CSE 8803 EPI: Data Science for Epidemiology, Fall 2024

Lecturer: B. Aditya Prakash
Scribes: Rohini Janivara*, Alejandro Danies-Lopez*

August 29, 2024
Lecture 4 : Models (II)

---

1 Summary of the lecture

The previous lecture motivated the need for epidemiological modeling in informing intervention policies by providing a framework for disease dynamic prediction in light of limited and noisy data. Model selection for epidemiological modeling, such as SIR models, is crucial for determining the accuracy of short-term forecasts and exploration of long-term behaviors through "what-if" scenarios. Basic assumptions guiding the parametrization of these models include transmission rates and contact patterns, which can vary based on demographic groups and spatial positions. Uncertainty is typically displayed using confidence intervals, and varying parameters like infection or recovery rates provide insights into different outcomes. When fitting the model to data, selecting relevant datasets is essential, as this directly impacts the model's applicability to specific populations or regions. These models help build a foundational understanding of disease spread, guiding data collection strategies while balancing the trade-offs between accuracy and flexibility. This balance is crucial for optimizing transparency, ensuring that the model is both interpretable and adaptable to new information or changing conditions. Further discussions on SIR and SIS models of disease spread delved into the mathematical derivation and possible extensions to fit the disease being modeled.

This lecture focused on two types of models: metapopulation models and network-based models. Metapopulation models build on the SIR ODE framework by incorporating spatial structure and heterogeneity across subpopulations. We explored a stochastic metapopulation model and discussed model calibration techniques, including common optimization methods used in this field.

Next, we discussed the basics of networks and network-based models. These models are more granular than SIR and metapopulation models because they incorporate structured human contact patterns. We define the structure of a network and explore properties of networks, including the friendship paradox. The simplest network-based epidemiological model is random trees, where a patient meets d others and infects them with probability q. We derive conditions under which the epidemic dies out or runs on forever.

2 Metapopulation Models

Metapopulation models discretize demographic and spatial information to model heterogeneity in disease spread without explicitly modeling the behaviour of the individuals. This reduces computational complexity and leverages sparse data to develop an intuition of population spread. For example, spatial occupation can be defined as districts, counties or states, according to the heterogeneity observed in the data. This granular heterogeneity can be informed by travel data highlighting the inflow and outflow in different regions. This discretization also depends on the time-frame of disease spread evaluation and the duration of infectiousness. Global epidemic behavior is typically governed by long-range traffic between regions more so than local traffic. The equation below shows that the expected level

of susceptible people in a region at time t ($X_i^{eff}(t)$) is composed of people present in the region at time t ($X_i(t)$), plus the summation of inflow of people($\sum_j X_j(t)\frac{\sigma_{ji}}{n_j}$), minus the sum of outflow of people($\sum_j X_i(t)\frac{\sigma_{ji}}{n_j}$) .

$$X_i^{eff}(t) = X_i(t) + [\sum_j X_j(t)\frac{\sigma_{ji}}{n_j} - \sum_j X_i(t)\frac{\sigma_{ji}}{n_j}] \tag{1}$$

Where σ_{ij} represents the flow of people from region i to j and vice versa. n_i is the population of city i which is assumed to be fixed. And $X_i(t)$, $Y_i(t)$, and $Z_i(t)$ are the number of people in Susceptible(S), Infected(I), and R(Removed) states in city i at time t. From the equation above, we can also write out similar equations for Y_i^{eff} and Z_i^{eff}.

The challenge of the metapopulation models includes discretization and time scale. In real life, the discretization level may not be clear. A model may be designed at the city or county level, while data collection might be at the state level. The time scale of disease and travel may be mismatched. For example, it is hard to take into account whether people come home at the end of timestep and to define a clear cutoff for what to consider as exposure, or duration of infectiousness. Regarding this, cross correlation with census data may be helpful.

2.1 Example: Stochastic metapopulation models

Stochastic metapopulation models enabled studies in global epidemic behavior. One of the examples is the Global epidemic and mobility model (GLEaM), which was motivated to model mobility between cities defined by airline commutes. Airline traffic data was used to derive effective passenger flow, which was then used to fit the models.

We then introduced the Susceptible-Latent-Infectious-Recovered (SLIR) model. The SLIR model contains compartmental scheme (Figure 1) and is typical for influenza-like illnesses (ILIs).

