Brown University (2023), , 1–26

BROWN

CLIMATE CHANGE AND MACHINE LEARNING

Natural Disasters and Human Migration in the US

Asghar, Zain,† Syrgkanis, Panos,† Lauerman, David,† and Lee, Mason†

Joint first authors

Abstract

Despite the profound alterations that climate change is inflicting upon our planet, the Intergovernmental Panel on Climate Change (IPCC) posits that its most consequential impact will be felt in the domain of human migration. The escalation of natural disasters has already spurred internal migration within the United States, a tendency that is projected to persist and intensify. The catalysts for such migrations can be both direct, stemming from a specific natural disaster, and indirect, driven by the looming threat of future disasters. Unfortunately, our grasp of post-disaster migration patterns remains tenuous, hampering our capacity to respond effectively. This deficiency is exemplified by the protracted states of emergency and sluggish policy implementations that have characterized numerous disaster responses, such as in New Orleans following Hurricane Katrina. By harnessing comprehensive governmental datasets pertaining to annual population movements between counties and county-level disaster declarations, we aim to elucidate the correlation between these phenomena and forecast future population shifts. This will enable states and counties to better anticipate and accommodate population influxes and outfluxes. We model the data with a graph structure, wherein counties are represented as nodes and population movements as edges. Our initial approach entails conducting network analyses on this expansive graph to discern baseline patterns and overarching trends. To generate the requisite population movement predictions, we must learn the inter-county relationships as well as the annual trends and their interactions with climate-related disasters. To accommodate these multifaceted factors, we employ three variations of Graph Neural Networks (GNNs): spatiotemporal message passing neural network (LSTM-MPNN), non-temporal MPNN, and temporal diffusion convolution, which facilitates the learning of patterns within the data and the formulation of predictions for the current or upcoming year. Through these methods we were able to establish qualitative migration differences and establish a pipeline which, with further work, could yield our desired results.

Keywords: Migration, Natural Disasters, Graph Neural Network

May 9th, 2023

1. Introduction

In the last few decades, climate-related migration has been increasing and projections have it only continuing to increase, with an estimated 200 million 'climate refugees' displaced by 2050 [1]. This includes not only the active displacement of people who are forced from their lands by extreme events, but also the people that are migrating because of decreased quality of life or uncertain future prospects with relation to the climate [2, 3]. These migration patterns are poorly understood and have been notoriously difficult to predict [1]. The real world implications of this poor understanding is seen in the prolonged 'states of emergency' and slow policy rollout for many natural disasters (e.g., New Orleans after Hurricane Katrina). If these migration patterns were well understood, especially if they could be predicted, areas could better prepare for the inward and outward flux of people thus allowing them to allocate resources and secure aid when necessary.

Prior research has investigated the relationship between climate change and migration, including Robinson et al.'s 2020 piece, Modeling migration patterns in the USA under sea level rise. Robinson was able to model migration relating to sea level rise to moderate success. Although they made progress in the right direction on the issue, they were still forced to admit that actual empirical estimates of the population movements remain 'elusive' [4]. That said, little work has been authored regarding natural disasters as they relate to empirical measures of population movements. This is the intention with our study– we do not intend to attribute specific natural disasters to migration patterns, but rather aim to investigate how natural disasters, broadly speaking, impact in and out migration.

To address this problem, we have partnered with the Brown Population Studies Center. They have access to an extensive IRS data set of annual migration at the US county level so we have decided to restrict our scope to purely internal migration in the US. To make our predictions on this domain we hope to use a Graph Neural Network [5]. Above all, a graph modeling of the problem is a natural fit. We are looking at a fixed number of counties that have certain fixed relationships with one another, and that each year either are or are not affected by climate disasters to some degree. These counties make up the nodes of our graph. The population flow from one county to the other will then constitute the edges of the graph, the value which we aim to model and predict. GNNs are also proven in this realm. They have been seen to be effective in modeling migration due to disease transmission. Ren et al. found that "GNN(s) can better learn the complex spatiotemporal dynamics of disease transmission." Given the complexity of the interplay between natural disasters and migration, we hope to have similar success.

There are a number of obstacles that we have kept in mind throughout the process. One concern with GNNs is a potential lack of extensive documentation, or that they may be generally difficult to work with. However, the hope has been that the natural fit of our problem to the framework would be enough to overcome any issues there. Another concern is the ability of our graph to pick up any meaningful patterns given our limited time scale. There are various compelling ways to deal with this issue but we were particularly drawn to a few. For one, we knew that our model should be recurrent so that it would be capable of remembering the complex and

evolving patterns while still staying open to new and emerging phenomena. This led us to an LSTM representation, which in turn brought us to the LSTM-MPNN and temporal diffusion convolution models. To further increase our chances of extracting meaningful patterns out of the limited data, we then segmented our graph into many subgraphs to give our model more time series to work with. We also decided to incorporate a non-temporal MPNN as a check to validate the contribution of the temporal elements of our other models.

