Epidemics Graph Neural Network Node Classification Project Definition

Jaykumar Patel
patel.jay4802@utexas.edu

Afnan Mir
afnanmir@utexas.edu

Abstract

The COVID-19 pandemic has shown that contact tracing is a key way to mitigate the spread of the disease. However, manual contact tracing is slow and can be inaccurate. Thus, this project aims to automate contact tracing by utilizing Graph Neural Networks (GNNs). This project will analyze the dataset to understand its properties, run simulations to see how the virus spreads through the network, and train GNNs to perform link prediction.

1. Introduction and Motivation

When COVID-19 first appeared, one of the key ways that its spread was mitigated was through contact tracing. Contact tracing is the process of tracking how the virus spread by identifying people who may have come in contact with an infected person.

However, the pandemic revealed that the COVID-19 disease can spread faster than manual contact tracing. Thus, the objective of this project is to automate contact tracing by incorporating machine learning using Graph Neural Networks (GNNs). This automation will allow quicker contact tracing, and possibly lead to a greater mitigation of the spread of COVID-19 when compared to manual contact tracing.

2. Previous Work

Methods pertaining to predicting the spread of COVID-19 include mathematical models, traditional ML models, and graph-based ML models.

One example of a mathematical model is the SEIRD model, which attempts to predict the change in Susceptible, Exposed, Infected, Recovered, and Deceased people over time through the use of differential equations. This model is used to simulate the spread of the virus over time [2].

SIR is simpler version of the SIERD model, which attempts to perform the same task [4].

Traditional ML models have also been used to predict the spread of COVID-19. For example, Long Short-Term Memory (LSTM) models have been used to predict the number of cases over time [2].

Another approach utilizes a hybrid of SIRD and LSTM to account for time dependent parameters of the SIRD model [1].

Furthermore, graph-based ML models, such as GNNs, have been used on mobility data to predict the number of cases and hospitalizations [3].

GNNs can also be used for link prediction [5], which is useful for contact tracing cite contact-tracing-GNN.

3. Approach

We used the "foursquare" dataset to build a contact tracing graph of Austin, Texas. Each entry in this dataset contains data such as a unique device ID, which represents a person, a location ID, which represents a location the person was at, UTC date and hour, which tell us when each person visited the given location, and finally, a dwell time, which tells us for how long this person was at the given location. Given this data, we could generate a contact tracing graph for people in the City of Austin.

We had data from the month of July, 2020, but we used only the first day of July ($\sim$19,000 entries) to create our sample network. Our nodes were all the unique device IDs that were present in the dataset. To create our edges, we used the following logic: We first did not consider devices whose dwell time was less than 60 minutes, as we our assumption was that if a person was at a location for less than an hour, they did not have enough time to make meaningful contact with anyone. Then, we used the UTC date and hour with the dwell time to determine an arrival and departure time for each entry. We then compared every entry with every other entry and checked if there was any overlap between their arrival and departure times. If there was, we considered this as a contact between the two people. We then added an edge between the two people in our graph.

In addition, we created a sample Susceptible-Infected simulation using the dataset. We did not include a re-

