Finding the Factors that Contribute Most to the Spread of COVID-19 Inside and Among US Counties

Mehdi Azabou
Georgia Tech
mazabou@gatech.edu

Tanya Churaman
Georgia Tech
tanyachuraman@gatech.edu

Kipp Morris
Georgia Tech
kmorris9@gatech.edu

1 ABSTRACT

Accurately and quickly predicting case counts for any epidemic disease is a difficult problem with a wide variety of approaches. Even harder is explaining how an outbreak took place and what the driving factor was. Being able to answer these questions would be extremely helpful in planning epidemic responses for a wide variety of stakeholders (public health officials, etc.). We decided to tackle COVID-19 case count prediction by improving on the performance of a spatio-temporal graph neural network model proposed by Kapoor et al. in [2], adding additional data features to the model to improve the performance, including more detailed data on mobility flow between counties, county-level population data, and county-level unemployment data. We were able to reduce the test RMSLE on the top 20 most populous counties from 0.0109 to 0.0080, and through a few ablation studies that we performed with the goal of finding out which features contribute most to the spread of COVID-19. We finally use [7] to closely study the dynamics of the spread within counties and between counties.

2 INTRODUCTION

The overall goal was to analyze different factors in the spread of COVID-19, such as the presence/absence of shelter-in-place orders, socioeconomic status, mobility patterns, etc., with the intent of disentangling these factors from each other and developing an idea of how much each factor contributes to the spread. We took mobility patterns between counties into account to see how the aforementioned factors combine with mobility to affect the spread. The primary result is a graph neural network that forecasts case counts, where the neural network's input is a graph with a node for each county that contains data about case counts and all or some of the factors mentioned in the previous paragraph. We did some work to determine which individual factors are most significant while also finding the model that makes the best predictions.

This project should be of significant interest to health professionals and public health officials. By understanding the factors that affect the transmission of COVID-19 within and between counties, better mitigation strategies can be put into place. A one-size-fits-all approach might not be the best course of action for certain areas. Awareness of the different factors that affect the infection rate at can help officials understand the efficacy of mobility restrictions and how to implement strategies that are best for both the citizens and the economy, helping save lives.

3 PROBLEM STATEMENT

Our goal was, at the county level, to determine the extent to which factors such as the presence/absence of shelter-in-place orders, socioeconomic status, and mobility patterns influence the spread of COVID-19. We did this by 1) predicting case counts using a graph neural network model that takes into account the aforementioned factors and 2) explaining the network's predictions by looking at neuron activations and identifying salient features causing COVID-19 spread.

4 RELATED WORK & SURVEY

4.1 Forecasting COVID-19 Cases in Italy with a Compartmental SEIIRD Model

In this paper, Russo et al. [6] discuss an SEIIRD compartmental model that they developed and successfully fit to COVID case counts from February 21st to March 8th. It also did a good job of approximating case counts up to May 4th. The second I state represents people who are infected but asymptomatic to take into account that people who fall into that category are more likely to recover.

Their idea is interesting, and the model appears to have performed very well. However, it does not take into account socioeconomic factors or mobility factors. For our project, we will consider implementing this model, using it to forecast case counts, and then using the results as a baseline of sorts to compare to the results we get with our graph neural network.

4.2 Social and Economic Factors

Mukherji [5] presents the idea of a "vulnerability index", a metric that represents how vulnerable a US county is to COVID based on socioeconomic factors, as a way to explain significant differences in COVID case occurrences between different states and counties.

She found that factors that positively correlate with case counts include median income and the degree of economic inequality. Certain demographic factors turned out to be positively correlated to case counts as well.

While this paper does not provide any results about case counts, it does gives some really interesting results regarding the influence of socioeconomic factors on the spread of COVID. When we are determining which socioeconomic factors to use as independent variables in our model(s), we can use the results from this paper as a guide and also as a baseline to compare our results to.

4.3 Mobility Restrictions

In early March, many mobility restrictions were put in place in response to COVID-19. The goal was to reduce the amount of cases. Very strict guidelines were shown to help contain the virus in Wuhan, China. However, that does not mean these guidelines were effective everywhere. Local governments may not have the resources to enforce these guidelines, meaning citizens must voluntarily implementing these guidelines in their everyday lives [1].

4.3.1 *Connectivity's Effect Upon Mobility Restrictions*

. To analyze the effectiveness of mobility restrictions, [1] analyzed social connectedness between US counties and foreign countries. A county-day panel of social distancing, local cases, and exposure to COVID-info via social connections were combined with county characteristics to analyze compliance of mobility restrictions.

The results depicted that counties with high social connectedness implemented social distance guidelines upon mobility restrictions more quickly than those with low social connectedness. Social connectedness increased this compliance by 50%.

Counties with older-aged residents were less compliant with mobility guidelines; however, counties with higher social connectedness had a higher compliance metric amongst this population. College-educated people had low responsiveness to mobility guidelines regardless of social connections. Conversely, less educated counties complied with restrictions in conjunction to the level of connectedness. Lastly, social connectedness did not have an impact upon the counties with health-conditions (e.g. obesity, diabetes, etc); these counties were more compliant in general. Overall, this study suggests that the high connectedness within counties can allow for the flow of information between people to result in compliance with mobility restrictions, thus giving direction on how to enforce these life-saving guidelines.

While we are not researching the connectedness of counties, this study provides inspiration on how to use the mobility data in conjunction with shelter-in-place guidelines, socioeconomic status, and other variables to understand the spread of the virus within and between counties.

4.3.2 *Efficacy Mobility Restrictions*

. [4] focuses upon the effect of these mobility restrictions. Health officials can prepare strategies for the potential second wave of the infection. The goal would be able to minimize the effect of the pandemic and the economic impact.

Networks consisting of 10,000 nodes (people) and edges that captured interactions between individuals emulated how COVID-19 would spread within the population under different mitigation strategies. Each mitigation simulation produced a probability of a second-wave happening and the number deaths occurring.

A longer lockdown period of 3 months was deemed the safest option; however, there was more of a negative economic impact. 2 months showed much higher chance of a second wave, thus a shorter period is not beneficial. 2 months of lockdown plus strict social distancing measures resulted in less deaths and a less chance of a second wave compared to the former. In addition, it does not have a dire economic impact compared to 3 months of lockdown, thus being the best option.

While it is important to understand the impacts of various shelter-in-place strategies, this study is a simulation. There are many assumptions, and it does not represent real-world scenarios and does not take into account of everyone following the mobility restrictions. Our study will focus on previous real demographic, mobility, and case count data to help predict future case counts in order to analyze the effectiveness of current COVID guidelines and determine what next steps should be taken based on a certain area.

4.4 Deep Graph Networks

Graphs are a representation which supports arbitrary relational structure, naturally lending themselves to the representation of multiple real-world systems. The power of deep networks has mainly come from leveraging the inherent structure of the data being manipulated. We can think of convolutional layers for images and recurrent layers for sequential data. Deep graph networks are used for a wide variety of tasks, including link prediction, graph classification, node segmentation, and recommendations. Kipf et al. [3] extended the existing deep graph networks into a common framework. These models are called Message Passing Neural Networks (MPNNs). The core idea is not new; each node and/or edge is represented by a set of features $\mathbf{x}_i$, and then information is propagated through the graph in the form of "messages", by iteratively applying equation 1: each node i is going to receive information from its immediate neighbors $\mathcal{N}(i)$ through a learned message function $\phi^{(k)}$. These messages are aggregated through $\square$ which denotes a differentiable, permutation invariant function, e.g., sum, mean or max, and finally its representation at iteration (k) is updated using $\gamma^{(k)}$.

$$\mathbf{x}_i^{(k)} = \gamma^{(k)} \left(\mathbf{x}_i^{(k-1)}, \square_{j \in \mathcal{N}(i)} \phi^{(k)} \left(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}, \mathbf{e}_{j,i} \right) \right) \quad (1)$$

Intuitively, MPNNs can be seen as a generalization of the convolution operator to irregular domains. One of the appealing properties of graphs networks is that they are permutation invariant and are more predisposed to interpretability which is usually difficult with deep neural networks.

4.5 COVID-19 Case forecasting

In the case of COVID-19 forecasting, there are multiple factors that we might want to consider. These include historical infection data, inter-region human mobility, the prevalence of wearing masks and shelter-in-place orders, socioeconomic factors, etc.

With mechanistic approaches, where compartmental models have predefined transmission dynamics, or time series learning approaches, like autoregression or deep learning, forecasting is usually done within a "closed-system" location, thus not capturing meaningful spatial dynamics. This is where graph neural networks become interesting.

Kapoor et al. [2] utilize the deep graph network by modeling counties as a spatio-temporal graph with different types of edges. Spatial weighted edges represent mobility flows between nodes, whereas temporal edges represent connections to past days. The regression task that is being learned takes the historical data from the past 7 days and forecasts the number of cases for the next day.

The model is trained using data aggregated from three different sources: the NYT COVID-19 dataset, the Google COVID-19 Aggregated Mobility Research dataset, and the Google Community Mobility Reports. The learned model is shown to outperform multiple baselines. Metrics used to evaluate the different models include the RMSLE and Pearson correlations for the case deltas.

While this paper only uses mobility and case count features, such models can be extended to account for other factors.

4.6 Model interpretation

Deep networks are usually described as black boxes, so interpreting and understanding the salient features the model is basing its prediction on can be complicated. In our case, we want to use the model learned through the deep graph framework to have a way of interpreting which features are responsible for COVID-19 spread.

In the case of Graph Networks, a lot of work was done along these lines. Ying et al. [7] introduce GNNExplainer which is a model-agnostic explanation framework that, given a node that needs to be "explained", learns a local model in the subgraph of the node, thus finding the most representative features as well as the most important edges that led to a given prediction. Within the message passing framework, each node aggregates information from its neighbors, so it is important to consider both external information as well as the node attributes in identifying the features responsible for a given outcome. GNNExpliner does this by solving an optimization task that maxmizes the mutual information between the network's prediction and the distribution of all possible subgraphs.

5 NETWORK IMPLEMENTATION

In this section, we present the architecture of our network. Our proposed network is an extension of [2], the specifics of this extension are made clear in the title of the paragraphs. There is no publicly available implementation of the network. We implement it ourselves. When making decisions about the architecture, we follow the information available in [2], if a piece of information is missing, we make a decision based on conventional deep learning practices and our own experience. There are two main parts to this. First we will explain how graphs are built, and then we will explain the network architecture.

