pdflatex main bibtex main pdflatex main pdflatex main

CSE 8803 EPI: Data Science for Epidemiology, Fall 2020

Lecturer: B. Aditya Prakash
Scribe: Clayton Horsfall

Lecture #7
September 16, 2020

1 Summary of Lecture Content

In this lecture we covered a general overview of the creation of Network Models. Where we had previously discussed common Epidemiological models (SIR, SIS, SEIR, etc.), this lecture focused on how to hypothesize node structure, edge creation, and the problems with clarifying models.

We reviewed common classes of networks, and how the networks change based on the composition of the nodes and the definitions of the contacts themselves. We reviewed the meaning of the Distribution of Contacts, which makes an effort to model large groups with stochastic behaviors and contacts. Lastly, we touched on First Principles for constructing social contact networks for disease spread.

2 How to construct a Network Model

The over-arching theme of general network construction is the modeling of contact patterns. Up until this lecture, we have viewed basic person-to-person networks, where the nodes of the network represent individuals and the edges represent their contacts with other individuals.

2.1 Common Classes of Networks for Disease Spread

In previous disease-spread models we have examined (SIR, SIS, SEIR, etc.), there is an underlying network model at play. We have seen how at given time intervals an infected node attacks another node via an edge, and infects that node with probability β. Subsequently the infected node recovering with probability δ. To visualize this, we need a network. In Figure 1 [1] we see four common classes of networks to model disease spread.

2.2 Defining the Nodes

As shown in Figure 1, we are provided four ways in which disease can spread on a small visual scale. Network A, the Person-to-Person network, is a basic contact work where edges are undirected and contact is reciprocal. In B, a Bipartite network connects nodes to location nodes vie undirected edges, where we can think of disease spread occurring when individuals make contact with door handles or communal tables. In C, a Semi-directed model shows a contact network in which some of the edges are directed and some are not, where we can consider a hospital setting in which patients can pass a disease to their caregiver, who then passes that disease to another patient. Model D shows a weighted undirected network, where the edges between some nodes are perhaps more frequented than others (consider the frequency of air traffic between Chicago and Minneapolis, versus Chicago to Cheyenne)., a node can take on the role of a number of "touch-points" within

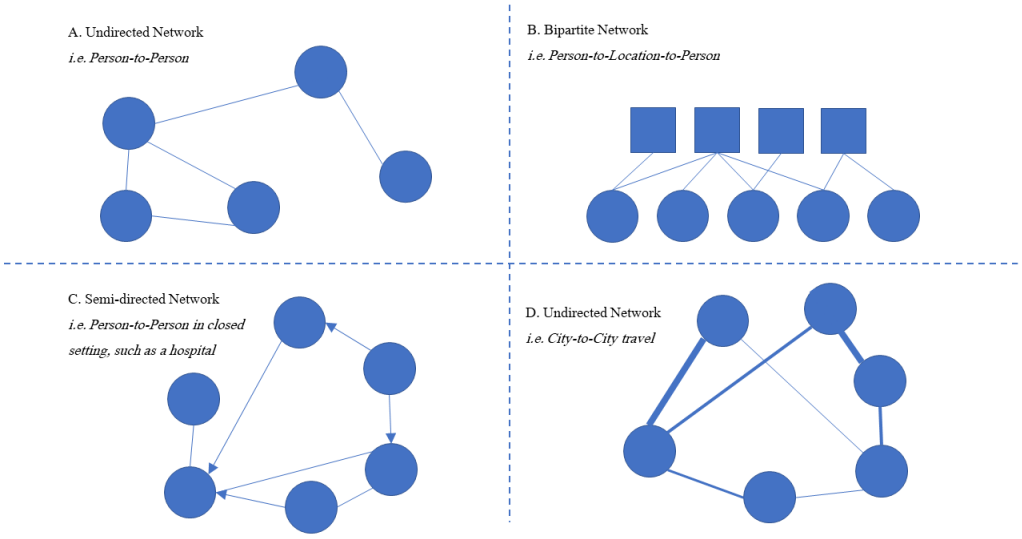

Figure 1: Common Classes of Networks used to model disease spread

the model. A node can represent an individual, a ZIP code, a hospital room, a city, or a table in a common room. Critically, defining the node in a network is dependent on what is being modeled and what is spreading disease (examples below). This, of course, highlights its own set of additional complexities in the network itself. More on that later.

