Statistical Network Analysis: A Review with Applications to the COVID-19 Pandemic

Joshua Daniel Loyal[1] and Yuguo Chen[1]

[1] *Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA*
E-mail: jloyal2@illinois.edu; yuguo@illinois.edu

Summary

As the COVID-19 outbreak evolves, statistical network analysis is playing an essential role in informing policy decisions. Therefore, researchers who are new to such studies need to understand the techniques available to them. As a field, statistical network analysis aims to develop methods that account for the complex dependencies found in network data. Over the last few decades, the area has rapidly accumulated methods, including techniques for network modeling and simulating the spread of infectious disease. This article reviews these network modeling techniques and their applications to the COVID-19 pandemic.

Key words: COVID-19; network models; network SEIR model; social networks.

1 Introduction

Coronavirus disease 2019 (COVID-19) is altering global society, policy, and economic activity in unprecedented ways. Policymakers are continuously modifying their actions based on statistical analyses of the novel coronavirus. For example, forecasts of short-term mortality already inform social distancing guidelines. Statistical analysis can also guide policy questions such as how to design sufficient contact tracing programs and implement effective vaccination strategies. Answering these difficult questions requires researchers to apply the appropriate statistical analysis to the data. In particular, the global scale of this pandemic means that any investigation must account for variations in disease spread among different communities.

Recent advances in network modeling make it possible to incorporate heterogeneity assumptions into statistical models of epidemic spread. This review draws on network science, statistical modeling, and mathematical epidemiology to demonstrate how to incorporate network effects into models of COVID-19. In Section 2, we introduce the basic

concepts of network data. Also, we summarize the characteristics of real-world networks that are essential to understanding the COVID-19 pandemic. Section 3 reviews various mathematical and statistical network models. In Section 4, we outline how to modify classical compartmental models from epidemiology to account for network structure. We also use the ideas developed in the previous sections to forecast the spread of a novel infectious disease. In Section 5, we provide a list of open-source software for statistical network analysis. Section 6 summarizes our conclusions and lists applicable methods not covered in this review.

2 An Introduction to Network Data

Network science is the study of phenomena governing relational data or networks. A network is composed of a set of actors (or nodes) and dyadic relationships (or ties) between actors. The COVID-19 pandemic highlights the necessity to understand network data and to cultivate a body of knowledge transferable to similar scenarios. Such an understanding is useful because network properties like an individual's "degree" and a network's "connectedness" dramatically affect the spread of an infectious disease. In this section, we summarize these fundamental network characteristics.

2.1 Network Representations

Real-world networks are often represented by graphs. Depending on the network, the graph may be undirected, directed, weighted, time-dependent, or various combinations of these. A graph G consists of a pair (V, E), where V is a set of vertices (or actors in the network context), and E is a set of dyads, $E \subseteq V \times V$, known as edges (or relations in the network context). A directed graph treats each edge pair (i, j) and (j, i) as distinct, while undirected graphs do not. Often we focus on parts of the full graph G called subgraphs. A graph $H = (V_H, E_H)$ is a subgraph of $G = (V, E)$ if $V_H \subseteq V$ and $E_H \subseteq E$. In network science, a graph is not synonymous with a network, but rather one possible mathematical representation.

An adjacency matrix is another popular representation for finite networks with n actors. In this case, the vertex set is taken as $V = \{1, \ldots, n\}$. The adjacency matrix $\mathbf{Y}$

corresponding to the graph $G = (V, E)$ is the $n \times n$ binary matrix with elements

$$Y_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E, \\ 0, & \text{otherwise.} \end{cases}$$

The adjacency matrix uniquely determines a finite graph. The matrix $\mathbf{Y}$ is symmetric for undirected graphs. Note that real-valued edge weights replace the binary entries of the adjacency matrix for relations represented by arbitrary real numbers, e.g., the amount of trade in dollars between nation i and nation j. We refer to such graphs as weighted graphs. Unless otherwise stated, we assume all graphs are undirected and unweighted in this work. As we show in the following sections, many network characteristics are expressible as functions of the adjacency matrix.

Although this work focuses on statistic networks, many real-world networks are dynamic (time-dependent) in nature. A dynamic network is a collection of time-indexed graphs $G_t = (V_t, E_t)$, where time t takes values in a continuous or discrete time interval $\mathcal{T}$. Here, V_t and E_t indicate the set of vertices and edges present in the network at time t, respectively. Many analyses transform a dynamic network into a static network by combining the vertex and edge sets over all time points, i.e., $G = (\bigcup_{t\in\mathcal{T}} E_t, \bigcup_{t\in\mathcal{T}} V_t)$. However, this conversion potentially removes temporal characteristics that affect disease spread. For this reason, we highlight a few results specific to dynamic networks throughout this work.

2.2 Characterizing Networks

When studying real-world networks, such as contact networks in a city, we are unlikely to measure the exact adjacency matrix. Furthermore, any two cities have different adjacency matrices. This heterogeneity might suggest that any attempt to transfer knowledge from one network to another is bound to fail. However, we expect networks generated by a common mechanism to share some characteristics. For example, Holland and Leinhardt (1976) observed a nontrivial prevalence of "triangles" in friendship networks. By identifying the essential structures in a specific type of network, we can better understand the processes behind how these networks formed. The development of such a taxonomy of networks has applications in modeling the spread of epidemics on networks since disease

spread is sensitive to network topology. In this section, we summarize characteristics commonly measured during network analyses.

2.2.1 Vertex Degree and Degree Distributions

Two of the most ubiquitous network characteristics are vertex degree and the degree distribution. The out-degree $d_{\text{out}}(i)$ of a vertex i is the number of edges starting from it, i.e., $\sum_{j=1}^{n} Y_{ij}$. Similarly, the in-degree $d_{\text{in}}(i)$ is the number of edges going into it, i.e., $\sum_{j=1}^{n} Y_{ji}$. For an undirected graph, the in-degree and out-degree are equal. In this case, this number is called the degree of the vertex $d(i)$. The degree sequence of an undirected graph is the collection of all vertex degrees $\mathbf{d} = (d(1), d(2), \ldots, d(n))$.

Given a graph G, we define f_d as the fraction of vertices with degree $d(i) = d$. The collection $\{f_d\}_{d \geq 0}$ is called the degree distribution of G. As we show in Section 4.2, the first two moments of the degree distribution inform whether an epidemic occurs on a network. We define the k-th degree moment of a graph as

$$\langle d^k \rangle = \frac{1}{n} \sum_{i=1}^{n} [d(i)]^k. \tag{2.1}$$

The average degree of a graph is $\langle d \rangle = 2(|E|/|V|)$. The degree variance, defined as $\langle d^2 \rangle - \langle d \rangle^2$, measures a network's degree heterogeneity. When every vertex has the same degree d, so zero degree variance, then we call the graph d-regular or homogeneous.

2.2.2 Assortative and Disassortative Mixing

Despite its popularity, the degree distribution does not capture all aspects of a network. To move beyond the degree distribution, we note that there are usually strong correlations between nodes of different degrees. In assortative networks, high degree nodes tend to connect with other high degree nodes. In disassortative networks, high degree nodes prefer to connect to low degree nodes. Determining whether a network is assortative or disassortative has major implications for the effectiveness of contact tracing strategies. For example, the removal of high degree nodes in disassortative networks massively limits the ability of a disease to spread (Kiss et al., 2008).

The degree correlation function is one way to assess whether a network is assortative or disassortative. Define $f_{d'|d}$ as the relative frequency with which edges from a vertex of

degree d connect to another vertex of degree d'. The degree correlation function is the mean of this conditional distribution:

$$\mathrm{Cor}(d) = \sum_{d'} d' f_{d'|d}. \tag{2.2}$$

As a function of d, the degree correlation function increases for assortative networks and decreases for disassortative networks. Other popular measures of degree correlation include visualization of the joint degree distribution and calculation of the degree correlation coefficient introduced in Newman (2002).

In general, assortative (disassortative) mixing refers to the propensity of nodes to connect with other nodes with similar (dissimilar) characteristics. Usually, aspects other than degree are the basis of these correlations. For example, students are often more likely to form friendships with other students in the same grade. Note that assortative mixing is known as homophily in the social network literature. Determining features that govern assortative or disassortative mixing in a statistically principled manner is an essential component of network modeling.

