CSE 8803 EPI: Data Science for Epidemiology, Fall 2022

Lecturer: B. Aditya Prakash
Scribe: Rigved Goyal

August 31, 2023
Lecture 4 : Models (II)

1 Summary of the lecture

In the previous lecture, we learned the importance of modeling. Modeling allows us to use otherwise limited or noisy data to forecast future behavior, and as a result guides decision making and prevention/intervention for various diseases and infections. This lecture covered two kinds of models: the metapopulation models and the Network-based models. While the metapopulation models extend from the SIR ODE model that they do not assume a completely homogenous population, they also do not assume complete heterogeneity. Instead, they contain spatial structures while not including all the complications by assuming homogeneity at sub-population levels. We also discussed the example stochastic metapopulation model and ways to calibrate the models, including commonly used optimizers in this area.

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 are models which assume a combination of homogeneous and heterogeneous populations. They assume that people living in regions of certain small granularity (such as a city, zip code, county, etc.) are the same, or homogeneous, but that people across different instances of this granularity (different cities, different zip codes, etc.) may be different, or heterogeneous. This granular heterogeneity could be modeled using inflow and outflow travel data of different regions, as global epidemic behavior is typically governed by long range traffic between regions moreso than the 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($\sum_j X_j(t)\dfrac{\sigma_{ji}}{n_j}$), minus the sum of outflow of people($\sum_j X_i(t)\dfrac{\sigma_{ji}}{n_j}$) .

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

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. Model may be designed for city/county level, but while data may be from 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 type of question, 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 [1]. 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).

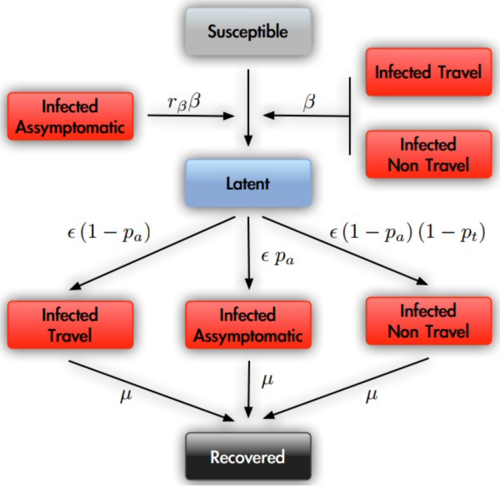


Figure 1: SLIR Model. Those in a population that are susceptible to an illness can be infected by others who are infected and asymptomatic, traveling, or not traveling. If they are infected, the infection can remain latent before it eventually spreads through the aforementioned subpopulations and the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

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

Calibration refers to the process of setting parameter values in mathematical models so the predictions derived from them are as accurate as possible. Parameters have a large influence on the accuracy of our models. If we are calibrating for a SIR model, we will calibrate for the rate of infection(β), rate of recovery(σ) as well as initial susceptible population(S_0), and the initial infected(I_0). We can write the equation as follow:

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

Typically available data for calibration will be previous infection data like time series of new cases from surveillance. These data have drawbacks including missing data, containing biases, and lags and are often high leveled data that are under-specified. In the case of COVID-19, using infected cases is unlikely to be robust due to delays and the quality of data. Instead, using mortality and hospitalization is likely to be more accurate.

Parameters are often motivated by biology and epidemiological data, they help our models fit the observed data. Our models should be able to model uncertainty in the data by going through multiple stochastic calibrations.

2.3 Optimizer

Optimizers applied in epidemiology are the same as optimizers for other subjects in science. Some most commonly used optimizers are Non-linear optimizers. Nelder-Mead is one of the most popular optimizers and is gradient-free [4]; Levenberg Marquardt solves nonlinear least squares and is very similar to gradient descent [3]. Powell optimizer performs direct-search along each direction until coverage [5]. Broyden-Fletcher-Goldfarb-Shanno algorithm also is a gradient descent optimizer and determines descent direction by conditioning gradient with curvature [2].

