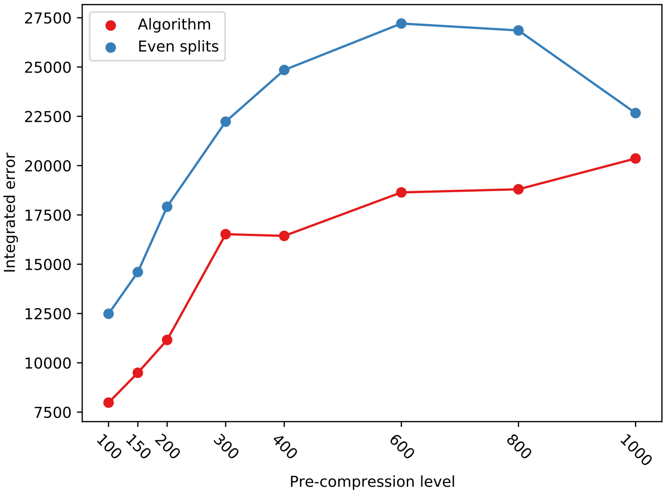

Figure 2.S8 : Integrated error (area between the solution curves for the compressed network compared to the fully temporal solution) of the SI process on the compressed snapshots using the algorithm using different levels of pre-compression. This error changes as a function of pre-compression level. As shown, the algorithm still results in generating a sequence of compressed temporal networks that perform better than even partitioning of the snapshots.

Table 2.S2 provides summary statistics about the pre-aggregated snapshots at each pre-compression level analyzed. Fig. 2.S7 shows the temporal boundaries of the resulting compressed snapshots when the algorithm is applied to compress the pre-compressed snapshots into just 10 snapshots remaining. Fig. 2.S8 shows how the overall integrated error (area between the solution curves for the compressed network compared to the fully temporal solution) changes as a function of pre-compression level.

Parameter Robustness Analysis

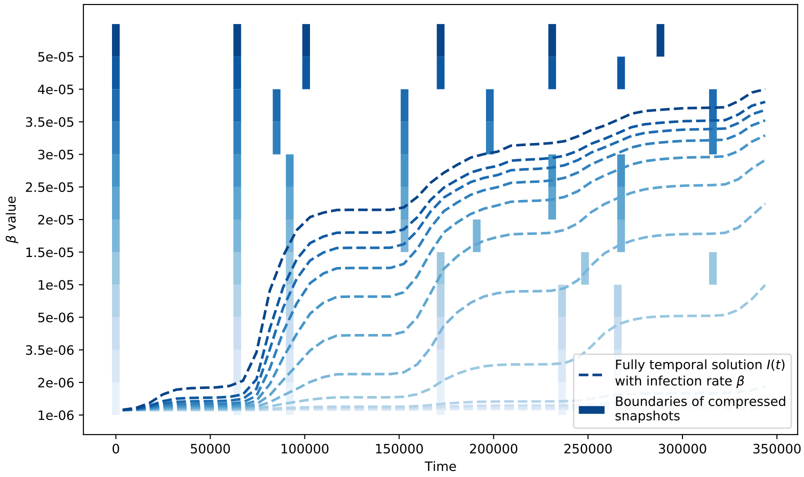

Figure 2.S9 : Locations of resulting compressed snapshot boundaries using the algorithm with different values of $\beta\delta_t$ for the empirical data set. Dashed lines show the resulting SI process. Notice how the temporal boundaries of the compressed snapshots are similar for similar values of β. We conclude that while β should be kept within a range that corresponds with fast enough spread on the network to activate dynamics, but slow enough to not saturate the network before the end of the epidemic process, but the precise value of β is not important for the compression algorithm.

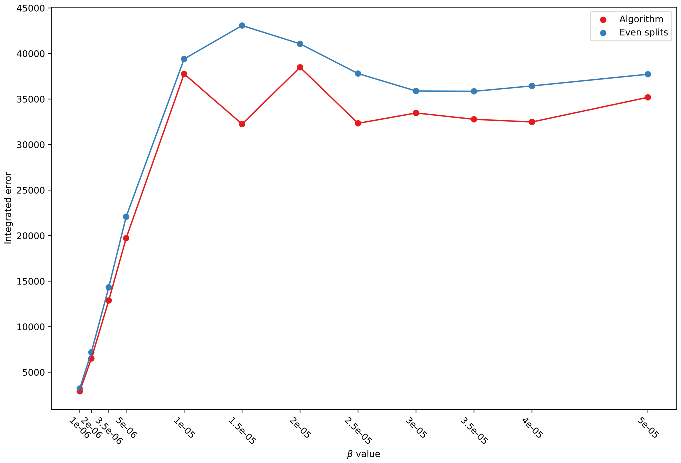

Figure 2.S10 : Integrated error of the SI process (area between the solution curves for the compressed network compared to the fully temporal solution) using different values of $\beta\delta_t$. We observe that regardless of the precise β value used for the compression algorithm, the algorithm produces a series of compressed networks that respect the temporal dynamics better than the sequence of evenly divided and compressed networks.

We also tested the sensitivity of the algorithm to variations in $\beta\delta_t$ values used to parameterize the process underlying the compression algorithm. We performed the same experiments as in the pre-compression sensitivity experiment, but as a function of increasing $\beta\delta_t$ values. Fig. 2.S9 shows the temporal boundaries of the resulting compressed snapshots when the algorithm is applied to compress 200 pre-compressed snapshots into just 10 snapshots remaining, as a function of increasing β. Along with the temporal boundaries, the temporal solution is shown, to illustrate how an increase in β affects the solution and therefore the choices of the compression algorithm. Fig. 2.S10 shows how the overall integrated error (area between the solution curves for the compressed network compared to the fully temporal solution) changes as a function of β. The algorithm still results in networks performing better than even partitioning of the snapshots.

Bibliography

[1] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS ONE*, 5(7):e11596, 2010.

[2] Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. Interaction data from the copenhagen networks study. *Scientific Data*, 6(1):1–10, 2019.

[3] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, October 2012.

[4] Naoki Masuda and Petter Holme. Detecting sequences of system states in temporal networks. *Scientific Reports*, 9(1):795, 2019.

[5] Tiago P Peixoto and Martin Rosvall. Modelling sequences and temporal networks with dynamic community structures. *Nature Communications*, 8(1):582–582, 2017. Place: England Publisher: Nature Publishing Group UK.

[6] Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore, and Leto Peel. Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks. *Physical Review X*, 6(3):031005, July 2016. Publisher: American Physical Society.

[7] Tiago P. Peixoto and Laetitia Gauvin. Change points, memory and epidemic spreading in temporal networks. *Scientific Reports*, 8(1):15511, October 2018.

[8] Petter Holme. Epidemiologically Optimal Static Networks from Temporal Network Data. *PLoS Computational Biology*, 9(7), July 2013.

[9] Ingo Scholtes, Nicolas Wider, and Antonios Garas. Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. *The European Physical Journal B*, 89(3):61, March 2016.

[10] Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Simulated Epidemics in an Empirical Spatiotemporal Network of 50,185 Sexual Contacts. *PLoS Computational Biology*, 7(3), March 2011.

[11] Mathieu Génois, Christian L. Vestergaard, Ciro Cattuto, and Alain Barrat. Compensating for population sampling in simulations of epidemic spread on temporal contact networks. *Nature Communications*, 6(1):8860, November 2015.

[12] Guangming Ren and Xingyuan Wang. Epidemic spreading in time-varying community networks. *Chaos*, 24(2):023116, June 2014. Publisher: American Institute of Physics.

[13] Eugenio Valdano, Luca Ferreri, Chiara Poletto, and Vittoria Colizza. Analytical Computation of the Epidemic Threshold on Temporal Networks. *Physical Review X*, 5(2):021005, April 2015. Publisher: American Physical Society.

[14] Stefano Boccaletti, D-U Hwang, Mario Chavez, Andreas Amann, Jürgen Kurths,

and Louis M Pecora. Synchronization in dynamical networks: Evolution along commutative graphs. *Physical Review E*, 74(1):016102, 2006.

[15] Daniel J Stilwell, Erik M Bollt, and D Gray Roberson. Sufficient conditions for fast switching synchronization in time-varying network topologies. *SIAM Journal on Applied Dynamical Systems*, 5(1):140–156, 2006.

[16] Yuanzhao Zhang and Steven H Strogatz. Designing temporal networks that synchronize under resource constraints. *Nature communications*, 12(1):1–8, 2021.

[17] Aming Li, Sean P Cornelius, Y-Y Liu, Long Wang, and A-L Barabási. The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046, 2017.

[18] Guillaume St-Onge, Jean-Gabriel Young, Edward Laurence, Charles Murphy, and Louis J. Dubé. Phase transition of the susceptible-infected-susceptible dynamics on time-varying configuration model networks. *Phys. Rev. E*, 97(2):022305, February 2018. Publisher: American Physical Society.

[19] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. *PLoS ONE*, 8(9):e73970, September 2013. Publisher: Public Library of Science.

Chapter 3

Predicting the diversity of early epidemic spread on networks

Abstract

The interplay of biological, social, structural and random factors makes disease forecasting extraordinarily complex. The course of an epidemic exhibits average growth dynamics determined by features of the pathogen and the population, yet also features significant variability reflecting the stochastic nature of disease spread. In this work, we reframe a stochastic branching process analysis in terms of probability generating functions and compare it to continuous time epidemic simulations on networks. In doing so, we predict the diversity of emerging epidemic courses on both homogeneous and heterogeneous networks. We show how the challenge of inferring the early course of an epidemic falls on the randomness of disease spread more so than on the heterogeneity of contact patterns. We provide an analysis which helps quantify, in real time, the probability that an epidemic goes supercritical or conversely, dies stochas-

tically. These probabilities are often assumed to be one and zero, respectively, if the basic reproduction number, or R_0, is greater than 1, ignoring the heterogeneity and randomness inherent to disease spread. This framework can give more insight into early epidemic spread by weighting standard deterministic models with likelihood to inform pandemic preparedness with probabilistic forecasts.

3.1 Introduction

By the time of this writing, the COVID-19 pandemic had reached every corner of the world. Public health efforts are now focused on identifying new clusters of outbreaks and their risk of causing new epidemic waves, much like they did at the beginning of the pandemic. As large outbreaks soared early on in a handful of countries, sporadic clusters of confirmed cases dotted regions in the United States. Data surrounding new clusters or waves tend to consist of low numbers of cases highly sensitive to noise, sparking concern and uncertainty at the expected progression of the epidemic.

The first confirmed case of COVID-19 in the US was reported on January 21st, 2020 in the state of Washington [1]. Three subsequent cases were later identified in Washington; two hospitalizations on February 19th [2], and two deaths on February 26th, one week later [3]. Then, on February 28th, a high school closed immediately after one of its students tested positive for a strain that had been associated with the January 21st case [4]. With limited knowledge of active cases, it was nearly impossible to predict the current and future severity of the outbreak.

One critical question in Washington after over a month with only a handful of detected cases, was whether this chain of events suggested a single tree of very few local transmissions, or multiple distinct introduction events from abroad. Despite decades of disease modeling, the community was ill-equipped to answer this question. The problem is challenging in part because of inadequate testing at the time, and also because well-established disease models often operate on deterministic mechanisms designed to describe the average behavior of large epidemics and not the random, discrete nature of small transmission chains. The looming question of whether a local COVID-19 outbreak would die off by itself or become a disaster, can only be modeled using tools capturing the stochasticity, or randomness, of person-to-person contact. To accurately model the potential outcomes of an epidemic based on limited case data, tools that capture the random nature of disease spread along with the structure of the population are required.

In this work, we analyze the diversity of early epidemic courses. In doing so, we also hope to provide analytical tools to inform disease forecasts by accounting for the heterogeneity and stochastic nature of disease transmission.

Since the introduction of mean-field epidemic models, deterministic models of disease spread have continued to evolve in complexity and detail. Kermack and McKendrick's early work [5, 6, 7] gave rise to compartmental models, in which the population under study is divided into two or more states. Perhaps the most widely known of these models is the Susceptible-Infectious-Recovered (SIR) model, where the population is divided into susceptible, infectious, and recovered states (or compartments) and the trajectory of the sizes of each compartment can be tracked analytically over time [8, 9]. The standard compartmental model assumes homogeneous mixing of the population and is deterministic, meaning that a given set of initial conditions and disease transmission rates always leads to the same expected outcome. A common extension to compartmental models is to relax the assumption of homogeneous mixing. One method for doing so is to derive mean-field equations for an epidemic process over contact networks, thereby introducing heterogeneous structure into the population [10]. Similarly, it is possible to partition the population based on traits such as age, risk behaviors, or location and define how these partitions mix [11, 12, 13, 14]. While these approaches introduce more realistic contact behavior into a model, they fail to account for the inherently stochastic nature of disease spread; something of particular importance early in an outbreak.

Models based on stochastic processes address the shortcoming of deterministic outcomes in the standard mean field compartment models. A commonly used approach is that of branching processes. Bienayme-Galton-Watson processes are one widely used example, as they provide a good approximation of more general stochastic epidemic models [15]. Beyond Bienayme-Galton-Watson processes, there exist a number of extensions such as including population structure, multiple types of hosts/pathogens, and considering time to be continuous rather than discrete [16, 17]. In these branching process models the basic reproduction number, R_0, the probability of an outbreak, and the final proportion of population infected (in a “supercriticial” model) are typically tractable to compute. While these are all important, a shortcoming of most branching models is the difficulty of tracking the trajectory of outbreaks through time and knowing whether it matches the continuous time dynamics of real epidemics. Stochastic differential equations are an alternative modeling approach that allow one to track outbreak trajectories, as well as often finding threshold conditions for the occurrence of an outbreak or the existence of an endemic equilibrium [18, 19, 20]. Like all models, stochastic differential equations have drawbacks;

the most relevant is standard formulations do not allow for stochastic extinction if $R_0 > 1$.