The GLEaM model found that global epidemic behaviour is governed more by long range traffic, and that neighboring regions demonstrate epidemic coupling. For example, an outbreak in Arizona may be caused by people traveling from California. It shows that spatial structure is important but neglecting local coupling if focusing on global pattern does not produce a dramatic effect.

Extending GLEaM model to study COVID-19, researchers found that in the initial stages of pandemics, the data collected are noisy and incomplete. Thus the group focused on studying imported cases and ignoring local transmission records. They discovered that at the level of countries, before the Wuhan travel ban, most cases are imported from Wuhan; Post travel ban, most cases are imported from other cities. This kind of study help understand policies like travel bans.

2.2 Calibration

In order to run simulations using such models, we first need values for our model parameters. Model calibration refers to the process of inferring parameter values from data to improve the accuracy of the model's disease dynamics and predictions. Though our models must attempt to approximate the ground truth, some degree of oversimplification is inevitable;

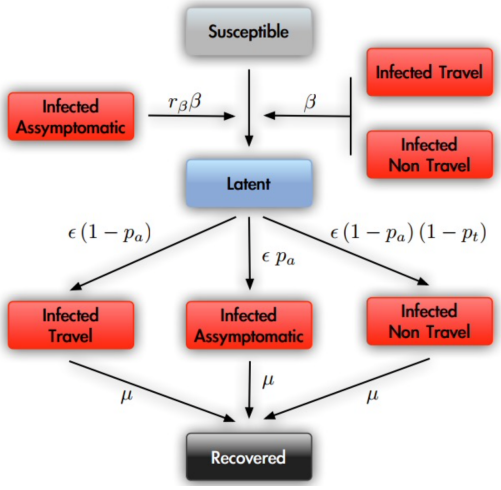


Figure 1: SLIR Model. Those in a population that are susceptible to an illness can be infected by others who are infected and asymptomatic, traveling, or not traveling. If they are infected, the infection can remain latent before it eventually spreads through the aforementioned subpopulations and the population eventually recovers.

an optimal set of parameters chosen through proper calibration will ensure that the error between our model and the ground truth is minimized.

For example, typical SIR models have two parameters: the rate of infection(β) and the rate of recovery(σ). Initial parameters include the fraction of the population that is susceptible (S_0), infected (I_0), and recovered (R_0, typically zero) as the disease spread is first noticed. To accurately model the speed of spread, surveillance data informing time-series of new cases can be used.

A useful measure to treat as ground truth is the number of deaths caused by the disease. The number of infected people isn't a very robust measure, considering the difficulty of obtaining accurate values (particularly at the start of an outbreak when tests haven't been developed and many cases aren't recognized as such), and the number of susceptible people is even more unreliable; meanwhile, deaths from a disease are much more likely to be recognized and recorded.

Given that we want to minimize the error between predicted and ground truth values, a good objective function to minimize is the squared difference between $R(t)$ and $R_{observed}(t)$. We can write the equation as follows:

$$\{\beta^*, \sigma^*\} = arg\ min(R(t) - R_{observed}(t))^2 \tag{2}$$

Here, β^* and δ^* are our optimal β and δ parameters, which minimize the objective function and capture ground truth disease dynamics with a higher predictive accuracy. This is typically done using machine learning methods such as backpropagation and gradient descent.

While performing model calibration, it is important to understand the inaccuracies in data during initial surveillance so as not to overinterpret the inferred transmission rates. These inaccuracies stem from several factors, such as lags in data collection, ascertainment biases, and inaccuracies in hospital reports. Calibration with unaccounted time lags in data

collection can lead to underestimating the radius and speed of disease spread. Ascertainment biases due to practical difficulties in estimating sources of disease spread can lead to misestimating the radius of spread as well. If counties of unreported infection are sectioned off together, infection rates might increase within those counties due to cases unaccounted for. Available data can also be restricted to particular sub-populations, limiting their generalizability. In addition, surveillance frameworks are often biased by economic and racial disparities, leading to underreporting of cases in specific neighborhoods.

Since disease dynamics are often non-linear, the effects of missing data are unbounded and can be catastrophic if ignored. For these reasons, choosing the right type of data to use for model calibration along with incorporating uncertainty in missing data can yield generalized predictions which can be fine-tuned for higher flexibility and accuracy once more data can be made available. As an example to illustrate this, the COVID-19 pandemic shows us that model robustness can be improved by using mortality data rather than infected data as the latter is often biased with underreporting.

