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

Presented By:
Mehdi Azabou, Tanya Churaman and Kipp Morris

Introduction

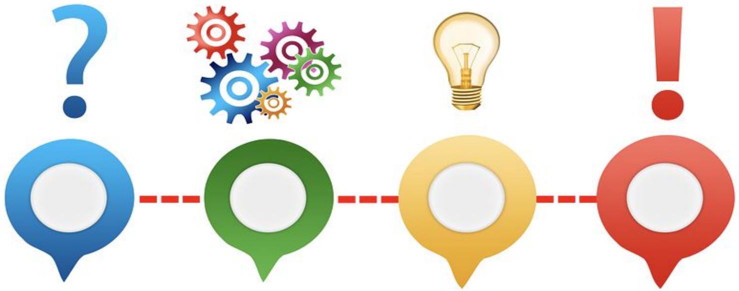
👨‍⚕️ Introduction and Problem Definitions 👩‍⚕️

- 💉 Analyze the contribution of various factors in the spread of COVID-19
  - Socioeconomic Status, Mobility Patterns. Demographics. Shelter-in-place Measures
- 💉 Determine the extent the presence/absence of these factors influence the spread of COVID-19.
  - Predicting case counts using a graph neural network model
  - Explain the network's predictions
    - Neuron activations
    - Identifying salient features
- 💉 Goal: A graph neural network to forecast case counts
  - Input: Graph in which the nodes represent the features
  - Based on aforementioned features

👩🏾‍⚕️ Importance 👨🏾‍⚕️

- 💉 **General Public**
  - Indirectly or Directly know someone affected by COVID-19
- 💉 **Health Professionals and Public Health Officials**
  - Development of better mitigation strategies
    - One-size-fit all approach may not work for all geographic area
  - Efficacy of mobility restrictions
  - Implementation of strategies best for both citizens and economy
  - **Save Lives**

Approach and Intuition



👩🏿‍⚕️ Our Approach 👨🏿‍⚕️

- 💉 Graph neural network
  - We have a baseline to compare to (Kapoor et al.)
  - Takes into account temporal and spatial relationships
  - Interpretable

- 💉 Determining which factors are most influential in the spread of COVID
  - GraphLIME: “Local Interpretable Model Explanations (Huang et al.)
  - Ablation
  - Does the importance of a given factor depend on the location? GNNs can help answer this

👩🏾‍⚕️ Previous work 👨🏾‍⚕️

- 💉 Kapoor et al. predict COVID-19 case counts using a variety of features from different data sources.
  - Historical case counts
  - Mobility data
  - Inter-county and Intra-county mobility flows.

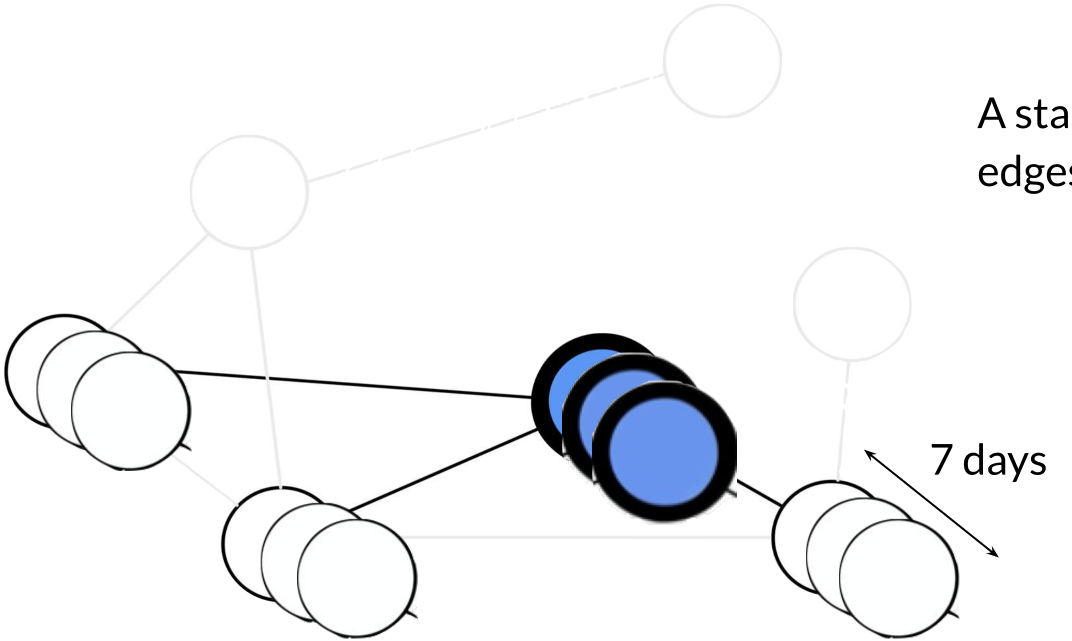
- 💉 They compute spatial and temporal features using a Spatio-Temporal Graph Neural Network.

- 💉 Message passing: Spatial nodes pass messages to each other through learned functions, which in our case would represent the spread of COVID between counties

👩🏻‍⚕️ Temporal convolutions 👨🏻‍⚕️

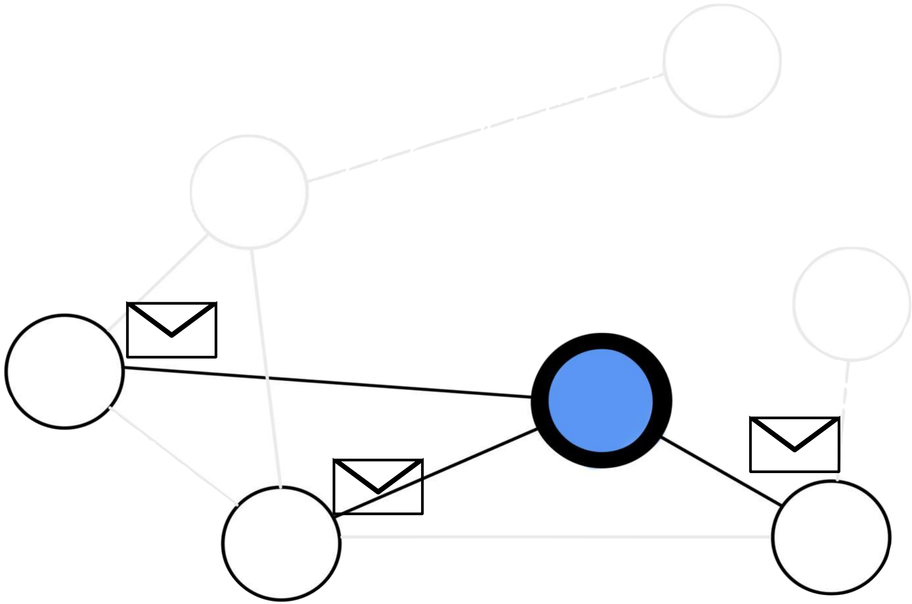
$$\mathbf{x}_i^{(1)} = \mathbf{mlp}(\mathbf{x}_{i,\mathbf{t}} | \mathbf{x}_{i,\mathbf{t-1}} | \ldots | \mathbf{x}_{i,\mathbf{t-d}})$$