2.2.3 Reciprocity

Reciprocity is a pairwise feature specific to directed networks. Reciprocity measures how often ties are reciprocated in a directed network, i.e., the frequency that both $(i, j) \in E$ and $(j, i) \in E$. The ratio between the number of reciprocated edges over the total number of edges measures the extent of reciprocity in a network. This ratio is zero for networks containing no mutual ties, and it is one for networks where all edges are reciprocated.

The level of reciprocity in a directed network provides insight into the nature of the relation. A high degree of reciprocity indicates that the relationship is often mutual between actors. Such an observation has important modeling implications. For example, it is common to treat a directed relation as undirected due to technical constraints. One can justify this decision for highly reciprocal relations.

2.2.4 Transitivity and Higher-Order Structures

Pairwise interactions do not uniquely characterize networks. The concept of transitivity, also known as clustering, pertains to triplets of actors. Transitivity quantifies the fre-

quency with which connected triplets (i.e., a subgraph of three vertices connected by two edges) close to form triangles. In a social network context, transitivity quantifies the 'a friend of my friend is my friend' phenomenon. The clustering coefficient (Newman, 2010; Kiss et al., 2017) measures the amount of transitivity in a network:

$$\mathrm{cl}(G) = \frac{6\tau_{\triangle}(G)}{\tau_3(G)} = \frac{\mathrm{tr}(\mathbf{Y}^3)}{\|\mathbf{Y}^2\| - \mathrm{tr}(\mathbf{Y}^2)}, \tag{2.3}$$

where $\tau_{\triangle}(G)$ is the number of triangles in the graph G, $\tau_3(G)$ is the number of connected triplets, and $\|\mathbf{Y}\| = \sum_{i=1}^{n}\sum_{j=1}^{n} Y_{ij}$. A non-zero entry Y_{ij}^k of $\mathbf{Y}^k$ indicates that nodes i and j are connected by Y_{ij}^k paths of length k. This means $\mathrm{tr}(\mathbf{Y}^3)$ in the numerator measures the amount of closed loops of length three in a graph and the denominator measures all non-backtracking paths of length two.

Higher-order connected subgraphs containing more than three nodes may be relevant in real-world networks. Such repeating subgraphs are called motifs. Common motifs include k-stars, k-cycles, and k-cliques. Depending on the relation of interest, one focuses on motifs that have interpretable meanings, such as feedback loops in biological networks. Determining whether meaningful higher-order motifs occur more often than by chance is a crucial aspect of network analysis.

2.2.5 Shortest Paths, Connectedness, and Components

The previous sections characterized networks based on local vertex-level properties; however, networks also contain distinguishable global structures. As we show in Section 4.5, interventions that alter the global topology of a network can drastically reduce the rate of disease spread.

A network's connectedness is an important global characteristic. A directed path connecting vertices $u, v \in V$ is a sequence of vertices $v_0, \ldots, v_m \in V$ such that $(v_i, v_{i+1}) \in E$ for all i and $v_0 = u, v_m = v$. We call a graph strongly connected if there is a direct path from any vertex to any other vertex. In the case of undirected graphs, we say the graph is connected. Many graphs are not connected; however, they often have connected components. A connected component is a maximally connected subgraph, which means that the addition of any other vertex to the subgraph would result in an unconnected subgraph. The giant component is a graph's largest connected component. In network

analysis, it is common to focus on the giant component.

When modeling the spread of infectious disease, we are often interested in the distance a disease can travel on a graph. The shortest path, geodesic distance, or distance between two vertices is the number of edges in the shortest directed path connecting the two vertices. In addition, the longest geodesic distance between two nodes in a graph is called the diameter. The diameter of real-world social networks is often quite small, scaling logarithmically in the number of vertices (Watts and Strogatz, 1998).

3 Network Models

A network model is a collection of probability distributions over graphs

$$\{\mathbb{P}_{\boldsymbol{\theta}}(G),\ G \in \mathcal{G} :\ \boldsymbol{\theta} \in \Theta\}, \tag{3.1}$$

where $\mathcal{G}$ is a collection of possible graphs, $\mathbb{P}_{\boldsymbol{\theta}}$ is a probability distribution on $\mathcal{G}$, and $\boldsymbol{\theta}$ is a vector of parameters that takes values in Θ. Network models allow for the rigorous quantification of many questions in network science. For example, the specification of a network model is the starting point for formal significance tests of network characteristics, goodness-of-fit tests for mechanisms leading to plausible network formation, models that predict the presence of ties between actors, as well as forecasts of disease spread in networks. We divide network models into two broad categories: (1) mathematical network models and (2) statistical network models. We review each model category in the remainder of this section.

3.1 Mathematical Network Models

Mathematical network models provide closed-form analytical statements about phenomena such as disease spread due to their mathematical tractability. However, their simplicity often makes them poor approximations to real-world networks. Nonetheless, the intuition they provide about disease spread is invaluable to understanding the COVID-19 pandemic. This section introduces the fundamental mathematical network models in the literature. We use these models to study the COVID-19 pandemic in Sections 4.2 and 4.3.

3.1.1 Random Graph Models

Random graph models are the simplest – although one of the most useful – classes of network models. A random graph model places a uniform distribution over a pre-specified collection of graphs. The Erdős-Rényi model is the earliest studied random graph model. There are two definitions of this model. The first definition coincides with the model proposed in the original work by Erdős and Rényi (1959). Under this definition, the Erdős-Rényi model places a uniform distribution over graphs with the same number of vertices and edges. The second definition corresponds to a model proposed around the same time by Gilbert (1959). This variant of the Erdős-Rényi model places a distribution over graphs with n vertices that can be obtained by independently connecting two vertices with probability $p \in (0, 1)$. In the network literature, the Gilbert (1959) variant is commonly referred to as the Erdős-Rényi model because it is easier to analyze mathematically. We adopt this definition in this article.

Beyond the Erdős-Rényi model, the most commonly used random graph model places a uniform distribution over graphs with a prescribed degree sequence $\mathbf{d} = (d(1), \ldots, d(n))$. This model often serves as the null distribution in statistical hypothesis tests. For example, Milo et al. (2002) determined specific network motifs are more prevalent in real-world networks than random graphs with the same degree sequence. Although there are a variety of Monte Carlo methods designed to sample uniformly from random graph models with a fixed degree sequence, a popular algorithm called the configuration model is useful for understanding disease spread. The configuration model begins by assigning $d(i)$ stubs to each vertex i. Then two stubs are connected uniformly at random to form an edge. This procedure proceeds on the remaining stubs until all stubs are connected. Note that this algorithm allows for self-loops and multi-edges; however, they represent a small proportion of edges. For this reason, one loses little by discarding them when the number of edges is large.

When studying the asymptotic properties of the configuration model, one often assumes that a known parametric distribution generates the degree sequence (Molloy and Reed, 1995, 1998). For example, the Erdős-Rényi model has a Binomial$(n-1, p)$ degree distribution, which is often approximated by a Poisson distribution when n is large. A line of network science research involves fitting parametric distributions to the degree sequence of real-world networks to determine more realistic random graph models.

3.1.2 Scale-Free Networks

In the network literature, one of the most pervasive degree distributions used to mimic real-world networks is the power-law distribution, i.e., $f_d \propto d^{-\gamma}$ where $d > 0$ and $\gamma > 0$. So much so that networks with a power-law degree distribution have garnered the name 'scale-free' networks. The scale-free term comes from the scale-free property of the power-law distribution: $f_{a \cdot d} = a^{-\gamma} f_d \propto f_d$. In other words, the shape of the degree distribution is the same regardless of the input's scale. Typically, γ is taken between two and three. In recent years, researchers have applied rigorous goodness-of-fit tests to networks that were conjectured to be scale-free (Khanin and Wit, 2006; Clauset et al., 2009). These studies found that a pure power-law is quite uncommon, and the degree distribution is typically just heavy-tailed. For this reason, one often restricts the power-law distribution to take values on a finite range of degrees where the power-law provides a good approximation (Pastor-Satorras and Vespignani, 2002a).