Parameters are often motivated by biology and epidemiological data, taking on values that help our models fit the observed data. Even beyond the initial calibration, different values for parameters should almost always be tested in order to compare their results and infer realistic parametric ranges. It is also good practice to perform multiple stochastic calibrations in order to understand the level of uncertainty inherent in the model.

2.3 Optimizer

Optimizers are algorithms that aid parameter range estimation for informing models of real-world dynamics. They employ different methods to minimize the optimization function described above in improving model predictions. Since disease spread often follows non-linear dynamics, gradient descent often fails to accurately estimate parameters accurately. For this reason, several non-linear optimizers have been employed in epidemiological studies. These are:

- Nelder-Mead: A direct-search optimizer which preserves a set of test points in a simplex and searches for a new test point to replace an older test point until convergence (Gao et al. 2012);
- Levenberg Marquardt: An algorithm similar to gradient descent that solves a nonlinear least squares problem.
- Powell optimizer: a derivative-free optimization algorithm that iteratively refines a solution by performing line searches along conjugate directions, optimizing non-linear functions without requiring gradient information.
- Broyden-Fletcher-Goldfarb-Shanno algorithm:an iterative optimization method for solving unconstrained nonlinear problems, which uses both gradient information and an approximated Hessian matrix to efficiently converge towards a local minimum.

Another class of optimizers is the Bayesian optimizer which uses probabilistic frameworks to perform optimization. This type of optimizer includes:

- Markov Chain Monte Carlo (MCMC): MCMC is a class of algorithms that generate samples from a probability distribution by constructing a Markov chain, allowing for efficient exploration of complex, high-dimensional spaces in Bayesian inference.

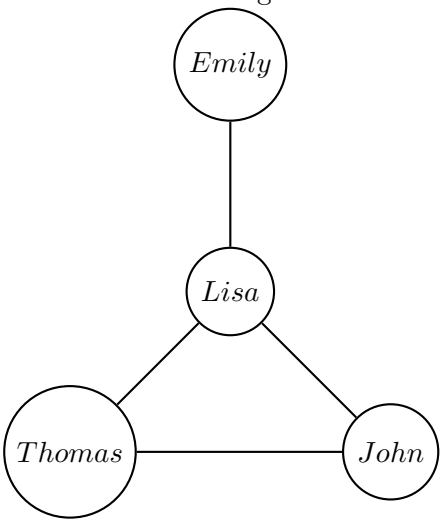
- Maximum Likelihood by Iterated Perturbed Bayes Maps: This method iteratively refines estimates of model parameters by applying perturbations to Bayesian posterior maps and maximizing the likelihood through successive approximations.
- Approximate Bayesian Computation (ABC): ABC is a computational technique for Bayesian inference that approximates the posterior distribution by simulating data and comparing it to observed data, avoiding the need for explicit likelihood calculations.
- Probe Matching: Probe matching is an optimization method that involves adjusting model parameters to match certain characteristics of the system being modeled, often used in physical simulations or biological systems.

3 Network-based Models

A limitation of the models we have discussed thus far is they assume that human contact patterns are homogeneous among a population or a sub-population. In reality, human contact patterns are very structured, and network-based models are able to incorporate these structures.

3.1 Friendship Paradox

We will first examine an interesting phenomenon that exists in networks. A recent Facebook study determined that an individual user's number of friends was less than the average friend count of their friends 93% of the time. Users had an average of 190 friends, while their friends averaged 635 friends. This phenomenon is almost always true in networks. This can be shown using a small example:



Here we will do 2 calculations: the average number of friends that each person has, and the number of friends of friends that each person has. The average number of friends per person here is:

$$\frac{1+3+2+2}{4} = \frac{8}{4} = 2 \tag{3}$$

Now we count the friends of friends. Emily has 3 friends of friends through Lisa, Lisa has

5 friends of friends through Emily, Thomas, and John, and Thomas and John each have 5 friends of friends through Lisa and each other. The average number of friends of friends here is:

$$\frac{3+5+5+5}{8} = 2.25 \tag{4}$$

Here we can see that even in this small example, the average number of friends of friends for each person is higher than each person's average number of friends. This can be further proved using the following calculations:

Assume there are N number of people in a network and each person has x_i friends, where $i = 1..N$. The average number of friends and variance is:

$$E[X] = \sum_{i=1}^{N} x_i/N \tag{5}$$

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \tag{6}$$

The average number of friends of friends is approximately:

$$\frac{E[X^2]}{E[X]} = E[X] + \frac{Var[X]}{E[X]} \geq E[X] \tag{7}$$

Therefore, we see that if there is any spread in the number of friends (i.e., $Var[X] > 0$) the average number of friends of friends is greater than the average number of friends.