Another common approach in disease modeling is times series analysis, more statistical in nature than mechanistic models. This theory can be applied to assist in estimating the parameters of compartmental models or to combine ensembles of compartmental models to increase prediction accuracy [21, 22]. Independently of compartmental models, time series analysis can be used to study covariates of disease occurrence (*e.g.,* weather), estimate the future variability in observed cases, or to make epidemic forecasts [23, 24, 25]. A necessary requirement for the effective use of many time series methods however is data. When facing sparse incidence numbers, and in the absence of historical data, the methods become problematic and thus are not suitable for emerging diseases.

Agent-based models are another family of models used for tracking epidemic progression, in which agents, or individuals in the population, are tracked throughout the course of the epidemic. Agents are parameterized with individual attributes, capturing the heterogeneity of the population and aspects from compartmental models are used to categorize the state of each agent [26, 27]. While there is great power in adjusting various attributes for different epidemic conditions and environmental factors, most of these models are computationally expensive and need a copious amount of information to generate the entire collection of agents [28, 29, 26, 30, 27, 31], making them ill-suited for modeling early epidemic spread with a handful of cumulative case counts and sparsely available data.

Early in an outbreak, we often face the unique challenge of modeling disease spread while taking into account the heterogeneity of the population and the stochastic nature of disease spread, including stochastic extinction, without substantial amounts of data. The heterogeneous contact structure found in populations is accounted for by network models, and a first approximation for a relevant contact structures in a novel outbreak can be taken from past outbreaks of similar diseases. Including a sufficient number of possible states will typically account for heterogeneity in host and pathogen type. The randomness of transmission is modeled with stochastic processes, many of which easily permit stochastic extinction.

The above considerations naturally lead to percolation theory, which can be used to analyze stochastic compartmental disease models on networks. Percolation models unite contact heterogeneity and stochasticity under a single modeling framework [32]. An underlying contact network acts as the substrate for disease to propagate through, resulting in a directed network of transmission [33, 34, 35]. The resulting epidemic percolation networks can be analyzed using branching process theory [36, 37] which model stochastic transmission between individuals using an underlying offspring distribution. Branching processes are especially useful for early epidemic modeling, as they allow for stochastic behavior of spread as well as stochastic extinction [38]. Specifically, the method of probability generating functions (PGFs) can be used to analyze branching processes on percolation networks [37, 39, 38]. Consequently, there have been many recent applications of this framework designed specifically for COVID-19 [40, 41, 42, 43, 44, 45].

The PGF formalism is traditionally used for estimating quantities that pertain to the predicted end of an epidemic — such as the probability of infecting a macroscopic fraction of the population and distribution of final outbreak sizes — but not how risk and outbreak sizes change dynamically over time. Kenah and Robins show how modified percolation models, epidemic percolation networks, has a final state isomorphic to a network-based SIR models [33]. Most bond percolation frameworks differ from SIR dynamics as SIR transmission events are correlated through the distribution of the infectious period of each infected individual whereas percolation models assume independent contacts and transmission events. More importantly, percolation models integrate over time to map transmission dynamics (which occur in continuous time) to discrete bond percolation (which occur in discrete time with a fixed probability of transmission).

In 2009, Noël *et al.* [46] offered a novel method for tracking the stochasticity of outbreak sizes by epidemic generations, allowing us to incorporate discrete time into the percolation-framework model. In this paper, we show how the generation-based PGF formalism also succeeds in tracking emerging epidemic size in *continuous* time, by validating the PGF approach with event-driven simulations on networks. This result allows us to use PGFs and early disease data to quantify epidemic risk and survival probability.

3.2 Theoretical Analysis and Simulations

3.2.1 Probability Generating Functions

PGFs succinctly encode a probability distribution in a power series representation so that the methods of power series analysis can be applied [47]. PGF theory naturally extends to disease modeling, where the distribution under study encapsulates a disease transmission network, framed as a bond percolation problem where the bond occupation probability T is the probability of an infected individual infecting one of their contacts over the course of the entire epidemic [37, 39]. Typically, this approach is used to solve for the average behavior of the system; we can solve for quantities such as the critical transmissiblity at which the entire connected population will become infected, or the distribution of outbreak sizes. However, an increasing necessity of disease modeling is to model early epidemic spread, analyzing early cases to predict whether an outbreak will become large before it actually happens. In 2009, Noël *et al.* [46] developed the epidemic PGF modeling theory further to model the sizes of progressive epidemic *generations*, demonstrated in Fig. 3.1.

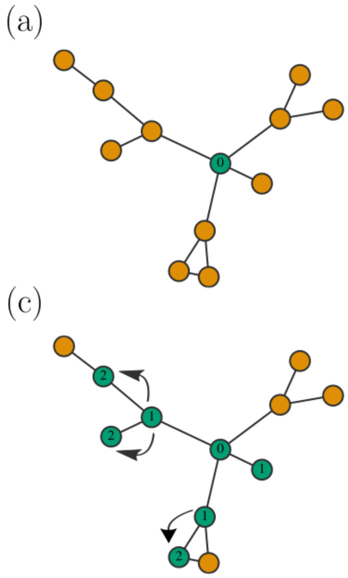

Figure 3.1: Schematic of generations of infection through a network. Each node's label corresponds to the epidemic generation in which it was infected. The initial infected node is in generation 0, any nodes they infect constitute generation 1, and so on.

The foundations for both aforementioned generating function methodologies are the same, beginning with the underlying contact network. In a contact network, we represent a collection of individuals as *nodes* and their contacts between each other with *edges*. We say that two nodes are *neighbors* if they are in contact, i.e. connected by an edge. A node's *degree* is how many neighbors it has. The *degree distribution* of a network is the probability distribution for the number of neighbors of one node. Under an SIR disease modeling framework, nodes begin as *susceptible*, and become *infectious* if it is infected by one of its neighbors, which occurs with probability T.

The framework introduced by Noël *et al.* uses PGFs to describe generations of infection as piece-wise generating function, which can then be studied using branching process techniques. First we introduce what an epidemic *generation* is. We say a node belongs to generation g if it became infected via a neighbor belonging to generation $g - 1$. Assuming an infinite-size network drawn from a specific degree distribution (known as a configuration model), each chain of infections stemming from an initial infected case, *patient zero*, can be considered essentially uncorrelated with an approximately 0 probability of interacting. This allows every subsequent case to treated as a node that was reached by following a random edge. This means each node in each generation can be treated as independent from all other nodes in its generation. Thus, for each node in generation g, the PGF describing the distribution of cases that node will cause over the course of the epidemic is given by

$$G_g(x;T) = \begin{cases} G_0(x;T) & (g = 0) \\ G_1(x;T) & (g > 0) \end{cases} \tag{3.1}$$

where $G_g(x;T)$ is the distribution, in PGF notation, of the *secondary* cases caused by a single node in generation g. Now, we will provide the derivations used to obtain this framework using the underlying network, generating function and branching process theory.

Using PGF notation, we will refer to the original underlying network degree distribution as $G_0(x)$, which we write as

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \tag{3.2}$$

The kth coefficient of Eq. (3.2), p_k, is the probability of randomly choosing a node with degree k from the network. The average degree of the network is denoted as $\langle k \rangle$, derived by the first derivative of the generating function as

$$G_0'(1) = \langle k \rangle = \sum_{k=0}^{\infty} k p_k. \tag{3.3}$$

To study the progression of an epidemic, we are interested in the distribution of infections from each subsequently infected node. Before introducing transmission probability, we work first with the aforementioned degree distribution to understand how many infections each node could cause through each generation. Assuming an initial infectious node, patient zero, we know $G_0(x)$ is the distribution of contacts for them, but that distribution is different for anyone patient zero infects. This phenomenon is known as the friendship paradox; the degree of a node chosen by following a random edge is on average, larger than the degree of the node selected at random whose edge we followed. In this context, patient zero has a degree distribution of $G_0(x)$, but the node who patient zero first infects has a degree distribution known as the *excess degree distribution*, denoted as $G_1(x)$ in PGF notation. To obtain $G_1(x)$, we are interested in the degree of nodes provided that we arrive there by following the edge from one of its neighbors. So, this means the resulting distribution will exclude that neighbor, reducing every node's degree by 1, and multiplied by the number of ways they could have been reached, which is the original degree. This algorithm surmounts to taking the derivative of $G_0(x)$, so that we have the excess degree distribution

$$G_1(x) = \frac{\sum_k (k+1)p_{k+1}x^k}{\sum_k (k+1)p_{k+1}} = \sum_{k=0}^{\infty} q_k x^k \tag{3.4}$$

and where the derivative is divided by the average degree of the network $\langle k \rangle$ in order to normalize the distribution tuned to the original node. The coefficients q_k represent the probability of reaching a node with degree k from a randomly chosen edge.

Returning to the percolation problem, we incorporate disease transmissibility T to transform the excess degree distribution into a *secondary case distribution.* The probability that a single infectious node infects l neighbors given it has degree k, or k neighbors, is given by

$$p_{l|k} = \binom{k}{l} T^l (1-T)^{k-l} \tag{3.5}$$

From this we can derive the PGF for the number of infections caused by "patient zero", which we denote $G_0(x;T)$ for short, given by

$$\begin{aligned} G_0(x;T) &= \sum_{l=0}^{\infty}\sum_{k=l}^{\infty} p_k p_{l|k} x^l \\ &= \sum_{k=0}^{\infty}\sum_{l=0}^{k} p_k \binom{k}{l} T^l (1-T)^{k-l} x^l \\ &= G_0(1+(x-1)T). \end{aligned} \tag{3.6}$$

From $G_0(x;T)$, $G_1(x;T)$ can be calculated in a parallel fashion as $G_1(x)$ is from $G_0(x)$. The PGF $G_1(x;T)$ is now the probability distribution of the number of infections caused by a single node, i.e., the secondary case distribution.

We now present how to study the evolution of the distribution of cumulative cases for the percolation model following Noël *et al.* Let s be the number of cumulative cases at generation g and let m be the number of infectious nodes strictly belonging to generation g. (Note that in this way, $s_g = \sum_0^g m_g$). We let the probability of having s total infections by the end of the g-th generation with m becoming infected (and thus being infectious) during that generation be denoted as ψ_{sm}^g [46]. This has an associated probability generating function, given by

$$\Psi_0^g(x, y) = \sum_{s,m} \psi_{sm}^g x^s y^m \tag{3.7}$$

over all s, m.

We know the distribution of infections following from a single infectious node in generation $g-1$ is generated by $G_{g-1}(1+(x-1)T)$ (from Eq. (3.6)). The PGF of a finite sum of independent processes is the product of their PGFs, and as discussed above, each node in generation $g-1$ can be treated independently. Thus, if we assume the state in generation $g-1$ is given by the pair (s', m'), then the probability of spawning m new infectious nodes in generation g is generated by

$$\sum_m P(m|s', m')x^m = [G_{g-1}(x; T)]^{m'} \tag{3.8}$$

where the equality occurs as a result of the right side describing the probability of m infectious nodes in generation g assuming m' such nodes at $g-1$ from branching process theory.

For a given state (s', m') in generation $g-1$, m new infections will result in $s'+m$ cumulative infections in generation g. So, having m new infections occurs with probability $\psi_{s'm'}^{g-1} P(m|s', m')$, where the $\psi_{s'm'}^{g-1}$ term is the probability of being in the state (s', m') at generation $g-1$. Now, we can re-write the entire PGF for the state space of (s, m) at generation g as

$$\begin{aligned}\Psi_0^g(x,y) =& \sum_{s,m} \psi_{sm}^g x^s y^m = \sum_{s',m} \psi_{sm}^g x^{s'} (xy)^m && (3.9)\\ =& \sum_{s'm'} x^{s'} \sum_m \psi_{s'm'}^{g-1} P(m|s',m')(xy)^m \\ =& \sum_{s',m'} \psi_{s'm'}^{g-1} x^{s'} \sum_m P(m|s',m')(xy)^m \\ =& \sum_{s'm'} \psi_{s'm'}^{g-1} x^{s'} [G_{g-1}(xy;T)]^{m'} \\ =& \Psi_0^{g-1}(x, G_{g-1}(xy;T)) && (3.10)\end{aligned}$$

This defines a recurrence relation when $g \geq 1$ and we have $\Psi_0^0 = xy$ as the assumption that there is only one initial infectious individual it must be that $\psi_{sm}^0 = \delta_{s1}\delta_{m1}$.

We also note that the probability of having s (cumulative) or m (current) infectious nodes in generation g can be computed marginally from $\Psi_{sm}^g(x, 1)$, given by

$$p_s^g = \sum_m \psi_{sm}^g \text{ and } p_m^g = \sum_s \psi_{sm}^g \qquad (3.11)$$

respectively. The distribution of p_s^g from Eq. (3.11) is our main quantity under study, which is shown in Fig. 3.2, along with event-driven simulations to validate the theory.

3.2.2 Simulations of Continuous SIR Dynamics

For a realistic model of the spread of disease in a population, we simulate a stochastic disease process of an SIR epidemic on synthetic contact networks in continuous time [48]. We use an event-driven framework, which is advantageous for epidemic modeling, because it is much faster compared to a brute-force time-step simulation due to its leveraging of the Markovian dynamics of infectious and recovery periods of individuals [49, 50, 51]. Recall in the SIR model that nodes inhabit the susceptible, infectious, and recovered states as the disease progresses, where nodes become infected if one of their infectious neighbors transmits to them. The standard SIR model is governed by two rate parameters; β, the rate per unit time of an infectious node transmitting to other nodes, and γ, the rate per unit time of an infected node recovering. In a continuous time event-driven simulation, infection and recovery are Poisson processes occurring at rates β and γ respectively, and relate back to the percolation framework by defining transmissibility $T = \beta/(\beta + \gamma)$.

We draw a random network from a given degree distribution, and begin the simulation algorithm by assuming a random initial infectious node, patient zero, with degree k_0. Patient zero could either recover before transmitting to any of its neighbors, or infect one or more of its neighbor nodes. The stochastic process governing the behavior of a single infected node is the superposition of $\hat{k}+1$ Poisson processes, where $\hat{k}$ is the number of susceptible neighbors, and with one extra process governing the time until recovery. Say patient zero infects Neighbor 1, who has k_1 neighbors. Then with two infectious nodes, the stochastic process encompassing all possible events is a Poisson process with rate $(\hat{k}_0 - 1)\beta + \hat{k}_1\beta + 2\gamma$, and so on as more nodes become infected.