Deep Learning Framework. With a convolutional graph network, we are doing computation on every node i and all of its neighbors $\mathcal{N}(i)$ to update its state x_i. Doing this using tensors can be tricky, especially because each node can have a different number of neighbors, so effectively, if we consider a batch of nodes and their neighbors, then each element in this batch has a different length; it can be tricky in terms of implementation. This is further aggravated when we think about sample batching. Usually, when training a deep network, we take an average over multiple samples, so we need to feed a batch of samples to the network at each iteration. In our case, each sample is a graph, so we would have batches of batches, also of different shapes. To solve this issue, we will be using one of the popular graph network frameworks called PyTorch Geometric. We will be using the graph data structure as well as the message passing framework, which would solve the issue of dynamic shapes.

Building the graph. We will have a graph for every day, covering all counties of the United States. This is how it is built: We will start with an edgeless graph, each node representing a county. Each node will be described using geographical features pos (latitude and longitude) as well as an initial representation vector $\mathbf{x} = \mathbf{x_t} \,|\mathbf{x_{t-1}}| \ldots | \,\mathbf{x_{t-d}}$ which contains features over the past 7 days (num_days × features_dim). Then, using k-Nearest Neighbors, and the geographic distance between counties to create edges between every node and their 32 closest counties, which gives us our initial graph.

We use 7-days worth of data for the graph. The network would predict the new case count corresponding to the next day.

Building the graph (Extension). We further extend the graph by adding demographics and socio-economic data as global, non-temporal attributes for each county as a node feature. We note these features $\mathbf{x}_i^{(g)}$.

Temporal convolution. First, we start by computing temporal features, which only requires using a multi-layer perceptron over the concatenated 7-day features for each node:

$$\mathbf{x}_i^{(1)} = \text{mlp}\left(\mathbf{x}_{i,\mathbf{t}} \left|\mathbf{x}_{i,\mathbf{t-1}}\right| \ldots | \,\mathbf{x}_{i,\mathbf{t-d}} \,|\, \mathbf{x}_i^{(g)}\right) \tag{2}$$

The mlp has one hidden layer of size 64. Notice that we also include the global features of the node as well, which are static, so don't need to be repeated multiple times for every day.

Spatial convolution. This is where spatial features are learned, we use message passing to iteratively propagate information from nodes to their neighbors. Each node in the graph is basically sending messages about their state to their neighbors. These messages are generated via a multi-layer perceptron $\phi^{(l)}$. Each node receives messages, aggregates them using summation, uses an activation function (ReLU), and finally concatenates the new spatial features to the temporal features. Each layer has one hidden layer of size 32.

$$\mathbf{x}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \phi^{(l)}\left(\mathbf{x}_i^{(l-1)}\right)\right) | \,\mathbf{x}_i^{(0)} | \,\mathbf{x}_i^{(g)} \tag{3}$$

Having temporal features is equivalent to using skip-connections between layers. The reason we keep the temporal features is to avoid diluting the self-node feature state. This is also why we repeat this step for 2 iterations only. Actually, some empirical experiments show that graph convolutional networks shouldn't be very deep, to avoid dilution: after a certain number of iterations, the same information will have propagated across the entire network and all nodes will gradually have closer and closer feature embeddings.

Spatial convolution (Extension). In [2], the edges are built using knn over the geographical features of counties, this for example doesn't take into account people flying between counties, and also might imply some equivalent relationship between counties which isn't necessarily true. To more accurately represent edges between counties, we will build weighted edges between counties based on the inter-county mobility flow. By default close counties (knn) will have an edge weight of 1. Then if there is a flow between county i and county j, we cumulate the edge weight if the edge exists or otherwise add the weighted edge between these two nodes. This means that now, the messages being passed between counties will also take into account this specific feature.

The mlp now takes both the features of the neighbor node as well as the features of the edge to generate the message:

$$\mathbf{x}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \phi^{(l)}\left(\mathbf{e}_{ij} * \mathbf{x}_i^{(l-1)}\right)\right) | \,\mathbf{x}_i^{(0)} | \,\mathbf{x}_i^{(g)} \tag{4}$$

Readout phase. After computing the spatial and temporal features, we predict the case count using a multi-layer perceptron Ψ with one hidden layer of size 32.

$$\mathbf{p}_i = \Psi\left(\mathbf{x}_i^{(s)}\right) \tag{5}$$

Interpretability. After implementing and training this network, we will use the GNNExplainer explanation model from [7] to start getting information about relevant features, which will help measure the impact of different features and their contributions as driving factors in COVID-19 spread. GNNExplainer, is model agnostic, so we can use it with our network out of the box. This method acknowledges that the driving factors can be different for each county. So we would be look at explaining the evolution of number of cases at the county level.

We would have a powerful network that predicts the number of new cases for the next day based on a variety of features. But what would be interesting for all stakeholders, is to be able to identify and understand what led to a sudden increase in number of new cases for example. Having used a graph network, it is easier to see whether the increase came from intra-county dynamics by looking a the node features or inter-county dynamics by looking at the messages passed between different counties. This would be a very powerful and flexible tool for all stakeholders.

6 DATA

Data preparation is an important aspect of the deep learning pipeline. In our case it is even more so crucial and difficult because we would be aggregating data from multiple different sources.

Getting the data together and ready to be passed into the network actually turned out to be the most time-consuming portion of the project by a long shot, and while we were somewhat caught off-guard, we were able to get the data prepared and train the network with no issue in the end. One thing that was particularly difficult to address was that C3.ai used their own set of unique county IDs that was different from the FIPS IDs used by every other data source, and C3.ai was missing the mappings from their IDs to the FIPS IDs for 299 counties. To shorten the amount of time it took to get the network running, we discarded those 299 counties from the data, leaving us with a total of 2942 counties. If we were to go back and look at this project again, we would want to manually enter the FIPS IDs for those counties so that we could join them with the other data and use them.

We combine data from multiple sources. One common issue we encountered was having missing values for many of these data sources. We replaced missing values by using a rolling average over the past 7 days for each county. We also use the same strategy to normalize all of the features. For counties for which features are not available for more than one week, we plan on using the average from either the corresponding state or from the adjacent counties. This method was used for the mobility and case count data.

Training data was collected from the 59th to 120th day in 2020 – February 28-April 29(62 days). Testing data was collected from the 121st day to 150th day in 2020 – April 30 -May 29 (30 days). In all the size of the data is 8 GB (5 GB training, 3 GB testing).

The sources we used were:

6.1 C3.ai COVID-19 API

The C3.ai API provides an easy and convenient way to access COVID-19 data from a variety of sources. We chose to use the New York Times COVID-19 Daily Case Counts, this is because we want to reproduce the work of [2] before expanding our scope to more features and sources. Our aim is also to familiarize ourselves with the data used in [2], and start thinking about how we can improve the current architecture.

6.2 County Location Data

We used the county location data found on a GitHub repo hosted by Benjamin Skinner from the University of Florida to match FIPS IDs to the latitude and longitude of the counties, since we needed the geographical locations of the counties to build the graphs.

6.3 Census Data

We also used total population and population of age 65 or older as static features of the county nodes. We used population estimates from 2019 from the US Census Bureau.

6.4 PlaceIQ movement data

In [2], The Google COVID-19 Aggregated Mobility Research Dataset, which is a non-public dataset, is used to obtain inter-county flows and intra-county flows for each county. As a substitute, we use data from the PlaceIQ movement dataset. This data source provides the mobility flows between counties: for each county i at day d, every device that pinged in county i at d, and also pinged in county j at least once during the previous 14 days is counted towards the flow from county j to county i.

The dataset provides csv files of the daily county-level location exposure index, in an approximately 2000-by-2000 square matrix, where rows and columns correspond to counties.

6.5 Google COVID-19 Community Mobility Reports

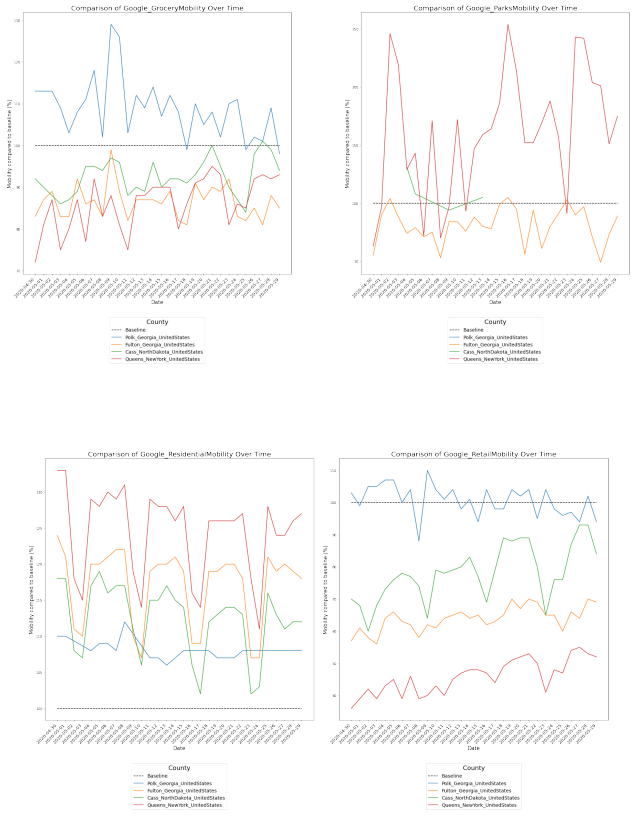
Normalized mobility trends were accessed via the C3.ai COVID-19 API. For each county, it gives a metric representing the mobility to/from six different categories of places, where the specific metric it gives is a percent increase or decrease in mobility relative to the mobility on a baseline day. The Excel file with the unique county IDS was used to extract the United States counties from the Google Mobility Trends. With these IDS, we were able to fetch the mobility trends for the training and testing periods.

This mobility data can reveal trends between counties. To elaborate below are 6 graphs (Figure 1) representing each of the 6 features mentioned above. To show the insights from this data, we have focused on two main county relationships:

6.5.1 **Highly Populous vs Lowly Populous county .** From the graphs, Polk County – the less populous county with approx 40 thousand residents – has a significantly higher Workplace, Retail and Grocery Mobility compared to the highly populous Fulton County (with approx. 1 million residents). These mobility patterns highlight that perhaps another factor, such as demographics, etc. might be affecting the mobility patterns.