A stack of graphs with temporal edges



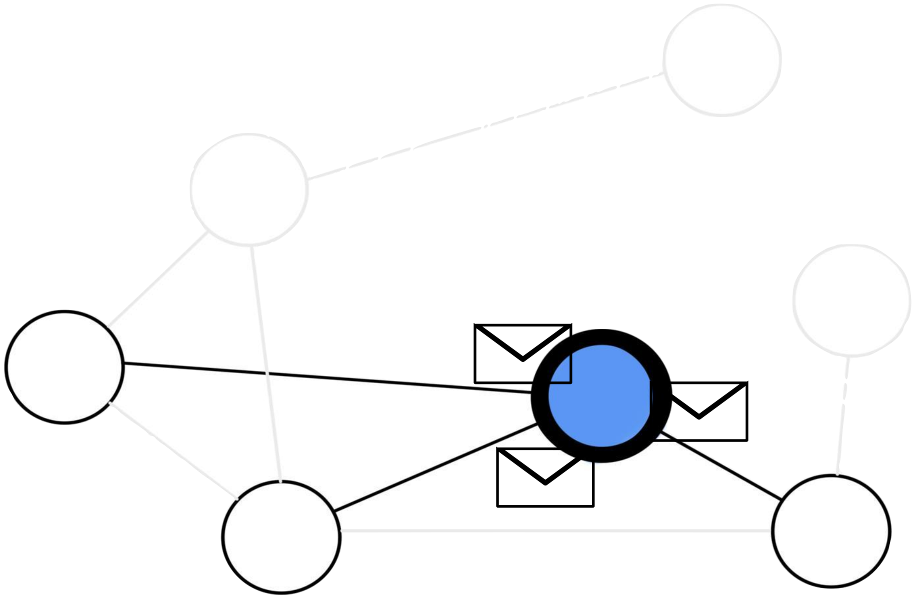
👩🏼‍⚕️ Message Passing 👨🏼‍⚕️

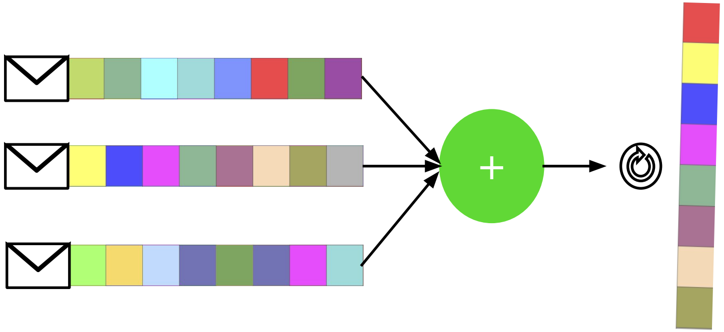
$$\phi^{(l)}\left(\mathbf{x}_i^{(l-1)}\right)$$



👨🏿‍⚕️ Message Passing 👩🏿‍⚕️

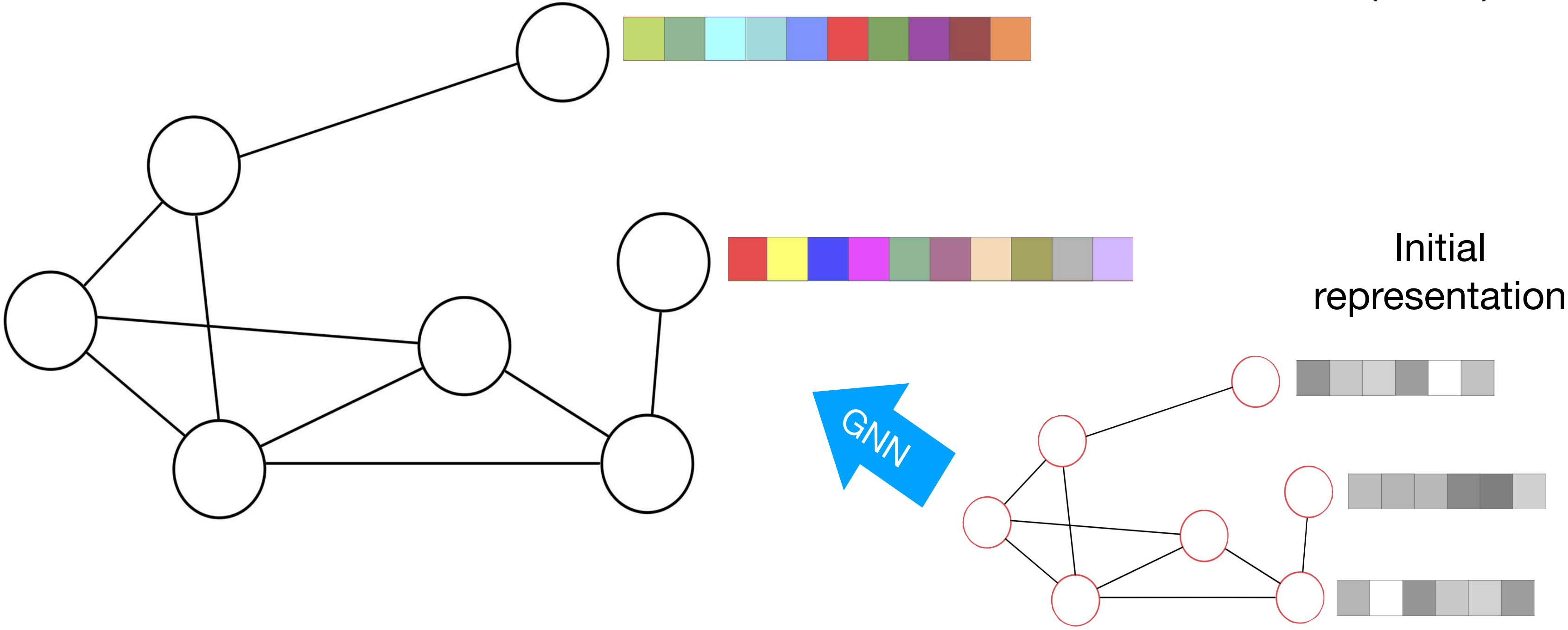
$$\mathbf{x}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \phi^{(l)}\left(\mathbf{x}_i^{(l-1)}\right)\right) \mid \mathbf{x}_i^{(0)}$$