Each possible event given by the sub-processes is the first to occur with probability $i/(\hat{k}\beta + \gamma)$ where $i \in \{\beta, \gamma\}$, with the Poisson process rate term from $\hat{k}$ reducing if an infection event occurs, and stopping entirely if the contagious node recovers. The disease process for the whole population is a natural extension of that described above, with each node assumed identical apart from degree. The evolution of the unmitigated disease process from here is intuitive, either eventually all the infectious nodes recover or the whole connected population becomes infected.

Computationally, the above process is simulated by generating a random network from a given degree distribution using a large enough number of nodes, N, that $k \ll N$. As we cannot simulate numerically on an infinite network, the best choice for N is the largest value the numeric simulation can support. A node is randomly selected to be patient zero, and the disease spread proceeds via stochastic event-driven simulation, often known as the Gillespie algorithm [52]. Continuous time is tracked using a random variable τ, known as the waiting time, which is exponentially distributed with parameter the sum of the rates of all the potential infection and recovery events. Each competing process is the first to occur with probability of its own rate divided by the sum of all rates of that process type, as described by the Poisson process above. The simulation is advanced via this algorithm until either there are no more infectious nodes or until there are no more susceptible nodes, and allows for obtaining the resulting evolution of the disease spread in terms of both generations of infection and continuous time.

3.3 Results

We employ the generational size distribution theory to explore the evolution of epidemic size on a variety of network structures, and compare the generating function theory against continuous-time simulations. We use the event-driven simulation framework so that we can track the progression of the epidemic in both continu-

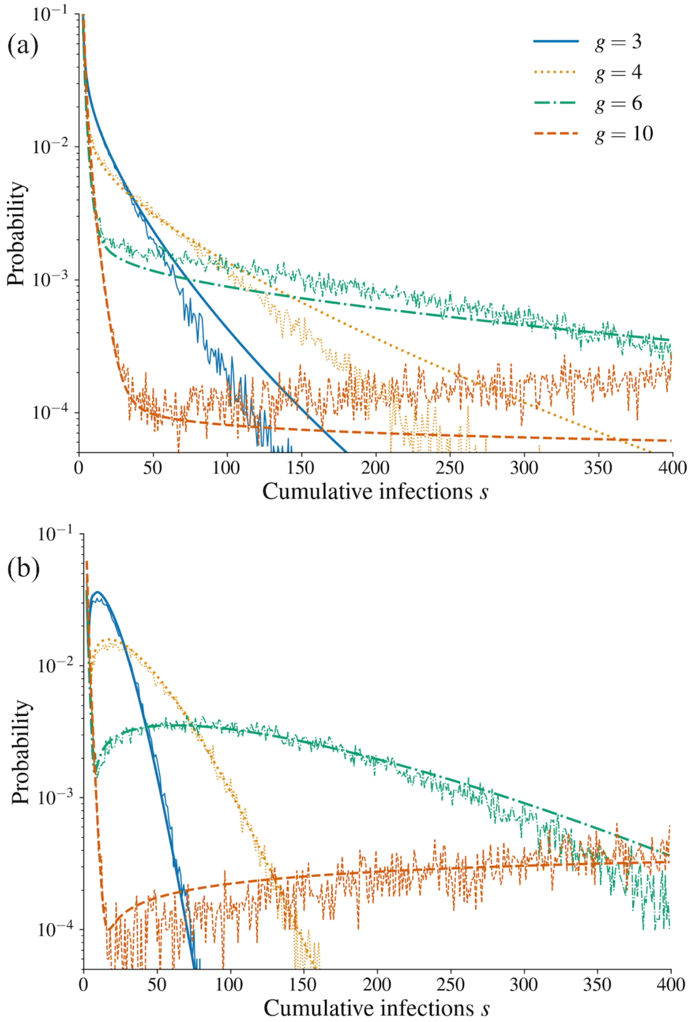

Figure 3.2: Time evolution of epidemics on homogeneous and heterogeneous networks. We show the probability of having s cumulative cases by and including generation g for select generations. Panel (a) shows the results on a modified power-law random networks with degree distribution given by $p_k = k^{-2}e^{-k/10}$ with average degree $\langle k \rangle = 1.79$, average excess degree $\langle q \rangle = 3.04$, $\beta = 0.004$ and $\gamma = 0.001$ such that $R_0 = T\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = \frac{\beta}{\beta+\gamma}\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = 2.44$. The smooth lines show the theoretical prediction for the probability distribution of cumulative infections. The distributions are validated by 75,000 simulations performed on 150 random network realizations with 10,000 nodes, following the process outlined in Section 3.2.2. Panel (b) shows the results of equivalent analysis and simulations on Erdős-Rényi random networks with $\langle k \rangle = 2.5$, $\beta = 0.004$ and $\gamma = 0.001$ such that $R_0 = 2.0$.

ous time as well as the generation sizes corresponding with the branching process, which allows us to validate the theoretical distributions, as well as introduce a preliminary prediction for the expected continuous time emergence of successive generations. Then, we use the PGF framework to measure the probability of an epidemic surviving, or continuing on, past an arbitrary generation, depending on the characteristics of the network and disease.

3.3.1 Time Evolution on Homogeneous and Heterogeneous Networks

In Fig. 3.2 we show the probability distributions of cumulative infections by the specific generation for two network models. It is noteworthy that this modeling method holds for configuration model networks with varying types of degree distributions. Here, we show the results on a modified power law network and an Erdős-Rényi (ER) network both used in Ref. [46]. The ER network has mean degree and excess degree $\langle k\rangle = \langle q\rangle = 2.5$, while the modified power law has mean degree $\langle k\rangle = 1.79$ and average excess degree $\langle q\rangle = 3.04$, a more heterogeneous distribution. We demonstrate that the distributions of outbreak size appear to be more a result of the stochastic nature of the disease spread, rather than the structure of the network, though the structure does play a role in the shape of the distribution.

Our results convey that there is not one clear trajectory of a typical large outbreak, in contrast to traditional results with deterministic modeling. Instead, the stochastic nature of epidemic size is captured by a long tail in the distribution of cumulative cases over each epidemic generation. One unique aspect of our work is that we validate this result using continuous-time simulations showing the same shape and long tail in outbreak size distributions as our analytical results. We do anticipate the simulated distributions and analytical distributions to vary from each other due to a few factors including the finite-size effects of simulated networks, and the fact that we compare a discrete analysis with a continuous-time process, but the general behavior appears consistent throughout the different generations.

We also find that on both the heterogeneous network and the homogeneous network, there is a high probability of an outbreak going extinct before growing large, however, if it does take off, the distribution levels off over the space of epidemic size. That is to say, if indeed an epidemic takes off and has arrived at generation six, via a transmission chain of length six, there is an almost equal probability of having anywhere from 50 to 500 cumulative cases by the time generation six is reached. We emphasize that these results display the unpredictability in early stages of epidemics, even ignoring the difficulty of estimating model parameters, it is near impossible to infer with much confidence how many infections there may actually be in the population.

3.3.2 Generations of Infection in Continuous Time

While the behavior of the epidemic in our formalism is described by generations of infection, most applications of disease models desire descriptions of the dynamics in continuous time. We find early agreement from our model of generational infections with a distribution in continuous time, described in terms of the expected time of emergence of an arbitrary generation g. The agreement is surprising since one might not expect a consistent relationship between a generation number and the expected

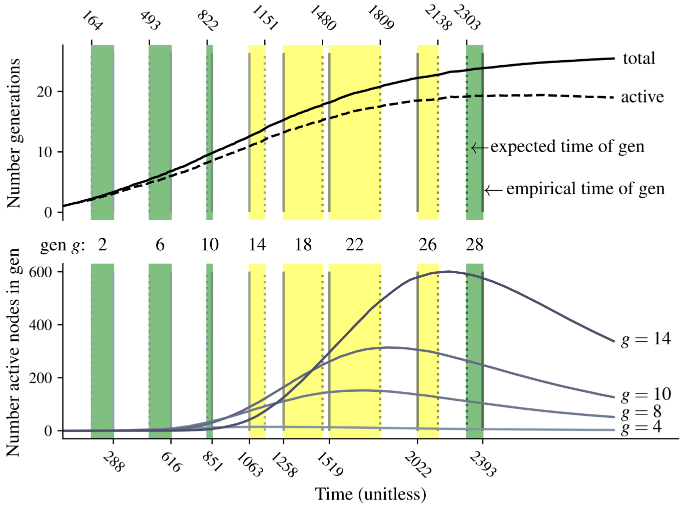

Figure 3.3: Time evolution of the active epidemic generations and emergence times. Top panel (curves): average number of total and *active* generations at time t for the modified power-law network with degree distribution $p_k = k^{-2}e^{-k/10}$. Bottom panel (curves): average number of active *nodes* belonging to each generation shown over time to accompany the top panel. The tick marks in the top panel (and dotted vertical lines) correspond to increments of $g/\langle q\rangle\beta$, the predicted generational emergence times, and the bottom tick marks (and solid vertical lines) correspond to the average empirical time at which that generation g emerged, for an example network. If the average time of emergence was greater than its respective $g/\langle q\rangle\beta$ value, that is, after the predicted time, the difference is highlighted in green. If the average empirical time was less than that predicted, the difference is highlighted in yellow.

time of its emergence given the observed heterogeneity of early spread in Fig 2. Yet, by defining the *emergence* of generation g as the time its first member is infected, we find a simple linear relationship that allows us to map the PGF framework to continuous time.

We can show that the expected time of emergence of an arbitrary generation g is given by

$$\mathbb{E}[t(g)] = \frac{g}{\langle q\rangle\beta}$$

where $\langle q\rangle = G_1'(1)$ is the average excess degree of the network. We arrive at this expression for $\mathbb{E}[t(g)]$ via a simple argument over the Poisson process governing how nodes in generation $g-1$ can lead to the first cases of generation g. Each node of generation $g-1$ can recover at rate γ but also has on average $\langle q\rangle$ neighbors they can infect at rate β. Therefore, the first event around them will occur at a combined rate $\alpha = \langle q\rangle\beta + \gamma$ and will lead to a case in generation g with probability $T_q = \langle q\rangle\beta/(\langle q\rangle\beta + \gamma)$. The first infectious node in generation $g-1$ can therefore lead to the emergence of generation g after $1/\alpha$ with probability T_q; if not, or the second node in generation $g-1$ could lead to the emergence of generation g with probability $T_q(1-T_q)$ after $2/\alpha$ (approximate delay between the first and second node of generation $g-1$ plus the expected time to generation g); and so on for the third node and beyond. This sequence of possibilities can be summarized by an arithmetico-geometric sum,

$$\begin{aligned}\mathbb{E}[t(g) - t(g-1)] &= \frac{T_q}{\alpha}\sum_{k=1}^{\infty}(1-T_q)^{k-1}k \\ &= \frac{T_q}{\alpha}\frac{1}{T_q^2} = \frac{1}{\langle q\rangle\beta}\ .\end{aligned} \tag{3.12}$$

In Fig. 3.3, we demonstrate in practice how the expected time of emergence of consecutive generations falls in line with the predicted time measure. To show intuitively why the we see this phenomenon, we show the time evolution of the active epidemic generations. We track time in two ways; in continuous time following the event-driven process discussed in Section 3.2.2, and also in terms of the expected time of emergence of each generation g, in the form $t = g/\langle q\rangle\beta$. We define a generation to be *active* if it contains one or more nodes who are not recovered and have susceptible neighbors at time t in the simulation. We illustrate the number of total and active generations over time, as well as the number of active *nodes* belonging to each generation, which helps clarify the roles each generation plays in causing the next wave of infection over a given interval in continuous time.

Having an understanding of the time at which a generation will emerge acts as a complement to the probabilities of extinction and cumulative cases discussed in Sections 3.3.1 and 3.3.3. Equipped with the distributions describing the stochasticity of outbreaks, the expected time mapping can be a tool for analysis of the dynamics of the worst-case scenarios when an outbreak does occur.

3.3.3 Probability of Pandemics or Stochastic Extinction

The PGF generational theory can also be used to measure the probability that an emerging epidemic has a chance of dying off on its own, or "surviving". Deterministic models always predict that an epidemic will occur if $R_0 > 1$, that is, if the average number of secondary infections caused by an infectious individual is more than one. In reality, there is a non-zero chance the outbreak will die off by chance, shown in Fig. 3.4. Branching process models have been used in theoretical epidemiology for estimating such probabilities [55, 56, 57]. However, simple branching process models are Markovian in the number of active infections, m. This is problematic in an applied setting as cumulative cases, s, is often the available data. Moreover, we show that conditioning on "reaching" generation g, the probability of the outbreak going extinct after generation g rather than becoming an epidemic is path dependent in the sense that the value of s at g changes the extinction probability, shown in Fig. 3.5.

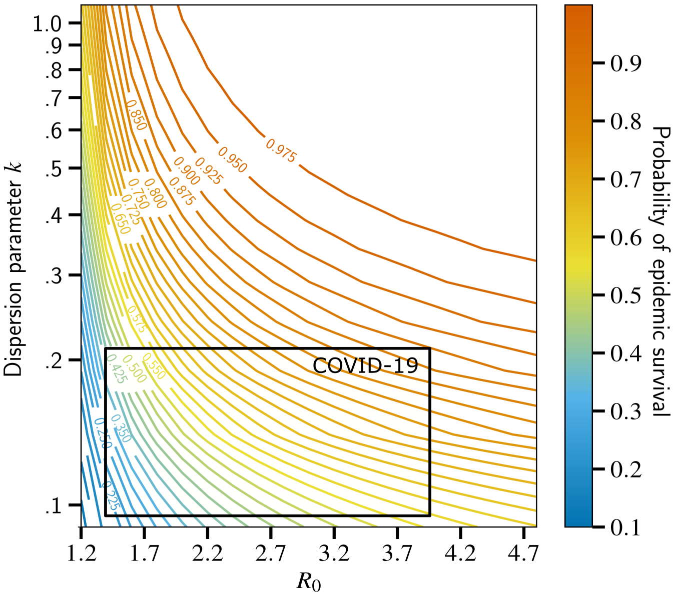

Figure 3.4: Probability of epidemic survival as a function of contact structure. The contour plot shows the initial probability of epidemic survival for negative binomial distributions of infections over a range of possible R_0 values (average transmissions per case) and dispersion parameter k (inverse of heterogeneity). The box highlights estimates for COVID-19 based on data from Wuhan, China [53]. We assume an epidemic generation of $g = 4$ and $s = 16$ cases which corresponds to the epidemic growing from 1 case to 16 over 4 generations. Using a serial interval of 4 days, the average of the estimated range for COVID-19 [54], this tracks to roughly over two weeks of spread. Similarly, in the state of Washington, the first recorded case of COVID-19 occurred on January 21st, 2020 but following cases were only identified on February 19th and increased to 18 by March 2nd. This figure illustrates how these cumulative case data could have been used in real time with our theoretical tools to estimate epidemic risk.