6.5.2 **Presence vs No COVID Restrictions.** From the graphs, Workplace, Retail, Transit, and Grocery Mobility is much higher from Cass County than Queens County. This relationship depicts the stark difference in mobility patterns due to COVID restrictions. Since Cass County had no guidelines, the mobility patterns are much higher for these categories compared to Queens County where there were strict guidelines. The opposite is seen, however, for Residential Mobility. This result suggests that another factor might be affecting the mobility patterns for this area.

From above, we wished to highlight that while mobility patterns might have an impact upon the spread of COVID-19 within each county, these mobility patterns themselves can be affected by other variables. Through the predictive neural net, we hope to pinpoint certain variables within certain counties that have a significant impact towards the spread of the virus.



6.6 Unemployment Rates from US Bureau of Labor

Unemployment rates were accessed via the C3.ai COVID-19 API. This metric was represented as the percent of unemployed population divided by the total labor force population. The unemployed population was characterized by individuals 16 or older that were willing to work but gained no employment. This data was available on a monthly basis.

Figure 2 has the unemployment rates of 4 counties graphed. The change of employment rates over time can be examined.

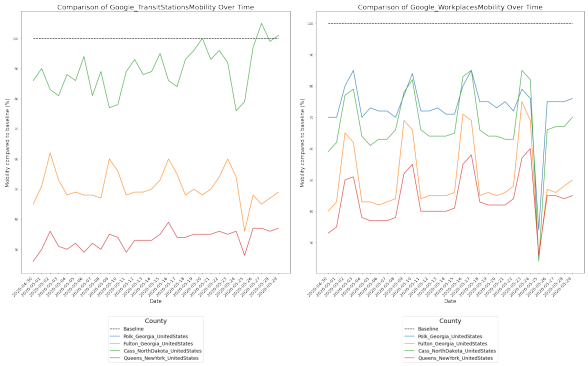


Figure 1: Mobility Pattern Graphs

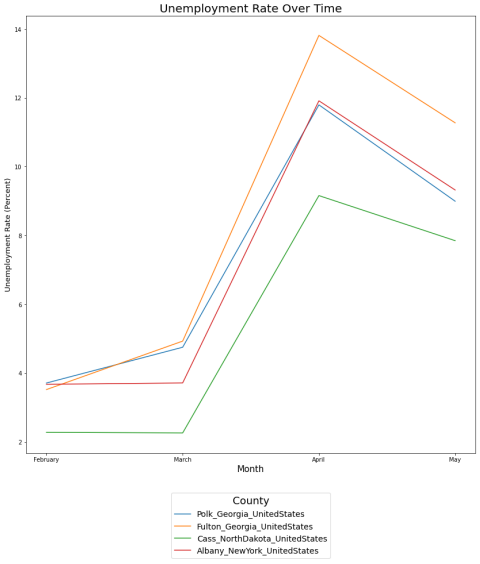


Figure 2: Unemployment Rate Graph

6.6.1 **Highly Populous vs Lowly Populous county.** Similar to Figure 1, Fulton County is the highly populous county while Polk County is the opposite. From the graph, Fulton County's unemployment spiked much higher than that of Polk County. Logically, this relationship makes sense. Since Fulton County has more people, the number of unemployed would be a larger value.

6.6.2 **Presence vs No COVID Restrictions.** Similar to Figure 1, Cass County has no COVID Restrictions and Albany County has strict guidelines. We switched from Queens to Albany due to other data sources not finding Queens County, NY. The graph depicts that Albany experienced a greater increase in unemployment compared to Cass County. Perhaps, this stark difference because of Cass County's lack of COVID restrictions; if there are no shelter-in-place restrictions then there wouldn't be as many layoffs due to businesses closing.

Unemployment data is meant to serve as a socioeconomic factor that could potentially help forecast the number of case counts. Perhaps a county with more unemployed people may have more

cases due to people trying to get some form of work, regardless of the circumstances.

7 EXPERIMENTS

After compiling, processing and exploring the data. We first train our network and then conduct experiment in the efforts of interpreting what information the model is using to make its predictions.

7.1 Experiment: New case prediction

In this experiment, we focus on training the network on the task of new case prediction (Regression task). To evaluate the performance of the network, we will follow [2] in using the RMSLE metric.

Root Mean Square Log Error. The formula for RMSLE is as follows:

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i+1) - \log(\hat{y}_i+1))^2}$$

Note that this is equivalent to:

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{\log(y_i+1)}{\log(\hat{y}_i+1)})^2}$$

The second formulation tells us that RMSLE does not take into account the magnitude of the error of a single prediction; it only considers the relative error of each prediction. This quality is desirable models that predict disease case counts because we want the predicted case count for each day to be relatively close to the ground truth value for that day.

Another key characteristic of RMSLE that makes it a good choice for our use case is that it punishes underpredictions much more harshly than overpredictions. This makes sense for us because we would rather overestimate future case counts and allocate too many ventilators, vaccines, etc. than too few.

Root Mean Square Log Error - Top 20. [2] reports the RMSLE for the top-20 most populous counties. Thus, we will report these metrics as well. They unfortunately do not provide the RMSLE for all counties in the US. We will be reporting ours, but will not be able to compare it with [2].

Training. We train the graph network using a batch size of 16 for stochasticity and the Adam optimizer with a learning rate of 10^{-3}. We train the network for $100k$ iterations, using the MSE loss. Training was done using an RTX 2080 GPU and took an average of 7 hours to train. The processed train and test data take respectively 5GB and 3GB of storage.

Baselines. As baselines to compare our results to, rather than running our own models and producing results, we have used the results summarized by Kapoor et al [2] that they compared their GNN's results to. Since we used the same data and we were tight on time at the end of our project, we decided not to duplicate the results ourselves for the baseline comparisons. They used two baselines: an ARIMA (Auto Regressive Integrated) model for time-series based predictions and an LSTM (Long Short Term Memory) model, another neural network model that can learn long-term dependencies. The results are summarized in Table 1. As they mention in their paper, their GNN performs better than the ARIMA and LSTM models on the top 20 most populous counties, and our GNN, trained on additional data features, improved on their GNN's performance by about 20% on the same set of counties.

Results. We present the results in Table 7.1. We find that our extension outperforms the baselines as well as the network from [2].

Model	RMSLE (Top 20)
ARIMA	0.0144
LSTM	0.0121
GNN (Kappor et al)	0.0109
Our Model	0.0080

Table 1: A table summarizing the RMSLE for the top 20 counties for different models. Our extension performs better than other models.

We find that the RMSLE over all counties is higher, which is to be expected. However, examine the predicted new case counts for Fulton county for example, which can be seen in Figure 3. While the prediction isn't fully accurate, we find that the network mostly overestimates the number of new cases.

7.2 Experiment: Ablation studies

As one method of attempting to interpret the predictions of our model, we did a couple of ablation studies. For each ablation study, we removed a particular predictor variable or some set of predictor variables from the model, trained the network with the partial dataset, and compared the testing results to those of the network trained with all of the predictor variables, the idea being that if one of the networks from the ablation studies performs worse than the network that had access to all of the data, it indicates that the variable(s) removed during the corresponding ablation study are significant to the predictions and important for us to pay attention to in real life. We performed three ablation studies: one where we removed the mobility flow data, one where we removed the two population-related features, and one where we removed the unemployment-related features. The results are summarized in the table below, the takeaway being that all of the features we tried removing are significant to the predictions. Taking any of the features away increased the RMSLE by around 200-300%. Specifically, we can see that at least on the top 20 most populous counties, the network performed better without the population features and that the mobility flow data looks to be particularly significant to the predictions among the features we looked at in the ablation studies.

	RMSLE (Top 20)	RMSLE
Baseline	0.0080	0.0130
No mobility flow	0.0096	0.0300
No population feats.	0.0077	0.0280
No unemployment feats.	0.0088	0.0220

Table 2: A table summarizing the RMSLE values on the testing data for the baseline model and the models from the ablation studies on a subset of the data made up of the 20 most populous counties and on all counties

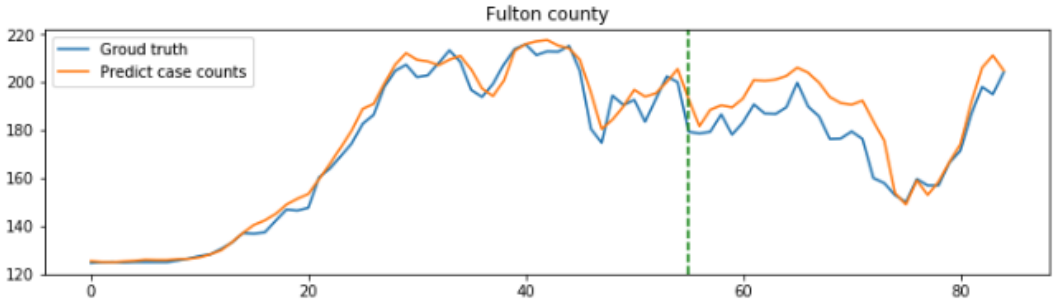


Figure 3: Predicted new cases for Fulton County

7.3 Experiment: Interpretability

We use GNNExplainer to identify compact subgraph structures and node features that play a crucial role in the graph network's predictions.

We select a county and a day in which there is a spike in new cases or a change of slope in the number of total cases and try to understand what happened, or how the network was able to predict the spike.

First, we look at Crawford County, Wisconsin, on day 20 of the test set. As can be seen in Figure 4, there is a spike in the next day, which is also predicted by the network. Crawford County has a small number of cases, thus, we hypothesize that the spike in cases might come from an inflow of cases from neighboring counties. We run the GNNExplainer and get the results in Figure 5. We find that indeed a lot of graph neighbors (2 counties in Oklahoma, 1 county in Texas) are contributing to the network's decision. This observation does not mean that the spike is specifically due to these counties. These counties might just be correlated in their dynamics as we can see in Figure 4, but it does indicate where the cases might have came from.