3.1.3 Preferential Attachment

The preferential attachment (PA) model is a famous mathematical network model used to generate scale-free networks. Unlike random graph models, the preferential attachment model is part of an area of mathematical network modeling that tries to construct realistic networks based on a plausible generation mechanism. The preferential attachment mechanism mimics the 'rich get richer' phenomenon. The PA model, proposed by Barabási and Albert (1999), builds up a network sequentially over time. Starting with a graph with node set V_t and edge set E_t at time t, one forms a new graph at time $t+1$ by adding a vertex with m new edges. These edges are connected to existing vertices in E_t with probability proportional to their degree. This procedure results in new vertices tending to attach to existing high-degree vertices. In the original work, the authors showed that as t goes to infinity, the degree distribution approaches a power-law with exponent $\gamma = 3$.

3.1.4 Small-World Models

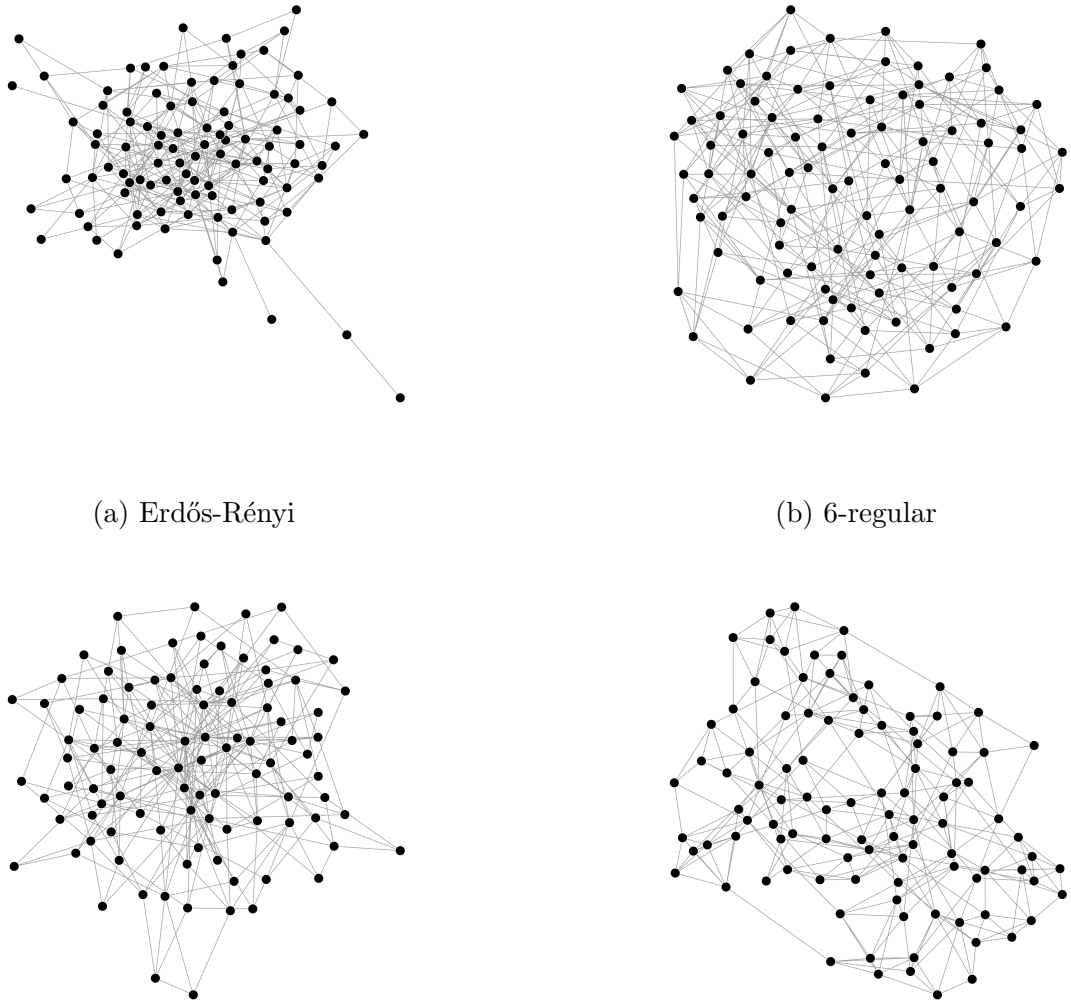
The previous models do not capture many high-order network characteristics prevalent in real-world networks. In particular, they lack the small-world property, which states that many networks have small average distances between nodes and high transitivity. For-

mally, a network model exhibits small-world behavior if its diameter scales as $\mathcal{O}(\log(n))$ while the clustering coefficient, $\mathrm{cl}(G)$, remains large. In particular, the Erdős-Rényi and preferential attachment models do not exhibit the small-world property since their clustering coefficients scale as n^{-1} (Kolaczyk, 2017; Bollobás and Riordan, 2003). Watts and Strogatz (1998) proposed a famous mathematical model that exhibits small-world behavior. In its basic form, to generate a graph, the model begins by arranging the nodes on a regular ring lattice, where each node is connected to r neighbors on each side. Then, for each edge, the model rewires the edge with probability p to another node chosen uniformly at random while avoiding self-loops and repeated edges. The resulting graphs tend to have high transitivity while maintaining small average distances.

3.1.5 An Empirical Comparison of Mathematical Network Models

We end this section by demonstrating how the network characteristics described in Section 2.2 can distinguish the previous mathematical network models. Figure 1 visualizes an (a) Erdős-Rényi, (b) 6-regular, (c) preferential attachment, and (d) small-world network with the Fruchterman-Reingold layout (Fruchterman and Reingold, 1991). To focus on structural differences, we fixed the number of nodes to $n = 100$ and chose the model parameters so that each had an average degree of roughly six. Although it is clear from the visualization that the networks are structurally different, it is difficult to pinpoint how they differ. For this, we use the network statistics described in Section 2.2.

Table 1 summarizes the median of five network statistics calculated from 100 simulations of an Erdős-Rényi, 6-regular, preferential attachment, and small-world network with the same settings as visualized networks but with $n = 10{,}000$ nodes to match the settings of Section 4.3. We also included the real-world primary school network ($n = 236$) analyzed in Section 4.4. The 6-regular and Erdős-Rényi models serve as baselines with small degree heterogeneity and little higher-order structure. The PA model has the most heavy-tailed degree distribution with a median maximum degree of 211. The small-world network is the only network model that contains a large amount of transitivity. Note that the distance between nodes is quite small for all networks. All of these models contain minimal degree assortativity, which is highly prevalent in the real primary school network. Ensuring that we capture the relevant structures found in the network under investigation is crucial for network modeling. In this regard, statistical network models



(a) Erdős-Rényi

(b) 6-regular

(c) Preferential attachment

(d) Small-world

Figure 1: Visualization of an (a) Erdős-Rényi, (b) 6-regular, (c) preferential attachment, and (d) small-world network. Each network has $n = 100$ nodes and an average degree of roughly six. The visualizations use the Fruchterman-Reingold layout (Fruchterman and Reingold, 1991).

provide a powerful tool, which we review in the next section.

	Max. Degree	Assortativity	Transitivity	Avg. Path Length	Diameter
Erdős-Rényi	17(1.1)	0.000(0.005)	0.006 (0.001)	5.35 (0.01)	10 (0.4)
6-Regular	6 (0)	NA	0.0042 (0.0001)	5.61 (0.0006)	8 (0)
PA	211 (35)	−0.006 (0.004)	0.003 (0.0002)	4.46 (0.02)	7 (0.2)
Small-World	11 (0.6)	−0.001 (0.006)	0.308 (0.003)	7.05 (0.03)	11 (0.5)
Primary School	36	0.21	0.47	2.94	7

Table 1: Summary statistics for four mathematical network models (Erdős-Rényi, 6-regular, preferential attachment, and small-world) as well as the real-world primary school network described in Section 4.4. To focus on topological differences, we fixed the number of nodes n = 10,000, the average degree to roughly 6, and the number of edges to the same order of magnitude (~30,000) for all simulated networks. For the simulations, the reported statistics are the median value over 100 different samples. Values in parentheses are the standard deviation over these 100 samples. Note that the assortativity of a 6-regular network is undefined since it has zero degree variance.

3.2 Statistical Network Models

Statistical network models posit that the adjacency matrix, $\mathbf{Y}$, and its elements are random variables with complex interdependencies. We differentiate statistical network models by the statistical dependencies they place on the adjacency matrix. As their name implies, statistical network models are suited for typical tasks of statistical analysis such as significance testing, model comparison, and uncertainty quantification. Furthermore, they are often highly flexible, which allows them to better capture the higher-order network structures described in Section 2.2. However, unlike mathematical network models, statistical network models are often analytically intractable. In the following sections, we provide an overview of the common types of statistical network models.