To utilize the extinction probabilities, we want to look specifically at the variable ρ_s^g, the probability that given s cumulative cases at generation g that the epidemic will go extinct, or die off, sometime afterwards. Given that the evolution of m occurs as a branching process with the offspring PGF given by Eq. (3.6), one can easily compute the probability of extinction of a single infection chain, p_e, as the solution of $p_e = G_1(p_e; T)$ using branching process theory [38]. The distribution of probabilities of *reaching* (s, m) in the state space for each g is given by ψ^g_{sm}, as discussed in Section 3.2.1. We define a new distribution, that of the probability of the outbreak still being in existence in generation g, by

$$\tilde{\psi}^g_{sm} = \begin{cases} \dfrac{\psi^g_{sm}}{\displaystyle\sum_{s',m'>0} \psi^g_{s'm'}} & m > 0 \\ & \\ 0 & \text{otherwise} \end{cases} . \tag{3.13}$$

Thus, ρ^g_s, the probability of the epidemic going extinct given it has arrived at s cases by generation g is given by

$$\rho^g_s = \sum_m \frac{\tilde{\psi}^g_{sm}}{\sum_{m'} \tilde{\psi}^g_{sm'}} p^m_e . \tag{3.14}$$

The probability of epidemic *survival* for an epidemic being active in generation g with s cumulative infections is then given by $1 - \rho^g_s$. We illustrate an example of how the survival probabilities change depending on the underlying network and disease parameters in Fig. 3.4.

3.3.4 Epidemic Probability and COVID-19 Data

We now apply the epidemic survival probability theory to early incidence of COVID-19 cases in the US. This allows us to look at the evolution over time of public health risk, while taking into account the stochastic elements of the early spread. We assume a distribution of secondary infections parameterized as a negative binomial with R_0, the basic reproductive number, and k, the dispersion parameter of the contact network [53]. Together, these parameters determine the average behavior of disease spread where k is responsible for the variation in secondary cases, in turn affecting the likelihood of superspreading events [58, 59, 60]. A low dispersion parameter k (high heterogeneity) means that a select few cases may cause the majority of secondary infections [61], which in our framework here might correspond to a single case leading to an extreme increase in cases in the next generation. For that reason, it is often assumed that the early spread of an epidemic is highly sensitive to superspreading events [62]. Yet, as shown in Fig. 3.2, heterogeneity in contact structure actually has less of an impact on the distribution of outcomes than the inherent stochasticity of transmission.

In Fig. 3.4 we show the probability of epidemic survival (that is, the probability of an epidemic continuing to grow) with a fixed generation $g = 4$ and fixed cumulative cases $s = 16$ over a range of R_0 and k values, highlighting parameter estimates for COVID-19 [61]. Despite the relatively low number of cases after several generations, clearly affected by the lack of testing resources at the time, the chances of the epidemic stochastically dying out were already close to a simple coin flip. In Fig. 3.5, we show the inverse problems: fixing disease parameters and varying temporal variables. We set $R_0 = 2.5$ and $k = 0.1$, falling within the range of values for COVID-19, and track seven US states over time to observe where their disease progression state falls in the probability space of epidemic survival.

Guided by the results shown in Fig. 3.3, we proceed knowing that our model predicts generations to emerge in linear increments of time. We use the serial interval of 4 days, taken from the window for COVID-19 [54] to correspond with successive generational emergence. We observe that several states hovered around a low probability of epidemic survival at low early cases, but very quickly crossed to a much higher bracket where natural extinction of the disease spread is virtually impossible. The states of Washington and Massachusetts each took only two generations to cross from sub to supercritical epidemic survival probability, even derived from limited data and poor testing at the time. The extraordinary leap in epidemic probability from just one generation to the next explain, in part, why it was so hard for public health systems to react and adapt to the spread of COVID-19.

3.4 Discussion

Temporal models of disease spread often fall in one of three categories. (i) Compartmental models that are deterministic in nature as they rely on ordinary differential equations. Therein, uncertainty only stems from our imperfect knowledge of model parameters, rather than from the inherent randomness of disease transmission. (ii) Complicated agent-based models that lose the tractability of analytical models, require significant amount of data to parameterize and do not produce explicit likelihood of outcomes. (iii) Time series analyses that can produce probabilistic forecasts. This last approach can produce useful predictions by ignoring transmission mechanisms or contact structure, but that perspective also precludes it from evaluating potential interventions that affect individual parameters or contact structure.

In this paper, we have shown that analysis of branching processes often used to only study the final state of epidemic models can actually combine the strengths of these different approaches by including stochasticity, contact heterogeneity and even individual characteristics [39, 33, 65]. The reason this framework is usually used to solely predict the probability and final size of an epidemic is that the mathematical

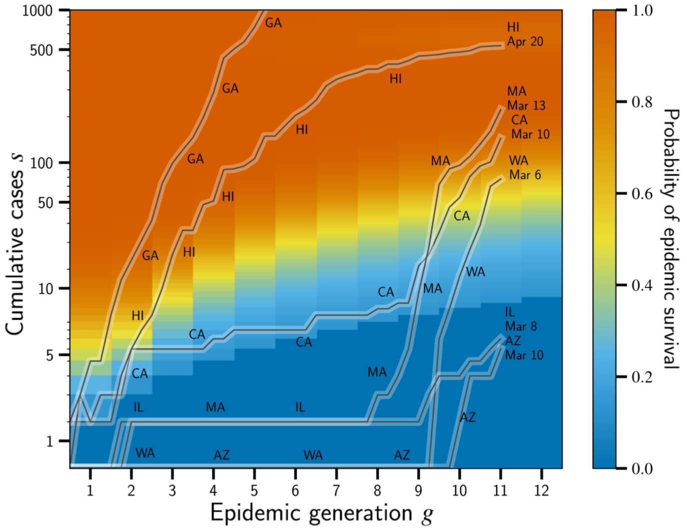

Figure 3.5: Probability of epidemic continuing on as a function of case counts and time. As a simple comparison, we use early data from the COVID-19 pandemic and show a selection of U.S. states following unique timelines from the first recorded case onward. This simple visualization is not meant as a validation but only to explore how quickly our predictions for the probability of the epidemic not dying off changes as an epidemic grows. To calculate these probabilities, we use a negative binomial distribution of secondary infections with $k = 10^{-1}$, $R_0 = 2.5$ along with data from the COVID-19 Repository by the CSSE at Johns Hopkins [63, 64]. The first data point for each state shown correspond to the first date on which 1 or more cases were recorded. Raw data of cumulative case counts are used, and plotted on the same range of epidemic generations for purposes of comparison, despite an evident variability in the duration of generation length. Using a serial interval of 4 days, progressive generations are shown along the horizontal axis (generation two corresponds to 8 days, for example). On the vertical axis, the cumulative case counts for each state are plotted. We see how a state's proclivity to the epidemic taking off changes over the course of successive generations. Several states such as California, Massachusetts, and Washington had a lower probability of epidemic survival early on, then crossed the band into a higher likelihood over a short time span. Although the data used in this figure does not take into account factors such as missing count data, it serves as a visualization of how sharply the interplay of generation of epidemic and cumulative cases demarcate the probability of the epidemic continuing.

treatment involves integrating over contacts and therefore time [53]. However, we provided a first demonstration that the predictions made over generations by the branching process are actually very close approximation of continuous time epidemic dynamics on equivalent contact networks. This result alone justifies a large body of work and creates a foundation for analytical, probabilistic, epidemic forecasts based on PGFs.

Our probabilistic and temporal forecasts allowed us to uncover the diversity of epidemic courses, in the form of an unusually broad distribution of potential transmission trees over time. We have also shown that these flat distributions emerge on both homogeneous (e.g. Erdős-Rényi graphs) and heterogeneous (e.g. scale-free) contact networks. This phenomenon is therefore driven by the stochasticity of disease transmission rather than by the complexity of the contact structure. This broad likelihood of early disease incidence justifies our use of a stochastic branching process, whereas deterministic models would typically track only the average or expected number of cases which is a poor description of flat distributions.

Our framework currently rests on a few assumptions, including that there are a finite number of active generations at any given time and that the distribution of contacts and transmission probability do not change over time. This first assumption was tested in Fig. 3.3 where we show that a simple network-based serial interval provides a great approximation for time of emergence of epidemic generations in the continuous dynamics. Explaining both why and how we can align the generation-based branching process with the underlying temporal dynamics.

Our assumption on the constant contact patterns and transmissibility provide a great road map for future work. In Eq. (3.1), we formulate our PGFs on a per generation basis, which would allow us to change these patterns over time to model adaptive behavior or top-down interventions (e.g. lockdowns limiting contacts or masks reducing transmissibility). Doing so would allow us to provide probabilistic forecasts not only of disease dynamics but also of the impact and timing of particular interventions.

Importantly, our results on the diversity of epidemic courses highlight how little information can actually be gathered from early incidence data. In Fig. 3.2, we see that the same disease in the same population can be roughly as likely to produce 40 or 400 cases after 10 epidemic generations.

Finally, our results on epidemic survival show how quickly a situation can move from an uncertain outbreak to supercritical exponential growth. Due to both the randomness of disease spread and the imperfect COVID-19 testing protocols from early 2020, most states in the US moved from below 20% survival probability of the epidemic to above 80% in about two epidemic generations (2 weeks or less).

Altogether, our results stress the danger of justifying a lack of intervention with slow trends in early disease spread data. Little can be learned about transmission mechanisms and dynamics from the first few epidemic generations. The distribution of epidemic courses is mostly driven by the inherent randomness of transmission, and the window in which the dynamics settle into their subcritical or supercritical behavior tends to be unfortunately narrow, which leaves little room for fast adaptive responses.

Faced now with emergence of variants of COVID-19 around the world, the current situation is often reminiscent of the scenario in the state of Washington during January of 2020 —sporadic clusters of cases with an unclear growth trajectory. We see from the data in Washington, as well as many other states and countries, how quickly cases explode and what that means for the likelihood of controlling the epidemic without external intervention efforts. Slow initial disease growth does not preclude a rapid increase shortly thereafter.

Acknowledgments

A.J.A. and L.H.-D. acknowledge financial support from the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award. M.C.B. is supported as a Fellow of the National Science Foundation under NRT award DGE-1735316 and N.J.R. is supported by the University of Vermont. A.A. acknowledges financial support from the Sentinelle Nord initiative of the Canada First Research Excellence Fund and from the Natural Sciences and Engineering Research Council of Canada (project 2019-05183).

Bibliography

[1] First Travel-related Case of 2019 Novel Coronavirus Detected in United States, 2020.

[2] D. Oxley and J. Ryan. “Volatile and unpredictable”: Life Care Center speaks publicly for the first time since COVID-19 outbreak. *KUOW News and Inf.*, 2020, March 7.

[3] O. Sullivan. “Coronavirus death toll rises to nine in Washington". *Kirkland Reporter*, 2020, March 3.

[4] A. Sundell. New coronavirus cases confirmed in Snohomish, King counties. *KING-TV*, 2020, February.

[5] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics-I. *Proc. R. Soc. Lond., A*, 115(772):700–721, 1927.

[6] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics—II. the problem of endemicity. *Proc. R. Soc. Lond., A*, 138(772):55–83, 1932.

[7] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics—III. further studies of the problem of endemicity. *Proc. R. Soc. Lond., A*, 141(772):94–122, 1933.

[8] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control.* Oxf. Univ. Press, Great Clarendon Street, Oxford OX2 6DP, 1991.

[9] M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals.* Princet. Univ. Press, 41 Williams St, Princeton, New Jersey 08540, 2007.

[10] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87(3):925, 2015.

[11] Ibana H. Threshold and stability results for an age-structured epidemic model. *J. Math. Biol.*, 28:411–434, 1990.

[12] W. Huang, K. L. Cooke, and C. Castillo-Chavez. Stability and bifurcation for a multi-group model for the dynamics of HIV/AIDS transmission. *SIAM J. Appl. Math*, 52:835–854, 1992.

[13] B. Bolker and B. Grenfell. Space, persistence and dynamics of measles epidemics. *Philos. Trans. R. Soc. B*, 348:309–320, 2015.

[14] A.L. Lloyd and V.A.A. Jansen. Spatiotemporal dynamics of epidemics: synchrony in metapopulation models. *Math. Biosci.*, 188:1–16, 2004.

[15] F. Ball and Donnelly. P. Strong approximations for epidemic models. *Stoch. Process. Their Appl.*, 55(1):1–21, 1995.

[16] F. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *Ann. Appl. Probab.*, 7(1):46–89, 1997.

[17] L.S.J. Allen. *Stochastic Population and Epidemic Models.* Springer, 2015.

[18] L.S.J. Allen. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infect. Dis. Model.*, 2(2):128–142, 2017.

[19] W. Wang, Y. Cai, Z. Ding, and Z. Gui. A stochastic differential equation sis epidemic model incorporating ornstein–uhlenbeck process. *Phys. A: Stat. Mech. Appl.*, 509(4), 2018.

[20] D.A. Gray, L. Greenhalgh, L. Hu, X. Mao, and J. Pan. A stochastic differential equation sis epidemic model. *SIAM J. Appl. Math.*, 71(3):876–902, 2011.

[21] B. F. Finkenstädt and B. Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. *Appl. Stat.*, 49:187–205, 2000.