Another county that we consider, which has a higher case count than Crawford County, is Fort Bend County, Texas, which has a spike at the end of the month (Figure 4, center). GNNExplainer (Figure 6) finds that there is some link to Fort Bend, which again might explain where this increase in cases is coming from.

In each of these experiments, we look at both node feature importance and edge importance, and in both, we find that edges are more important, so they might be the leading factor for spread. Now we look at a more populous county: LA County. We interestingly find that edge importance is very low, while in terms of nodes, we find that both unemployment rates and previous day case count are most important, with importance scores over 0.8. The unemployment rate in LA is around 20%, which is higher than the national average. We suspect that the unemployment rate is just correlated with the case count, but it is not clear that it is a driving factor for COVID-19 spread. Unfortunately, no further experiments were conducted to analyse this further.

8 DISCUSSION

Deep neural networks networks can be viewed as a black box – we do not necessarily know how the network is forecasting the case counts. In addition, deep learning was a fairly new topic for some members of the group. In addition, the team was trying to analyze a wide range different factors that contributed in the spread of COVID-19.

We were able to establish a good predictor and show how our proposed method improves over the previous method. We also were able to find some inconsistency with the proposed method in [2] where the choice for the reported metric being specific to 20 out of 3000 plus counties, wasn't really justified. We also found that training the network was hard as we didn't have enough resources to conduct hyperparameter tuning or make more extensive and statistically accurate ablation studies.

Our experiments on interpretability show the power of graph networks and how they allow for such insightful explanations of the dynamics of COVID-19 spread. We find that while our results are promising, they don't tell the full story. Due to most of the time being spent on data acquisition and processing and then training the network, we weren't able to conduct further interpretability experiments and we leave that for future work.

9 DATA SOURCES

Listed here are links to all of the data sources we used.

- C3.ai COVID-19 API
- County Location Data
- Census Population by County by Age
- Census Total Population by County
- PlaceIQ Movement Data
- Google COVID-19 Community Mobility Data

REFERENCES

[1] Ben Charoenwong, Alan Kwan, and Vesa Pursiainen. 2020. Social connections to COVID-19-affected areas increase compliance with mobility restrictions. *Science Advances* (2020). https://doi.org/10.1126/sciadv.abc3054

[2] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O'Banion. 2020. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. arXiv:cs.LG/2007.03113

[3] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR* abs/1609.02907 (2016). arXiv:1609.02907 http://arxiv.org/abs/1609.02907

[4] Parul Maheshwari and Réka Albert. 2020. Network model and analysis of the spread of Covid-19 with social distancing. arXiv:physics.soc-ph/2006.09189

[5] Nivedita Mukherji. 2020. The Social and Economic Factors Underlying the Incidence of COVID-19 Cases and Deaths in US Counties. arXiv:2020.05.04.20091041

[6] Lucia Russo, Cleo Anastassopoulou, Athanassios Tsakris, Gennaro Nicola Bifulco, Emilio Fortunato Campana, Gerardo Toraldo, and Constantinos Siettos.

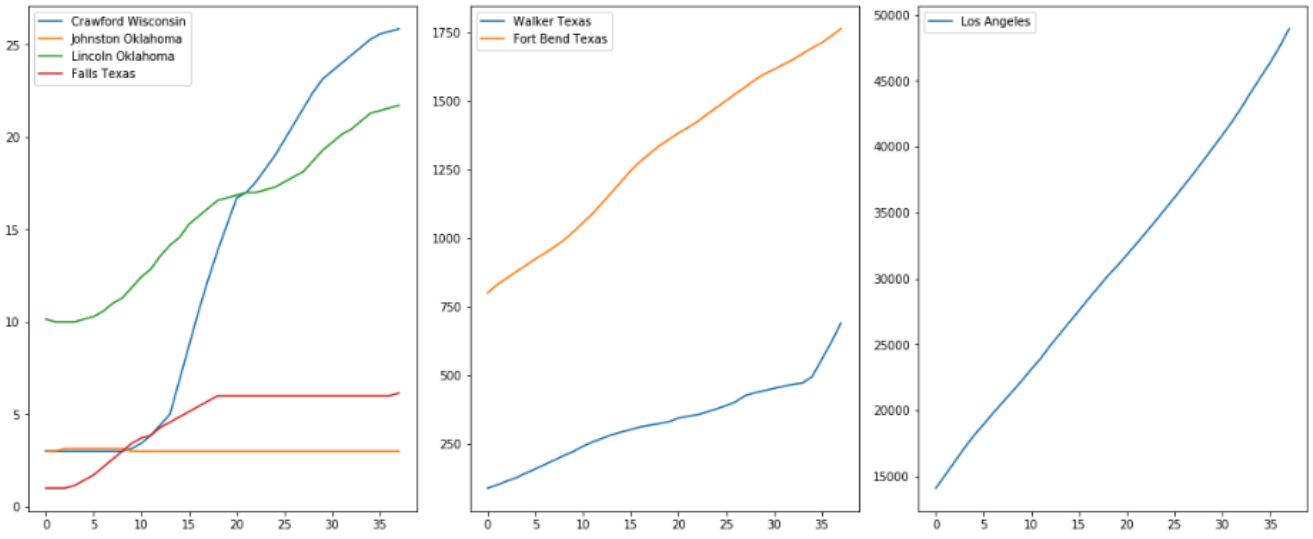


Figure 4: Case counts for different counties.

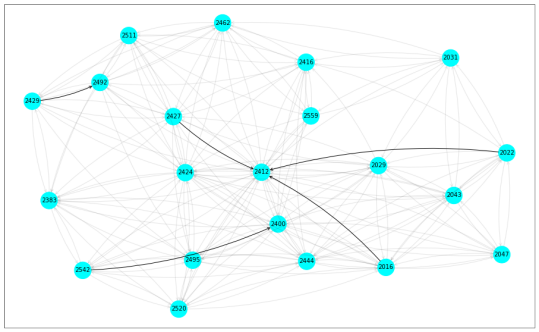


Figure 5: Subgraph explaining the predicted spike for Crawford County, Wisconsin (node id: 2412)

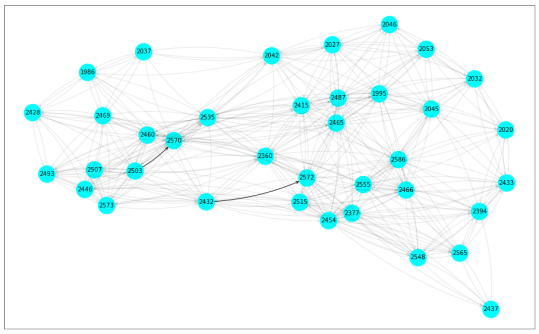


Figure 6: Subgraph explaining the predicted spike for Fort Bend, Texas (node id: 2572)

2020. Tracing DAY-ZERO and Forecasting the COVID-19 Outbreak in Lombardy, Italy: A Compartmental Modelling and Numerical Optimization Approach. arXiv:2020.03.17.20037689

[7] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. GNN Explainer: A Tool for Post-hoc Explanation of Graph Neural Networks. *CoRR* abs/1903.03894 (2019). arXiv:1903.03894 http://arxiv.org/abs/1903.03894

Milestone

Finding the Factors that Contribute Most to the Spread of COVID-19 Inside and Among US Counties

Mehdi Azabou
Georgia Tech
mazabou@gatech.edu

Tanya Churaman
Georgia Tech
tanyachuraman@gatech.edu

Kipp Morris
Georgia Tech
kmorris9@gatech.edu

1 ABSTRACT

Our goal is, at the county level, to determine the extent to which factors such as the presence/absence of shelter-in-place orders, socioeconomic status, and mobility patterns influence the spread of COVID-19. We will do this by 1) predicting case counts using a graph neural network model that takes into account the aforementioned factors and 2) explaining the network's predictions by looking at neuron activations and identifying salient features causing COVID-19 spread.

2 INTRODUCTION

The overall goal is to analyze different factors in the spread of COVID-19, such as the presence/absence of shelter-in-place orders, socioeconomic status, mobility patterns, etc., with the intent of disentangling these factors from each other and developing an idea of how much each factor contributes to the spread. We also plan to take mobility patterns between counties into account to see how the aforementioned factors combine with mobility to affect the spread. The primary result we are aiming for is a graph neural network that forecasts case counts, where the neural network's input is a graph with a node for each county that contains data about case counts and all or some of the factors mentioned in the previous paragraph. If possible, we would also like to determine which individual factors are most significant while also finding the model that makes the best predictions.

This project should be of significant interest to health professionals and public health officials. By understanding the factors that affect the transmission of COVID-19 within and between counties, better mitigation strategies can be put into place. A one-size-fits-all approach might not be the best course of action for certain areas. Awareness of the different factors that affect the infection rate at can help officials understand the efficacy of mobility restrictions and how to implement strategies that are best for both the citizens and the economy, helping save lives.

3 RESPONSE TO COMMENTS ON PROPOSAL

The following are the comments that were made on the proposal and the corresponding changes we have made:

- **Rewrite abstract as a summary of the project in prose.** Done.
- **Structure the report more conventionally like a paper.** We removed a lot of things that were in the proposal that are not really necessary for the milestone report.
- **Fix incomplete citations.** We noticed that two of the citations were missing DOIs, so we added those in.
- **Shorten and contextualize the literature survey.** We made sure to shorten it and also to clearly state at the end of the section for each paper how that paper is relevant to our work.
- **Check out other data sources.** We made great use of C3.ai!
- **Elaborating on graph neural network interpretability.** Addressed in the section toward the end about future work.

4 RELATED LITERATURE

4.1 Forecasting COVID-19 Cases in Italy with a Compartmental SEIIRD Model

In this paper, Russo et al. [7] discuss an SEIIRD compartmental model that they developed and successfully fit to COVID case counts from February 21st to March 8th. It also did a good job of approximating case counts up to May 4th. The second I state represents people who are infected but asymptomatic to take into account that people who fall into that category are more likely to recover.

Their idea is interesting, and the model appears to have performed very well. However, it does not take into account socioeconomic factors or mobility factors. For our project, we will consider implementing this model, using it to forecast case counts, and then using the results as a baseline of sorts to compare to the results we get with our graph neural network.

4.2 Social and Economic Factors

Mukherji [6] presents the idea of a "vulnerability index", a metric that represents how vulnerable a US county is to COVID based on socioeconomic factors, as a way to explain significant differences in COVID case occurrences between different states and counties.

