Epidemics Graph Neural Network Node Classification and Link Prediction

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). In our preliminary work on network analysis, we found that the contact network is a combination of an exponential and scale-free network. Also, our simulation showed that during the first half of the day, the infection does not spread much, but with some time, it starts spreading steadily.

1. Introduction and Motivation

When COVID-19 first appeared, manual tracing was deployed to mitigate the initial outbreak. Contact tracing is the process of tracking how the virus spreads by identifying people who may have come in contact with an infected person, and then asking them to isolate and get tested.

However, the pandemic revealed that the COVID-19 disease can spread faster than manual contact tracing [2]. Thus, the objective of this project is to automate contact tracing by incorporating machine learning using GNNs to hopefully increase the mitigation of the spread of COVID-19 when compared to manual contact tracing.

2. Previous Work

Methods for 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 use of differential equations. This model is used to simulate the spread of the virus over time [3]. SIR is simpler version of the SIERD model that attempts to perform the same task [7].

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 [3]. 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 [5]. GNNs can also be used for link prediction, which is useful for contact tracing [6].

3. Approach

We used the foursquare dataset to build a contact graph of Austin, TX [4]. Each entry contains a device ID, a location ID, UTC date and hour, and a dwell time, which tell us when and how long a person visited a location. Given this data, we generated a contact graph of Austin.

We used data only from July 1st, 2020 to generate a sample network. Our nodes were all the unique device IDs in the dataset, which correlate to people. For our edges, we used the following logic: we ignored entries with a dwell time less than 60 minutes, as we assumed this not enough time to make significant contact with others. Then, we used the UTC date and hour with the dwell time to determine the arrival and departure time interval for each entry. We then compared every entry with every other entry. If the entries' locations were the same and if intervals overlapped by at least 60 minutes, we considered this as a contact between the two people and added an edge between them.

Separately, we also created a sample Susceptible-Infected simulation using the data from July 1st, 2020. We did not include a recovery/death or incubation period in this model mainly because we only performed the simulation over one day. To perform the simulation, we did the following: at hour 0, we randomly selected 20% of the nodes in the graph to be infected. Then for the first hour, we looked at all contacts between people, and if an infected person came into contact with a susceptible person, we infected the susceptible person with a probability of 1.0. We iterated this process for 24 hours.

Figure 1: Contact network after one day

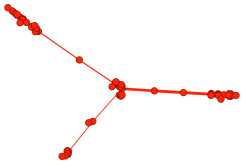

Figure 2: Infected graph after one day

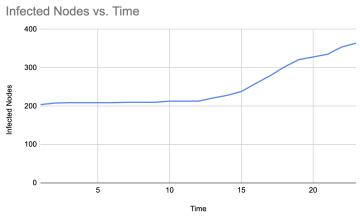

Figure 3: Number of infected nodes over 24 hours

4. Experimental Setup and Results

In Figure 1, we can see the contact network after one day. Network properties for this network were also calculated. The average node degree is 10.41, the network diameter is 14, the average clustering coefficient is 0.687, and the average path length is 4.784. In addition to this, the degree distribution was mainly an exponential distribution with subtle hints at a power-law. This can also be seen from the network 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, but we do not have all the data points, so it is not fully scale-free on the sample network.

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 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 Figure 3, we can see the number of infected nodes over time. We can see that the infection doesn't spread much until the latter half of the day, after which it spreads steadily. We expect the rate of spread to be exponential when we simulate the spread over more days. 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 combination of an exponential and scale-free network. Some people came into contact with many other people whereas others stayed within their cliques. The simulation showed that the number of infected people initially stayed relatively constant, but after 12 hours, infections began increasing steadily.

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 realistic simulations by accounting for the incubation period and deceased/recovered people. 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, allowing us to predict the spread of the virus.

For this milestone, Afnan spent most of the time performing the network generation and analysis for the one day sample network, and Jay focused on generating the simulation of the spread of the infection.

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] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, and et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in european countries: technical description update, 2020.

[3] 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.

[4] C. D. Lab. Foursquare Community Mobility Data with Basemap (US), 2020.

[5] 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.

[6] C. W. Tan, P.-D. Yu, S. Chen, and H. V. Poor. Deeptrace: Learning to optimize contact tracing in epidemic networks with graph neural networks, 2023.

[7] R. S. Yadav. Mathematical modeling and simulation of sir model for covid-2019 epidemic outbreak: A case study of india, May 21 2020.