[22] Z. Zhan, W. Dong, Y. Lu, P. Yang, Q. Wang, and P. Jia. Real-time forecasting of hand-foot-and-mouth disease outbreaks using the integrating compartment model and assimilation filtering. *Sci. Rep.*, 9:1–9, 2018.

[23] R. Allard. Use of time-series analysis in infectious disease surveillance. *Bull. World Health Organ.*, 76:327–333, 1998.

[24] B. Lopman, B. Armstrong, C. Atchinson, and J. J. Gray. Host, weather and virological factors drive norovirus epidemiology: Time-series analysis of laboratory surveillance data in England and Wales. *PLoS One*, 4:e6671, 2009.

[25] W. Hu, S. Tong, K. Mengersen, and B. Oldenburg. Rainfall, mosquito density

and the transmission of Ross River virus: A time-series forecasting model. *Ecol. Model.*, 196:505–514, 2006.

[26] P. C. L. Silva, P. V. C. Batista, H. S. Lima, M. A. Alves, F. G. Guimarães, and R. C. P. Silva. Covid-abs: An agent-based model of covid-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos Solitons Fractals*, 139:110088, 2020.

[27] N. M. Gharakhanlou and N. Hooshangi. Spatio-temporal simulation of the novel coronavirus (COVID-19) outbreak using the agent-based modeling approach (case study: Urmia, Iran). *Inform. Med. Unlocked*, 20:100403, 2020.

[28] E. Cuevas. An agent-based model to evaluate the COVID-19 transmission risks in facilities. *Comput. biol. med.*, 121:103827, 2020.

[29] N. Hoertel, M. Blachier, C. Blanco, M. Olfson, M. Massetti, F. Limosin, and H. Leleu. Facing the COVID-19 epidemic in NYC: a stochastic agent-based model of various intervention strategies. *MedRxiv*, pages 1–34, 2020.

[30] A. Staffini, A. K. Svensson, U.-I. Chung, and T. Svensson. An agent-based model of the local spread of SARS-CoV-2: Modeling study. *JMIR med. inform.*, 9(4):e24192, 2021.

[31] V. Srikrishnan and K. Keller. Small increases in agent-based model complexity can result in large increases in required calibration data. *Environ. Model. Softw.*, 138:104978, 2021.

[32] L. Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bull New Ser. Am Math Soc.*, 44(1):63–86, 2007.

[33] E. Kenah and J. M. Robins. Second look at the spread of epidemics on networks. *Phys. Rev. E*, 76(3):036113, 2007.

[34] J. C. Miller. Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. *Phys. Rev. E*, 76(1):010101, 2007.

[35] E. Kenah and J. C. Miller. Epidemic percolation networks, epidemic outcomes, and interventions. *Interdiscip. perspect. infect. dis.*, 2011:1–13, 2011.

[36] K. B. Athreya and P. E. Ney. *Branching Processes.* Springer-Verlag Berlin Heidelberg, New York, 1972.

[37] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001.

[38] J. C. Miller. A primer on the use of probability generating functions in infectious disease modeling. *Infect. Dis. Model.*, 3:192–248, 2018.

[39] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, 2002.

[40] J. Levesque, D. W. Maybury, and R.H.A. D. Shaw. A model of COVID-19 propagation based on a gamma subordinated negative binomial branching process. *J. Theor. Biol.*, 512:110536, 2021.

[41] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge. The challenges

of modeling and forecasting the spread of COVID-19. *Proc. Natl. Acad. Sci.*, 117(29):16732–16738, 2020.

[42] I. A. Mitrofani and V. P. Koutras. A branching process model for the novel coronavirus (Covid-19) spread in Greece. *Int. J. Model. Optim.*, 11(3), 2021.

[43] L. Zhang, H. Wang, Z. Liu, X. F. Liu, X. Feng, and Y. Wu. A heterogeneous branching process with immigration modeling for COVID-19 spreading in local communities in China. *Complexity*, 2021:1–11, 2021.

[44] M. Akian, L. Ganassali, S. Gaubert, and L. Massoulié. Probabilistic and mean-field model of COVID-19 epidemics with user mobility and contact tracing. *arXiv*, pages 1–23, 2020.

[45] Sadamori Kojaku, Laurent Hébert-Dufresne, Enys Mones, Sune Lehmann, and Yong-Yeol Ahn. The effectiveness of backward contact tracing in networks. *Nature Physics*, 17(5):652–658, 2021.

[46] P.-A. Noël, B. Davoudi, R. C. Brunham, L. J. Dubé, and B. Pourbohloul. Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E*, 79:026101, 2009.

[47] H. S. Wilf. *generatingfunctionology.* CRC press, Boca Raton, Florida, 2005.

[48] andrea-allen/epintervene: Initial Release v1.0.0, 2021.

[49] I. Z. Kiss, J. C. Miller, and P. L. Simon. *Mathematics of Epidemics on Networks: from exact to approximate models.* Springer, 2019.

[50] J. C. Miller and T. Ting. EoN (Epidemics on Networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. *J. Open Source Softw.*, 4(44):1731, 2019.

[51] P. Bauer, S. Engblom, and S. Widgren. Fast event-based epidemiological simulations on national scales. *Inter, J. High Perform. Comput. Appl.*, 30(4):438–453, 2016.

[52] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, 1977.

[53] L. Hébert-Dufresne, B. M Althouse, S. V Scarpino, and A. Allard. Beyond R0: heterogeneity in secondary infections and probabilistic epidemic forecasting. *J. R. Soc. Interface*, 17(172):20200393, 2020.

[54] Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling, and L. A. Meyers. Serial interval of COVID-19 among publicly reported confirmed cases. *Emerg Infect Dis.*, 26(6):1341, 2020.

[55] N. G. Becker. Estimation for discrete time branching processes with applications to epidemics. *Biometrics*, 33:515–522, 1977.

[56] N. G. Becker. On parametric estimation for mortal branching processes. *Biometrika*, 61:393–399, 1974.

[57] O. Diekmann, H. Heesterbeek, and T. Britton. *Mathematical Tools for Understanding Infectious Disease Dynamics.* Princet. Univ. Press, 41 Williams St,

Princeton, New Jersey 08540, 2013.

[58] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005.

[59] C. L. Althaus. Ebola superspreading. *Lancet Infect. Dis.*, 15(5):507–508, 2015.

[60] A. J. Kucharski and C. L. Althaus. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. *Eurosurveillance*, 20(25), 2015.

[61] J. Riou and C. L. Althaus. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance*, 25(4):2000058, 2020.

[62] B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hébert-Dufresne, and H. Hu. Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLoS Bio.*, 18(11):e3000897, 2020.

[63] Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, 2021.

[64] E. Dong, H. Du, and Gardner L. An interactive web-based dashboard to track covid-19 in real time. *Lancet Infect. Dis.*, 20(5):533–534, 2020.

[65] A. Allard, B. M. Althouse, S. V. Scarpino, and L. Hébert-Dufresne. Asymmetric percolation drives a double transition in sexual contact networks. *Proc. Natl. Acad. Sci. U.S.A.*, 114(34):8969–8973, 2017.

Chapter 4

Application of both models

Abstract

Stochastic models for disease spread on networks are useful because they quantify the uncertainty associated with spread. One drawback is that branching process methods which are powerful stochastic analysis tools for disease spread do not traditionally provide solutions for the temporal dynamics of disease spread, and are more concerned with quantities relating to the final state. In contrast, Chapter 2 introduced a temporal network model for predicting how the disease spread process would change depending on the temporal network. Here, we show a proof of concept for how temporal network theory could be used in conjunction with a stochastic branching process to estimate probability distributions for the early generations of epidemic spread in a network that is dynamic over continuous time.

4.1 Probability Generating Functions on Temporal Networks

The probability generating function framework for the generational evolution of disease spread on networks is useful for understanding the stochastic nature of the process. It provides probability distributions for generations of infection to understand the variability the course of the disease may take.

As useful as it is, the PGF framework requires a static network, in which the generating functions are derived using power series representations of the degree distribution of the underlying contact network [1, 2]. As discussed in Chapter 2, networks often evolve over time as well, and these structural changes are important to the overlying spreading process.

Interventions can be applied to the generational PGF framework by changing something about the structure or dynamics at assigned generations. But to modify the network at key continuous times would require a way to map from continuous time to discrete generations, if possible. This is not a trivial task, as the relationship between stochastic analysis models using branching process theory does not have a well-defined relationship with continuous time.

Fortunately, we showed in Chapter 3 that we can define the expected continuous time value of emergence of the first member of a generation roughly as $E[t_g]$ as a function of the spreading rate β and $\langle q \rangle$, the average excess degree of the network. This is different than the estimation for the size of the generation, just a continuous time estimate of when the first member of generation g will be infected.

Therefore, we can use the continuous time partition of a given temporal network to estimate for which generation g will be emerging at the start of that snapshot, and then use the corresponding snapshot's degree distribution to solve for the infection distribution for the size of that generation.

The challenge that arises using given snapshot boundary times is that we end up deriving generation values that are between integers. The problem here is that the formalism is defined upon the generations having discrete values, otherwise the branching process framework does not make sense. So, for the purpose of demonstrating how the framework could be applied in theory, we reverse-engineer an example to reduce complications in this proof of concept.

Here we show a preliminary approach that seeks to bridge generational temporal evolution with continuous time evolution, though it warrants much further research as there are many parametric and structural considerations that are overlooked in this simple application.

4.2 Direct Relationship between Generations and Continuous Time

We take a temporal network following the structure discussed in the results of Chapter 2 and derive the generational distributions to predict the early epidemic spread on the temporal network by altering the network structure used in the model for deriving each subsequent generation.

Some notation for this: Let g_s be a continuous variable refer to the nearest generation at which snapshot s begins. g_s might take on a value between generations, which can be corrected for later. Let t_s be the starting time of snapshot s. Let d_s be defined by

$$d_s = g_{s+1} - g_s \tag{4.1}$$

representing the number of generations spanned by snapshot s. We have

$$t_1 = \frac{d_0}{q_0\beta} \tag{4.2}$$

$$t_2 = \frac{d_1}{q_1\beta} + \frac{d_0}{q_0\beta} \tag{4.3}$$

$$t_S = \sum_{s=0}^{S-1} \frac{d_s}{q_s\beta} \tag{4.4}$$

Solving for d_S, we have

$$d_S = \left(t_{S+1} - \sum_{s=0}^{S-1} \frac{d_s}{q_s\beta}\right) q_S\beta \tag{4.5}$$

where $g_{t_{s=0}} = 0, g_{t_{s=1}} = d_0, g_{t_{s=2}} = d_1 + d_0$, etc. Simply put,

$$g_{t_S} = \sum_{s=0}^{S-1} d_s. \tag{4.6}$$

Also note that re-arranging Eq. (4.5) produces the simple relationship

$$t_{S+1} = \frac{d_S}{q_S\beta} + \sum_{s=0}^{S-1} \frac{d_s}{q_s\beta} = \sum_{s=0}^{S} \frac{d_s}{q_s\beta} \tag{4.7}$$

where g_{t_S} is essentially the generation emerging at the time of the beginning of Snapshot S.

The problem is, if given an arbitrary $\langle q\rangle$ and β, then some d_s may be a non-integer, making each g_{t_S} also a non-integer value. Unfortunately, we can only build the generating functions for consecutive integer values, or else cannot use the combinatorial branching process approach.

For this example and proof of concept, we instead reverse engineer an example where we set each desired duration d_S, and each boundary time t_S, and derive $\langle q \rangle$ from the above equations. Then we create network snapshots with the derived q values. Using this new temporal network with 5 snapshots, we show changing the degree distribution to compute each new epidemic generation produces size distributions that agree well with continuous time event-driven Gillespie simulations, in which the network is switched at continuous time values t_S the boundaries of each snapshot.

To generate 5 synthetic snapshots, we derive $\langle q \rangle$ for each one. Starting with Eq. (4.5) again, we isolate q_S, as

$$q_S = \frac{d_S}{\left(t_{S+1} - \sum_{s=0}^{S-1} \frac{d_s}{q_s \beta}\right) \beta} = \frac{d_S}{\left(t_{S+1}\beta - \sum_{s=0}^{S-1} \frac{d_s}{q_s}\right)} \tag{4.8}$$

which is noticeably a recursive equation.

For the proof of concept, we set all $d_S = 1$ with the idea that each temporal snapshot should span a single generation. The variable t_{S+1} represents the boundary of each snapshot, beginning with Snapshot $S = 0$ with $t_0 = 0$. Then we use Eq. (4.8) to recursively define each excess degree for each snapshot q_S. Beginning with $S = 0$, we simply have

$$q_0 = \frac{d_0}{t_1 \beta} \tag{4.9}$$

with $d_0 = 1$ (though it could be set to as many generations we expect to appear in snapshot 0). which is simply the inverse of the standard equation introduced in

Chapter 3 for $E[t_{g=1}]$ which was

$$E[t_{g=1}] = \frac{1}{q\beta} \tag{4.10}$$

We then define an Erdős-Rényi network for each temporal snapshot with average degree (which in the Erdős-Rényi case is the same as average excess degree) of q_S, by defining each snapshot network as

$$G(n, p) := G(n = 5000, p = q_S/5000) \tag{4.11}$$

since the average degree of an Erdős-Rényi graph is $E[k] = N(p)$. A table of network statistics for the five networks created in shown in Tab. 4.2 .

4.3 Numerical Approach for Solving the Generating Functions

In the original method in Chapter 3, to solve for Eq. (4.12) we followed this mathematical framework to develop $\Psi_0^g(x, y)$. To do so computationally, we used an approximation of the infinite network assumption and solved for the phase space Ψ recursively.

In order to apply a new degree distribution at each generation g, we modify the numerical computation process at each recursive step. Here we describe the numerical approximation process for solving for the generating functions for both the original problem and the temporal one where we intervene with new networks at each generation.

Recall that the generating function for each generation g is defined as follows:

$$G_g(x;T_g) = \begin{cases} G_0(x;T) & g = 0 \\ G_1(x;T) & 0 < g < I \\ G_1(x;T_I) & g \geq I \end{cases} \tag{4.12}$$

and that ultimately, we solve for the phase space

$$\Psi_0^g(x,y) = \sum_{s,m} \psi_{sm}^g x^s y^m \tag{4.13}$$

and marginalize over m to obtain distributions over cumulative number of infections s for a given generation g.