She found that factors that positively correlate with case counts include median income and the degree of economic inequality. Certain demographic factors turned out to be positively correlated to case counts as well.

While this paper does not provide any results about case counts, it does gives some really interesting results regarding the influence of socioeconomic factors on the spread of COVID. When we are determining which socioeconomic factors to use as independent variables in our model(s), we can use the results from this paper as a guide and also as a baseline to compare our results to.

4.3 Mobility Restrictions

In early March, many mobility restrictions were put in place in response to COVID-19. The goal was to reduce the amount of cases. Very strict guidelines were shown to help contain the virus in Wuhan, China. However, that does not mean these guidelines were effective everywhere. Local governments may not have the resources to enforce these guidelines, meaning citizens must voluntarily implementing these guidelines in their everyday lives [1].

4.3.1 Connectivity's Effect Upon Mobility Restrictions

. To analyze the effectiveness of mobility restrictions, [1] analyzed social connectedness between US counties and foreign countries.

A county-day panel of social distancing, local cases, and exposure to COVID-info via social connections were combined with county characteristics to analyze compliance of mobility restrictions.

The results depicted that counties with high social connectedness implemented social distance guidelines upon mobility restrictions more quickly than those with low social connectedness. Social connectedness increased this compliance by 50%.

Counties with older-aged residents were less compliant with mobility guidelines; however, counties with higher social connectedness had a higher compliance metric amongst this population. College-educated people had low responsiveness to mobility guidelines regardless of social connections. Conversely, less educated counties complied with restrictions in conjunction to the level of connectedness. Lastly, social connectedness did not have an impact upon the counties with health-conditions (e.g. obesity, diabetes, etc); these counties were more compliant in general. Overall, this study suggests that the high connectedness within counties can allow for the flow of information between people to result in compliance with mobility restrictions, thus giving direction on how to enforce these life-saving guidelines.

While we are not researching the connectedness of counties, this study provides inspiration on how to use the mobility data in conjunction with shelter-in-place guidelines, socioeconomic status, and other variables to understand the spread of the virus within and between counties.

4.3.2 *Efficacy Mobility Restrictions*

. [5] focuses upon the effect of these mobility restrictions. Health officials can prepare strategies for the potential second wave of the infection. The goal would be able to minimize the effect of the pandemic and the economic impact.

Networks consisting of 10,000 nodes (people) and edges that captured interactions between individuals emulated how COVID-19 would spread within the population under different mitigation strategies. Each mitigation simulation produced a probability of a second-wave happening and the number deaths occurring.

A longer lockdown period of 3 months was deemed the safest option; however, there was more of a negative economic impact. 2 months showed much higher chance of a second wave, thus a shorter period is not beneficial. 2 months of lockdown plus strict social distancing measures resulted in less deaths and a less chance of a second wave compared to the former. In addition, it does not have a dire economic impact compared to 3 months of lockdown, thus being the best option.

While it is important to understand the impacts of various shelter-in-place strategies, this study is a simulation. There are many assumptions, and it does not represent real-world scenarios and does not take into account of everyone following the mobility restrictions. Our study will focus on previous real demographic, mobility, and case count data to help predict future case counts in order to analyze the effectiveness of current COVID guidelines and determine what next steps should be taken based on a certain area.

4.4 Deep Graph Networks

Graphs are a representation which supports arbitrary relational structure, naturally lending themselves to the representation of multiple real-world systems. The power of deep networks has mainly come from leveraging the inherent structure of the data being manipulated. We can think of convolutional layers for images and recurrent layers for sequential data. Deep graph networks are used for a wide variety of tasks, including link prediction, graph classification, node segmentation, and recommendations. Kipf et al. [4] extended the existing deep graph networks into a common framework. These models are called Message Passing Neural Networks (MPNNs). The core idea is not new; each node and/or edge is represented by a set of features $\mathbf{x}_i$, and then information is propagated through the graph in the form of "messages", by iteratively applying equation 1: each node i is going to receive information from its immediate neighbors $\mathcal{N}(i)$ through a learned message function $\phi^{(k)}$. These messages are aggregated through $\square$ which denotes a differentiable, permutation invariant function, e.g., sum, mean or max, and finally its representation at iteration (k) is updated using $\gamma^{(k)}$.

$$\mathbf{x}_i^{(k)} = \gamma^{(k)} \left(\mathbf{x}_i^{(k-1)}, \square_{j \in \mathcal{N}(i)} \phi^{(k)} \left(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}, \mathbf{e}_{j,i} \right) \right) \quad (1)$$

Intuitively, MPNNs can be seen as a generalization of the convolution operator to irregular domains. One of the appealing properties of graphs networks is that they are permutation invariant and are more predisposed to interpretability which is usually difficult with deep neural networks.

4.5 COVID-19 Case forecasting

In the case of COVID-19 forecasting, there are multiple factors that we might want to consider. These include historical infection data, inter-region human mobility, the prevalence of wearing masks and shelter-in-place orders, socioeconomic factors, etc.

With mechanistic approaches, where compartmental models have predefined transmission dynamics, or time series learning approaches, like autoregression or deep learning, forecasting is usually done within a "closed-system" location, thus not capturing meaningful spatial dynamics. This is where graph neural networks become interesting.

Kapoor et al. [3] utilize the deep graph network by modeling counties as a spatio-temporal graph with different types of edges. Spatial weighted edges represent mobility flows between nodes, whereas temporal edges represent connections to past days. The regression task that is being learned takes the historical data from the past 7 days and forecasts the number of cases for the next day.

The model is trained using data aggregated from three different sources: the NYT COVID-19 dataset, the Google COVID-19 Aggregated Mobility Research dataset, and the Google Community Mobility Reports. The learned model is shown to outperform multiple baselines. Metrics used to evaluate the different models include the RMSLE and Pearson correlations for the case deltas.

While this paper only uses mobility and case count features, such models can be extended to account for other factors.

4.6 Model interpretation

Deep networks are usually described as black boxes, so interpreting and understanding the salient features the model is basing its prediction on can be complicated. In our case, we want to use the

model learned through the deep graph framework to have a way of interpreting which features are responsible for COVID-19 spread.

In the case of Graph Networks, a lot of work was done along these lines. Huang et al. [2] introduce GraphLIME, where LIME stands for Local Interpretable Model Explanations. GraphLIME is a model-agnostic explanation framework that, given a node that needs to be "explained", learns a local model in the subgraph of the node, thus finding the most representative features. Within the message passing framework, each node aggregates information from its neighbors, so it is important to consider both external information as well as the node attributes in identifying the features responsible for a given outcome.

A kernel-based nonlinear feature selection algorithm (Hilbert-Schmidt Independence Criterion Lasso) is applied on a N-hop neighborhood of the target node.

5 DATA

We combine data from multiple sources:

5.1 C3.ai COVID-19 API

(URL: https://c3.ai/covid-19-api-documentation/section/Quickstart-Guide/R-Quickstart)
The C3.ai API provides an easy and convenient way to access COVID-19 data from a variety of sources. We chose to use the New York Times COVID-19 Daily Case Counts, this is because we want to reproduce the work of [3] before expanding our scope to more features and sources. Our aim is also to familiarize ourselves with the data used in [3], and start thinking about how we can improve the current architecture. We also used the C3.ai API.

As a brief aside, we also used the county location data found at https://github.com/btskinner/spatial to match FIPS IDs to the latitude and longitude of the counties, since we need the geographical locations of the counties as part of the neural network input.

5.2 PlaceIQ movement data

(URL: github.com/COVIDExposureIndices/COVIDExposureIndices)
In [3], The Google COVID-19 Aggregated Mobility Research Dataset, which is a non-public dataset, is used to obtain inter-county flows and intra-county flows for each county. As a substitute, we use data from the PlaceIQ movement dataset. This data source provides the mobility flows between counties: for each county i at day d, every device that pinged in county i at d, and also pinged in county j at least once during the previous 14 days is counted towards the flow from county j to county i. We aggregate the incoming flow from all counties to get the aggregated inter-county flow for each county.

5.3 Google COVID-19 Community Mobility Reports

(URL: https://www.google.com/covid19/mobility/)

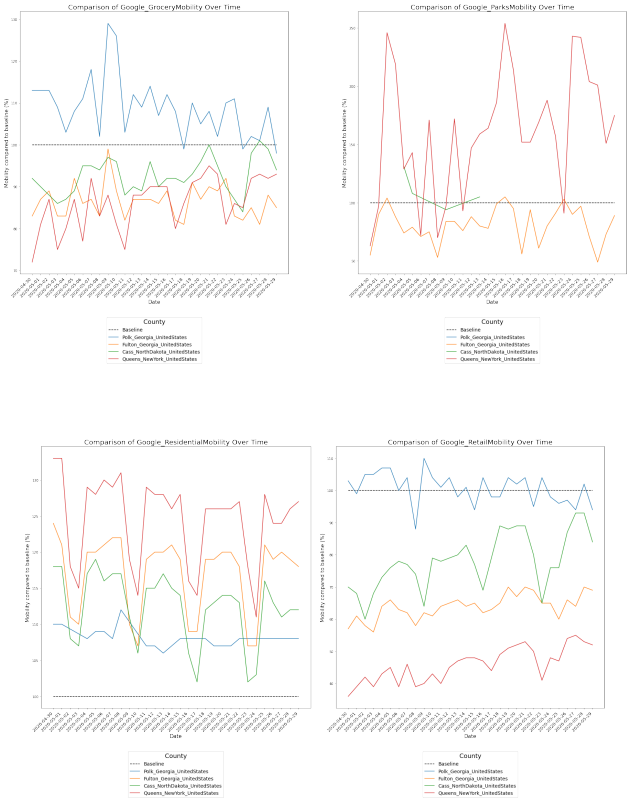
Normalized mobility trends were accessed via the C3.ai COVID-19 API. For each county, it gives a metric representing the mobility to/from six different categories of places, where the specific metric it gives is a percent increase or decrease in mobility relative to the mobility on a baseline day.

This mobility data can reveal trends between counties. To elaborate below are 6 graphs (Figure 1) representing each of the 6 features mentioned above. To show the insights from this data, we have focused on two main county relationships:

Highly Populous vs Lowly Populous county. From the graphs, Polk County – the less populous county with approx 40 thousand residents – has a significantly higher Workplace, Retail and Grocery Mobility compared to the highly populous Fulton County (with approx. 1 million residents). These mobility patterns highlight that perhaps another factor, such as demographics, etc. might be affecting the mobility patterns.