2.3 Defining the Edges

In a network model, the nodes are the "touch-points" whatever is being modeled. Similarly, we have to define our edges as per the goal of the model as well. And, similarly, there are complexities within edge definition as well. The purpose of defining edges is to create a distinction of what the **Epidemiological Contact** is. In defining *"How does the disease in question pass from node to node?"*, consider the examples below.

- For **COVID-19** or **Influenza**, disease is spread via close proximity over time with some probability.
- For **HIV**, the virus is spread via sexual contact, needle-sharing, and contaminated medical equipment.

Once the nodes and edges of the model have been identified within the context of the problem to be solved, the next critical component is defining the distribution of the contacts and relationships themselves.

3 Distribution of Contacts

Once we've defined our edges and we've defined our nodes, we should have a general idea of what the social contact network might look like. The next critical component is finding the distribution of the contacts: we need data about the contacts, to determine what each contact means. The inherent difficulty with constructing accurate and useful network

models is applying it to scale; to be sure, finding every individuals' exact contacts within a network is more reasonable on a small scale, but it becomes impractical very quickly to expect exact results for large networks. This is why we need a distribution to reasonably estimate those contacts. Even given advances in modern technology, including GPS, passive location tracking, an wireless network hosting, distributions are necessary to model the observed behavior.

3.1 Surveys

One of the most common ways to get a distribution is through a survey. At a small scale, surveys are excellent at providing exact (or near-exact) depictions of networks. Creating a network of contacts in a small high-school, for example, can be done with a survey [2], resulting in an accurate model of who is connected with whom. However, scaling such a method to, say, an entire city, becomes problematic. The solution becomes surveying groups of people, and generating a distribution of their typical movements and contacts throughout the network.

The distribution of contacts describes a *mean* behavior of a node. This makes intuitive sense when the node is defined as a group of "similar" individuals. Take, for example, a subset of young adults in a city. Their mean behavior may involve a typical day of starting at home, going to work, heading to the gym, and then returning home in the evening. While deviations from this pattern surely exist, remember that is this the mean behavior.

3.2 Mobility

With a distribution, we can develop a model of **mobility**, from which we can *infer* a contact network (in essence, the contact network becomes the outcome of how we model people moving). This method captures the periodicity of the "mean" behavior mentioned above. Acquiring mobility presents its own unique set of decisions to be made, once again based on the context of the model itself.

Mobility data is typically split into two subsets. **Survey** data, which we've already discussed, can be simple activity surveys, mobility statistics, and the Census survey. The other subset, **Tracing** data, can be active or passive tracking of node location by means of GPS location, wireless network usage, Bluetooth connections, social application, and transportation check-ins. Common types of mobility data are shown in Figure 2, with their own unique pros and cons [3].

Based on the context of the model, the desired output of the model, and the observed movement of a population, learning how people come into contact with each other is critical to constructing a contact network.

3.3 Considerations at Scale for Mobility Data

The source of the mobility data, be it from a small survey to large-scale GPS tracking, presents plenty of caveats to the mode that need to be addressed. Coverage across **socioeconomic backgrounds** can be limited if the source of data is expensive technology like a smartphone or wearable, which inherently provides a barrier to entry of anyone without the means to purchase and use one. **Privacy** concerns present themselves when the data required, such as demographics, is unavailable due to privacy and anonymity laws. **Security** concerns present themselves when the integrity of the data-set required may have been compromised, either intentionally and maliciously, or accidentally. Lastly, the **Speed of**

Comparison of Mobility Data Methods

Method	Advantages	Disadvantages
Surveys, Direct Search	At small scale, can be highly accurate to exact; can serve multiple purposes; can capture multiple correlations	Can be expensive to collect observation; can be time-consuming at large-scale
WiFi Localization	Accurate; Half the energy expenditure of GPS	Must provide access point, which can be expensive
GPS Localization	Precise to within 5m; can distinguish between transportation nodes	High energy usage; expensive; sampling bias (e.g. cost barrier to entry); less effective indoors
Passive Cellular Network Localization (Call Records)	Automatically Generated	Large time gaps; low special resolution (175m+); requires filtering; sampling bias; proprietary
Active Cellular Network Localization	More accurate than Passive Cellular Localization; less expensive than other methods	More expensive than Passive Cellular Localization; sampling bias; proprietary