3.2.1 ERGMs

Exponential random graph models (ERGMs) or p^* models (Holland and Leinhardt, 1981; Frank and Strauss, 1986; Wasserman and Pattison, 1996; Robins et al., 2007) posit that the adjacency matrix is drawn from a canonical exponential family, i.e.,

$$\mathbb{P}\left(\mathbf{Y} = \mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}\right) = k(\boldsymbol{\theta}, \mathbf{X})^{-1} \exp(\boldsymbol{\theta}^{\mathrm{T}} g(\mathbf{y}, \mathbf{X})), \tag{3.2}$$

where $\mathbf{X}$ is a collection of exogenous covariates, $g(\mathbf{y}, \mathbf{X})$ is a p-dimensional vector of statistics, $\boldsymbol{\theta}$ is a p-dimensional vector of coefficients, and $k(\boldsymbol{\theta}, \mathbf{X})$ is the normalizing constant.

The main benefit of ERGMs are their ability to test the prevalence of specific characteristics in a network. These tests require choosing network statistics $g(\mathbf{y}, \mathbf{X})$ that plausibly describe the network's structure. For example, to test the presence of transitivity in a social network, one often includes some form of triangle counts. Other commonly used statistics include the number of edges, the degree sequence, and the number of k-stars. More recently, Hunter and Handcock (2006) found that geometrically weighted degree, edgewise shared partner, and dyadwise shard partner statistics adequately fit real-world networks. See Snijders et al. (2006) or Hunter et al. (2008a) for a comprehensive overview of commonly used network statistics.

The coefficients, $\boldsymbol{\theta}$, have a simple interpretation as a log-odds. To see this, one defines the so called 'change statistics':

$$\delta_g(y_{ij}) = g(\mathbf{y}_{ij}^+, \mathbf{X}) - g(\mathbf{y}_{ij}^-, \mathbf{X}), \tag{3.3}$$

where $\mathbf{y}_{ij}^+$ and $\mathbf{y}_{ij}^-$ represent the network realized by fixing $y_{ij} = 1$ or $y_{ij} = 0$, respectively, while keeping the rest of the observed network unchanged. In other words, $\delta_g(y_{ij})$ is the change in the network statistics when we flip $y_{ij} = 0$ to $y_{ij} = 1$ while keeping the rest of the observed network and covariates fixed. It can be shown that

$$\text{logit}\left[\mathbb{P}\left(Y_{ij} = 1 \mid \mathbf{Y}_{ij}^c = \mathbf{y}_{ij}^c\right)\right] = \boldsymbol{\theta}^{\mathrm{T}} \delta_g(y_{ij}), \tag{3.4}$$

where $\mathbf{Y}_{ij}^c$ represents the rest of the network other than the single random variable Y_{ij}. This expression shows that each coefficient in $\boldsymbol{\theta}$ can be interpreted as its statistic's contribution, per unit increase in the statistic, to the log-odds of an individual tie conditioned on the rest of the network. For additional exposition on the interpretation of ERGM coefficients, see Goodreau et al. (2008).

Despite their interpretability, ERGMs can have trouble fitting observed networks due to model degeneracy (Handcock, 2003). Model degeneracy refers to the phenomenon where an inferred distribution places most of its probability mass on a small subset of the probability space. In particular, ERGM estimates often place most of their mass on

either empty or complete networks. Such degeneracy can lead to poor estimates or issues of convergence of the estimating algorithms.

Since their introduction, various extensions of ERGMs have been proposed. Krivitsky and Morris (2017) extended the ERGM to networks sampled under an egocentric design. Extensions to dynamic networks include the temporal ERGM (TERGM) (Hanneke et al., 2010) and the separable temporal ERGM (STERGM) (Krivitsky and Handcock, 2013).

3.2.2 Latent Space Models

Latent space models (LSMs) are an alternative to ERGMs designed to capture the high levels of transitivity, reciprocity, and assortativity found in real-world networks. They do not suffer from the same model-degeneracy of ERGMs while still providing a highly flexible statistical network model. Introduced by Hoff et al. (2002), LSMs embed a network's actors into a p-dimensional latent metric space. The two most popular choices of metric space are the Euclidean space (the distance model) or a hypersphere (the projection model). The likelihood of the model assumes that closeness in the latent space increases the probability that two actors form an edge in the observed network. For this reason, one interprets the proximity of two actors in the latent space as an indication that they have similar latent characteristics. This property of the LSM mimics the assortative nature of observed characteristics in real-world networks.

Formally, the latent space model assumes edges form independently when conditioned on the actor's latent positions $\{\mathbf{Z}_i\}_{i=1}^n$ and the exogenous dyadwise covariates $\{\mathbf{X}_{ij}\}$:

$$\mathbb{P}\left(\mathbf{Y} = \mathbf{y} \mid \{\mathbf{X}_{ij}\}, \{\mathbf{Z}_i\}_{i=1}^n, \alpha, \boldsymbol{\beta}\right) = \prod_{i<j} \mathbb{P}\left(Y_{ij} = y_{ij} \mid \mathbf{X}_{ij}, \mathbf{Z}_i, \mathbf{Z}_j, \alpha, \boldsymbol{\beta}\right), \tag{3.5}$$

where

$$\text{logit}\left[\mathbb{P}\left(Y_{ij} = 1 \mid \mathbf{X}_{ij}, \mathbf{Z}_i, \mathbf{Z}_j, \alpha, \boldsymbol{\beta}\right)\right] = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij} - \|\mathbf{Z}_i - \mathbf{Z}_j\|_2. \tag{3.6}$$

In the previous expression, α is a scalar intercept and $\boldsymbol{\beta}$ is a vector of coefficients. From this, one can see that the latent positions act as multiplicative random effects in a dyadwise logistic regression model. Note that Equation (3.6) uses the distance model formulation. The equivalent expression for the projection model replaces the negative Euclidean distance with the projection operator $\mathbf{Z}_i^{\mathrm{T}}\mathbf{Z}_j/\|\mathbf{Z}_j\|_2$. The projection model is suited to modeling reciprocity in directed networks because of the asymmetry in the metric. Also,

all models exhibit transitivity since the latent metric space incorporates the triangle inequality.

There have been numerous extensions to the original latent space model. The latent position clustering model (LPCM) of Handcock et al. (2007) added community structure to the latent space through a Gaussian mixture model. Individual additive random effects were incorporated to explicitly model degree heterogeneity in Hoff (2005) and Krivitsky et al. (2009). Young and Scheinerman (2007) introduced the random dot product graph, which removes the logit-link function and utilizes the dot-product, i.e., $\mathbb{P}\left(Y_{ij} = 1 \mid \mathbf{Z}_i, \mathbf{Z}_j\right) = \mathbf{Z}_i^{\mathrm{T}}\mathbf{Z}_j$. Many asymptotic results and applications exist for random dot product graphs, see Athreya et al. (2018) for a thorough survey. Sarkar and Moore (2006) and Sewell and Chen (2015) proposed extensions of the LSM to dynamic networks.

3.2.3 Stochastic Block Models

Stochastic block models (SBMs) (Holland et al., 1983; Nowicki and Snijders, 2001) are among the most well-studied network models in the statistics literature. The SBM posits that each actor belongs to one of K communities. Conditioned on these community memberships, the entries of the adjacency matrix are independent Bernoulli random variables:

$$Y_{ij} \sim \text{Bernoulli}(B_{pq}), \tag{3.7}$$

where actor i is a member of community p, actor j is a member of community q, and B is a $K \times K$ matrix with entries indicating the probability of forming an edge between communities p and q. Note that the community memberships are unknown, so they must be inferred from the data. It is possible to analyze the SBM as a mixture of Erdős-Rényi graphs. Therefore, many of the properties of the random networks sampled from an SBM are analytically expressible in terms of the underlying model parameters (Daudin et al., 2008). This analytical tractability makes the SBM a compelling alternative to standard mathematical network models.

Various extensions to the SBM have been proposed. The mixed membership stochastic block model (MMSBM) (Airoldi et al., 2008) and the overlapping stochastic block model (OSBM) (Latouche et al., 2011) allow vertices to be a member of more than one community. The class of degree corrected stochastic block models aims to produce SBMs with more heterogeneous degree distributions (Karrer and Newman, 2011). Finally, the

SBM was extended to dynamic networks by Matias and Miele (2017).