👩🏾‍⚕️ Readout 👨🏾‍⚕️

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



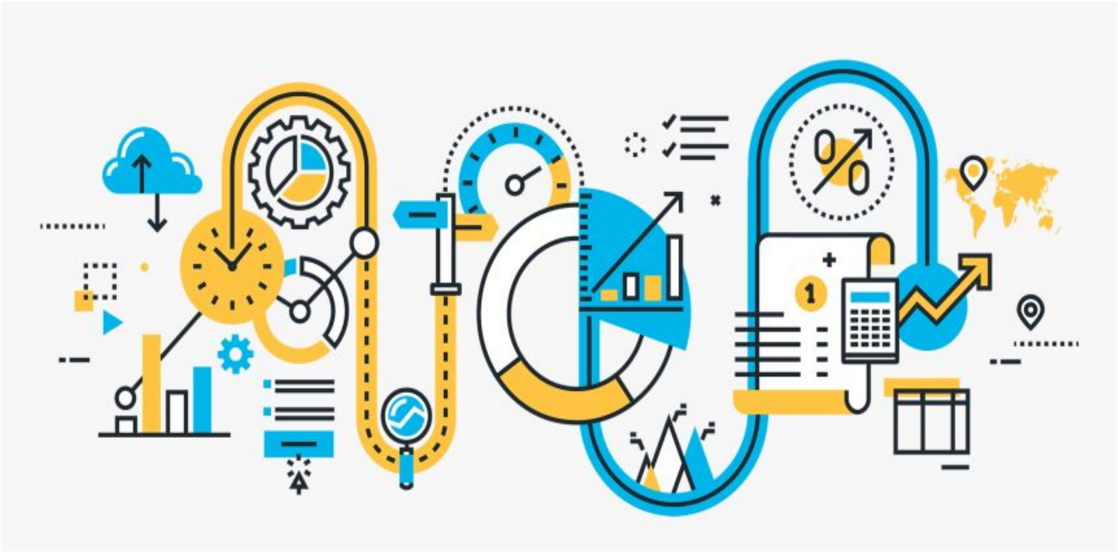
👨🏻‍⚕️ Our extension 👩🏻‍⚕️

- 💉 Adding better edges
  - Kapoor et al. use geographical proximity to define edges.
  - We use flow data between counties, thus accounting for aerial travel. We add weights to the edges as well.

- 💉 Adding static attributes
  - In addition to spatial and temporal features, we'd have static features, including population and other demographic data.

- 💉 Adding other features
  - Monthly unemployment rates.

Data Collection



👨‍⚕️ Data Overview 👩‍⚕️

- 💉 Training Dates: February 28th, 2020 to April 29th, 2020 (62 Days)
- 💉 Testing Dates: April 30th, 2020 to May 30th, 2020 (30 Days)
- 💉 Number of Counties: 2942 Counties
  - Originally: 3345 Counties
  - FIPS IDS to match
- 💉 Size: 5GB train / 3GB test

United States™ Census Bureau

C3.ai

👩🏾‍⚕️ Case Count Data 👨🏾‍⚕️

- 💉 Used C3ai's COVID-19 Data Lake API to pull the New York Times case count data

- 💉 Used FIPS county ids to join with our other data (kind of a headache; this is also where we lost those 299 counties)

- 💉 Predicted using the **rolling average** of the cumulative case counts

👨🏿‍⚕️Using the Rolling Average of the Case Counts 👩🏿‍⚕️

💉 The case count feature we used in prediction is the rolling average of the cumulative case counts over a 7-day window <u>by county</u>

Ex.