Figure 2: Mobility Data Methods Pros and Cons

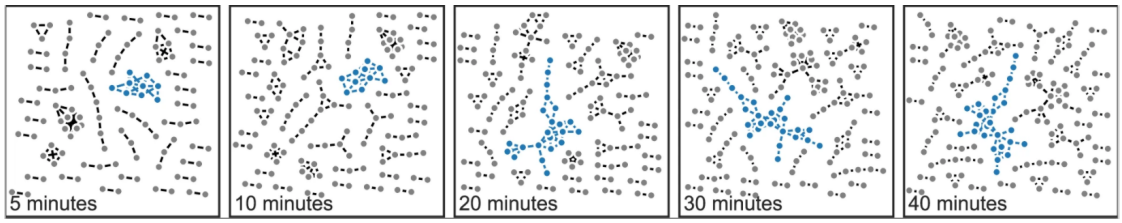

Figure 3: Aggregation Window Snapshots

Updating is a time-based caveat of the data, where the most recent mobility data available doesn't accurately describe the current state of mobility.

As an example of the Speed of Updating caveat, consider COVID-19, where communities have gone from "normal" mobility, to full quarantine, to partial quarantine, potentially back to full-quarantine, etc. Modeling mobility during COVID-19 at any one moment would require very recent mobility data, where data that isn't updated frequently quickly becomes obsolete. Further, the granularity of the observed time-frames can significantly change the model. Mobility between nodes can change depending on the frequency in which snapshots in time are taken. In a study [4] of university students in Copenhagen, physical proximity data was aggregated via Bluetooth at five-minute intervals. If the aggregation window becomes too large, then important contact network structures become lost (the largest connected component, shown in Figure 3, becomes larger as the aggregation window increases).

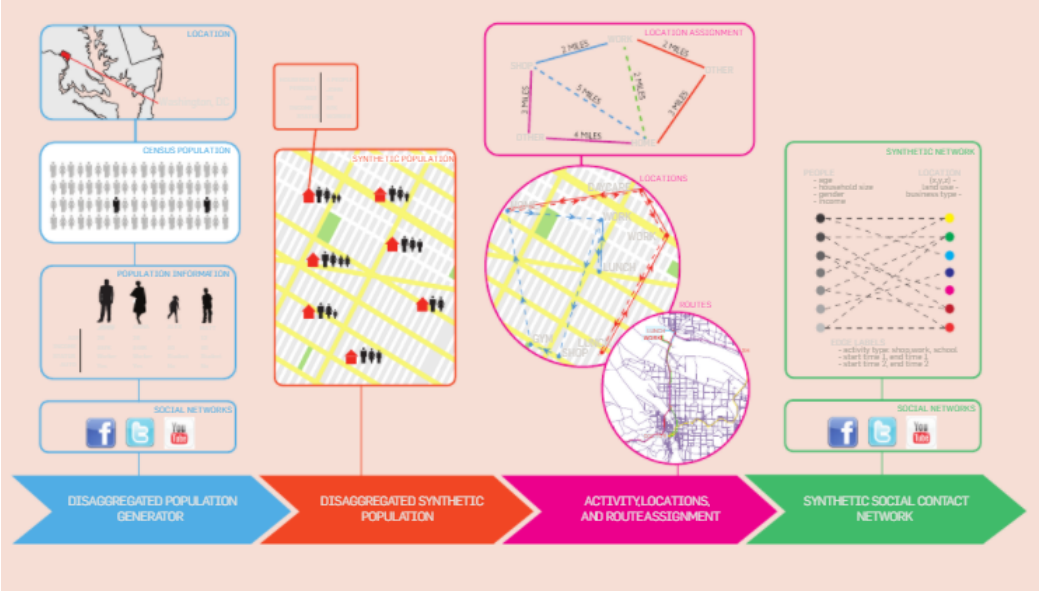

Figure 4: Creating a Synthetic Social Contact Network

4 First Principles Approach for Constructing Social Contact Networks

When it comes to scoping a social contact network with mobility data, defining certain aspects of the participants and their movements is at the core of the First Principles approach for constructing social contact networks, with the objective being to find:

- The **who**: the demographics of the individuals
- The **what**: the sequence of the individuals' activities
- The **when**: the times of those activities
- The **where**: the locations of those activities
- The **why**: the reasons for those activities

It is important to cover these aspects, because the behaviors of individuals change in response to both disease spread and intervention responses. First Principles approach will try to build a network by collecting data that addresses the above objectives, and modeling the contact distribution accordingly. Unfortunately, for the most part, datasets containing all of this information comprehensively do not exist. Rather, synthesizing data across multiple sources and platforms is required to build the First Principles. Once aggregated, the model can then begin to observe patterns in movement and mean behaviors, thus creating a "synthetic" social contact network.

In Figure 4 [5], we have an example road-map of the synthesizing of multiple data streams into a social contact network. The 'who' data comes from social media and census surveys, the 'where', 'what', and 'when' from various location and movement tracking sources, and the 'why' is a derivative of the 'where'. Together, the various sources form the synthetic social contact network.

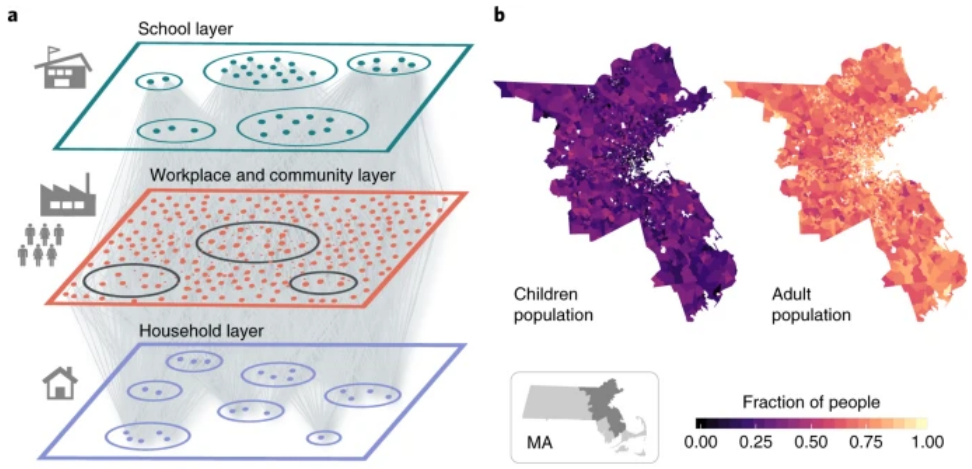

Figure 5: Mobility Layers in Boston

As previously mentioned, modeling the movement of people after an intervention is also required, because behavior and mobility change as a response to disease spread. If we consider mobility in terms of layers, as this recent mobility study in the Boston area shows in Figure 5 [6], we can visualize how different groups of people (adults versus children, for example) move throughout their day, from home to work, or home to school, or work to recreation, etc.

In an intervention, like what we've seen in COVID-19 precautions, most schools have closed to in-person learning, so the entire school layer of a child's movement data may have been removed from the overall model, and our social contact network must reflect this change in order for it to be an accurate reflection of mobility.

References

[1] N. B. Dimitrov and L. A. Meyers, "Mathematical approaches to infectious disease prediction and control," *Informs Tutorials in Operations Research*.

[2] P. S. Bearman, J. Moody, and K. Stoval, "Chains of affection: The structure of adolescent romantic and sexual networks," *AJS*, vol. 110, no. 1, pp. 44–91, 2004.

[3] M. Marathe and A. K. Vullikanti, "Tutorial on computational epidemiology," 08 2014. ACM Knowledge Discovery and Data Mining Conference.

[4] P. Sapiezynski, A. Stopczynski, D. D. Lassen, and S. Lehmann, "Interaction data from the copenhagen networks study," *Scientific Data*, vol. 6, no. 1, pp. 97–111, 2019.

[5] M. Marathe and A. K. Vullikanti, "Computational epidemiology," *Communications of the ACM*, vol. 56, pp. 88–96, 07 2013.

[6] A. Aleta, D. Martín-Corral, and A. P. y Piontti et al, "Modelling the impact of testing, contact tracing and household quarantine on second waves of covid-19," *Nature Human Behaviour*, vol. 4, pp. 88–96, 09 2020.