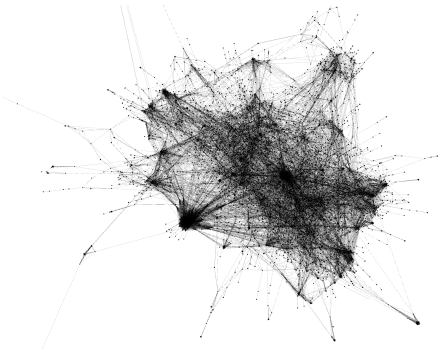

Figure 1. Network after one day

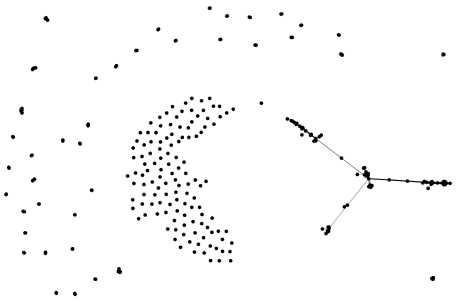

Figure 2. Infected graph after one day

covery/death or incubation period in this simulation model mainly because we only performed the simulation over one day, so there would be no time for either of those. To perform the simulation, we did the following: We first randomly selected 20% of the nodes in the graph to be infected. Next, with each time step (1 hour), we looked at all the neighbors of each infected node and infected them with a probability of 1.0, so we assumed any contact led to an infection. We did this for a 24 hour time range.

4. Experimental Setup and Results

In Figure 1, we can see the contact tracing network after one day. Network properties for this network were also calculated. The average node degree was 10.41, the network diameter was 14, the average clustering coefficient was 0.687, and the average path length was 4.784. In addition to this, the degree distribution was that of a power law, which tells us that this is a scale-free network. This can also be seen from the graph itself, as we can see the presence of a few hubs in the network. This makes sense, as we should expect a social network to be scale-free. In addition, we can see the resulting infected graph from our simulation in Figure 2. In this network, it is important to note that all of the visible nodes are infected, and the edges represent how the infection has been spread. We can see that

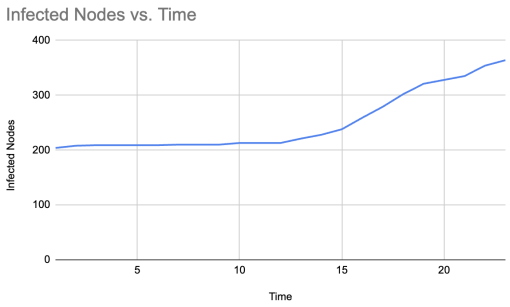

Figure 3. Number of infected nodes over time

not a lot of infected nodes have spread the disease, but the ones that have created one connected component that show the beginnings of a scale-free network. In 3, we can see the number of infected nodes over time. We can see that the infection doesn't really start to spread until the latter half of the day, and then looks as though the rate of infection is about to increase exponentially before the day ends. This lines up with our intuition, as we would expect the infection to spread faster as more people get infected.

5. Conclusion and Short-Term Plans

Through the analysis of the network, we were able to determine the network of contacts is a scale-free network. There were some people who came into contact with many other people where as others stayed within their cliques.

The simulation showed that the virus initially spread slowly, but the rate of infection increased rapidly. As people with many connections became infected, the virus was able to spread significantly faster.

For M2, we plan to do a more complete analysis of the network by considering both July and August data. We also plan to run more simulations. Specifically, we want to make them more realistic by accounting for the incubation period, deaths from the virus, and the fact that people may not be able to get infected again after they have recovered. We also plan to implement GNNs at a small scale to perform link prediction. Specifically, we will use GNNs to predict the people that come into contact with previously infected people. By predicting contact between people, we will be able to predict the spread of the virus.

References

[1] A. Bousquet, W. H. Conrad, S. O. Sadat, N. Vardanyan, and Y. Hong. Deep learning forecasting using time-varying parameters of the sird model for covid-19, February 22 2022.

[2] T. Geroski, A. Blagojevic, D. M. Cvetković, A. M. Cvetković, I. Lorencin, S. B. Šegota, D. Milovanovic, D. Baskic, Z. Car, and N. Filipovic. Epidemiological predictive modeling of covid-19 infection: Development, testing, and implementation on the population of the benelux union, October 28 2021.
[3] K. Skianis, G. Nikolentzos, B. Gallix, R. Thiebaut, and G. Exarchakis. Predicting covid-19 positivity and hospitalization with multi-scale graph neural networks, March 31 2023.
[4] R. S. Yadav. Mathematical modeling and simulation of sir model for covid-2019 epidemic outbreak: A case study of india, May 21 2020.
[5] M. Zhang and Y. Chen. Link prediction based on graph neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.