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/)

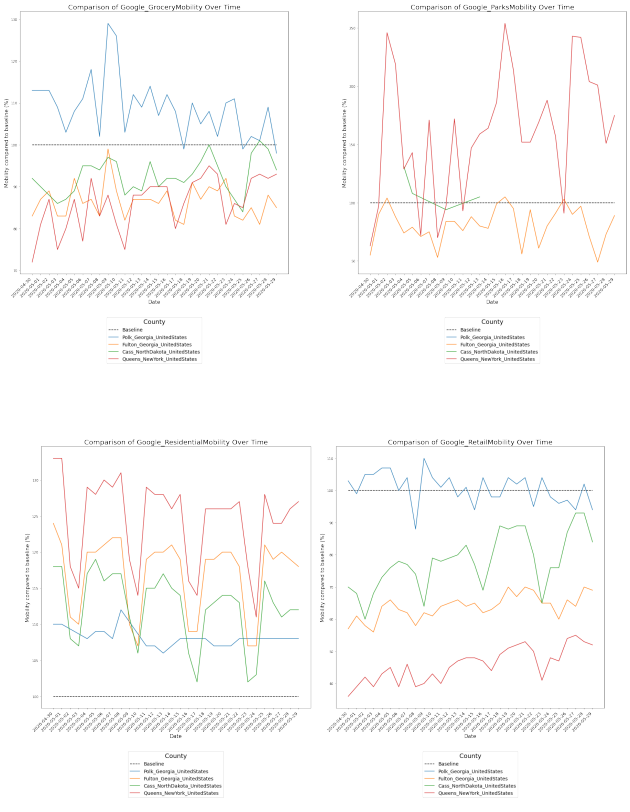
Normalized mobility trends were accessed via the C3.ai COVID-19 API. For each county, it gives a metric representing the mobility to/from six different categories of places, where the specific metric it gives is a percent increase or decrease in mobility relative to the mobility on a baseline day.

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.



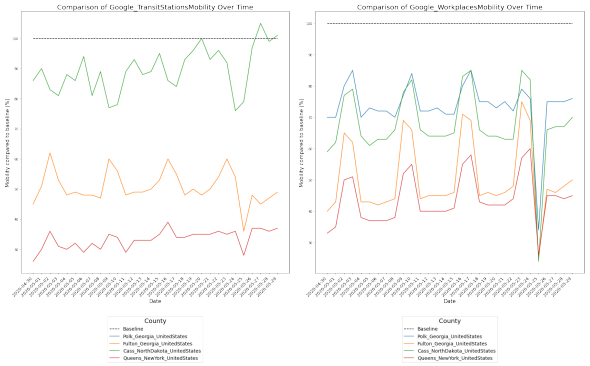


Figure 1: Mobility Pattern Graphs

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