The friendship paradox is an interesting phenomenon of human interaction, but it has practical implications in epidemiology. For example, you would like to immunize a subset of a population and target those with a large number of friends. Rather than randomly selecting individuals to be immunized, it is more effective to randomly select individuals and to immunize one of their friends. This strategy is called "acquaintance immunization".

3.2 Network Basics

A network is a structure of nodes with relationships connecting the nodes. These nodes (N), or vertices, are connected by edges (E), or links, to form a graph G(N,E). This graph is a mathematical representation of a real network system.

3.3 Which representation?

There are several different kinds of representation that can be chosen from depending on the network's use case. For instance, a professional network can be used to connect people who work together, or a co-author network can be used to connect authors with their respective research papers. The formulation of a real system into a mathematical graph is up to the modeler, who has the choice of how to model the system. The choice of formulation is important, and should be based around what outcomes the modeler wishes to produce.

3.4 Undirected/Directed graphs

A graph may be undirected, where the edges are symmetrical across the two nodes they are connecting. Examples of these types of connections are friendships on Facebook, collaborators, or meetings. A graph may also be directed, where the edges are directed from

one node to another. Examples of these types of connections are followers on Twitter, or a phone call.

An undirected graph may also be connected (Figure 2), where there exists a path between any two nodes. We call the largest connected component of a graph the giant component. A directed graph may have strong or weak connectivity. Weak connectivity indicates that, if edge directions are disregarded, the graph is connected (Figure 3), while strong connectivity indicates there exists a path between any two nodes (Figure 4).

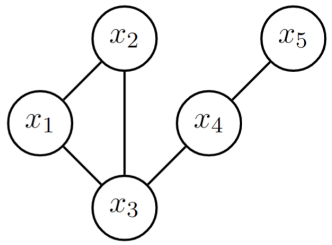


Figure 2: Undirected, Connected Graph. All edges are bidirectional, and there exists a path between any two nodes.

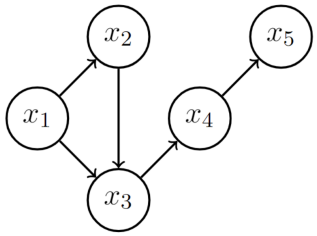


Figure 3: Directed, Weakly Connected Graph. All nodes are connected with monodirectional edges, but there doesn't exist paths between all pairs of nodes.

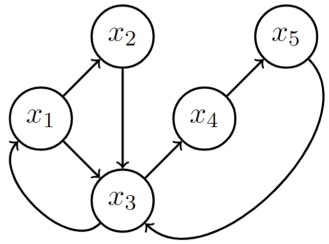


Figure 4: Directed, Strongly Connected Graph. All nodes are connected with monodirectional edges, and there exists a path to and from any node.

3.5 Classical Network Models

Network models capture simple local properties such as degree sequencing and clustering coefficients. These models provide analytical tractability in which bounds and theorems can be proven, baselines (null models) to compare against, and realistic network generation.

3.5.1 Erdos-Renyi Model

The Erdos-Renyi model, G(n,p), is a model in which each edge, $e = (u, v)$, is selected independently and with probability p.

3.5.2 Chung-Lu Model

The Chung-Lu model, G(w), is a model in which each node $v_i \in V$ has an associated weight w_i for $i = 1..n$. Each edge (v_j, v_k) is selected independently with probability proportional to $w_j * w_k$.

3.5.3 Generative/Incremental Models

Generative, or incremental, models are models in which a new node v connects to earlier nodes u with probability proportional to the degree of node u, where the degree of a node is the number of edges connected to that node.

3.6 Random Trees

The simplest type of epidemiological network-based model is an epidemic on random trees. In this model, a patient meets d other people and infects each one with probability $q > 0$. The epidemic spreads if any of the d other people are successfully infected, and they in turn meet d more people and infect them with the same probability, $q > 0$.