As the distribution $\Psi_{s,m}$ is a distribution across two variables, it can be viewed as a matrix where each entry of the matrix as probability of having the ordered pair (s,m), which represents that phase space state of having s current infections in generation g, m of which belong to generation g. Note this implies the rows give total infections while the columns account for active infections in the given generation. Using Ψ_{sm}^g as a matrix representing the coefficients of Ψ_0^g, we have

$$\Psi_{sm}^0 = \begin{bmatrix} 0 & 0 & .... & 0 \\ 0 & \psi_{s=1,m=1}^0 = 1 & \psi_{s=1,m=2}^0 = 0 & .... \\ 0 & \psi_{s=2,m=1}^0 = 0 & .... & 0 \end{bmatrix} \tag{4.14}$$

One can see that for generation $g = 0$, as the formalism is defined, the only non-zero entry in Ψ_{sm}^0 is entry $(s,m) = (1,1)$. Then we need to find Ψ_{sm}^1 from this, beginning the recursion. We present only this one case as the algorithm applies for future g in a similar fashion.

Recall that we can write

$$\sum_{s,m} \psi^{g}_{sm} x^{s} y^{m} = \sum_{s'm'} \psi^{g-1}_{s'm'} x^{s'} [G_{g-1}(1 + (xy-1)T_g)]^{m'} \tag{4.15}$$

where

$$\sum_{m} P(m|s', m') x^{m} = [G_{g-1}(1 + (xy-1)T_g)]^{m'}. \tag{4.16}$$

In words, the left hand side of the equation describes the probability distribution over m new infections in generation g given the prior state (s', m') in generation $g-1$. This is intuitively derived from the right hand side, in which we know that $G_{g-1}(1 + (xy-1)T_g)$ describes the generating function for the number of infections caused by a *single* node in generation $g-1$. Since there are m' infectious nodes in generation $g-1$, the convolution of m' copies of the generating function generates a PGF describing the probability of all infections caused in the next generation g by all the infected nodes together.

Given

$$G_{g-1}(1 + (xy-1)T) = a_1 x^1 + a_2 x^2 + a_3 x^3 + ... + a_l x^l$$

we let l index the number of infections caused by a single node, and let a_l is the probability of that happening.

The outcome of the convolution

$$[G_{g-1}(1 + (xy-1)T)]^{m'}$$

for a fixed m' will be its own generating function, where we index the powers in terms of m, the number of infections caused in generation g by m' infected nodes of generation $g-1$. This new generating function looks like

$$\sum_m P(m|s', m')(xy)^m = \rho_1 x^1 + \rho_2 x^2 + ... + \rho_m x^m$$

where each ρ_m can be expressed as

$$\rho_m = \prod_{i=1}^{m'} a_i$$

and the a_i's come from the equation above for G_{g-1}. Further, each index power m will be the sum of exactly m' l's so that we have

$$\sum_m P(m|s', m')(xy)^m = \rho_1 x^{(l_1+...+l_{m'})=1} + \rho_2 x^{(l_1+...+l_{m'})=2} + ... + \rho_m x^{(l_1+...+l_{m'})=m}.$$

With s' fixed, the equation above describes a probability generating function over two variables, m, and m', the number of new infected nodes in generation g and $g-1$ respectively. We store the generating function coefficients over m for each value of m' in a matrix, M, where the rows are in m' and the columns in m. Thus, each row represents the coefficients of the generating function for a given s', m' for

$\sum_m P(m|s', m')(xy)^m$. M looks like

$$M = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \rho_{10} = a_0 & \rho_{11} = a_1 & \rho_{12} = a_2 & \dots & \rho_{14} = a_m \\ \rho_{20} & \rho_{21} & \rho_{22} & \cdots & \rho_{2m} \\ \vdots & & & & \\ \rho_{m'0} & \rho_{m'1} & \rho_{m'2} & \cdots & \rho_{m'm} \end{bmatrix} \tag{4.17}$$

In theory this matrix would be infinite, as our degree distributions are unbounded and the formalism is for an infinite network. For computational purposes, we use an appropriate truncation, to 400 coefficients.

Returning to solving for Ψ^g_{sm}, we have

$$\Psi^1_{sm} = \begin{bmatrix} 0 & 0 & .... \\ \psi^1_{s=1,m=0} & \psi^1_{s=1,m=1} & .... \\ \psi^1_{s=2,m=0} & \psi^1_{s=2,m=1} & .... \\ \vdots & & \end{bmatrix} \tag{4.18}$$

The problem reduces to finding ψ^1_{sm} for each pair (s, m). We use the following algorithm, which holds for all $g \geq 1$.

The forward recurrence for Ψ is given by

$$\Psi^g(x, y) = \sum_{s',m'}^{g-1} \psi^{g-1}_{s'm'} x^{s'} [G_{g-1}(1 + (xy - 1)T)]^{m'}$$

To generate each (s, m) matrix entry of Ψ^g_{sm}, we

- Assign s and m of interest.

- Derive s' such that $s = s' + m$

- The s'th row of $\Psi^{g-1}_{s,m}$ (a whole generating function) corresponds to the contributions $\psi^{g-1}_{s'm'}$ from (4.18)

- The mth column of M corresponds to all possible states of m' in gen $g-1$ for which m became infected in gen g.

- Thus, the contribution for $(x^s y^m)$ term in $\Psi^g_{s,m}$ is given uniquely by the fixed $s' = s'_f$ and fixed m_f over all possible m''s from the previous state,

$$\langle \psi^{g-1}_{s'_f,1}, \psi^{g-1}_{s'_f,2}, ... \psi^{g-1}_{s'_f,m'} \rangle \cdot \langle \begin{array}{c} M[1,m_f] \\ M[2,m_f] \\ \vdots \\ M[m',m_f] \end{array} \rangle = \sum_{m'=1}^{\max m'} \psi_{s'_f,m'} M[m',m_f] \tag{4.19}$$

This is the algorithm used to compute the PGF's recursively for each generation. In this case, where we are dealing with a temporal network that switches networks at each generation, at each step we re-compute $G_{g-1}(1+(xy-1)T)$ using the new excess degree distribution from the new snapshot. Then we re-compute M according to the process outlined above, and from there can solve for Ψ^g_{sm} using Ψ^{g-1}_{sm} and the new M defined by the new network snapshot. The idea here is that the m' infectious nodes exist, but switch who their contacts are at the emergence of the next generation, causing the next m infections by using the links in the new network to infect the next generation.

4.4 Results on Erdős-Rényi Networks

We defined the following parameters for the disease model in Tab. 4.1 . Transmissibility T is equal to $\beta/(\beta+\gamma)$. Approximation length is the length of the vectors used to hold the generating functions and derived matrices, which represent infinite-length distributions but since the $P(k \to \infty) \to 0$, an approximation of 800 yields relatively good approximations of the resulting distributions. High transmissibility was used to consider recovery time much slower than the rate of contagion, and to match the values used in Chapter 3.

Infection rate β	Recovery rate γ	Transmissibility T	Approximation length
0.05	0.001	0.98	800

Table 4.1 : Model parameters for temporal PGF application

We set the temporal network to have five snapshots. For simplicity in this proof of concept, we say that each snapshot will be the substrate for one generation of infection. That is, $d_S = 1, \forall S$. Then we derived $\langle q \rangle$ for each snapshot using the set of durations d_S, t_S, the start times for each snapshot, and β. After deriving each $\langle q \rangle$, a random Erdős-Rényi network was generated using Python's NetworkX package and the five networks are summarized in Tab. 4.2 . Since the networks are all ER, they have low clustering and low degree assortativity, because nodes are connected to one another at random. While the ER networks are not very realistic of real populations or contact patterns, we use them for this simple example.

We compare the resulting theoretical distributions of generations of infection to simulated distributions of generational infection as in Chapter 3. In each simulation, at each snapshot time t_S from $S = [0, 1, 2, 3, 4]$, the network used in the simulation

Snapshot number	N	$\langle q \rangle$	Assortativity	Clustering	t_S	d_S
0	5000	3.99	-0.023789	0.000631	0	1
1	5000	1.32	0.031216	0.000000	5	1
2	5000	4.03	-0.022286	0.000579	20	1
3	5000	1.35	-0.011554	0.000000	25	1
4	5000	2.01	0.016646	0.000120	40	1

Table 4.2 : Temporal network snapshot metadata

switches to the next snapshot. At each change, the status of any currently infected or recovered node is preserved, but their contacts change according to the adjacency list of the new snapshot network. The results comparing the theoretical and simulated distributions are compared in Fig. 4.1. A depiction of the time series for the spreading process is shown in Fig. 4.2. The standard deviation of the distribution of infected nodes for each continuous time point (approximated into 1,000 time points within $t = 0$ to $t = 50$) are shown as shaded regions around the mean. Also shown are the time boundaries of each temporal network time boundary t_S, and with them the red vertical lines show the average time that the corresponding generation from $g = 1$ to 5 emerged. It can be seen that generations 1 through 3 emerge very close to the switching of the first 3 snapshots, however, generation 4 and 5 emerge much earlier than expected. This may be because the estimation of $E[t_g]$ may be more complex than anticipated and warrants further work.

4.5 Challenges and Future Work

A number of challenges arose while attempting to apply the generating function formalism to a temporal network with time boundaries defined in continuous time. First, if the disease spreading rate parameters make it so the generations emerge over

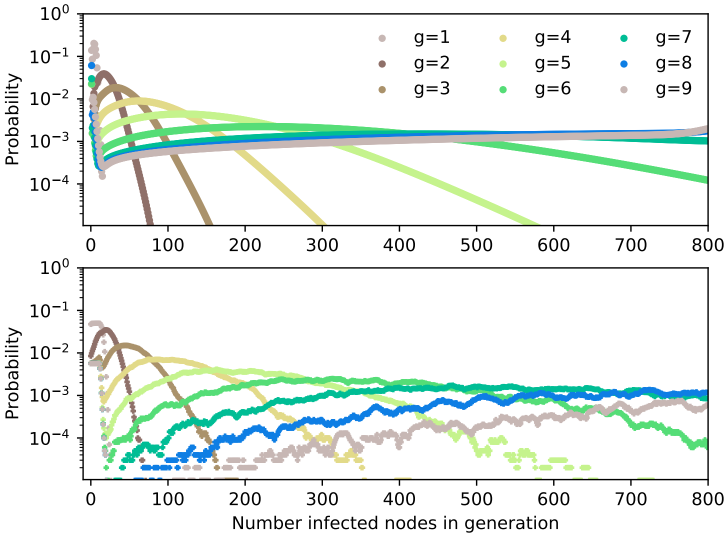

Figure 4.1: Temporal PGF validation against simulations: Top panel shows the generational probability distributions, bottom panel shows accompanying simulations. Networks had 5000 nodes and 10,000 simulations were performed. See Chapter 3 for simulation details.

multiple snapshots, then without further aggregating snapshots together there is no way to account for both (or several) snapshots in the computation of the generation size distribution, which takes in only one static degree distribution. Therefore, β needs to be high enough or the other parameters modified enough that generations are either one-to-one with snapshots at minimum, or there are multiple generations per snapshot.

The main difficulty with the approach overall is that we had to reverse engineer an example in which discrete generation values corresponded with the snapshot times so that exactly one generation was expected to emerge per snapshot, by designing specific networks with $\langle q \rangle$ that were derived from the desired durations. With empirical temporal networks, solving for the corresponding generation g_S that emerges when

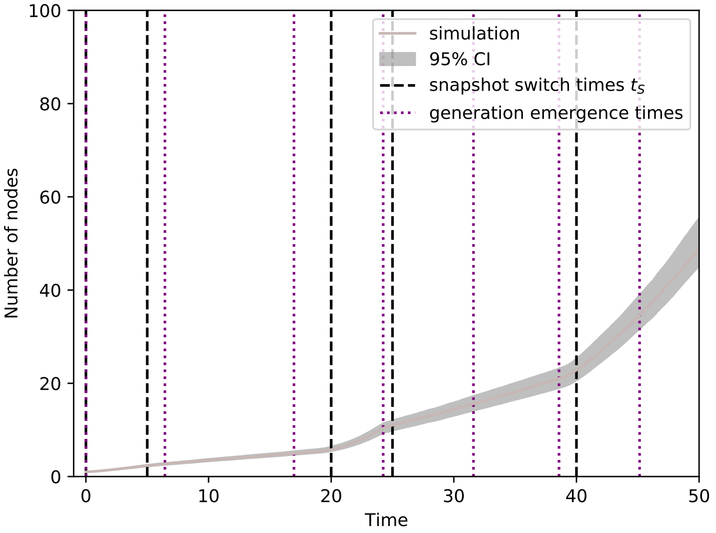

Figure 4.2: Average time-series of 5000 simulations and average generation emergence times. Green shading implies the simulated generation emerged after the prediction, yellow shading implies the simulated generation emerged before the prediction.

the network switches to a new snapshot likely will be a non-integer value. This presents an open question into how best to approximate which snapshots go with which generation, when generation values are integers by the nature of the framework itself.

The approach was also tested on other classes of networks with heterogeneous distributions, like a Barabási-Albert model [3] which is generated using a preferential attachment algorithm, which had similar results and also the same challenges of needing to reverse engineer an example such that generations emerged alongside the snapshot switch times.

4.6 Discussion

As discussed, heterogeneities exist on both the population structural level as well as the temporal level, with networks changing structures over time. Chapter 2 presented a way to handle temporal changes in the network, but still yielded a deterministic estimate for disease spread. Chapter 3 showed how it can be useful to use a stochastic framework in order to obtain probability distributions for the early spread of disease, instead of point estimates representing the expected state of the system over time. We used a generational framework to discretize the evolution of the disease and present distributions for the sizes of generations.

We have examined how to find distributions of generations of infection for a disease spreading process on a temporal network. This concept is extremely useful because it addresses uncertainty in the disease modeling process in two main ways: First, using temporal networks as the substrate allows for more realistic dynamic analysis as many real-world networks, particularly human and animal contact networks, display temporal dynamics.