Although we hope to make predictions on the data, we recognize the difficulty of achieving this. Accordingly, we implemented network analysis techniques to probe our data. Following similar network analysis studies, the work presented is largely qualitative, though it can inform further analyses or inspire future models [6].

Network analysis is a common practice to extract meaning from graph structures. It has been used to analyze transportation, social, and, more recently, migration networks. As Bilecen et al. have noted, recently, scholarly literature has shifted to recognizing the network-like aspects of migration patterns and applying network analysis accordingly [6].

Researchers have used network analysis to understand international and interregional migration patterns. Kemper analyzed the migration patterns between East and West Germany post-war [7]; Tranos et al. studied global migration patterns [8]; and Pitoski et al. studied the internal migration of Croatia in the last decade [9]. However, as Charyyev et al. have noted, though many international migration patterns have been studied, "US migration patterns have not been analyzed as a complex network." Charyyev's study was the first to do so–it used network analysis to investigate the differences in migration patterns between periods of economic boom and bust. Our work intends to build on this body of work by completing an analysis similar to Charyyev et al. but focusing on the impact of natural disasters on migration patterns [10].

2. Data

This project was based on two datasets. The first, the contribution of the Brown Population Studies Center is a thorough collection of inter-county migration from 1990-2011. This dataset has been compiled by the IRS by comparing tax returns from one year to the next since the 1970s, and has been released in conjunction with the US Census Bureau since 1990 [14]. It is uniquely valuable for its comprehensiveness and its geographic specificity, recording migration on a county scale. It is estimated that this dataset covers up to and potentially exceeding 87 percent of all households, approximately 95-98% of the tax-filing universe. The overwhelming majority of US householders file taxes, making this dataset very useful [15]. It is, however, worth mentioning that because the data is based solely on tax returns, those less likely to file tax returns are disproportionately underrepresented in the data, particularly the elderly, the poor, and to a lesser extent, undocumented individuals [16]. This data continues to be collected up to this day, but changed hands beginning in the 2011-2012 year from US Census Bureau management to IRS management. Systemic problems have since been noted in this portion of the data and are yet unaddressed, leading many to recommend that the data not be used, at least for the time being [17].

Once cleaned and organized, it is of the following structure for each of the

individual population exchanges that occurred between the 3143 counties in the US each year.

Dataset Structure:

- **i_fips**: refers to the sending county's full fips number, its unique 5 digit identifier composed of both the county identifier, the three digit **cfips**, and the state identifier, the 2 digit **sfips**
- **j_fips**: refers to the receiving county's fips.
- **mig**: the number of migrants, calculated by the IRS based on tax returns.

This data had to be cleaned and prepared before it could be studied and combined with our second data set. The data had to be combined by year to ensure no split entries and pruned to remove empty or incomplete entries. We also added a column for net migration, migrants in minus migrants out, for our county level dataframe. Altogether, there are $\sim 173,000$ rows of data for each year in the format listed above, a number which varies somewhat from year to year based on the variety of migrant routes.

The second dataset composed the natural disaster side of our investigation. We have decided to use the OpenFEMA Disaster Declarations Summaries dataset [18]. This FEMA dataset is composed of successful disaster declarations made at both the state and county level, along with some other information about the disasters themselves, including name and type. This data is ideal for our purposes because it is set up by county, and also uses the county 'fips', the same location identifying system as used by our IRS data. This creates a nice one to one for minimal information loss in the merging. It has data going back all the way to the 1960s, but because of our other data's more limited time scale we have separated out the data from 1990-2011. In total this comprises approximately 30,000 disaster declarations, generally about 800-2500 per year, with many of these declarations owing to just one disaster that affected multiple neighboring counties- recall that the dataset is split by county and not by disaster. Each disaster can affect anywhere from 1 to 20 or more counties, so the true number of separate disasters represented is actually a fraction of the total number of 'declarations', 3,117 to be exact [19]. This has no direct impact on our analyses but is important to bear in mind.