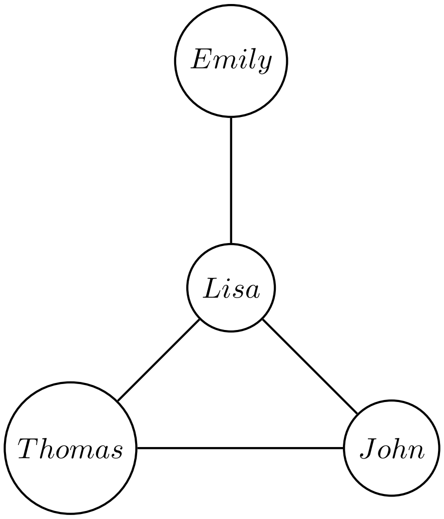
Another class of optimizers is the Bayesian optimizer which uses probability to perform optimization. This type of optimizer includes Markov chain Monte Carlo, maximum likelihood by iterated perturbed Bayes maps, approximate Bayesian computation, and probe matching. All of the above optimizers are available as python packages.

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).

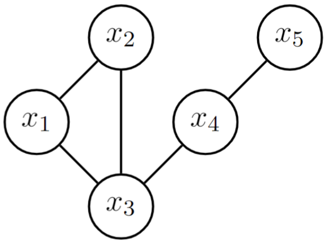


Figure 2: Undirected, Connected Graph. All edges are bidirectional, and there exists a path between any two nodes.

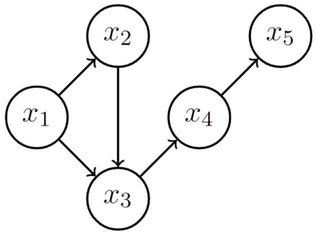


Figure 3: Directed, Weakly Connected Graph. All nodes are connected with monodirectional edges, but there doesn't exist paths between all pairs of nodes.

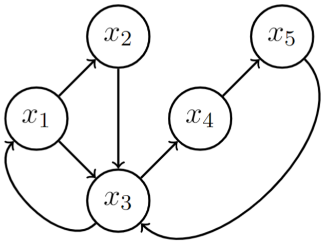


Figure 4: Directed, Strongly Connected Graph. All nodes are connected with monodirectional edges, and there exists a path to and from any node.

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, $q > 0$.

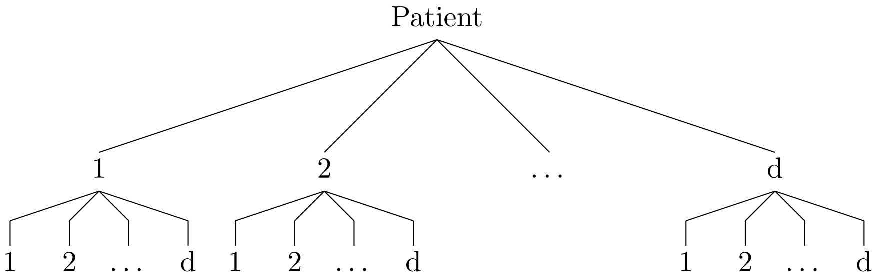


Figure 5: Each patient has a set of d people that have probability q of being infected, and each of those people have their own set of d people who they may infect with q probability.

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

This model is reasonable as the start of a pandemic where we can assume very few people are infectious. We can find values of d and q for which the epidemic will not die out. An epidemic will not die out 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 before, 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 that a node will get infected at the next depth, thus < 1
3. $f'(x) = q * d(1 - qx)^{d-1}$. This derivative indicates whether the probability of infected nodes at subsequent depths is increasing or decreasing. Essentially, this indicates whether the spread of infection is speeding up or slowing down.

Thus, $f'(0) = q * d$ so $f'(x)$ is monotone decreasing on $[0, 1]$.