4 Applications to the COVID-19 Pandemic

4.1 Stochastic Compartmental Models in Epidemiology

Models of infectious disease spread on networks have a long history in the literature. For this case study, we focus on the SEIR epidemic model, which has been widely used to understand the spread of COVID-19 (Branas et al., 2020; Li, 2020). The SEIR model is a compartmental model (Kermack and McKendrick, 1927) for which there are many standard references (Anderson and May, 1991; Daley and Gani, 2001).

The SEIR model breaks the population up into four non-overlapping compartments: susceptible individuals (S), exposed individuals (E), infectious individuals (I), and recovered/deceased individuals (R). As depicted in Figure 2, the SEIR model assumes the epidemic evolves as follows: individuals are first susceptible, then they are exposed but cannot spread the disease, after an incubation period they become infectious, and after some time they recover and are immune. The three positive rate parameters, $\{\tau, \alpha, \gamma\}$, determine the nature of the disease's spread. The infection rate τ is the rate at which infectious individuals make infection-transmitting contacts. The incubation rate $\alpha = 1/D$, where D is the incubation period of the disease, i.e., the time between contracting the disease and becoming infectious. The recovery rate γ is the time it takes an infectious individual to recover either through developed immunity or death.

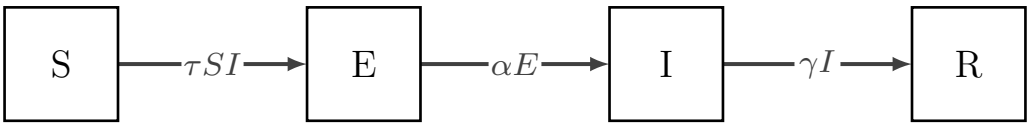


Figure 2: Flow diagram showing the flux between compartments for the stochastic SEIR model. The compartments include susceptible (S), exposed (E), infectious (I), and recovered (R) individuals.

Standard stochastic compartmental models assume that individuals are homogeneously mixed in the population. This property states that any pair of individuals in the population is equally likely to interact. In real-life scenarios, this assumption is rarely justified. Heterogeneous contact patterns often emerge due to social structures such as friend or work relations, household structures, and geographical structures. By encoding a popu-

lation's heterogeneous contact patterns in a network, network models provide a powerful tool for understanding the spread of disease in the presence of heterogeneous mixing.

4.2 Epidemic Spread on Networks

The SEIR epidemic on a network G with n vertices is modeled as a continuous-time Markov chain with state vector $\mathbf{X}(t) \in \{s, e, i, r\}^n$, an n-tuple with entries in the set $\{s, e, i, r\}$. The ith entry of $\mathbf{X}(t)$ corresponds to the state of the ith vertex in the network at time t. The assumptions are that

- transmission from an infectious node (i) to a susceptible node (s) occurs across an edge as a Poisson process with rate τ,
- an exposed node (e) becomes infectious (i) as a Poisson process with rate α, and
- an infectious node (i) recovers as a Poisson process with rate γ.

All these events are assumed to be independent. A consequence of these independence assumptions is that a susceptible node with k infectious neighbors becomes exposed as a Poisson process with rate $k\tau$. Monte Carlo methods are used to simulate SEIR epidemics on a network. In this work, we adopt the Gillespie algorithm (Gillespie, 1977) to simulate the SEIR stochastic process. See Kiss et al. (2017) and Andersson and Britton (2000) for further details on network epidemic models and other approaches to simulation.

Whether an infectious disease becomes an epidemic is a well-studied question in regard to disease spread. The basic reproduction number, R_0, which is (loosely) defined as the expected number of new infections caused by a typical infectious individual during the early stages of the epidemic, is crucial to characterize the disease. The importance of R_0 derives from its role in certain threshold theorems, which state that an epidemic occurs when $R_0 \geq 1$. For the configuration model with a prescribed degree distribution, Andersson (1998) showed that the basic reproduction number is

$$R_0 = \frac{\tau}{\tau + \gamma}\left[\frac{\mathbb{E}\left[d^2\right]}{\mathbb{E}\left[d\right]} - 1\right] = \frac{\tau}{\tau + \gamma}\left[\mathbb{E}\left[d\right] + \frac{\mathrm{Var}(d)}{\mathbb{E}\left[d\right]} - 1\right], \tag{4.1}$$

where d is the vertex-degree, and expectations are taken with respect to the degree distribution. Equation (4.1) has an intuitive meaning. The first term, $\tau/(\tau + \gamma)$, is the probability that an infectious node transmits the infection to an arbitrary neighbor before

recovering. The second term is the infectious node's expected number of neighbors during the early stages of the epidemic minus one. We subtract one because an infectious node cannot re-infect the neighbor that initially transmitted the infection. Note that a node's expected number of neighbors is not the expected degree of a node, but larger by a factor that depends on the variance of the degree distribution. In the social network literature, this phenomenon is known as the 'friendship paradox': your friends have more friends than you do (Feld, 1991). Under the configuration model, the friendship paradox occurs because a node of degree k is k times as likely to be your neighbor than a node of degree one.

The expression in Equation (4.1) has significant consequences for epidemics on networks. In particular, for a fixed expected degree, R_0 is at a minimum for homogeneous (d-regular) networks, and increases as the degree distribution becomes more heterogeneous. Intuitively, this occurs because a disease that infects a few high degree nodes in a heterogeneous network creates 'super-spreaders' that can quickly spread the disease throughout the network. In the remainder of this work, we use network models along with network epidemic models to better understand the COVID-19 pandemic.

4.3 The Effect of Network Properties on an SEIR Epidemic

The network properties discussed in Section 2.2 have significant effects on disease spread in a network. In particular, we know from Equation (4.1) that networks with different degree distributions lead to epidemics with different behaviors. To demonstrate these effects, we simulate an SEIR epidemic on four mathematical network models: a scale-free network generated through preferential attachment, an Erdős-Rényi model, a d-regular network, and a small-world model. We generated each network with n = 10,000 vertices. We chose the parameters of these models so that the networks roughly shared an average degree of six and an edge count of 30,000. These constraints allowed differences in disease spread to result from different higher-order characteristics in the networks. We summarize various network statistics related to these networks in Table 1.

We chose the SEIR model parameters to match estimated rates for the spread of COVID-19 in Wuhan, China early in the pandemic (Prem et al., 2020). We set the incubation period to 6.4 days so that $\alpha = 1/6.4$. The recover period was set to 5 days so that $\gamma = 1/5$. Using Equation (4.1), we set the edge infection rate $\tau = 0.157$ so that

$R_0 = (6-1)\,\tau/(\tau+\gamma) = 2.2$ on a 6-regular network. The simulations began with 1% of the population infected. Furthermore, we started time $t = 0$ at the day prevalence (infectious plus recovered) reached 2.5% of the population. For each network, we ran 25 simulations of the epidemic. The results of the stochastic simulations are displayed in Figure 3.

