CSE 8803 EPI: Data Science for Epidemiology, Fall 2020

Lecturer: B. Aditya Prakash
Scribe: Shishir Bhat

Lecture # 4
September 2, 2020

---

1 Summary of the Lecture

This lecture is an overview of two main classes of models: metapopulation models and network-based models. In both of these models we take away the assumption that the population is completely homogeneous and instead look at ways to model different levels of interaction between people/groups.

We start by describing metapopulation models that can take as parameters that different travel rates between regions. These models can effectively differentiate between homogeneous groups, so that there isn't an assumption that a population is completely mixing consistently.

We then go into talking about network-based models. These are models that can represent every person/group in a community as nodes on a graph, and then represent disease spread as probabilistic events along the connections between the nodes. We go into the basics of networks and different well-studied models of networks. In the end, we talk about random trees, and what disease spread would look like when modeled through a random tree.

2 Metapopulation Models

Recall the SIR ODE model from before. It is governed by the three equations below to demonstrate the change in the susceptible, infected, and recovered populations. In addition to the initial populations in each category, this model has two variable parameters, β and δ, which are the rate of infection and the rate of being cured.

$$\frac{dS}{dt} = -\beta SI \tag{1}$$

$$\frac{dI}{dt} = \beta SI - \delta I \tag{2}$$

$$\frac{dR}{dt} = \delta I \tag{3}$$

Although SIR models can be used to model the spread of disease, they make a large assumption that the population is completely homogeneously mixed. Metapopulation models aim to improve this by only assuming that populations are homogeneously mixed at the "right levels" (eg. counties, cities). The heterogeneity between these different regions can be modeled as travel data, by examining the inflow and outflow between different regions.

Given that σ_{ij} is the daily passenger flow between region i and region j, that n_i is the fixed population of region i, and that $X_i(t)$ is the number of people in the S state in region i, we can use an equation to model $X_i(t)$:

$$X_i^{\text{eff}}(t) = X_i(t) + \left[\sum_j X_j(t) \frac{\sigma_{ji}}{n_j} - \sum_j X_i(t) \frac{\sigma_{ij}}{n_i} \right] \tag{4}$$

In this equation, we are just modeling the susceptible population of a region as the expected level of susceptible people, plus the inflow of people from every other region, minus the outflow of people to the other regions. Similar equations can be made for $Y_i(t)$ and $Z_i(t)$, which correspond to the I and R states respectively.

To create a time update equation for this model, we can do something very similar to the SIR update step:

$$X_i(t+1) = X_i(t) + \sum_j X_i^{\text{eff}}(t) \beta \frac{I_j^{\text{eff}}(t)}{N_j} \tag{5}$$

Notice how the summation is effectively still just the βSI step of the SIR model, just split up for the different statistics of different regions.

Although a metapopulation model like this can be used for more accurate modeling, it still has downsides. It can get hard to figure out a good "discretization" level for this, especially when trying to find out how small to make the regions and how small to make the timestep. Even with small enough values, you can still run into trouble since the model doesn't account for situations like people coming back home during the timestep or for the infection lasting longer than a single timestep.

2.1 Calibrating Model Parameters

In order to get accurate models, we need to find the correct parameters. For the SIR ODE model, this would be β and δ, along with the initial S_0 and I_0. We know we have good parameters when the difference between the expected removed population and the actual removed population is minimized. More formally, this can be written as:

$$\{\beta^*, \delta^*\} = \arg\min (R(t) - R_{observed}(t))^2 \tag{6}$$

Oftentimes, we can estimate these parameters from observed data, such as new case count and observed symptomatic period. These estimates are not always perfect, since there are always problems with data collection such as lags, biases, and missing data. For more advanced data, we can look into chains of transmission such as contact tracing. Hospitalization/mortality rates are generally less noisy data, since they are very prominent and hard to miss.

In real life, there can often be many sets of parameters that allow a model to fit the observed data. When this is the case, it is important for presented models to present the ranges of uncertainty for completeness.

3 Network-based Models

By taking the idea of metapopulation models to the extreme, we come up with network-based models. Network-based models treat every person as an individual interacting with their world, since human contact patterns are not random. People generally have structured patterns of interactions with others, and we can use that information to our advantage.

3.1 Friendship Paradox

If one were to query a group of people and see how many friends each of them has, versus how many friends each of *their* friends have, a surprising result comes up. A randomly picked person is likely to have less friends on average than one of their randomly picked friends. This result is backed up empirically by a recent study on the Facebook Social Graph, that showed that the average user has 190 friends, even though their friends had an average of 635 friends [1].