Fulton County cumulative case counts for each day in range April 1 - 7

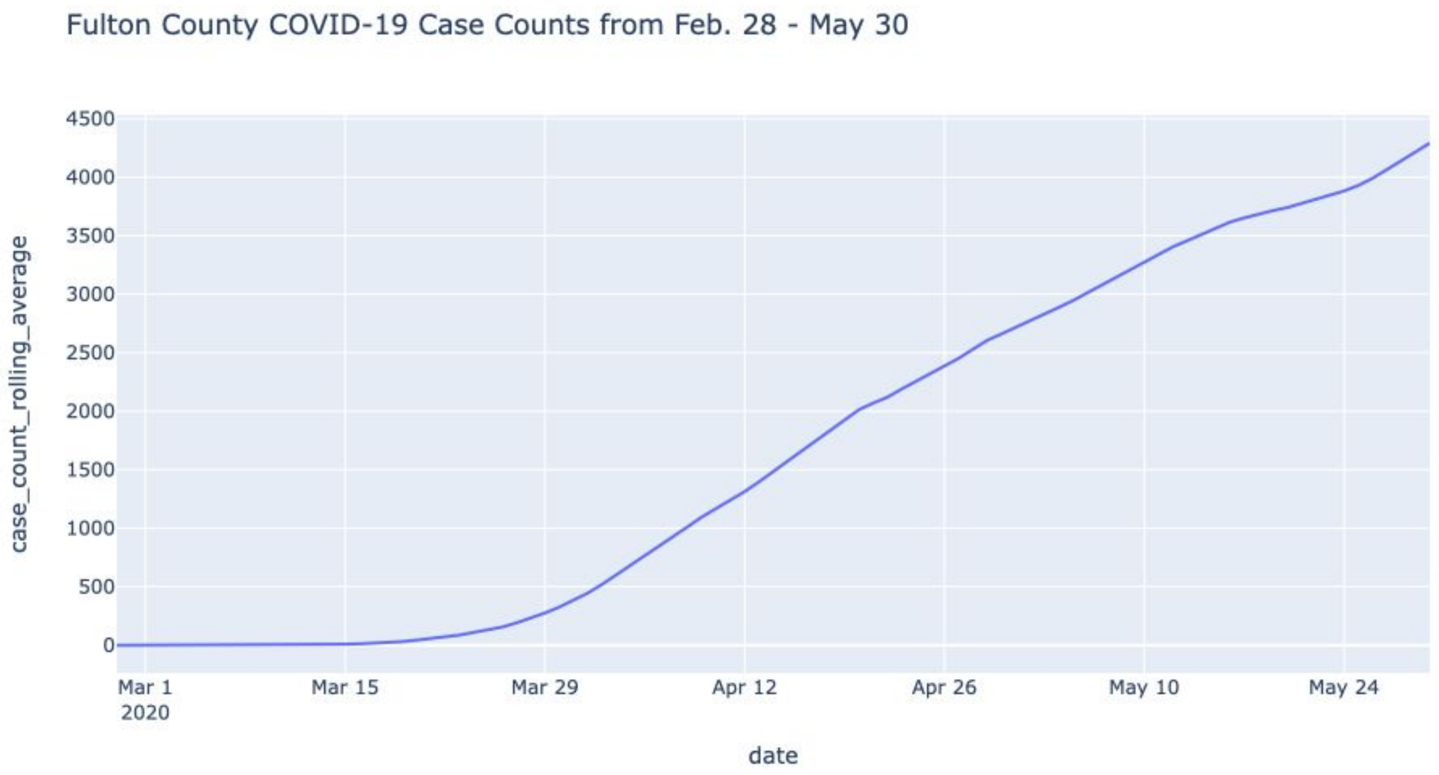
100
113
•
•
•
•
224

Average $\Longrightarrow$

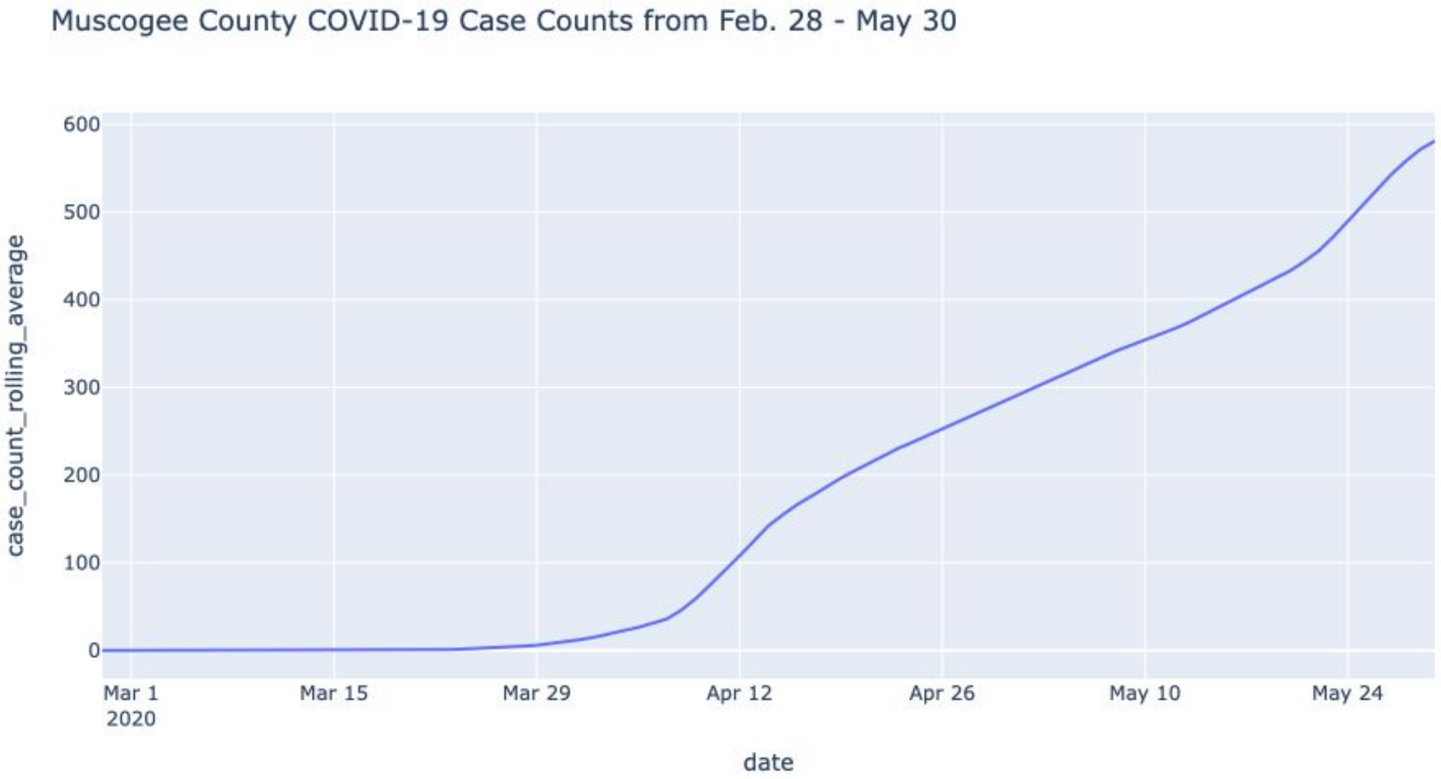
Prediction feature for Fulton County for April 7

154.5

👩🏼‍⚕️ Case Count Data Quick Look: Fulton



Case Count Data Quick Look: Muscogee County



👨‍⚕️ Census Data 👩‍⚕️

- 💉 Two population related features:
  - Total population
  - Population of age 65 or older (significant because they are at high risk)

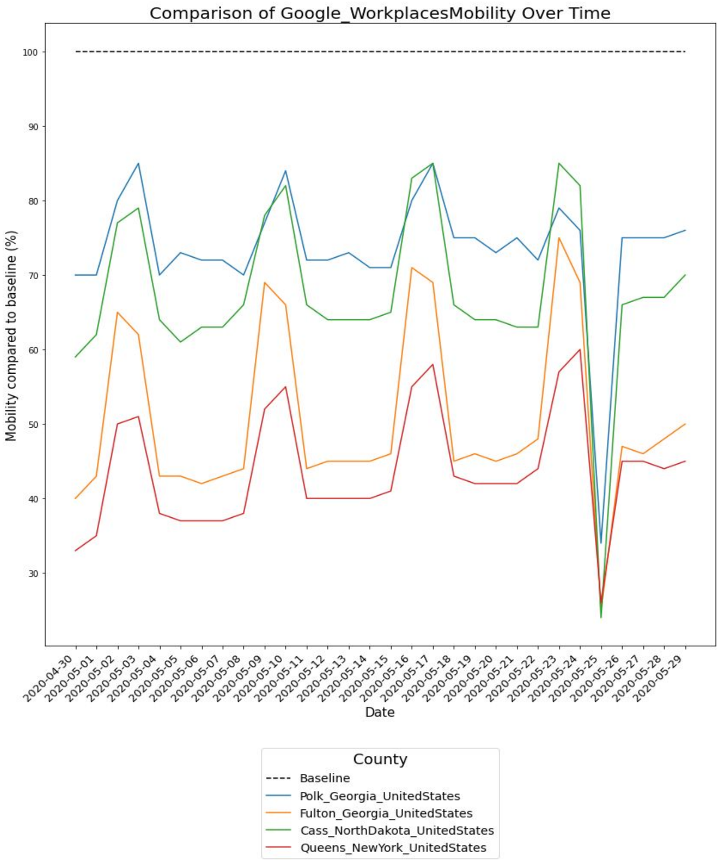
- 💉 Used as county-level static features