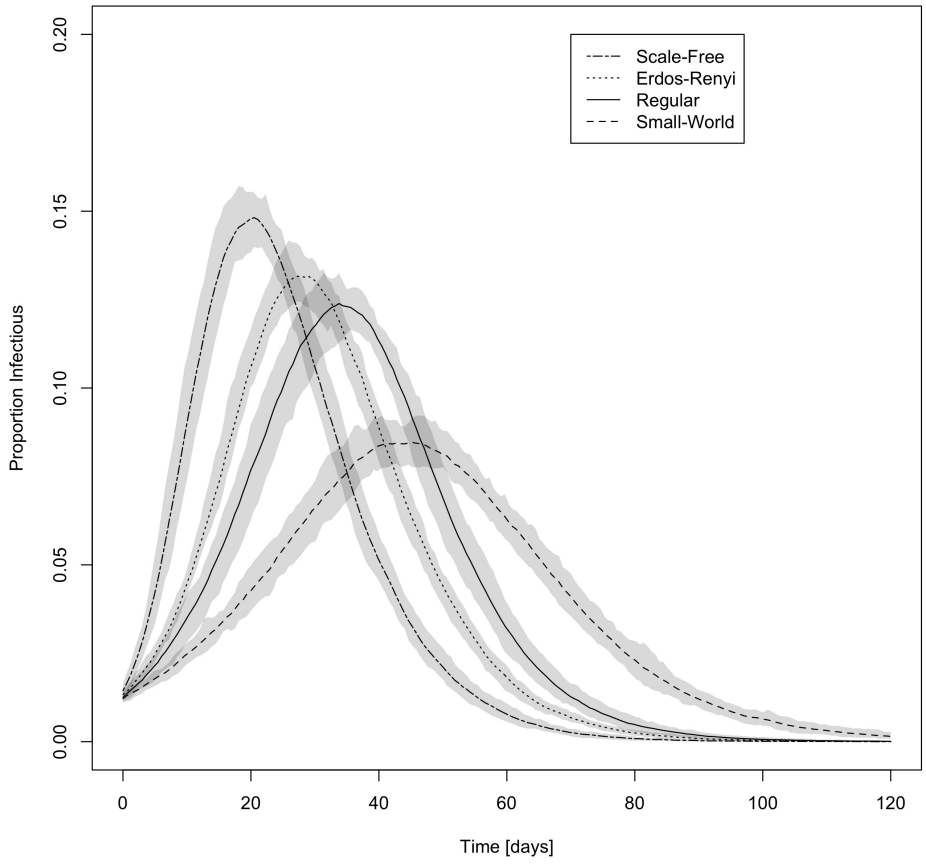


Figure 3: A comparison of an SEIR epidemic on four different mathematical network models each with roughly an average degree of six. Each network had $n = 10{,}000$ vertices. The parameters of the SEIR epidemic were set to match estimated values for COVID-19 in Wuhan, China. The simulations began with 1% of the population infected. The time $t = 0$ corresponds to the day prevalence (infectious plus recovered) reached 2.5% of the population. The curves indicate the fraction of the total population who are infectious at a given time. The black lines correspond to the mean infection curve over all simulations. The shaded gray regions correspond to the 95% simulation quantile regions over the 25 simulations. These curves demonstrate that degree heterogeneity strongly influences the course of an epidemic.

We summarize the simulation results with infection curves, which indicate the fraction of the total population infectious at a given time. The black lines correspond to the mean infection curve over all simulations. The shaded gray regions correspond to the 95%

simulation quantile regions over the 25 simulations. In this case study, we highlight two characteristics of infection curves relevant to the COVID-19 response: (1) the growth rate of the epidemic and (2) the peak infectious rate. A high growth rate quickly overwhelms local health care systems if the peak infectious rate is above their capacity for treatment. Policies that modify appropriate aspects of people's contact networks can mitigate this outcome. Any differences in the infection curves of the four mathematical network models allow us to pinpoint these characteristics.

The infection curves associated with the four mathematical network models are different in significant ways. Focusing on the networks with known degree distributions (scale-free, Erdős-Rényi, and the 6-regular network), we notice a few common trends. The infection curves peak at lower values as we move from power-law to a homogeneous degree distribution. Furthermore, the growth rate of the epidemic increases as the variance in the degree distribution increases. These results highlight the danger of highly heterogeneous networks wherein an individual with a large number of neighbors can quickly disseminate the disease. On the other hand, the small-world network demonstrates the smallest growth rate and peak. The small-world network's higher average path length and diameter may contribute to this behavior since these properties slow the spread of the infection. This result emphasizes the need to spread out people's contacts.

When modeling epidemics, we have no way to rigorously determine whether mathematical network models capture all of the essential intricacies found in real-world networks. As we demonstrate in the next section, we can use statistical network models for such an analysis.

4.4 Case Study: Infectious Disease Spread in a Primary School

This case study's goal is to demonstrate how one can use statistical network models to transfer uncertainty about disease spread in one network to other real-world networks with similar characteristics. We use a real-world network constructed from contact patterns observed on a single day at a primary school (grades 1 - 5) in Lyon, France. The data was collected by the SocioPatterns collaboration (`http://www.sociopatterns.org`) and initially analyzed in Stehlé et al. (2011).

Figure 4 displays the full network. An edge between two actors indicates that they shared at least one interaction that lasted more than two minutes on October 1st, 2009.

RFID sensors worn by each individual measured the interactions. The RFID sensor registered a contact when two actors were within 1 to 1.5 meters during a 20-second interval. Stehlé et al. (2011) chose a distance of 1 to 1.5 meters to correspond to the range at which a communicable infectious disease could spread. There is a total of 236 actors (226 students and 10 teachers) in the network. Metadata such as grade, gender, and section was recorded for each actor.

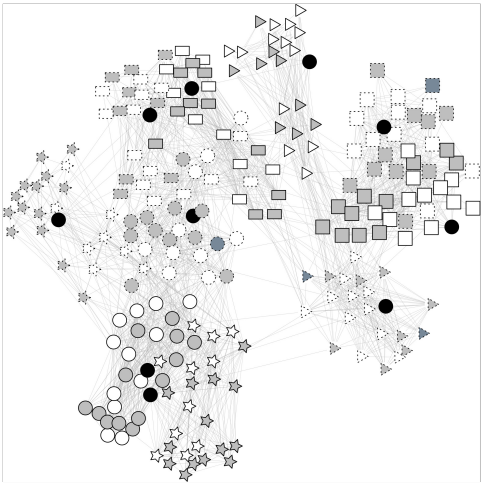


Figure 4: Face-to-face contact network measured at a French primary school during October 1st, 2009 (Stehlé et al., 2011). Two individuals are connected by an edge if they shared at least one interaction that lasted more than two minutes. The node's shapes (circle, star, rectangle, triangle and square) denote grades 1 through 5, respectively. Dashed nodes correspond to individuals from section A, while solid nodes correspond to individuals in section B. Male students are colored in white, female students are colored in light gray, and students with an unknown sex are colored in dark gray. Black circles are the teachers.

Our analysis aims to quantify the range of infection curves we expect to see during an SEIR epidemic at the school. Furthermore, although the contact network corresponds to a single day, we would like to account for variations in the network's edges on other days. For this purpose, we make the reasonable assumption that the same underlying mechanisms drive the contact networks on other days. Under this assumption, we can quantify possible network variability with a statistical network model.

For our statistical network model, we fit an ERGM to the primary school contact network. We chose geometrically weighted edgewise shared partner, dyadwise shared

partner, and degree with a fixed decay rate of 0.25 as our network statistics. In Figure 4, we see a large amount of homophily between nodes sharing the same grade, section, or sex. We incorporated this structure into the model through homophily statistics, which count the number of edges where actors share a given trait. We use homophily statistics for grade, section, and sex. Also, we added a statistic counting the number of edges sharing a specific difference in grade, e.g., the number of edges whose actors are separated by two grade levels. For a more thorough introduction on how to apply ERGMs to school networks, see Hunter et al. (2008a). The goodness-of-fit diagnostics displayed in Figure 5 demonstrate that the ERGM adequately describes the primary school contact network.

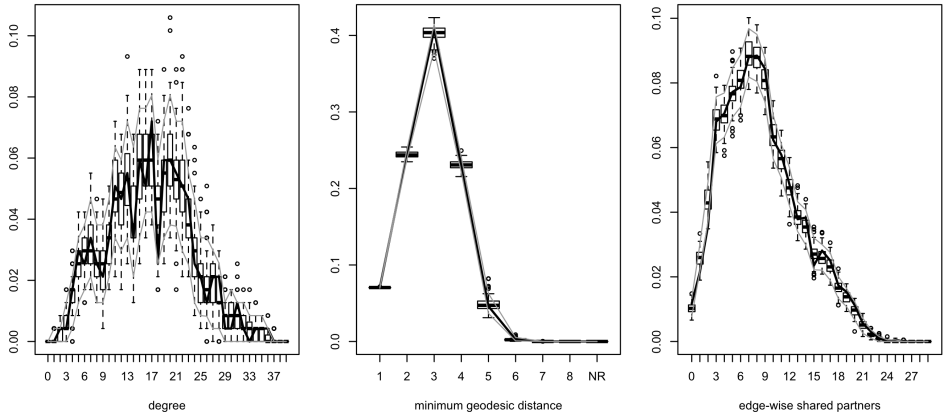


Figure 5: Goodness-of-fit diagnostics for the ERGM fit to the primary school network in Section 4.4. Black lines correspond to the statistics calculated on the observed network. Boxplots contain draws from the distribution inferred by the ERGM. The observed network is consistent with the simulations, which indicates the ERGM fits the data well.