Presence vs No COVID Restrictions. From the graphs, Workplace, Retail, Transit, and Grocery Mobility is much higher from Cass County than Queens County. This relationship depicts the stark difference in mobility patterns due to COVID restrictions. Since Cass County had no guidelines, the mobility patterns are much higher for these categories compared to Queens County where there were strict guidelines. The opposite is seen, however, for Residential Mobility. This result suggests that another factor might be affecting the mobility patterns for this area.

From above, we wished to highlight that while mobility patterns might have an impact upon the spread of COVID-19 within each county, these mobility patterns themselves can be affected by other variables. Through the predictive neural net, we hope to pinpoint certain variables within certain counties that have a significant impact towards the spread of the virus.



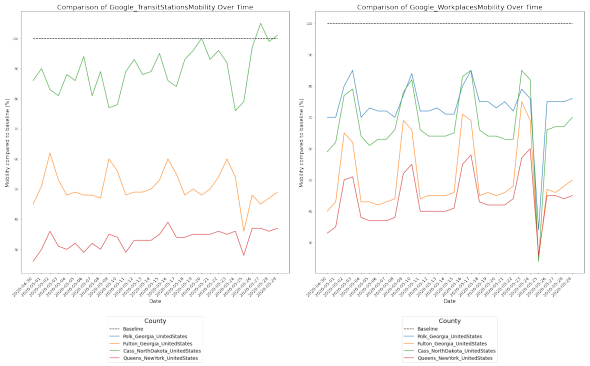


Figure 1: Mobility Pattern Graphs

6 NETWORK IMPLEMENTATION

After acquiring the data, our next step was to reproduce the network proposed in [3]. There are two main parts to this. First we will explain how graphs are built, and then we will explain the network architecture.

With a convolutional graph network, we are doing computation on every node i and all of its neighbors $\mathcal{N}(i)$ to update its state x_i. Doing this using tensors can be tricky, especially because each node can have a different number of neighbors, so effectively, if we consider a batch of nodes and their neighbors, then each element in this batch has a different length; it can be tricky in terms of implementation. This is further aggravated when we think about sample batching. Usually, when training a deep network, we take an average over multiple samples, so we need to feed a batch of samples to the network at each iteration. In our case, each sample is a graph, so we would have batches of batches, also of different shapes. To solve this issue, we will be using one of the popular graph network frameworks called PyTorch Geometric. We will be using the graph data structure as well as the message passing framework, which would solve the issue of dynamic shapes.

Building the graph. We will have a graph for every day, covering all counties of the United States. This is how it is built: We will start with an edgeless graph, each node representing a county. Each node will be described using geographical features pos (latitude and longitude) as well as an initial representation vector $\mathbf{x} = \mathbf{x_t} \,|\mathbf{x_{t-1}}| \ldots | \,\mathbf{x_{t-d}}$ which contains features over the past 7 days (num_days × features_dim). Then, using k-Nearest Neighbors, and the geographic distance between counties to create edges between every node and their 32 closest counties, which gives us the final graph we will use to train the network.

Temporal convolution. First, we start by computing temporal features, which only requires using a multi-layer perceptron over the concatenated 7-day features for each node:

$$\mathbf{x}_i^{(1)} = \text{mlp}\left(\mathbf{x}_{i,\mathbf{t}} \left|\mathbf{x}_{i,\mathbf{t-1}}\right| \ldots | \,\mathbf{x}_{i,\mathbf{t-d}}\right) \tag{2}$$

The mlp has one hidden layer of size 64.

Spatial convolution. This is where spatial features are learned, we use message passing to iteratively propagate information from nodes to their neighbors. Each node in the graph is basically sending messages about their state to their neighbors. These messages are generated via a multi-layer perceptron $\phi^{(l)}$. Each node receives messages, aggregates them using summation, uses an activation function (ReLU), and finally concatenates the new spatial features to the temporal features. Each layer has one hidden layer of size 32.

$$\mathbf{x}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \phi^{(l)}\left(\mathbf{x}_i^{(l-1)}\right)\right) | \,\mathbf{x}_i^{(0)} \tag{3}$$

Having temporal features is equivalent to using skip-connections between layers. The reason we keep the temporal features is to avoid diluting the self-node feature state. This is also why we repeat this step for 2 iterations only. Actually, some empirical experiments show that graph convolutional networks shouldn't be very deep, to avoid dilution: after a certain number of iterations, the same information will have propagated across the entire network and all nodes will gradually have closer and closer feature embeddings.

Readout phase. After computing the spatial and temporal features, we predict the case count using a multi-layer perceptron Ψ with one hidden layer of size 32.

$$\mathbf{p}_i = \Psi\left(\mathbf{x}_i^{(s)}\right) \tag{4}$$

7 PLAN OF ACTION

7.1 What we have done thus far

Case Count Data Compilation. The process of compiling the case count data was straightforward. C3.ai provides an Excel file that gives the unique IDs they use to represent the counties. We downloaded that file, extracted the United States county IDs, and then used the county IDs to fetch the case and death counts for the date ranges we wanted from the C3.ai EvalMetrics endpoint. We have designated the data from February 28th to April 29th for training data and the data from April 30th to May 30th for testing data.

County Location Data Compilation. To represent the geographic relationships between counties in the input to the neural network we also needed the latitude and longitude for each county in our data. We got this data by using C3.ai's Fetch endpoint to match the C3.ai county IDs to FIPS IDs and then matching the FIPS IDs to location data using the data in county_centers.csv from https://github.com/btskinner/spatial.

Mobility Flow Compilation The dataset provides csv files of the daily county-level location exposure index, in an approximately 2000-by-2000 square matrix, where rows and columns correspond to counties. The intra-flow is found on the diagonal of the matrix, while the inter-flow is computed by summing over out-of-diagonal values.

Mobility Trends Compilation. To capture the mobility trends in various contexts (grocery, workplace, retail, etc.), the C3.ai was used to capture the available data for each of the counties. The Excel file with the unique county IDS was used to extract the United States counties from the Google Mobility Trends. With these IDS, we were able to fetch the mobility trends for the training and testing periods.

Issues with Getting Data. One common issue we encountered was having missing values for many of these data sources. We tried replacing missing values by using a rolling average over the past 7 days for each county. We also use the same strategy to normalize all of the features. For counties for which features are not available for more than one week, we plan on using the average from either the corresponding state or from the adjacent counties. We also found that each data source, has a different number of counties, which was very confusing and we are still investigating why this is happening and trying to come up with a strategy to work around this.

During Dr. Bin Yu's lecture as part of the IDEaS-TRIAD lecture series on October 23rd, we learned about the COVID-19 data repository that was curated by their lab, as well as the baselines that were made available in that same repository. Although this would have been the perfect data source for us to use, we weren't able to load it properly. We spent a considerable amount of time trying to get around bugs and errors in the code due to python and OS incompatibilities.

Network implementation. To implement the network, we had to extensively use the documentation of PyTorch Geometric, we implemented a dataset class, to load, process and generate graphs from the raw data, as described earlier. We also had to implement a network module that inherits from the MessagePassing class. The network is almost done, we only need to write a training script and train it once the data is fully compiled and processed.

7.2 Future work

Finish formatting the data. For case counts and location data, we have the raw data that we need, but we need resolve the issues we've encountered.

Training the network. We will train the network, try to reproduce the same results in [3] and then try to improve the model's performance, using different ideas that we've came up with, while working on this data.

Adding edge features. In the current representation of our graph, we build edges using knn over the geographical features of counties, this for example doesn't take into account people flying between counties, and also might imply some equivalent relationship between counties which isn't necessarily true. In To more accurately represent edges between counties, we will build weighted edges between counties based on the inter-county mobility flow. So if there is a flow between county i and county j, then we will add an edge between these two nodes, we will also associate features to this edge (mobility flow), which means that the messages being passed between counties will also take into account that specific feature between nodes.

The mlp now takes both the features of the neighbor node as well as the features of the edge to generate the message:

$$\mathbf{x}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \phi^{(l)}\left(\mathbf{x}_i^{(l-1)}, \mathbf{e}_{ij}\right)\right) | \, \mathbf{x}_i^{(0)} \tag{5}$$

Adding non-temporal attributes. We will add demographics and socio-economic data as global, non-temporal attributes for each county.

So now we will have temporal features, spatial features and global non-temporal features:

$$\mathbf{x}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \phi^{(l)}\left(\mathbf{x}_i^{(l-1)}, \mathbf{e}_{ij}\right)\right) | \, \mathbf{x}_i^{(0)} | \, \mathbf{x}_i^{(g)} \tag{6}$$

Interpretability. After implementing and training this network, we will use the explanation models from [2] to start getting information about relevant features, which will help measure the impact of socioeconomic factors across different counties. To evaluate our learned graph model, we will be following the same evaluation procedure as [3], i.e. we would use RMSLE and Pearson correlations for the case deltas. Measuring the success of the explanation model is more tricky, as we would have to rely on the consistency of these explanations across different counties. As we will work on this aspect of the project, we will have to come up with a way of showing how the representation we are extracting are meaningful.

8 DISCUSSION

Using deep neural networks is a difficulty we will encounter. These networks can be viewed as a black box – we do not necessarily know how the network is forecasting the case counts. In addition, deep learning is a fairly new topic for some members of the group. There is the risk of trying to learn more introductory principles while completing this project. In addition, the team is trying to analyze different factors that contributed in the spread of COVID-19. We will have to carefully analyze the relationship of each variable in conjunction with the infection and other variables as well.

However, there are benefits. This project could help health officials come up with better strategies to contain the virus. By containing the virus, we are protecting the essential workers, giving time to help those who are currently ill without having influxes of new patients, and giving time to other health professionals to develop a vaccine. By the end of the semester, the team expects to have a neural net that can determine which individual factors have a more significant impact upon the virus transmission in various areas – characterized by counties. In addition, we aim to find a model that most accurately represents the current COVID-19 scenario.