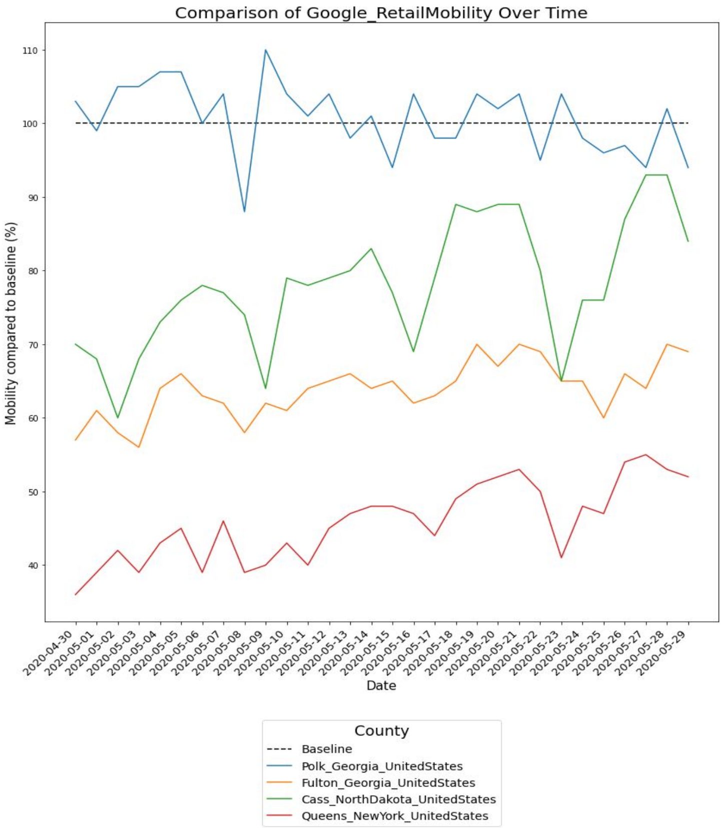
- 💉 Used 2019 population estimates

👨‍⚕️ Data Collection: Mobility Data 👩‍⚕️

- 💉 Normalized Mobility Trends from C3.ai COVID-19 API
- 💉 Features
  - Parks Mobility -- parks, beaches, plazas, gardens, etc.
  - Residential Mobility -- residences
  - Grocery Mobility -- grocery/farmer markets, drug stores, etc.
  - Transit Stations Mobility -- public transit, trains, etc.
  - Retail Mobility -- shops, restaurants, libraries, entertainment, etc.
  - Workplaces Mobility -- workplaces
- 💉 Metric value: percent increase/decrease in mobility relative baseline day
- 💉 Temporal: Daily Mobility for Counties
- 💉 Rolling Average Calculation

👩🏿‍⚕️ Data Collection: Mobility Data Cont 👨🏿‍⚕️

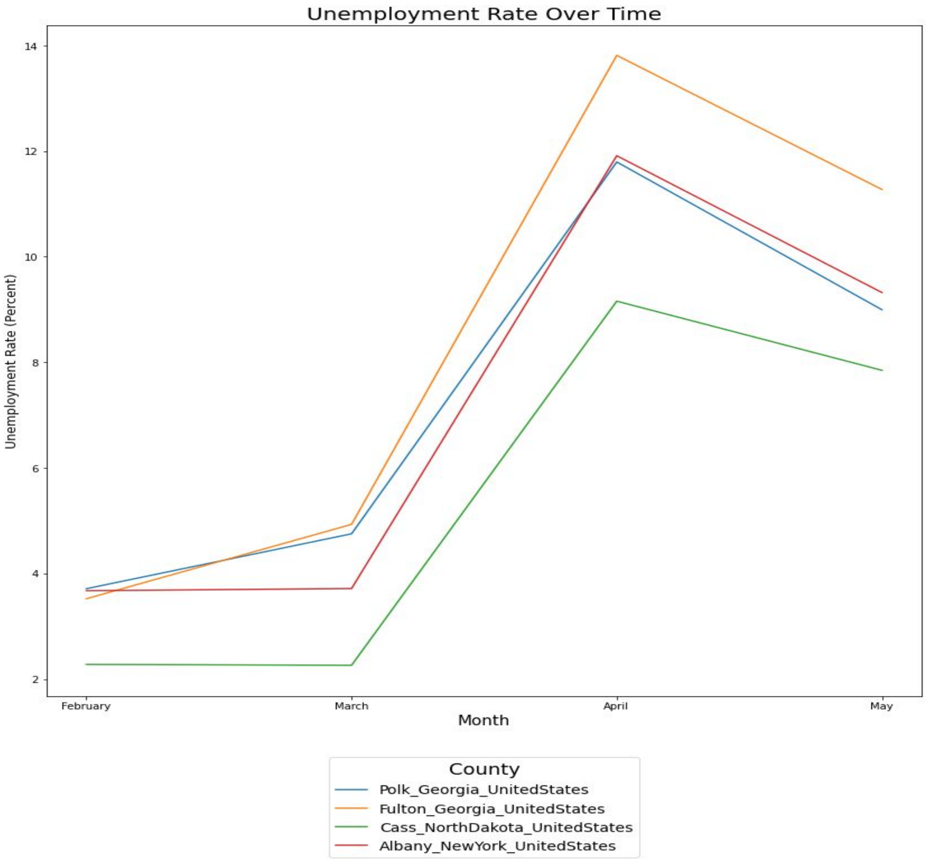




👨‍⚕️ Data Collection: Unemployment Data 👩‍⚕️

- 💉 Unemployment Rate from C3.ai COVID-19 API
  - Example of Socioeconomic Variable
- 💉 Metric value: Percent of unemployed population/total labor force population
  - Unemployed individuals -- ≥16 who had no employment and were seeking
- 💉 Temporal: Monthly Mobility for Counties
  - Train: February, March, April
  - Test: May