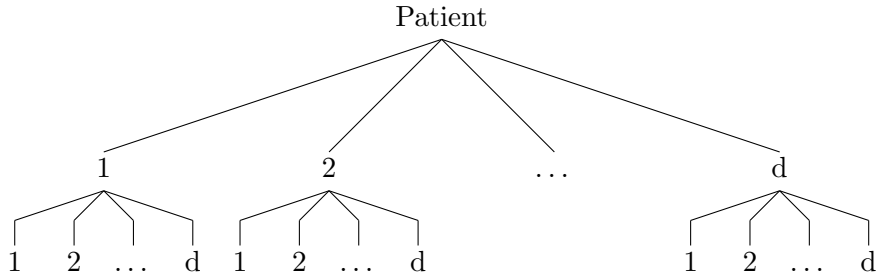


Figure 5: Each patient has a set of d people that have probability q of being infected, and each of those people have their own set of d people who they may infect with q probability.

This highly simplified model assumes that the patients will eventually run out of new people to meet and infect. It also assumes that there are no interconnected links.

Despite these limitations, this model is meaningful at the start of a pandemic, where we can assume very few people are infectious. By analyzing the model recursively, we can find values of d and q for which the epidemic will not die out, namely if:

$$\lim_{h\to\infty} P[\text{infected node at depth h}] = \lim_{h\to\infty} p_h > 0 \tag{8}$$

Where,

$$p_h = 1 - \underbrace{(1 - q * p_{h-1})^d}_{\text{prob no child at depth } h \text{ gets infected}} \tag{9}$$

To solve for this limit, we can find the fixed point p_∞ where

$$\lim_{h\to\infty} p_h = p_\infty = 1 - (1 - q * p_\infty)^d = f(p_\infty) \tag{10}$$

Some properties of this $f(x) = 1 - (1 - q * x)^d$, where $f(x)$ is the probability of an infected node at a certain depth and x the probability of an infected node at the depth immediately before it, are:

1. $f(0) = 0$. If the probability of there being an infected node at any given depth is 0, the infection can't be passed down to further depth.

2. $f(1) = 1 - (1 - q)^d < 1$. Even if the probability of there being an infected node at a certain depth is 1, indicating that there's a guarantee that someone is infected, there's no guarantee in this scenario that a node will get infected at the next depth. Thus, $f(1) < 1$

3. $f'(x) = q * d(1 - qx)^{d-1}$. This derivative indicates whether the probability of infected nodes at subsequent depths increases or decreases. Essentially, it indicates whether the spread of infection is speeding up or slowing down.

In order for the epidemic to die out, we need $f'(0) = q * d < 1$, meaning that the probability of a node being infected increases as you move down the tree. In other words, when $q * d < 1$, $\lim_{h\to\infty} p_h = 0$. Note that $q * d$, the expected number of people each patient infects, is equivalent to the reproductive number $R_0 = q * d$.

This fixed point can be understood more intuitively through the following plot:

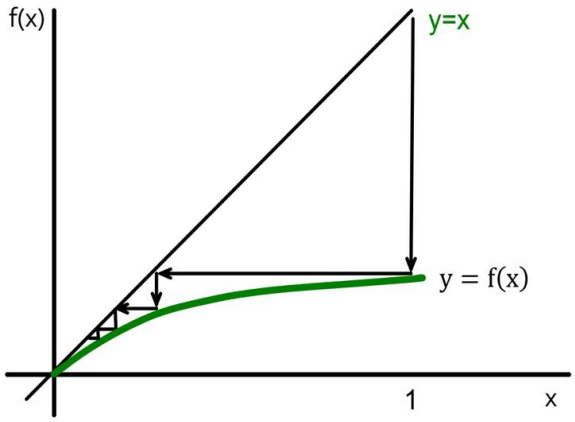


Figure 6: Fixed point of a branching tree

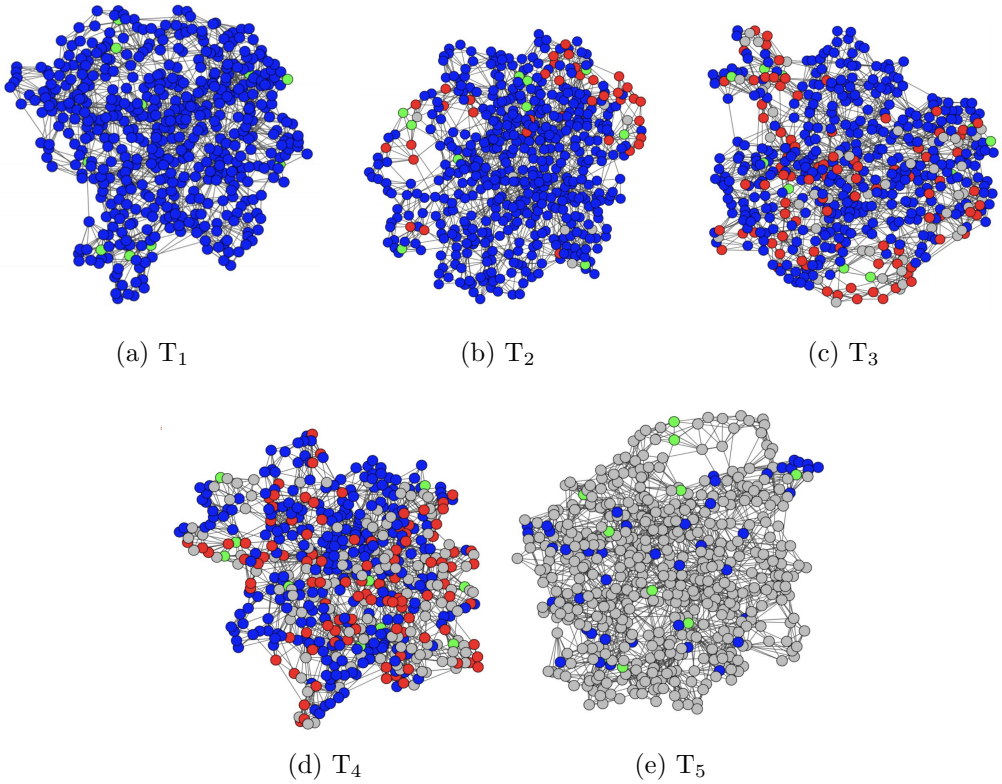
In this example, we can see that $f'(x)$ is monotonically decreasing throughout the plot. Since $q * d < 1$, the system is below the fixed point and the probability of finding an infected person decreases at each layer of the tree. If the R_0 was equal to one, then the plot would follow the $f(x) = x$ line perfectly, and the system would be in a fixed point where the probability of finding an infected person at each depth doesn't change. Finally, for an R_0 greater than one, the plot would be above the $f(x) = x$ line, and the probability of finding an infected person would increase at each depth.

3.7 SIR Network Models

We can generalize nodes to have three possible statuses: Susceptible, Infected, and Removed. Thus, as in the simple SIR model, a node may become infected by a connected infectious node with probability β and an infectious node may recover with probability δ.

Incorporating networks into SIS and SIR models can more closely model contact patterns, resulting in higher accuracy in spread dynamics and predictions. Several scenarios, depending on the type of contact and frequency of interactions, can result in differing dynamics. Networks built for the specific disease can highlight differences in the predisposition

of certain individuals to these infections. For example, a disease network developed modeling the spread of HIV would be very different from one built for modeling the spread of H1N1 on the same set of individuals. The network topology can be inferred based on the interaction type modeled. For the example of modeling the spread of COVID-19 using SIR methods, an initial fraction of the nodes are randomly assigned to one of the three states S (Susceptible), I (Infected), and R (Recovered). In a discrete-time model, the infection spread can be modeled to spread to neighbors of infected nodes at each step using computed transmission probabilities. Similarly, based on the recovery rate, individual infected nodes can recover with a certain probability at each time step.



(a) T_1 (b) T_2 (c) T_3

(d) T_4 (e) T_5

Figure 7: Example of SIR model. Blue nodes represent susceptible individuals, green nodes represent the initially infected individuals, red nodes represent subsequently infected individuals, and grey nodes represent recovered individuals. With time, blue nodes gradually diminish while red nodes multiply initially and diminish later, converting to grey nodes.

By modeling the spread of the disease through the network structure, we can anchor predictions of the speed of disease spread based on network connectivity. Sparsely connected networks have subgraphs that can be largely protected, while highly connected networks have a higher risk of an epidemic. In addition, identifying nodes that can be protected or removed can help isolate sections of individuals.