REFERENCES

[1] Ben Charoenwong, Alan Kwan, and Vesa Pursiainen. 2020. Social connections to COVID-19-affected areas increase compliance with mobility restrictions. *Science Advances* (2020). https://doi.org/10.1126/sciadv.abc3054

[2] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Dawei Yin, and Yi Chang. 2020. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. arXiv:cs.LG/2001.06216

[3] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O'Banion. 2020. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. arXiv:cs.LG/2007.03113

[4] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR* abs/1609.02907 (2016). arXiv:1609.02907 http://arxiv.org/abs/1609.02907

[5] Parul Maheshwari and Réka Albert. 2020. Network model and analysis of the spread of Covid-19 with social distancing. arXiv:physics.soc-ph/2006.09189

[6] Nivedita Mukherji. 2020. The Social and Economic Factors Underlying the Incidence of COVID-19 Cases and Deaths in US Counties. arXiv:2020.05.04.20091041

[7] Lucia Russo, Cleo Anastassopoulou, Athanassios Tsakris, Gennaro Nicola Bifulco, Emilio Fortunato Campana, Gerardo Toraldo, and Constantinos Siettos. 2020. Tracing DAY-ZERO and Forecasting the COVID-19 Outbreak in Lombardy, Italy: A Compartmental Modelling and Numerical Optimization Approach. arXiv:2020.03.17.20037689

Proposal

Finding the Factors that Contribute Most to the Spread of COVID-19 Inside and Among US Counties

Mehdi Azabou
Georgia Tech
mazabou@gatech.edu

Tanya Churaman
Georgia Tech
tanyachuraman@gatech.edu

Kipp Morris
Georgia Tech
kmorris9@gatech.edu

1 ABSTRACT

Q1. What are you trying to do?
Our goal is, at the county level, to determine the extent to which factors such as the presence/absence of shelter-in-place orders, socioeconomic status, and mobility patterns influence the spread of COVID-19.
The other 7 Heilmeier questions (labelled Q2, Q3, etc.) are provided out of numerical sequence within the proposal's structure: introduction, literature survey, data, and plan of action. The instructor's additional questions (labelled E1, E2, etc.) are also within the proposal structure.

2 INTRODUCTION

E1. What is the problem you are solving?

The overall goal is to analyze different factors in the spread of COVID-19, such as the presence/absence of shelter-in-place orders, socioeconomic status, mobility patterns, etc., with the intent of disentangling these factors from each other and developing an idea of how much each factor contributes to the spread. We also plan to take mobility patterns between counties into account to see how the aforementioned factors combine with mobility to affect the spread. The primary result we are aiming for is a graph neural network that forecasts case counts, where the neural network's input is a graph with a node for each county that contains data about case counts and all or some of the factors mentioned in the previous paragraph. In addition to training the neural network on a combination of factors, we could try using only one factor (e.g. attempt to forecast using only the past case counts and some sort of socioeconomic metric). This would help determine which individual factors are most significant while also finding the model that makes the best predictions.

Q4. Who cares? What difference will it make?

In the grand scheme of the pandemic, everyone cares. This pandemic has had an effect upon everyone. For almost every person, they either know someone directly (e.g. family, friends) or indirectly (e.g. friends of friends, celebrities) who has gotten COVID-19. More specifically, health professionals and officials care. By understanding the different factors that affect the transmission of COVID-19 within and between counties, better mitigation strategies can be put into place. A one-size-fits-all approach might not be the best course of action for certain areas. Cognizance of the varying factors that affect the infection rate at a county level can allow for officials to understand the efficacy of mobility restrictions and how to implement better strategies that are best for both the citizens and the economy. Most importantly, lives can be saved.

3 LITERATURE SURVEY

Q2. What is the current practice, and what are its limits?

3.1 Forecasting COVID-19 Cases in Italy with a Compartmental SEIIRD Model

In this paper, Russo et al. [7] discuss an SEIIRD compartmental model that they developed and successfully fit to COVID case counts from February 21st to March 8th. It also did a good job of approximating case counts up to May 4th. The second I state represents people who are infected but asymptomatic to take into account that people who fall into that category are more likely to recover. The assumption in their model is that while people who are infected and have severe symptoms will either die or recover, those who are asymptomatic or have light symptoms will definitely recover. Their idea of using two I states to factor in an observed characteristic of the COVID spread (the fact that a moderate fraction of those who contract it will have light symptoms or none) is really interesting, and the model appears to have performed very well. However, it does not take into account socioeconomic factors or mobility factors, which is likely a major reason that it performed as well as it did for the time period that they used it for; Italy was under lockdown for the time period for which they were able to successfully approximate the case counts. For our project, we will consider implementing this model, using it to forecast case counts, and then using the results as a baseline of sorts to compare to the results we get with our graph neural network.

3.2 Social and Economic Factors

Mukherji [6] presents the idea of a "vulnerability index", a metric that represents how vulnerable a US county is to COVID based on socioeconomic factors, as a way of explaining the significant differences in COVID case occurrences between different states and counties. She calculated the vulnerability indexes using a dynamic panel regression over data from a 20-day period that she chose because it represents the most significant period of rapid COVID spread in the US.

She found that factors that positively correlate with case counts include median income (more economic activity would imply less strict social distancing measures) and the degree of economic inequality (represented using the Gini coefficient). Certain emographic factors turned out to be positively correlated to case counts as well.

While this paper does not provide any results about case counts, it does gives some really interesting results regarding the influence of socioeconomic factors on the spread of COVID. When we are determining which socioeconomic factors to use as independent variables in our model(s), we can use the results from this paper as a guide and also as a baseline to compare our results to.

3.3 Mobility Restrictions

In early March, many mobility restrictions were put in place in the US in response to COVID-19. Some of the more known restrictions are shelter-in-place, school closures, workplace closures, and social distancing guidelines. The goal was to reduce the amount of COVID cases. Very strict guidelines were shown to help contain the virus in Wuhan, China. However, that does not mean these social distancing guidelines were effective everywhere. Local governments may not have the resources to enforce these guidelines, meaning citizens must voluntarily take part in implementing these guidelines in their everyday lives [2].

3.3.1 Connectivity's Effect Upon Mobility Restrictions
. To analyze the effectiveness of mobility restrictions, [2] analyzed social connectedness between US counties and foreign countries by combining Facebook's county-level Social Connectedness Index (SCI) and SafeGraph's county-level mobility data via anonymized phone location statistics. A county-day panel of the social distancing, local cases, and exposure of COVID-information via social connections were combined with county characteristics to analyze the compliance of mobility restrictions. Social distancing metric was created to measure how much social distancing a county participates in per day (t represents the day, i represents the country).

$$Social\ Distancing = \frac{Completely\ Home_{it}}{Total\ Device\ Count_{it} - Working_{it}}$$

The results depicted that Italian and Chinese counties with high social connectedness implemented social distance guidelines upon mobility restrictions more quickly than those with low social connectedness. Social connectedness increased this compliance by 50%.

Counties with older-aged residents were less compliant with mobility guidelines; however, the counties with higher social connectedness had a higher compliance metric amongst this population. College-educated people had low responsiveness to mobility guidelines regardless of social connections. Conversely, less educated counties complied with restrictions in conjunction to the level of connectedness. Lastly, social connectedness did not have an impact upon the counties with health-conditions (e.g. obesity, diabetes, etc). These high-risk populations were more compliant with guidelines regardless of the level of social connectedness. Overall, this study suggests that the high connectedness within counties can allow for the flow of information between people to result in compliance with mobility restrictions, thus giving direction on how to enforce these life-saving guidelines.

While we are not researching the connectedness of counties, this study provides inspiration on how to use the mobility data in conjunction with shelter-in-place guidelines, socioeconomic status, and other variables to understand the spread of the virus within and between counties.

3.3.2 Efficacy Mobility Restrictions
. [5] focuses upon the effect of these mobility restrictions. Health officials can prepare strategies for the potential second wave of the infection. The goal would be able to minimize the effect of the pandemic and the economic impact.

Networks consisting of 10,000 nodes were created. Nodes represented people while edges depicted the non-zero probability of COVID being transmitted. Edges also captured the interactions resulting from essential activities and interactions. In each network 5 random nodes were initialized as infected; the rest were in the susceptible state. Via the simulation, nodes can take on one of these states: susceptible, infected, immune, and dead. The simulations emulated how COVID-19 would spread within the population under different mobility restrictions/mitigation strategies. Each mitigation simulation produced a probability of a second-wave happening and the number deaths occurring.

A longer lockdown period of 3 months was deemed the safest option – less of a chance of a second wave and less deaths; however, there was more of a negative economic impact. However, 2 months showed much higher chance of a second wave. Thus, a shorter period of lockdown is not beneficial. Therefore, a scenario with 2 months of lockdown plus strict social distancing measures was emulated. This scenario resulted in less deaths and a less chance of a second wave compared to the former. In addition, it does not have a dire economic impact compared to 3 months of lockdown. While it is important to understand the impacts of various shelter-in-place strategies, this study is a simulation. There are many assumptions – not including children, representing the population as a scale free network, etc. This model does not represent real-world scenarios and does not take into account of everyone following the mobility restrictions. A 3 month lockdown might not be as successful as the model determined if people are not following the mitigation guidelines – as shown what happened in the US.

3.4 Deep Graph Networks

Graphs are a representation which supports arbitrary relational structure, naturally lending themselves to the representation of multiple real-world systems. The power of deep networks has mainly come from leveraging the inherent structure of the data being manipulated. We can think of convolutional layers for images and recurrent layers for sequential data. Deep graph networks are used for a wide variety of tasks, including link prediction, graph classification, node segmentation, and recommendations. Gilmer et al. [1] extended the existing deep graph networks into a common framework. These models are called Message Passing Neural Networks (MPNNs). The core idea is not new; each node and/or edge is represented by a set of features $\mathbf{x}_i$, and then information is propagated through the graph in the form of "messages", by iteratively applying equation 1: each node i is going to receive information from its immediate neighbors $\mathcal{N}(i)$ through a learned message function $\phi^{(k)}$. These messages are aggregated through $\square$ which denotes a differentiable, permutation invariant function, e.g., sum, mean or max, and finally its representation at iteration (k) is updated using $\gamma^{(k)}$.

$$\mathbf{x}_i^{(k)} = \gamma^{(k)} \left(\mathbf{x}_i^{(k-1)}, \square_{j \in \mathcal{N}(i)} \, \phi^{(k)} \left(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}, \mathbf{e}_{j,i} \right) \right) \quad (1)$$

Intuitively, MPNNs can be seen as a generalization of the convolution operator to irregular domains. One of the appealing properties of graphs networks is that they are permutation invariant and are more predisposed to interpretability which is usually difficult with deep neural networks.