👩🏿‍⚕️ 👨🏿‍⚕️ Data Collection: Unemployment Data Cont



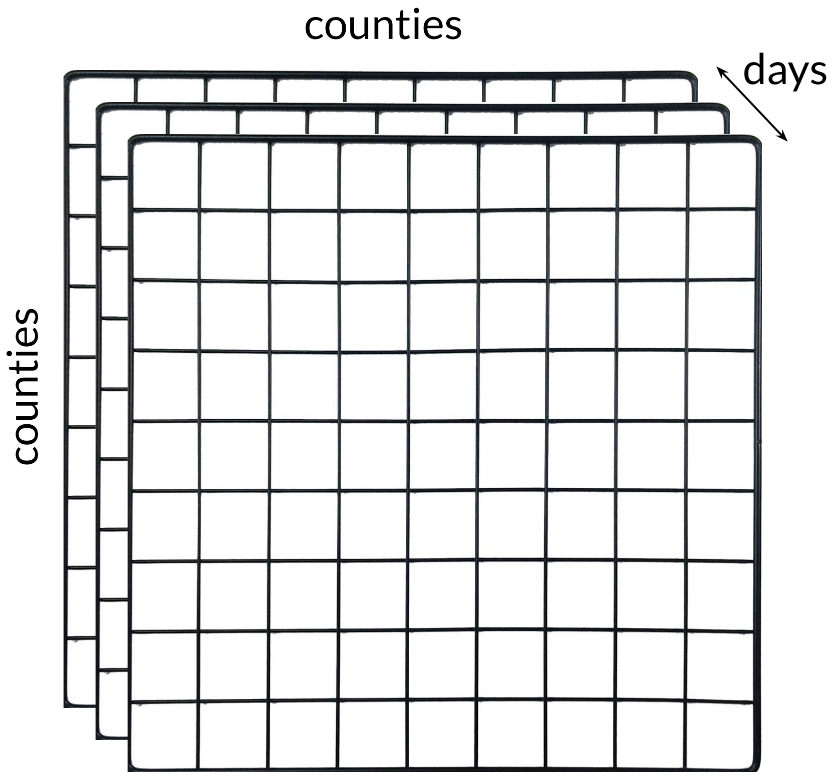
👩‍⚕️ Location Data 👨‍⚕️

- 💉 Mapping of county FIPS ids to latitude/longitude pairs of the spatial centers of the counties
- 💉 From a GitHub repo hosted by Benjamin Skinner, assistant professor at University of Florida
- 💉 Used to augment the mobility flow data we had from Google

👨🏾‍⚕️ Mobility Flow 👩🏾‍⚕️

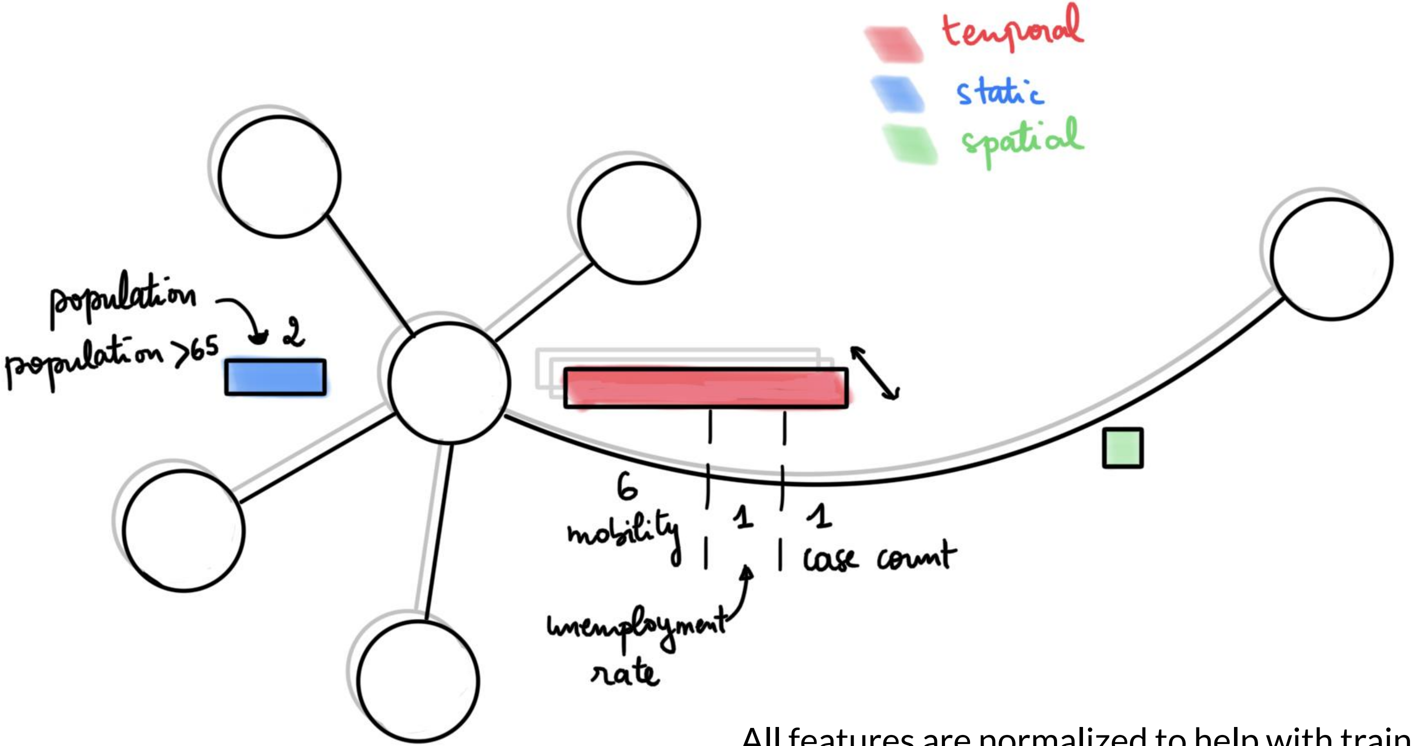
- 💉 Using the location data, we create edges between each county and its 16 closest neighbors.

- 💉 We add the daily County-level location exposure index: “Among smartphones that pinged in a given county today, what share of those devices pinged in each county at least once during the previous 14 days?”