To quantify the uncertainty in the spread of the epidemic, we combined the results from the statistical network model with SEIR epidemic simulations. We used the same SEIR rate parameters given in Section 4.3. To incorporate uncertainty due to the network into the epidemic simulations, we ran 50 SEIR epidemics over 10 different networks sampled from the fitted ERGM. We combined the resulting 500 epidemic trajectories to infer a distribution of infection curves. As a comparison, we estimated another infection curve based on 50 epidemic simulations on the observed contact network. The results of this experiment are displayed in Figure 6.

The infection curve's shape indicates that a significant fraction of the primary school's population, 15% to 25%, is infected at the peak of the epidemic. Furthermore, this peak occurs only two weeks after the disease reached 2.5% prevalence in the population. This

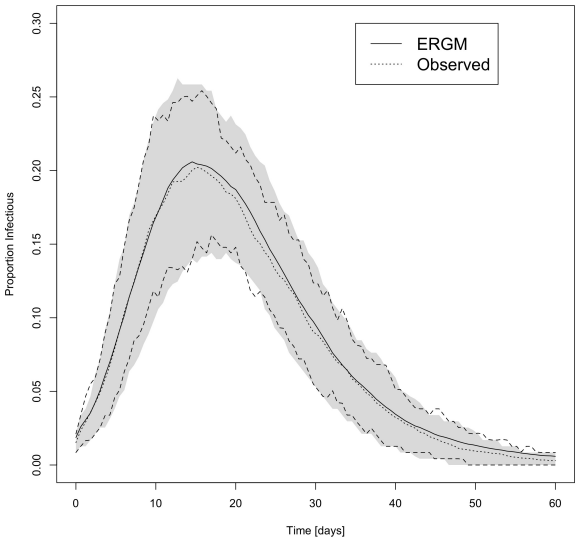


Figure 6: Estimated infection curve for the primary school network. Dashed lines bound the 95% simulation quantile regions when only accounting for the epidemic process on the observed primary school network. The dotted curve is the mean of these simulations. The shaded gray area denotes the 95% simulation quantile region that incorporates variability due to the network model. Similarly, the solid black curve is the mean of these simulations. Notice how the gray region is higher and broader than the region contained in the dashed lines.

quick rate of spread illustrates the danger of an outbreak in such a highly connected community. Furthermore, the inclusion of network uncertainty inflates the 95% quantile intervals near the peak of the epidemic compared to the estimates based on only the observed network. However, in this case, the increase in uncertainty is small compared to the uncertainty due to the SEIR simulation based on only the observed network. This behavior is primarily because the network is small (n = 236). Regardless, these simulations provide insight into the course of the epidemic at the primary school. The results invite the question of what interventions we could put into place to effectively 'flatten' the infection curve.

4.5 Case Study: The Effect of a Stratified School Opening

In this case study, we explore the effect of a stratified school opening on flattening the infection curve. Testing such social distancing intervention strategies is a unique ability of network epidemic models compared to their homogeneous counterparts described in Section 4.1. To do this, we again utilize the same primary school contact network described

in Section 4.4. However, we assume the school puts an intervention in place that cuts off contacts between actors in section A and actors in section B. For example, by alternating the days in which individuals from section A and section B attend school. Figure 7 displays the resulting network. Notice that the network has two connected components coinciding with the two sections.

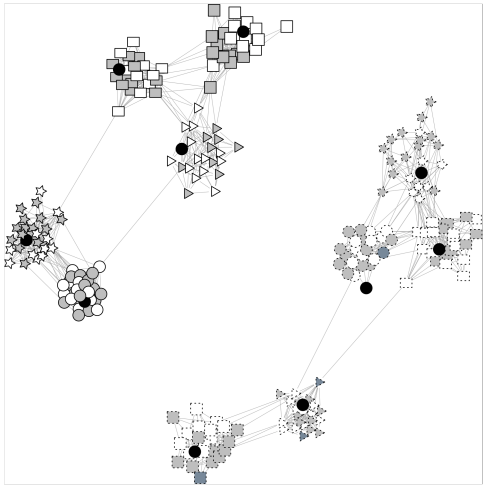


Figure 7: Face-to-face contact network at a French primary school during October 1st, 2009 (Stehlé et al., 2011). Edges between individuals in different sections were removed to mimic a stratified opening intervention. See Figure 4 for the meaning of the various symbols.

We compare the intervention with a non-intervention baseline using simulation. The non-intervention baseline corresponds to the original network analyzed in Section 4.4. For each network, we ran 100 SEIR simulations using the same parameters described in Section 4.3. Figure 8 displays the resulting infection curves.

The intervention dramatically reduces the spread of the infection. With the intervention in place, the infected population comprises only 5% to 15% of the school during the peak of the epidemic. This proportion is almost half the amount as the non-intervention curve. With no intervention, 15% to 25% of the actors are infected at the epidemic's peak. Also, the growth rate of the epidemic is much slower with the intervention in place. Both observations support the claim that a stratified school opening dramatically reduces the magnitude of the epidemic. While the intervention strategy we explored here is rather simplistic, it is the starting point for exploring more complicated and realistic

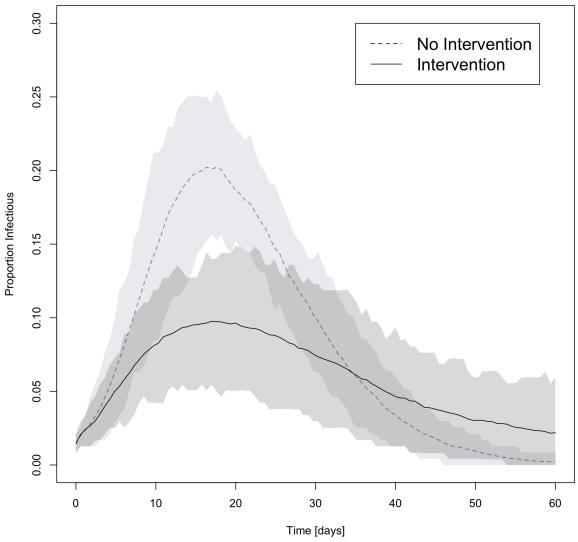


Figure 8: Estimated infection curve for the primary school network with and without a stratified opening in place. The dashed black line corresponds to the mean of the process with no intervention in place. The shaded light gray area denotes the 95% simulation quantile region for the network with no intervention. The solid black line is the mean of the process with an intervention in place. Similarly, the shaded dark gray area denotes the 95% simulation quantile region for the network with an intervention in place. The intervention significantly flattens the infection curve.

approaches to flattening the curve.

4.6 Disease Spread on Dynamic Networks

In recent years, there has been rapid progress in understanding disease spread in dynamic networks. This task is challenging because one must account for two coupled stochastic processes: (1) the disease dynamics on the network and (2) the dynamics of the network itself (Gross and Blasius, 2008). Specifically, susceptible individuals tend to cease contact with infectious individuals. However, accounting for this coupling is only necessary when the two processes operate on the same time scale, i.e., the time scales of contact duration and disease transmission are comparable. Otherwise, an equivalent analysis performed on an appropriately weighted static network provides a good approximation for the spread of the disease (Pastor-Satorras et al., 2015).

Epidemic spread on dynamic networks contains a broader range of behavior than those found in static networks. The following outcomes are possible: (a) epidemic extinction where the disease dies out, (b) endemic equilibrium where the disease is always present in

a fraction of the population, and (c) oscillations in the number of infectious individuals. For example, Rogers et al. (2012) analyzed a SIRS epidemic with node birth and death, edge rewiring based on proximity to an infectious node, and random edge rewiring. They observed oscillatory behavior in the infection curve, where the oscillation's frequency increased with the infection rate τ and its amplitude increased with the rate of rewiring between susceptible and infectious individuals.