Second, in studying the evolution of spread over a network, it can be helpful to know more than just the average trajectory of the number of infected individuals. The average is not always the best description of the system, as shown in Chapter 3, because there can be a wide array of possible states at each generation and this is helpful for epidemic preparedness and mitigation.

Third, it can be helpful to model population changes along with modeling the temporal variations in spread. Combining these two modeling approaches outlined in the prior chapters could lead to a powerful epidemic modeling tool that takes into account the temporal structure of both the underlying network and the disease propagation process, capturing uncertainty, probability, and precision of the network and disease dynamics.

Bibliography

[1] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):016128, July 2002. Publisher: American Physical Society.

[2] P.-A. Noël, B. Davoudi, R. C. Brunham, L. J. Dubé, and B. Pourbohloul. Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E*, 79:026101, 2009.

[3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002. Publisher: APS.

Chapter 5

Conclusion

The entirety of this body of work was conducted during the ongoing COVID-19 pandemic beginning in early 2020. While the nature of the models presented is primarily theoretical, the need for infectious disease models that go beyond temporal averaging and end-state summary statistics became clear as the pandemic unfolded. This thesis presented two such models that go beyond some of the assumptions made in traditional disease models, highlighting the importance of heterogeneous contact patterns both in terms of dynamic changes in contact patterns and structural differences.

All models of complex systems must make assumptions about some components of the underlying system, or else they risk having any practical usability and interpretability. That is why it is important to have a suite of disease models at hand, as there is no perfect model to answer every question. In this thesis, two models were introduced that break away from two common assumptions that are commonly used in idealized disease models. While removing these assumptions make for more complex models, they produce realistic results when applied to appropriate problems.

First, Chapter 2 presented a novel framework for analyzing temporal network data. This framework diverges from the common assumption of a static contact network underlying a disease model. Specifically, most established network-based disease models assume an annealed network with a fixed degree distribution, allowing for interactions to be considered between random individuals who's degrees are drawn from a fixed degree distribution, but specific node identities are arbitrary. In contrast, this model uses a quenched network, where the adjacency matrix of each network is fixed. Using a deterministic framework for a spreading process, the model determines the viability of compressing temporal network data into static, aggregate networks, to reduce the dimension of a data set while maintaining the temporal dynamics that are integral to the spreading process.

This work presented challenges that warrant further research. Namely, the compression algorithm developed and presented in Chapter 2 is useful, but does not have a clear way to determine when to stop compressing data. In that sense, an optimal stopping point has not been determined.

In Chapter 3, we discuss results based on a relatively recent stochastic disease model, which uses probability generating functions to derive probability distributions for the sizes of epidemic generations as the disease progresses through the network. This model relieves the assumption that the typical, or average, trajectory of a disease is a good descriptor of the disease dynamics. Sometimes, as we showed, there is actually a very broad distribution of possible trajectories, which is crucial knowledge for preparing for the initial stages of disease spread.

One of the primary challenges that stands out in the probability generating function model is that it is difficult to develop a rigorous link between continuous time evolution of disease spread under a stochastic framework. In Chapter 3, it was shown that roughly the expected time of emergence for an epidemic generation has a relationship to continuous time intervals defined by properties of the network and the disease dynamics, but the relationship is not well defined. Chapter 4 provides a proof of concept for how temporal networks may be incorporated into the stochastic framework by deriving the relationship between epidemic generations and continuous time, but this attempt involved challenges such as the inevitable difficulty of dealing with discrete-time generations rooted in branching process theory and continuous time dynamics. There is much room for further investigation for how the parameters of a network and disease parameters relate to the continuous time evolution of a stochastic process.

Overall, this thesis introduced novel ways of looking at disease spread on networks with dynamic and structural heterogeneity. The temporal network compression framework can be used to further the development of data compression efforts that respect the network data itself by focusing on how heterogeneous contact patterns affect an overlying process. The generating function model contributes to the field of stochastic network models, focusing on the variability in the early temporal evolution of disease spread instead of focusing on properties of the theoretical final state. Both models offer tools to deal with modeling emerging disease spread when large amounts of data are available but long-term trajectory is unknown, a situation that is all too familiar now and necessitates the development of reliable modeling frameworks.

Bibliography

[1] Abbey, H. (1952). An examination of the reed-frost theory of epidemics. *Hum. Biol.*, 24(3):201–233.

[2] Akian, M., Ganassali, L., Gaubert, S., and Massoulié, L. (2020). Probabilistic and mean-field model of COVID-19 epidemics with user mobility and contact tracing. *arXiv:2009.05304*.

[3] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47. Publisher: APS.

[4] Albert, R., Jeong, H., and Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature*, 401(6749):130–131. ISBN: 1476-4687 Publisher: Nature Publishing Group.

[5] Aleta, A., Ferraz de Arruda, G., and Moreno, Y. (2020). Data-driven contact structures: from homogeneous mixing to multilayer networks. *PLoS Comp. Bio.*, 16(7):e1008035. ISBN: 1553-734X Publisher: Public Library of Science San Francisco, CA USA.

[6] Allard, A., Althouse, B. M., Scarpino, S. V., and Hébert-Dufresne, L. (2017). Asymmetric percolation drives a double transition in sexual contact networks. *Proc. Natl. Acad. Sci. U.S.A.*, 114(34):8969–8973.

[7] Allard, A., Hébert-Dufresne, L., Noël, P.-A., Marceau, V., and Dubé, L. J. (2012). Bond percolation on a class of correlated and clustered random graphs. *J. Phys. A Math. Theor.*, 45(40):405005.

[8] Allard, A., Hébert-Dufresne, L., Young, J.-G., and Dubé, L. J. (2015). General and exact approach to percolation on random graphs. *Phys. Rev. E*, 92(6):062807–062807.

[9] Allard, R. (1998). Use of time-series analysis in infectious disease surveillance. *Bull. World Health Organ.*, 76:327–333.

[10] Allen, A. (2021). andrea-allen/epintervene: Initial Release v1.0.0.

[11] Allen, L. (2015). *Stochastic Population and Epidemic Models.* Springer.

[12] Allen, L. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infect. Dis. Model.*, 2(2):128–142.

[13] Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, 8(6):450–461. ISBN: 1471-0064 Publisher: Nature Publishing Group.

[14] Althaus, C. L. (2015). Ebola superspreading. *Lancet Infect. Dis.*, 15(5):507–508.

[15] Althouse, B. M., Wenger, E. A., Miller, J. C., Scarpino, S. V., Allard, A., Hébert-Dufresne, L., and Hu, H. (2020). Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLoS Bio.*, 18(11):e3000897.

[16] Amaral, L. A. N., Scala, A., Barthelemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proc. Natl. Acad. Sci. U.S.A.*, 97(21):11149–11152. ISBN: 0027-8424 Publisher: National Acad Sciences.

[17] Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxf. Univ. Press, Great Clarendon Street, Oxford OX2 6DP.

[18] Andersson, H. and Britton, T. (2012). *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media.

[19] Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*. Springer-Verlag Berlin Heidelberg, New York.

[20] Bailey, N. T. (1975). *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

[21] Bajardi, P., Barrat, A., Natale, F., Savini, L., and Colizza, V. (2011). Dynamical patterns of cattle trade movements. *PloS one*, 6(5):e19869. ISBN: 1932-6203 Publisher: Public Library of Science San Francisco, USA.

[22] Bakhta, A., Boiveau, T., Maday, Y., and Mula, O. (2021). Epidemiological forecasting with model reduction of compartmental models. application to the covid-19 pandemic. *Biology*, 10(1):22.

[23] Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Ann. Appl. Probab.*, 7(1):46–89.

[24] Ball, F. and P., D. (1995). Strong approximations for epidemic models. *Stoch. Process. Their Appl.*, 55(1):1–21.

[25] Bansal, S., Read, J., Pourbohloul, B., and Meyers, L. A. (2010). The dynamic nature of contact networks in infectious disease epidemiology. *J. Biol. Dyn.*, 4(5):478–489. ISBN: 1751-3758 Publisher: Taylor & Francis.

[26] Barabási, A.-L., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Phys. A: Stat. Mech. Appl.*, 272(1-2):173–187. ISBN: 0378-4371 Publisher: Elsevier.

[27] Bauer, P., Engblom, S., and Widgren, S. (2016). Fast event-based epidemiological simulations on national scales. *Inter, J. High Perform. Comput. Appl.*, 30(4):438–453.

[28] Becker, N. G. (1974). On parametric estimation for mortal branching processes. *Biometrika*, 61:393–399.

[29] Becker, N. G. (1977). Estimation for discrete time branching processes with applications to epidemics. *Biometrics*, 33:515–522.

[30] Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B., and Sledge, D. (2020). The challenges of modeling and forecasting the spread of COVID-19. *Proc. Natl. Acad. Sci.*, 117(29):16732–16738.

[31] Boccaletti, S., Hwang, D.-U., Chavez, M., Amann, A., Kurths, J., and Pecora, L. M. (2006). Synchronization in dynamical networks: Evolution along commuta-

tive graphs. *Phys. Rev. E*, 74(1):016102.

[32] Bolker, B. and Grenfell, B. (2015). Space, persistence and dynamics of measles epidemics. *Philos. Trans. R. Soc. B*, 348:309–320.

[33] Bollobás, B. (1998). Random graphs. In *Modern graph theory*, pages 215–252. Springer.

[34] Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier.

[35] Braha, D. and Bar-Yam, Y. (2009). Time-dependent complex networks: Dynamic centrality, dynamic motifs, and cycles of social interactions. In *Adaptive Networks*, pages 39–50. Springer.

[36] Burgio, G., Steinegger, B., and Arenas, A. (2022). Homophily impacts the success of vaccine roll-outs. arXiv:2112.08240.

[37] Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2000). Network Robustness and Fragility: Percolation on Random Graphs. *Phys. Rev. Lett.*, 85(25):5468–5471. Publisher: American Physical Society.

[38] Camacho, A., Kucharski, A., Aki-Sawyerr, Y., White, M. A., Flasche, S., Baguelin, M., Pollington, T., Carney, J. R., Glover, R., Smout, E., Tiffany, A., Edmunds, J. W., and Funk, S. (2015). Temporal change in Ebola transmission in Sierra Leone and implications for control requirement: a real-time modelling study. *PLoS Currents*, 7:1–18.

[39] Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.-F., and Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PloS one*, 5(7):e11596. ISBN: 1932-6203 Publisher: Public Library of Science San Francisco, USA.

[40] Centers for Disease Control and Prevention. (2020). First Travel-related Case of 2019 Novel Coronavirus Detected in United States.

[41] Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., and Faloutsos, C. (2008). Epidemic Thresholds in Real Networks. *ACM Trans. Inf. Syst. Secur.*, 10(4). Place: New York, NY, USA Publisher: Association for Computing Machinery.

[42] Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. (2000). Resilience of the Internet to Random Breakdowns. *Phys. Rev. Lett.*, 85(21):4626–4628. Publisher: American Physical Society.

[43] Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. (2001). Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86:3682–3685.

[44] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (2021).

[45] Cox, D. R. (1958). Renewal Theory. Methuen & Co., London, 1967. *Science Paperback Edition.*

[46] Cuevas, E. (2020). An agent-based model to evaluate the COVID-19 transmission risks in facilities. *Comput. biol. med.*, 121:103827.

[47] Diekmann, O., Heesterbeek, H., and Britton, T. (2013). *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princet. Univ. Press, 41 Williams St, Princeton, New Jersey 08540.

[48] Diekmann, O. and Heesterbeek, J. A. P. (2000). *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons.

[49] Diekmann, O., Heesterbeek, J. A. P., and Metz, J. A. (1995). The legacy of Kermack and McKendrick. *Publications of the Newton Institute*, 5:95–115. Publisher: Cambridge University Press.

[50] Dietz, K. and Heesterbeek, J. A. P. (2002). Daniel Bernoulli's epidemiological model revisited. *Math Biosci*, 180(1-2):1–21. ISBN: 0025-5564 Publisher: Elsevier.

[51] Dong, E., Du, H., and L., G. (2020). An interactive web-based dashboard to track covid-19 in real time. *Lancet Infect. Dis.*, 20(5):533–534.

[52] Du, Z., Xu, X., Wu, Y., Wang, L., Cowling, B. J., and Meyers, L. A. (2020). Serial interval of COVID-19 among publicly reported confirmed cases. *Emerg Infect Dis.*, 26(6):1341.

[53] Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

[54] Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184. ISBN: 1476-4687 Publisher: Nature Publishing Group.

[55] Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140.

[56] Finkenstädt, B. F. and Grenfell, B. (2000). Time series modelling of childhood diseases: a dynamical systems approach. *Appl. Stat.*, 49:187–205.

[57] Fosdick, B. K., Larremore, D. B., Nishimura, J., and Ugander, J. (2018). Configuring random graph models with fixed degree sequences. *SIAM review*, 60(2):315–355.

[58] Gharakhanlou, N. M. and Hooshangi, N. (2020). Spatio-temporal simulation of the novel coronavirus (COVID-19) outbreak using the agent-based modeling approach (case study: Urmia, Iran). *Inform. Med. Unlocked*, 20:100403.

[59] Ghasemian, A., Zhang, P., Clauset, A., Moore, C., and Peel, L. (2016). Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks. *Phys. Rev. X*, 6(3):031005. Publisher: American Physical Society.

[60] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361. ISBN: 0022-3654 Publisher: ACS Publications.

[61] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12):7821–7826. ISBN: 0027-

8424 Publisher: National Acad Sciences.

[62] Grassberger, P. (1983). On the critical behavior of the general epidemic process and dynamical percolation. *Math Biosci*, 63(2):157–172.

[63] Gray, D., Greenhalgh, L., Hu, L., Mao, X., and Pan, J. (2011). A stochastic differential equation sis epidemic model. *SIAM J. Appl. Math.*, 71(3):876–902.

[64] Génois, M., Vestergaard, C. L., Cattuto, C., and Barrat, A. (2015). Compensating for population sampling in simulations of epidemic spread on temporal contact networks. *Nature Communications*, 6(1):8860.

[65] H., I. (1990). Threshold and stability results for an age-structured epidemic model. *J. Math. Biol.*, 28:411–434.