Experiment & Results

👨🏽‍⚕️ Summary of graph 👩🏽‍⚕️



All features are normalized to help with training the network

👨‍⚕️ Model training 👩‍⚕️

PyTorch geometric

- 💉 We implement the dataset loader and network in PyTorch geometric which advantages include:
  - Support for graph batching
  - Support for message passing networks
- 💉 We train the network:
  - Using a batch size of 16
  - Using the Adam optimizer with a learning rate of 1e-3
  - Over 100k iterations

👩🏿‍⚕️ Baselines 👨🏿‍⚕️

- 💉 LSTM
  - Long Short Term Memory
  - RNNs that can learn long-term dependencies
- 💉 ARIMA
  - Auto Regressive Integrated Moving Average
  - Time Series Forecasting
    - Using the past to predict the future

👩🏻‍⚕️ What we are predicting 👨🏻‍⚕️

💉 Utilized on Test Data

- For each county independent over a 30-day period
- For each day, use 7 days prior to predict the case counts

👨‍⚕️Our Error Metric: RMSLE 👩‍⚕️

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

Key properties:

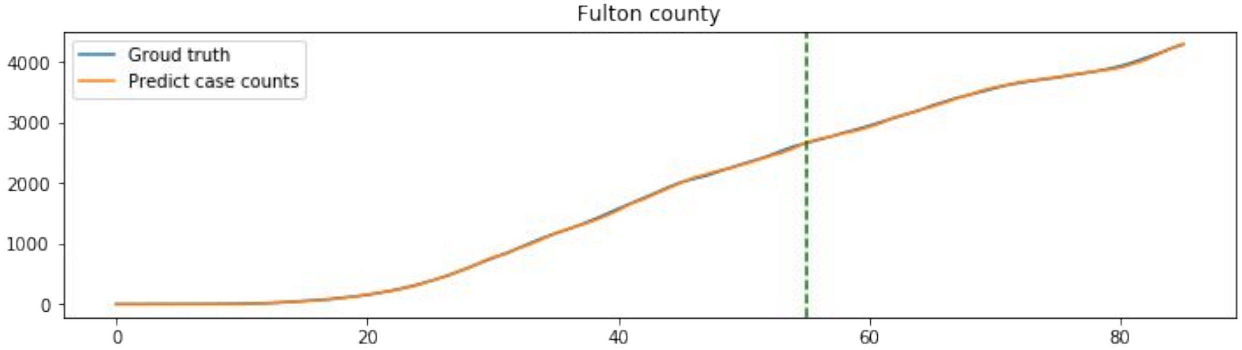
- 💉 Unlike RMSE, represents the relative error; the magnitude of a single error is not significant
- 💉 Punishes underestimation more than overestimation
  - Desirable for our use case!
  - Better to have too many ventilators, vaccines, etc. ready than too few

👨🏽‍⚕️ Results 👩🏽‍⚕️

	RMSLE (top 20)
ARIMA	0.0144
LSTM	0.0121
Kapoor et al.	0.0109
Our method	0.0080

👩🏽‍⚕️ Results 👨🏽‍⚕️

Cumulative Case Counts



New cases

Interpretability

👨🏿‍⚕️Ablation Studies 👩🏿‍⚕️

We remove features and re-train the model, to evaluate how important that feature was to the predictive power of the network.

	RMSLE (top 20)	RMSLE
Baseline	8.0e-3	0.013
No edge weights (mobility flow)	9.6e-3	0.030
No population features	7.7e-3	0.028
No unemployment features	8.8e-3	0.022

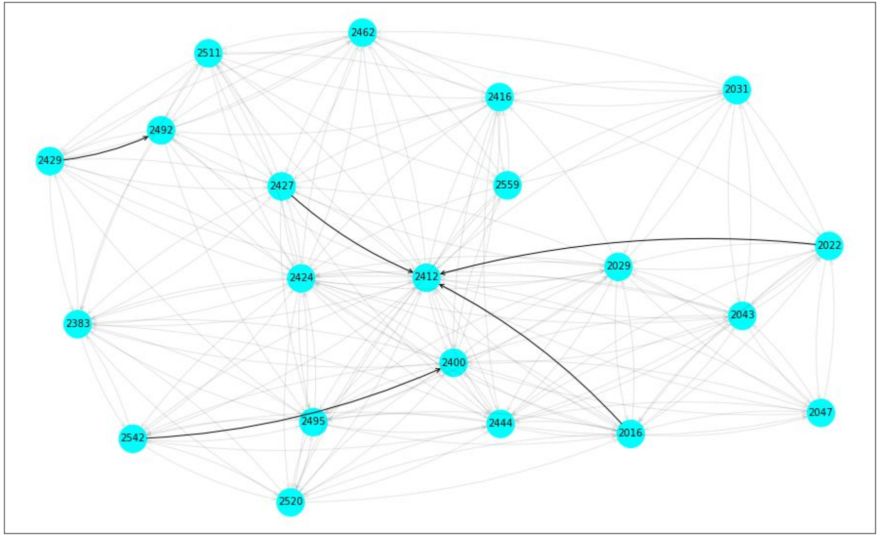
👩‍⚕️ GNN Explainer 👨‍⚕️

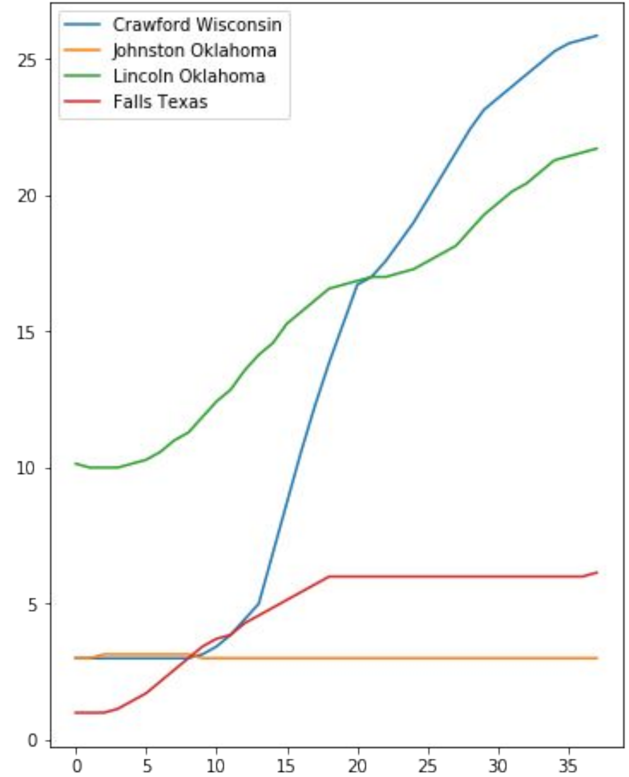
- 💉 GNNExplainer, is model agnostic, so we can use it with our network out of the box.
- 💉 It identifies compact subgraph structures and small subsets node features that play a crucial role in the network's predictions.
- 💉 For example, we can see whether the increase came from intra-county dynamics by looking a the node features or inter-county dynamics by looking at the messages passed between different counties.