5 Software

Various software packages are available for statistical network analysis. For simplicity, we focus on implementations in the `R` programming language. The `igraph` package (Csardi and Nepusz, 2006) provides tools that calculate the network statistics described in Section 2 and generate the mathematical network models described in Section 3. Also, the `igraph` package processes large networks efficiently. The `ergm` package (Hunter et al., 2008b) provides functions for estimating ERGMs. The ERGM estimation procedure is computationally expensive, so it can only handle networks with hundreds of nodes. Latent space models are available in the `latentnet` package (Krivitsky and Handcock, 2008). These algorithms suffer from the same computational complexity problems as ERGMs; however, an efficient variational approximation is available in the `lvm4net` package (Gollini, 2019). Algorithms to estimate stochastic block models are available in the `mixer` package (Daudin et al., 2008; Zanghi et al., 2008, 2010; Latouche et al., 2012). Finally, `EpiModel` (Jenness et al., 2018) is an excellent software package for the mathematical modeling of infectious disease. This package contains tools for modeling disease spread that are similar to the analysis described in Section 4.4.

6 Conclusion

Statistical network analysis consists of a comprehensive set of tools for drawing inferences from complex relational data. These tools are especially relevant in light of the COVID-19 pandemic, where networks play a crucial role in understanding the spread of this deadly virus. Luckily, research in the network sciences has identified characteristics shared by many real-world networks that are relevant to disease spread. These include the realiza-

tion that real-world networks often have high levels of degree heterogeneity, assortativity, and transitivity, along with relatively small diameters. Of course, when studying dynamic processes such as the spread of disease on a network, it is important to develop models for the network itself. These models allow one to understand the properties of disease spread even when the real contact network is unavailable. We split these models into two categories: (1) mathematical network models such as random graph models, preferential attachment models, and small-world models, and (2) statistical network models such as ERGMs, latent space models, and stochastic block models. Mathematical network models are particularly useful when making closed-form analytical statements about epidemics. In contrast, statistical network models play a role in quantifying uncertainty when measurements from observed networks are available. As we demonstrated, all of the tools in statistical network analysis play a role in adequately modeling the spread of infectious disease.

This review only covered the basics of network modeling. As such, there are several ways to develop more realistic network models relevant to understanding COVID-19. One approach is to incorporate models for how networks change over time into epidemic simulations. For example, in Section 4.4, one could use a STERGM instead of an ERGM to model network dynamics (Jenness et al., 2018). Furthermore, we omitted a few areas of research. For example, when modeling networks, one should take into account how the network was sampled. The two most common sampling schemes for networks are link-tracing and egocentric designs. For an example of a network model used to understand HIV spread under an egocentric design, see Krivitsky and Morris (2017). We also omitted the concept of node centrality, whose goal is to identify important nodes in a network. Such identification is an essential component in the development of effective vaccination strategies, see Pastor-Satorras and Vespignani (2002b) and Wang et al. (2016) for details. Regardless, it is clear that network analysis plays a crucial role in understanding the COVID-19 pandemic. For this reason, we hope that this review empowers researchers to utilize statistical network analysis techniques in their work.

Acknowledgements

This work was supported in part by National Science Foundation grant CCF-1934986.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*, volume 1. Oxford University Press, Oxford.

Andersson, H. (1998). Limit theorems for a random graph epidemic model. *The Annals of Applied Probability*, 8(4):1331–1349.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Springer.

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., Qin, Y., and Sussman, D. L. (2018). Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Bollobás, B. and Riordan, O. M. (2003). Mathematical results on scale-free random graphs. In Bornholdt, S. and Schuster, H. G., editors, *Handbook of Graphs and Networks: From the Genome to the Internet*, chapter 1, pages 1–34. Wiley-VCH Verlag GmbH & Co. KGaA.

Branas, C. C., Rundle, A., Pei, S., Yang, W., Carr, B. G., Sims, S., Zebrowski, A., Doorley, R., Schluger, N., Quinn, J. W., and Shaman, J. (2020). Flattening the curve before it flattens us: hospital critical care capacity limits and mortality from novel coronavirus (SARS-CoV2) cases in US counties. *medRxiv*. 2020.04.01.20049759.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Daley, D. J. and Gani, J. (2001). *Epidemic Modelling: An Introduction*, volume 1. Cambridge University Press, Cambridge.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):151–171.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6(290):290–297.

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164.

Gilbert, E. N. (1959). Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Gollini, I. (2019). *lvm4net: Latent Variable Models for Networks*. R package version 0.3.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., and Morris, M. (2008). A statnet tutorial. *Journal of Statistical Software*, 24(9):1–26.

Gross, T. and Blasius, B. (2008). Adaptive coevolutionary networks: a review. *Journal of the Royal Society Interface*, 5(20):259–271.

Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. Technical Report 39, Center for Statistics and the Social Science, University of Washington.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering of social networks. *Journal of the Royal Statistical Society, Series A*, 170(2):301–354.

Hanneke, S., Fu, W., and Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

Holland, P. W. and Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology*, 7:1–45.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008a). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(408):248–258.

Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008b). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29.

Jenness, S. M., Goodreau, S. M., and Morris, M. (2018). EpiModel: An R package for mathematical modeling of infectious disease over networks. *Journal of Statistical Software*, 84(8):1–47.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society. Section A. Mathematics.*, 115(772):700–721.

Khanin, R. and Wit, E. (2006). How scale-free are biological networks. *Journal of Computational Biology*, 13(3):810–818.

Kiss, I. Z., Green, D. M., and Kao, R. R. (2008). The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. *Journal of the Royal Society Interface*, 5(1):791–799.

Kiss, I. Z., Miller, J. C., and Simon, P. L. (2017). *Mathematics of Epidemics on Networks*. Springer.

Kolaczyk, E. D. (2017). *Statistical Analysis of Network Data*. Springer.

Krivitsky, P. N. and Handcock, M. S. (2008). Fitting position latent cluster models for social networks with latentnet. *Journal of Statistical Software*, 24(5):1–23.

Krivitsky, P. N. and Handcock, M. S. (2013). A separable model for dynamic networks. *Journal of the Royal Statistical Society, Series B*, 76(1):29–46.

Krivitsky, P. N., Handcock, M. S., and Raftery, A. E. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213.

Krivitsky, P. N. and Morris, M. (2017). Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US. *Annals of Applied Statistics*, 11(1):427–455.

Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336.

Latouche, P., Birmelé, E., and Ambroise, C. (2012). Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling, SAGE Publications*, 12(1):93–115.

Li, M. L. (2020). Overview of DELPHI model v2.0. Technical report, Massachusetts Institute of Technology.

Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society, Series B*, 79(4):1119–1141.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827.

Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6(2-3):161–180.

Molloy, M. and Reed, B. (1998). The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7(3):295–305.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20):208701.

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

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Pastor-Satorras, R., Castellano, C., Mieghem, P. V., and Vespignani, A. (2015). Epidemic processes on complex networks. *Reviews of Modern Physics*, 87(3):925–979.

Pastor-Satorras, R. and Vespignani, A. (2002a). Epidemic dynamics in finite size scale-free networks. *Physical Review E*, 65(3):035108(R).

Pastor-Satorras, R. and Vespignani, A. (2002b). Immunization of complex networks. *Physical Review E*, 65(3):036104.

Prem, K., Liu, Y., Russel, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Jit, M., and Klepac, P. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Health*, 5(5):E261–E270.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29:173–191.

Rogers, T., Clifford-Brown, W., Mills, C., and Galla, T. (2012). Stochastic oscillations of adaptive networks: application to epidemic modelling. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08018.

Sarkar, P. and Moore, A. W. (2006). Dynamic social network analysis using latent space models. In *Advances in Neural Information Processing Systems*, pages 1145–1152.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153.

Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., den Broeck, W. V., Régis, C., Lina, B., and Vanhems, P. (2011). High-resolution measurements of face-to-face contact patterns in primary school. *PLoS ONE*, 6(8):e23176.

Wang, Z., Bauch, C. T., Bhattacharyya, S., d'Onforio, A., Manfredi, P., Perc, M., Perra, N., Salathé, M., and Zhao, D. (2016). Statistical physics of vaccination. *Physical Reports*, 664(9):1–113.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika*, 61(3):401–425.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

Young, S. and Scheinerman, E. (2007). Random dot product graph models for social networks. In *Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph*, pages 138–149.

Zanghi, H., Ambroise, C., and Miele, V. (2008). Fast online graph clustering via Erdős-Rényi mixture. *Pattern Recognition*, 41:3592–3599.

Zanghi, H., Picard, F., Miele, V., and Ambroise, C. (2010). Strategies for online inference of model-based clustering in large and growing networks. *Annals of Applied Statistics*, 4(2):687–714.