[66] Hébert-Dufresne, L. and Allard, A. (2019). Smeared phase transitions in percolation on real complex networks. *Phys. Rev. Res.*, 1:013009.

[67] Hébert-Dufresne, L., Althouse, B. M., Scarpino, S. V., and Allard, A. (2020). Beyond R0: heterogeneity in secondary infections and probabilistic epidemic forecasting. *J. R. Soc. Interface*, 17(172):20200393.

[68] Heesterbeek, J. A. P. (2002). A brief history of R0 and a recipe for its calculation. *Acta Biotheor*, 50(3):189–204. Place: Netherlands.

[69] Henkel, M., Hinrichsen, H., Lübeck, S., and Pleimling, M. (2008). *Non-equilibrium phase transitions*, volume 1. Springer.

[70] Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, 42(4):599–653. ISBN: 0036-1445 Publisher: SIAM.

[71] Hinch, R., Probert, W. J., Nurtay, A., Kendall, M., Wymatt, C., Hall, M., Lythgoe, K., Cruz, A. B., Zhao, L., Stewart, A., Ferretti, L., Montero, D., Warren, J., Mather, N., Abueg, M., Wu, N., Finkelstein, A., Bonsall, D., Abeler-Dörner, L., and Fraser, C. (2020). OpenABM-Covid19-an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *medRxiv*, pages 1–23.

[72] Hiraoka, T., Rizi, A. K., Kivelä, M., and Saramäki, J. (2022). Herd immunity and epidemic size in networks with vaccination homophily. arXiv:2112.07538.

[73] Hoertel, N., Blachier, M., Blanco, C., Olfson, M., Massetti, M., Limosin, F., and Leleu, H. (2020). Facing the COVID-19 epidemic in NYC: a stochastic agent-based model of various intervention strategies. *MedRxiv*, pages 1–34.

[74] Holme, P. (2003). Network dynamics of ongoing social relationships. *EPL (Europhysics Letters)*, 64(3):427. ISBN: 0295-5075 Publisher: IOP Publishing.

[75] Holme, P. (2013). Epidemiologically Optimal Static Networks from Temporal Network Data. *PLoS Comp. Bio.*, 9(7).

[76] Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125.

[77] Hu, W., Tong, S., Mengersen, K., and Oldenburg, B. (2006). Rainfall, mosquito density and the transmission of Ross River virus: A time-series forecasting model.

Ecol. Model., 196:505–514.

[78] Huang, W., Cooke, K. L., and Castillo-Chavez, C. (1992). Stability and bifurcation for a multi-group model for the dynamics of HIV/AIDS transmission. *SIAM J. Appl. Math*, 52:835–854.

[79] Hébert-Dufresne, L., Allard, A., Young, J.-G., and Dubé, L. J. (2013). Percolation on random networks with arbitrary k-core structure. *Phys. Rev. E*, 88(6):062820–062820.

[80] Jo, H.-H., Karsai, M., Kertész, J., and Kaski, K. (2012). Circadian pattern and burstiness in human communication activity. *New J Phys*, 14(1):013055.

[81] Karrer, B. and Newman, M. E. J. (2010). Random graphs containing arbitrary distributions of subgraphs. *Phys. Rev. E*, 82(6 Pt 2):066118–066118.

[82] Keeling, M. J. and Grenfell, B. T. (2000). Individual-based Perspectives on R0. *J. Theor. Biol.*, 203(1):51–61.

[83] Keeling, M. J. and Rohani, P. (2007). *Modeling Infectious Diseases in Humans and Animals*. Princet. Univ. Press, 41 Williams St, Princeton, New Jersey 08540.

[84] Kenah, E. and Miller, J. C. (2011). Epidemic Percolation Networks, Epidemic Outcomes, and Interventions. *Interdisciplinary Perspectives on Infectious Diseases*, 2011:543520. Publisher: Hindawi Publishing Corporation.

[85] Kenah, E. and Robins, J. M. (2007). Second look at the spread of epidemics on networks. *Phys. Rev. E*, 76(3):036113. Publisher: APS.

[86] Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics-I. *Proc. R. Soc. Lond., A*, 115(772):700–721. Publisher: The Royal Society London.

[87] Kermack, W. O. and McKendrick, A. G. (1932). A contribution to the mathematical theory of epidemics—II. The problem of endemicity. *Proc. R. Soc. Lond., A*, 138(772):55–83. Publisher: The Royal Society London.

[88] Kermack, W. O. and McKendrick, A. G. (1933). A contribution to the mathematical theory of epidemics—III. Further studies of the problem of endemicity. *Proc. R. Soc. Lond., A*, 141(772):94–122. Publisher: The Royal Society London.

[89] Kiss, I. Z., Miller, J. C., and Simon, P. L. (2019). *Mathematics of Epidemics on Networks: from exact to approximate models*. Springer.

[90] Kojaku, S., Hébert-Dufresne, L., Mones, E., Lehmann, S., and Ahn, Y.-Y. (2021). The effectiveness of backward contact tracing in networks. *Nature Physics*, 17(5):652–658.

[91] Kucharski, A. J. and Althaus, C. L. (2015). The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. *Eurosurveillance*, 20(25).

[92] Levesque, J., Maybury, D. W., and Shaw, R. D. (2021). A model of COVID-19 propagation based on a gamma subordinated negative binomial branching process. *J. Theor. Biol.*, 512:110536.

[93] Li, A., Cornelius, S. P., Liu, Y.-Y., Wang, L., and Barabási, A.-L. (2017). The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046.

[94] Lloyd, A. and Jansen, V. (2004). Spatiotemporal dynamics of epidemics: synchrony in metapopulation models. *Math. Biosci.*, 188:1–16.

[95] Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., and Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359.

[96] Lopman, B., Armstrong, B., Atchinson, C., and Gray, J. J. (2009). Host, weather and virological factors drive norovirus epidemiology: Time-series analysis of laboratory surveillance data in England and Wales. *PLoS One*, 4:e6671.

[97] Ludwig, D. (1975). Final size distribution for epidemics. *Math Biosci*, 23(1):33–46.

[98] Malmgren, R. D., Stouffer, D. B., Campanharo, A. S., and Amaral, L. A. N. (2009). On universality in human correspondence activity. *science*, 325(5948):1696–1700. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

[99] Marro, J. and Dickman, R. (1999). Nonequilibrium Phase Transitions in Lattice Models. *Nonequilibrium Phase Transitions in Lattice Models*, page 344. ISBN: 0521480620.

[100] Masuda, N. and Holme, P. (2019). Detecting sequences of system states in temporal networks. *Sci. Rep.*, 9(1):795.

[101] Melnik, S., Hackett, A., Porter, M. A., Mucha, P. J., and Gleeson, J. P. (2011). The unreasonable effectiveness of tree-based theory for networks with clustering. *Phys. Rev. E*, 83(3 Pt 2):036112–036112.

[102] Meyers, L. (2007). Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bull New Ser. Am Math Soc.*, 44(1):63–86.

[103] Miller, J. C. (2007). Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. *Phys. Rev. E*, 76(1):010101.

[104] Miller, J. C. (2009). Percolation and epidemics in random clustered networks. *Phys. Rev. E*, 80(2):020901.

[105] Miller, J. C. (2018). A primer on the use of probability generating functions in infectious disease modeling. *Infect. Dis. Model.*, 3:192–248.

[106] Miller, J. C. and Ting, T. (2019). EoN (Epidemics on Networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. *J. Open Source Softw.*, 4(44):1731.

[107] Mitrofani, I. A. and Koutras, V. P. (2021). A branching process model for the novel coronavirus (Covid-19) spread in Greece. *Int. J. Model. Optim.*, 11(3).

[108] Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.3240060204.

[109] Moore, C. and Newman, M. E. J. (2000). Epidemics and percolation in small-world networks. *Phys. Rev. E*, 61(5):5678. Publisher: APS.

[110] Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

[111] Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):016128. Publisher: American Physical Society.

[112] Newman, M. E. J. (2009). Random graphs with clustering. *Phys. Rev. Lett.*, 103(5):058701–058701.

[113] Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118.

[114] Noël, P.-A., Davoudi, B., Brunham, R. C., Dubé, L. J., and Pourbohloul, B. (2009). Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E*, 79:026101.

[115] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.*, 104(18):7332–7336. ISBN: 0027-8424 Publisher: National Acad Sciences.

[116] Oxley, D. and Ryan, J. (2020, March 7). “Volatile and unpredictable”: Life Care Center speaks publicly for the first time since COVID-19 outbreak. *KUOW News and Inf.*

[117] P. Peixoto, T. and Gauvin, L. (2018). Change points, memory and epidemic spreading in temporal networks. *Sci. Rep.*, 8(1):15511.

[118] Parshani, R., Carmi, S., and Havlin, S. (2010). Epidemic threshold for the susceptible-infectious-susceptible model on random networks. *Phys. Rev. Lett.*, 104(25):258701. Publisher: APS.

[119] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87(3):925. Publisher: APS.

[120] Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200. Publisher: APS.

[121] Peixoto, T. P. and Rosvall, M. (2017). Modelling sequences and temporal networks with dynamic community structures. *Nature Communications*, 8(1):582–582. Place: England Publisher: Nature Publishing Group UK.

[122] Perra, N., Gonçalves, B., Pastor-Satorras, R., and Vespignani, A. (2012). Activity driven modeling of time varying networks. *Sci. Rep.*, 2(1):469.

[123] Ren, G. and Wang, X. (2014). Epidemic spreading in time-varying community networks. *Chaos*, 24(2):023116. Publisher: American Institute of Physics.

[124] Riou, J. and Althaus, C. L. (2020). Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance*, 25(4):2000058.

[125] Rizi, A. K., Faqeeh, A., Badie-Modiri, A., and Kivelä, M. (2022). Epidemic

spreading and digital contact tracing: Effects of heterogeneous mixing and quarantine failures. arXiv:2103.12634.

[126] Rocha, L. E. C., Liljeros, F., and Holme, P. (2011). Simulated Epidemics in an Empirical Spatiotemporal Network of 50,185 Sexual Contacts. *PLoS Comp. Bio.*, 7(3).

[127] Ross, S. M. (1996). Stochastic Processes. John Wiley & Sons. *New York*.

[128] Sapiezynski, P., Stopczynski, A., Lassen, D. D., and Lehmann, S. (2019). Interaction data from the copenhagen networks study. *Scientific Data*, 6(1):1–10.

[129] Scholtes, I., Wider, N., and Garas, A. (2016). Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. *Eur Phys J B*, 89(3):61.

[130] Silva, P. C. L., Batista, P. V. C., Lima, H. S., Alves, M. A., Guimarães, F. G., and Silva, R. C. P. (2020). Covid-abs: An agent-based model of covid-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos Solitons Fractals*, 139:110088.

[131] Srikrishnan, V. and Keller, K. (2021). Small increases in agent-based model complexity can result in large increases in required calibration data. *Environ. Model. Softw.*, 138:104978.

[132] St-Onge, G., Young, J.-G., Laurence, E., Murphy, C., and Dubé, L. J. (2018). Phase transition of the susceptible-infected-susceptible dynamics on time-varying configuration model networks. *Phys. Rev. E*, 97(2):022305. Publisher: American Physical Society.

[133] Staffini, A., Svensson, A. K., Chung, U.-I., and Svensson, T. (2021). An agent-based model of the local spread of SARS-CoV-2: Modeling study. *JMIR med. inform.*, 9(4):e24192.

[134] Stilwell, D. J., Bollt, E. M., and Roberson, D. G. (2006). Sufficient conditions for fast switching synchronization in time-varying network topologies. *SIAM Journal on Applied Dynamical Systems*, 5(1):140–156.

[135] Sullivan, O. (2020, March 3). “Coronavirus death toll rises to nine in Washington". *Kirkland Reporter*.

[136] Sundell, A. (2020, February). New coronavirus cases confirmed in Snohomish, King counties. *KING-TV*.

[137] Valdano, E., Ferreri, L., Poletto, C., and Colizza, V. (2015). Analytical Computation of the Epidemic Threshold on Temporal Networks. *Phys. Rev. X*, 5(2):021005. Publisher: American Physical Society.

[138] Van Kampen, N. G. (1981). *Stochastic processes in chemistry and physics*. North Holland, Amsterdam.

[139] Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., Kim, B.-a., Comte, B., and Voirin, N. (2013). Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. *PLoS ONE*,

8(9):e73970. Publisher: Public Library of Science.

[140] Vazquez, A., Racz, B., Lukacs, A., and Barabasi, A.-L. (2007). Impact of non-Poissonian activity patterns on spreading processes. *Phys. Rev. Lett.*, 98(15):158702. Publisher: APS.

[141] Vázquez, A. and Moreno, Y. (2003). Resilience to damage of graphs with degree correlations. *Phys. Rev. E*, 67(1 Pt 2):015101–015101.

[142] Wang, W., Cai, Y., Ding, Z., and Gui, Z. (2018). A stochastic differential equation sis epidemic model incorporating ornstein–uhlenbeck process. *Phys. A: Stat. Mech. Appl.*, 509(4).

[143] Wang, Y., Chakrabarti, D., Wang, C., and Faloutsos, C. (2003). Epidemic spreading in real networks: an eigenvalue viewpoint. In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pages 25–34.

[144] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442. ISBN: 1476-4687 Publisher: Nature Publishing Group.

[145] Wilf, H. S. (2005). *generatingfunctionology*. CRC press.

[146] Zhan, Z., Dong, W., Lu, Y., Yang, P., Wang, Q., and Jia, P. (2018). Real-time forecasting of hand-foot-and-mouth disease outbreaks using the integrating compartment model and assimilation filtering. *Sci. Rep.*, 9:1–9.

[147] Zhang, L., Wang, H., Liu, Z., Liu, X. F., Feng, X., and Wu, Y. (2021). A heterogeneous branching process with immigration modeling for COVID-19 spreading in local communities in China. *Complexity*, 2021:1–11.

[148] Zhang, Y. and Strogatz, S. H. (2021). Designing temporal networks that synchronize under resource constraints. *Nature communications*, 12(1):1–8.

[149] Zhao, Q., Tian, Y., He, Q., Oliver, N., Jin, R., and Lee, W.-C. (2010). Communication motifs: a tool to characterize social communications. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1645–1648.