To convey some degree of disaster severity, we have built a one-hot encoding out of the four possible assistance programs that FEMA can declare after a disaster. These programs are the Individuals and Households program (IH), the Individual Assistance program (IA), the Public Assistance program (PA), and the Hazard Mitigation program (HM), and can only be declared a maximum of one time per disaster. Fortunately these programs seem to be relatively representative of severity, composing a good spread in their frequency of invocation- for instance, the IA program is rarely ever declared whereas the PA program is almost always. The hope is that these programs will serve as proxies for disaster severity. A disaster declaration with no program declarations is, relatively, barely registered, whereas a declaration that triggered all four programs implies extensive damage and an earnest response, or a serious disaster. These one-hots were summed up column wise for each county, in addition to one

extra column that is always one to simply tally disaster declarations per county per year. This makes up what would be each county's node representation in the graph, encoding the relative degree to which that county was affected by disaster in the year.

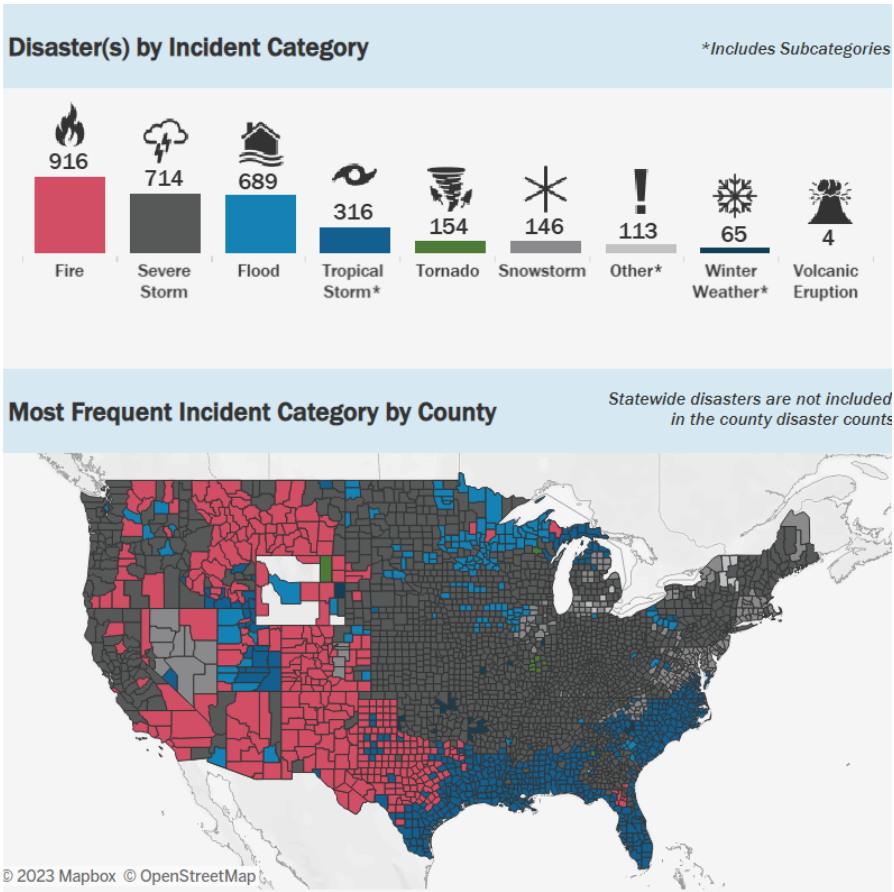

Figure 1. Credit to OpenFEMA dataset visualizer, [19]

3. Methodology

3.1 Spatio-Temporal Graphs

We represent our data using a graph structure $G(V, E)$, where nodes V represent counties with a series of features and edges E are directed edges connecting the nodes. Spatiotemporal graphs pose a variety of especially challenging problems where multiple features of the graph evolve differently throughout time. Additionally, the overall structure of the graph may change over time, necessitating models that can cope with such changes.

To model the spatio-temporal relationships between counties, we use a graph structure where each county is represented as a node with directed edges connecting

to other counties representing the migration. The node features x_i include a representation of the aggregated natural disasters for a county over the course of a given year, represented as a sum of one-hot arrays, describing the number of disasters as well as the number of times various disaster programs have been declared for the county (refer to Data section). Each directed edge (i, j) has only one feature, the number of migrants $\gamma_{i,j}$ that followed that particular path. The node features represent the actual migration values, not fluxes between years.