GNNExplainer

In Crawford Wisconsin, on day 20, we notice a spike in case counts.

We find that this is mostly caused by the flow from other counties.

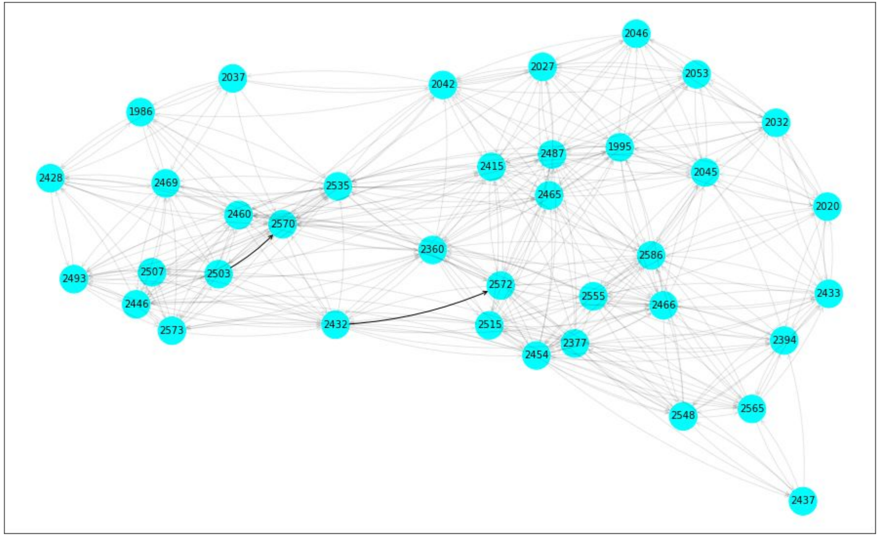


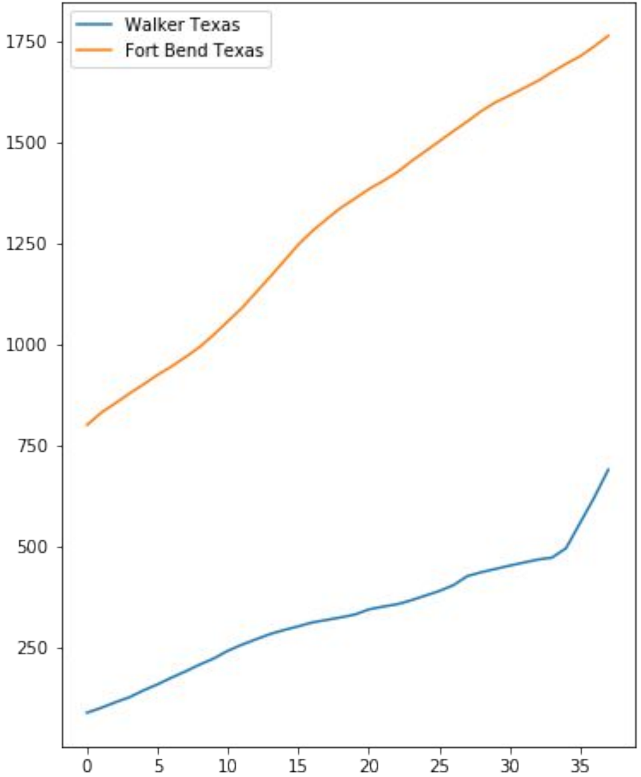


👩🏾‍⚕️ GNNExplainer 👨🏾‍⚕️

In Walker Texas, on day 35, we notice a spike in case counts.

The network predicts this spike using information from Fort Bend, Texas





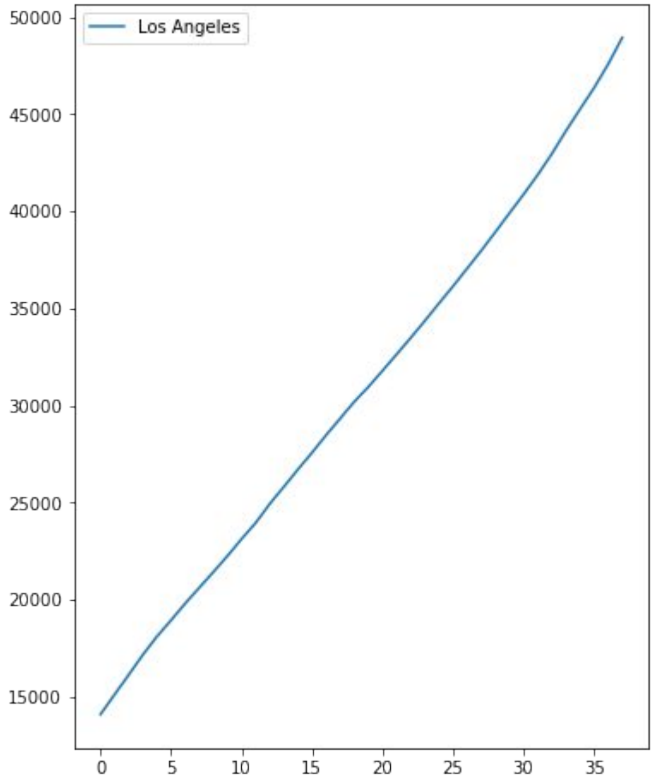
👨🏿‍⚕️ GNN Explainer 👩🏿‍⚕️

What is the feature that best explains the increase in cases in LA county?

The most important features are:

- Unemployment rate: 0.8479 (importance score)
- Previous day case count: 0.8323

There seems to be a high correlation between unemployment rate and case count.



👩🏾‍⚕️ Conclusions 👨🏾‍⚕️

- 💉 Improved upon the results from Kapoor et al.'s spatio-temporal graph neural network by about 20%
- 💉 Could spend more time looking at interpretability
- 💉 Would want to add in the missing 299 counties if we were to come back to this