3.5 COVID-19 Case forecasting

In the case of COVID-19 forecasting, there are multiple factors that we might want to consider. These include historical infection data, inter-region human mobility, the prevalence of wearing masks and shelter-in-place orders, socioeconomic factors, etc.

With mechanistic approaches, where compartmental models have predefined transmission dynamics, or time series learning approaches, like autoregression or deep learning, forecasting is usually done within a "closed-system" location, thus not capturing meaningful spatial dynamics. This is where graph neural networks become interesting.

Kapoor et al. [4] utilize the deep graph network by modeling counties as a spatio-temporal graph with different types of edges. Spatial weighted edges represent mobility flows between nodes, whereas temporal edges represent connections to past days. The regression task that is being learned takes the historical data from the past 7 days and forecasts the number of cases for the next day. A visual of the graph can be seen in Fig 1.

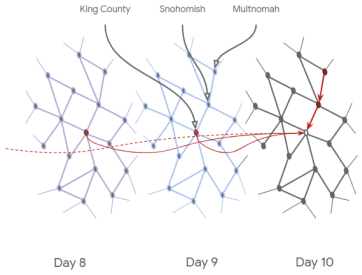


Figure 1: Spatio-Temporal graphs COVID-19 graph showing spatial and temporal edges across three days. (Figure taken from [4])

A skip-connections Model is then introduced for graph convolutions. The equations for this model [4] are shown in Eq. 2.

$$\begin{aligned} \mathbf{H_0} &= \text{mlp}\left(\mathbf{x_t} \,|\mathbf{x_{t-1}}| \ldots | \,\mathbf{x_{t-d}}\right) \\ \mathbf{H_{l+1}} &= \sigma\left(\hat{\mathbf{A}}\mathbf{H_l}\mathbf{W_l}\right) \,|\, \mathbf{H_0} \\ \mathbf{P} &= \text{mlp}\left(\mathbf{H_s}\right) \end{aligned} \tag{2}$$

where H represents the hidden state at layer l, $\hat{A}$ is the spectral normalized adjacency matrix, W is the learned weight matrix at layer l, $|$ is the concat operator, and σ is a nonlinearity.

The model is trained using data aggregated from three different sources: the NYT COVID-19 dataset, the Google COVID-19 Aggregated Mobility Research dataset, and the Google Community Mobility Reports. The learned model is shown to outperform multiple baselines. Metrics used to evaluate the different models include the RMSLE and Pearson correlations for the case deltas.

While this paper only uses mobility and case count features, such models can be extended to account for other factors.

3.6 Model interpretation

Deep networks are usually described as black boxes, so interpreting and understanding the salient features the model is basing its prediction on can be complicated. In our case, we want to use the model learned through the deep graph framework to have a way of interpreting which features are responsible for COVID-19 spread.

In the case of Graph Networks, a lot of work was done along these lines. Huang et al. [3] introduce GraphLIME, where LIME stands for Local Interpretable Model Explanations. GraphLIME is a model-agnostic explanation framework that, given a node that needs to be "explained", learns a local model in the subgraph of the node, thus finding the most representative features. Within the message passing framework, each node aggregates information from its neighbors, so it is important to consider both external information as well as the node attributes in identifying the features responsible for a given outcome.

A kernel-based nonlinear feature selection algorithm (Hilbert-Schmidt Independence Criterion Lasso) is applied on a N-hop neighborhood of the target node.

4 DATA

E4. What data will you use (how will you get it)?

The two primary datasets we plan to use are:

University of Maryland's COVID-19 Impact Analysis Platform County-Level Data (URL: https://data.covid.umd.edu) This data, which can be viewed for different time ranges through an interactive dashboard at the given link, would be our primary source of county-level case count data and socioeconomic metrics. The socioeconomic metrics in this dataset include % African American (in a county), % Hispanic American, median income, and more. It also includes some mobility metrics that we could use, such as trips/person, % of people staying home, etc.

Mehdi has emailed to ask for raw data files, which they state on their website that they are willing to provide, but we have not heard back yet. The website unfortunately does not provide the file format or the size of the file(s) that we would receive the data in, so we will have to wait to hear back before we know those details. If Dr. Prakash is able and willing to help us hear back faster or provide access to the data (since he might already have the data or may have a connection at UMD), we would appreciate it. Our backup plan is to use a different dataset as county-level data is widely available.

Google COVID-19 Community Mobility Reports
(URL: https://www.google.com/covid19/mobility/)
The UMD dataset contains limited mobility metrics, so this would be our main source of mobility data. For each county, it gives a metric representing the mobility to/from six different categories of places, where the specific metric it gives is a percent increase or decrease in mobility relative to the mobility on a baseline day.

In addition to CSV files containing the raw data, there are also PDF reports that summarize the data, allowing us to get an idea of what the data is like even before opening the raw data. The CSV data for the United States is about 43 MB in size.

5 PLAN OF ACTION

Q3. What's new in your approach and why do you think it will be successful?

The contribution of our approach is three fold. First we would be extending the model in [4] by including more features and more data

sources and possibly improving the accuracy of case forecasting. Second, we would be adding a layer of interpretability to the model where we would be disentangling all these different factors that might be driving factors in COVID-19 spread in a given county. The third contribution would be exploring the impact of socioeconomic disparity on COVID-19 infection rates. The fact that a deep graph architecture is already established makes this less daunting, as we would be building on solid grounds.

E2. Which algorithms/techniques/models do you plan to use/ develop?

First we will be conducting an exploratory data analysis study to familiarize ourselves with the data. Our goal will also be to establish simple baselines and highlight the shortcomings of these methods in accurately demixing and measuring the impact of different latent factors. No public code is available from [4], so we will implementing the network using the PyTorch, using information available in the paper about the training procedure and the optimal hyperparameters. We will extend the model to use other features specific to each county.

After implementing and training this network, we will use the explanation models from [3] to start getting information about relevant features, which will lead us to measuring the impact of socioeconomic factors across different counties.

E3. How will you evaluate your method? How will you test it? How will you measure success?

To evaluate our learned graph model, we will be following the same evaluation procedure as [4], i.e. we would use RMSLE and Pearson correlations for the case deltas. Measuring the success of the explanation model is more tricky, as we would have to rely on the consistency of these explanations across different counties. As we will work on this aspect of the project we will have to come up with a way of showing how the representation we are extracting are actually meaningful.

Q5. What are the risks and the payoffs?

Using deep neural networks is a risk itself. These networks can be viewed as a black box – we do not necessarily know how the network is forecasting the case counts. In addition, deep learning is a fairly new topic for some members of the group. There is the risk of trying to learn more introductory principles while completing this project. In addition, the team is trying to analyze different factors that contributed in the spread of COVID-19 – another daunting task. We will have to carefully analyze the relationship of each variable in conjunction with the infection and other variables as well. One payoff would be that it could help health officials come up with better strategies to contain the virus. By containing the virus, we are protecting the essential workers, giving time to help those who are currently ill without having influxes of new patients, and giving time to other health professionals to develop a vaccine.

Q8. What are the midterm and final "exams" to check for success? How will progress be measured?

Our midterm goal is to have implemented and trained the graph neural network, that includes having reproduced the results from [4] and training the model using additional features. Our final goal is "explaining" what the learned model is learning and conducting an analysis on the different features and how they contribute to COVID-19 spread.

E6. You must describe what portion of the project each team member will be expected to do. Include an expected timeline of activities.

The team has agreed to simultaneously work on the phases together – in order for each person to gain experience. For the first month, Kipp will lead the data preparation phase as well as establishing the baselines we will be working with. Mehdi will be responsible for network implementation and training. Tanya will be responsible for the exploratory data analysis.

The primary tasks for the second month will be training the network, analyzing the results (to go back to our goal of determining which factors play a role in the spread), and comparing the results to results from existing approaches (the baselines we discussed in the previous paragraph). Mehdi will be in charge of the network training, Tanya will be in charge of analyzing the results, and Kipp will be in charge of comparing our results to the baselines.

Q7. How long will it take?

The phases for this project are: Phase 1 – preparing the data and exploratory analysis of county data, Phase 2 – Preparing and Training the network, and Phase 3 – Interpreting and fixing network. Phase 1 should take 1-1.5 weeks. Phase 2 will take 3 weeks, and Phase 3 will take 2 weeks. The remaining time will be used to prepare project documents. Some steps might be worked on simultaneously.

E5. What do you expect to accomplish by the end of the semester?

By the end of the semester, the team expects to have a neural net that can determine which individual factors have a more significant impact upon the virus transmission in various areas – characterized by counties. In addition, we aim to find a model that most accurately represents the current COVID-19 scenario.

Q6. How much will it cost?

The cost is zero dollars. The team plans to use tools and data that are freely available – Google Colab and Google Docs. In terms of time cost, the computing power needed to train the neural networks might be significant.

REFERENCES

[1] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. arXiv:cs.LG/1806.01261

[2] Ben Charoenwong, Alan Kwan, and Vesa Pursiainen. 2020. Social connections to COVID-19-affected areas increase compliance with mobility restrictions. *Available at SSRN* (2020).

[3] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Dawei Yin, and Yi Chang. 2020. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. arXiv:cs.LG/2001.06216

[4] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O'Banion. 2020. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. arXiv:cs.LG/2007.03113

[5] Parul Maheshwari and Réka Albert. 2020. Network model and analysis of the spread of Covid-19 with social distancing.

[6] Nivedita Mukherji. 2020. The Social and Economic Factors Underlying the Incidence of COVID-19 Cases and Deaths in US Counties. arXiv:2020.05.04.20091041

[7] Lucia Russo, Cleo Anastassopoulou, Athanassios Tsakris, Gennaro Nicola Bifulco, Emilio Fortunato Campana, Gerardo Toraldo, and Constantinos Siettos. 2020. Tracing DAY-ZERO and Forecasting the COVID-19 Outbreak in Lombardy, Italy: A Compartmental Modelling and Numerical Optimization Approach. arXiv:2020.03.17.20037689