If $f'(0) = q * d < 1$ the epidemic will die out, thus when $q * d < 1$, $\lim_{h\to\infty} p_h = 0$. Note that $q * d =$ the expected number of people each patient infects, which is equivalent to the reproductive number $R_0 = q * d$.

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.

At the beginning of an outbreak, the susceptible population drops as they become infected, leading to an increase in the infected population. The infected then start to either recover, or pass away, leading to the increase in the removed population while the infected population eventually decreases.

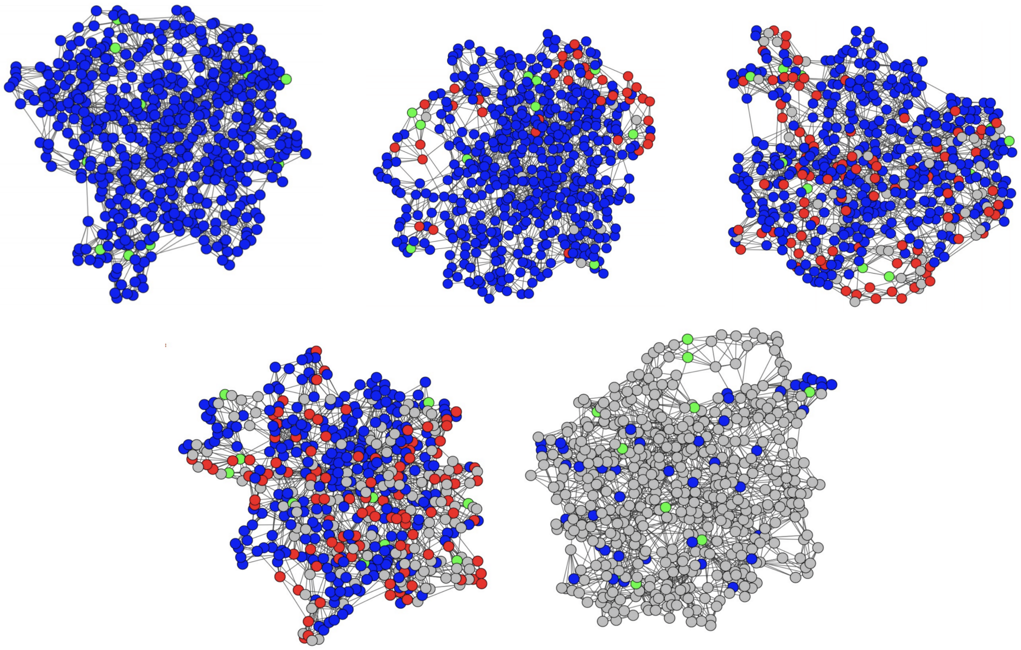


Figure 6: Example of SIR model. Blue nodes represent susceptible population, green nodes represent the initial infected population with red nodes representing subsequent infected people, and grey nodes representing those who have recovered. As time goes on blue nodes continually disappear and red nodes start to become more prominent. However, towards the end, red nodes have also disappeared in place of grey nodes that have recovered

References

[1] Duygu Balcan and Vittoria Colizza and Bruno Gonçalves and Hao Hu and José J. Ramasco and Alessandro Vespignani . Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[2] J. D. Head and M. C. Zerner. A broyden—fletcher—goldfarb—shanno optimization procedure for molecular geometries. *Chemical Physics Letters*, 122(3):264–270, 1985.

[3] J. J. Moré. The levenberg-marquardt algorithm: Implementation and theory. In G. Watson, editor, *Numerical Analysis*, volume 630 of *Lecture Notes in Mathematics*, pages 105–116. Springer Berlin Heidelberg, 1978.

[4] D. M. Olsson and L. S. Nelson. The nelder-mead simplex procedure for function minimization. *Technometrics*, 17(1):45–51, 1975.

[5] V. S. Vassiliadis and R. Conejeros. *Powell methodPowell Method*, pages 2001–2003. Springer US, Boston, MA, 2001.