This is a general principle in Network Science known as the Friendship Paradox, finding that implies that friends of randomly selected people in a network likely have a higher degree of centrality in the network than the selected person.

$$E[X] = \sum \frac{x_i}{N} \tag{7}$$

$$Var[X] = E[X^2] - E[X]^2 \tag{8}$$

$$\frac{E[X^2]}{E[X]} = E[X] + \frac{Var[X]}{E[X]} \tag{9}$$

As a proof of this, let us consider a statistical approach. We can treat the average number of friends for person as a random variable X. In doing so, the expected value of X represents the average number of friends for a randomly picked person as seen in Equation 7. By the definition the variance of the random variable in Equation 8, after Rearranging we can get the equality in Equation 9. Since $\frac{E[X^2]}{E[X]}$, the ratio of expected value of X^2 and expected value of X represents the average number of friends of friends for a random person. In Equation 9 it has been show that $\frac{E[X^2]}{E[X]}$ is greater than $E[X]$ if the $Var[x]$ is not zero, in other words, the average number of "friends of friends" is always going to be greater than the average number of friends if the variance among people for number of friends is not zero.

In public health, this effect can be used to pick who to immunize. By randomly picking people and then immunizing a random friend of each of those people, the vaccine distribution is likely to be more effective, since the vaccine is likely to be going out to people who interact with more people, compared to a randomly selected group of people.

This effect can also be used for tracking the spread of a disease. By again choosing a random set of "friends of friends" and tracking them, we can likely get an earlier idea of disease spread in a community, since these people likely have interactions with more people and would get the disease sooner.

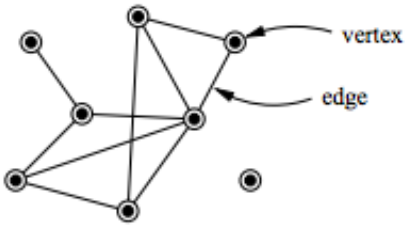

Figure 1: a sample network [2]

3.2 Network Basics

A network is a series of vertexes/nodes N and corresponding edges between the nodes E described as a graph $G(N, E)$ that represents a real system. As seen in Figure 1, all of the nodes in a network do not necessarily share an edge and there is not necessarily a path between any two nodes.

By modeling every person as a node and every potential contact as an edge, we can create a rich model of how disease spread is happening, instead of just a time series. By choosing how we define "potential contact", we can change our model depending on the type of disease and the environment we're trying to model (eg. airborne disease, spread through contact, spread through water infection).

3.2.1 Network properties

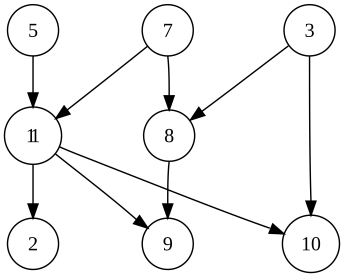

Figure 2: a directed graph [3]

Edges on a network can also be directed, to only point one-way, as seen in Figure 2. A network with directed edges is normally used to model situations where there is asymmetrical contact between two people, such as mapping out follows on Twitter. In contrast, an undirected graph may be more useful in mapping out situations where mutual contact has to happen, like becoming friends on Facebook.

A network is said to be *connected* if there is exists a path between any two nodes in the graph, as seen in Figure 3. In directed graphs, the graph is said to be *strongly connected* if a path still exists between every two nodes, and *weakly connected* if the graph only becomes connected when treating every edge as undirected.

Oftentimes, the largest connected component of a graph in a network is called the GIANT component.

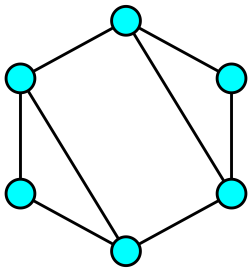

Figure 3: a connected graph [4]

3.3 Different Models of Networks

There are different ways we can try to create our network model:

- Erdős–Rényi (ER) model [5]: Each edge between two nodes is selected independently with probability p. This is a very well studied modeled.
- Chung-Lu model [6]: Given a weight w_n for every node, select an edge between nodes i and j with probability $w_i * w_j$.
- Generative/incremental models are created incrementally with every new node more likely to attach to high degree nodes. This is known as a "rich get richer" model, and these models follow a power law around the number of edges each node has [7].

By trying to model disease spread with different types of networks, we can analyze what the best/worst case bounds on the disease spread can be, along with getting a baseline of what we can probably expect from the disease.

3.4 Network-based Disease Models

3.4.1 Branching Processes

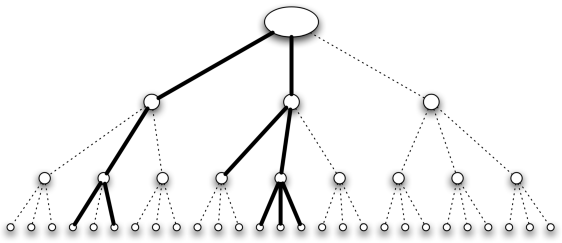

Figure 4: a tree where $d = 3$ [8]

If we assume that every infected patient meets d others and infects each of them with probability q, we can model the disease spread as a random tree, with the root as the index case. In Figure 4, we can see a such tree where $d = 3$ and q is relatively high. In this case, the disease will continue to spread due to the high probability of infection.

Although this is not a realistic model, since populations are not infinite, it is realistic at smaller scales, especially where the disease is first introduced to a population and most people are not infected.

The reproductive number R_0 for this case would be $q \cdot d$, meaning that $q \cdot d$ can be seen as the average number of people someone infects [8]. Only when $R_0 > 1$ is the disease able to successfully spread in a new population. And the following shows the proof.

First, let p_h represents the probability that there is an infected node at depth h, then we know that $1 - q \cdot p_{h-1}$ represents the probability that the parent node in depth h-1 doesn't infect one child that it meets. Since every patient meet d others, so the parent node will meet d children, and the probability that the parent node in depthh-1 doesn't infect any of d child it meets is $(1 - q \cdot p_{h-1})^d$. And we can get the probability of a parent node in depth $h - 1$ will affect at least a child in depth h, which is the equation 10.

$$p_h = 1 - (1 - q \cdot p_{h-1})^d \tag{10}$$

$$1 - (1 - q \cdot \lim_{h\to\infty} p_h)^d = \lim_{h\to\infty} p_h \tag{11}$$

$$f(x) = 1 - (1 - q \cdot x)^d \tag{12}$$

$$f'(x) = q \cdot d(1 - q \cdot x)^{d-1} \tag{13}$$

Now, if the epidemic doesn't die out, which means the probability $p_h > 0$ for big enough h, that is $\lim_{h\to\infty} p_h$ exists, also notice that if $\lim_{h\to\infty} p_h$ exists, then $\lim_{h\to\infty} p_{h+1} = \lim_{h\to\infty} p_h$. After swap the position of limit, we can get Equation 11. To guarantee Equation 11 has a solution, we only need to make sure Equation 12 has a fixed point solution in interval $(0, 1]$. Let $g(x) = x$, Since $f(0) = 0 = g(0)$, and $f(1) = 1-(1-q)^d < 1 = g(1)$, then we consider derivative of $f(x)$, that is Equation 13, and $g'(x) = 1$. If $qd \leq 1$, then $f'(x) < g'(x)$ for all the interval (0,1], thus no fixed point solution; if $qd > 1$, then $f'(x) > g'(x)$ at the begin of interval (0,1] thus there must be a fixed point solution.

3.4.2 Other Models

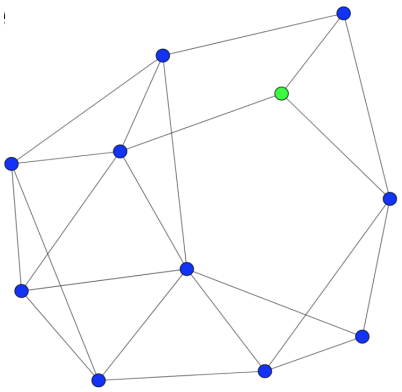

Figure 5: initial configuration of an SIR network model [8]

We can use network-based models to make the same generalizations as the SIR models, to allow every node in the graph to be in one of three states (S, I, and R). These network-based

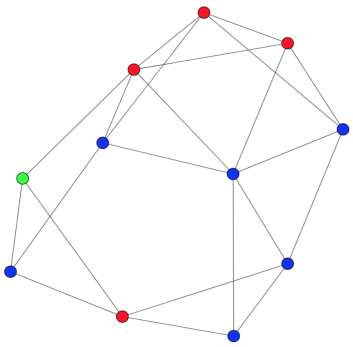

Figure 6: the same model after 200 iterations [8]

models are stochastic, meaning that every iteration a neighboring node can get infected with probability β and infected nodes can be cured with probability δ.

As an example, take Figure 5. In this model, the blue nodes are susceptible, the green nodes are infected, and the red nodes are removed. After 200 iterations of the model, we end up with the state in Figure 6. Notice how this network-based model can account for individual contact between people, and doesn't assume that the remaining infected person in the population has contact with every other individual.

Although a network-based SIR model can be more accurate, since it can account for individual contact patterns, you have to keep track of much more information than the ODE models, since you also need to have some model of the network itself alongside the β and δ parameters. One observation to be made is that network-based SIR models perform roughly the same as the ODE SIR models when there is an edge between every two nodes, but the proof of this is left as an exercise to the reader.

For an example SIR network to mess around with and test the effects of different parameters, here is web application that can simulate runs of the SIR model on a borough map of London: http://epirecip.es/epicookbook/chapters/london-boroughs-network/r

References

[1] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint arXiv:1111.4503*, 2011.

[2] Wikipedia, the free encyclopedia, "Network theory," 2020. [Online; accessed September 12, 2020].

[3] Wikipedia, the free encyclopedia, "Directed graph," 2020. [Online; accessed September 12, 2020].

[4] Wikipedia, the free encyclopedia, "k-edge-connected graph," 2020. [Online; accessed September 12, 2020].

[5] P. Erdos and A. Renyi, "On random graphs i," *Publ. math. debrecen*, vol. 6, no. 290-297, p. 18, 1959.

[6] W. Aiello, F. Chung, and L. Lu, "A random graph model for power law graphs," *Experimental Mathematics*, vol. 10, no. 1, pp. 53–66, 2001.

[7] A.-L. Barabási, E. Ravasz, and T. Vicsek, "Deterministic scale-free networks," *Physica A: Statistical Mechanics and its Applications*, vol. 299, no. 3-4, pp. 559–564, 2001.

[8] D. Easley and J. Kleinberg, "Networks, crowds, and markets: Reasoning about a highly connected world," *Cambridge Univ. Press*, 2012.