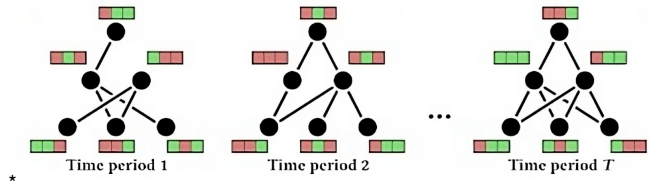

Figure 2. Overview of the Spatio Temporal Graph

To capture changing node features over time, we define the node feature matrix $X_t \in \mathbb{R}^{N \times F}$ at time t, where N is the number of nodes and F is the number of features. Similarly, we define the edge feature matrix $Y_t \in \mathbb{R}^{N \times N}$ at time t, where $Y_{i,j}$ represents the number of migrants from node i to node j at time t. The set of edges E_t can change between time steps as the number of migrants between counties can vary. Therefore, the set of edges at time t is defined as:

$$E_t = (i, j) \mid \gamma_{i,j} \neq 0 \tag{1}$$

where $\gamma_{i,j}$ is the number of migrants from node i to node j at time t. We use this edge set to define the adjacency matrix $A_t \in 0, 1^{N \times N}$ at time t as:

$$A_{t,i,j} = \begin{cases} 1 & \text{if } (i, j) \in E_t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

We chose to include both distance between counties and number of migrants as edge features as we expect distance to be important in predicting migration. Further, using distance for the nodes instead of absolute location enables an interpretation which does not depend on the exact location of the counties (which is necessary for the subgraph sampling, below).

3.2 Network Analysis

Given time constraints, we aim to use network analysis as a proof of concept. We narrowed our scope to compare state-level migration flux (normalized migration) after a particularly destructive natural disaster, 2006 migration post Hurricane Katrina, versus a baseline migration. Existing literature utilizes various practices to establish a baseline. Since we are following the work of Charyyev et al., we initially intended to use their method, the mean migration; however, to make our baseline more robust against outliers, we used median migration [10]. Put short, we aim to establish a

qualitative difference between our baseline of median migration from 1990–2011 and the migration in 2006. Following the work of Charyyev et al., this was explored via centrality analysis, community detection, and a gravity model [10]. For centrality analysis and clustering, networkX and iGraph Python packages were used. For the gravity model, sklearn's OLS was used.

3.2.1 Centrality Analysis

Centrality Analysis illustrates which nodes are central or important. There are several measures for calculating node centrality. Following the work of Charyyev et al. we used five centrality metrics to gain insight into the state-level migration patterns: (i) degree; (ii) hub; (iii) eigenvector; (iv) PageRank; and (v) betweenness centralities [10]. Following is a brief description of each metric and what they represent.

- *Degree centrality* is the number of edges a node has, divided by the number of nodes minus one. Given its inability to capture edge weights, it can be limited.
- *Betweenness centrality* ranks nodes based on how frequently they are on the shortest path among node pairs. Betweenness centrality captures dynamics where one node connects multiple subgraphs. In the context of this study, a high betweenness centrality implies that the state is on the path of most migrations.
- *Hub centrality* ranks nodes highly if they are connected to authority nodes, which receive links from hubs. An authority node has many high-ranked nodes linked to it, meaning it will have a high in-migration. A hub node links to many high-ranked nodes, meaning it will have a high out-migration. A high hub centrality corresponds to high out-migration, and vice versa for a low hub centrality.
- *Eigenvector centrality* is often considered a transitive metric for both directed and undirected graphs where high ranks transfer across adjacent nodes. A high rank implies that the node is adjacent to many other high-ranked nodes. In the context of migration, the patterns relevant to a high eigenvector centrality are unclear. Instead, a high eigenvector centrality means that the node is simply influential for the network whether that be high in-migration or out-migration.
- *PageRank* is a variant of eigenvector centrality that uses the in-degrees of directed graphs. In the context of state migration, a higher PageRank represents greater in-migration.

For a technical description of these metrics, the reader should check the networkX documentation [11]; for a deeper explanation of the importance of these metrics, the authors point the readers to Charyyev et al.

3.2.2 Community Detection

In accordance with the community detection methods that were used by Pitoski et al. [9] we used Louvain, FastGreedy, and InfoMap. Another modularity based approach was used, named Leading Eigenvector [12]. Other information theoretical community detection algorithms were explored such as Label Propagation and Walktrap [13]. Each of the algorithms was applied on our two datasets, the median migration and the 2006 migration, in order to observe any notable qualitative differences between the constructed communities.

Both Louvain and FastGreedy are modularity-based algorithms. Namely, they identify communities within networks by optimizing a modularity score, which quantifies the density of links within communities compared to the expected density of links in a random network. Both the Louvain and Fast Greedy algorithms iteratively optimize modularity. In each iteration, they first operate greedily, then construct a new network where nodes represent the communities formed in the previous step. This process is repeated until modularity converges to a max. Louvain differs in that its motivation is to identify hierarchical structures in large networks. The Leading Eigenvector algorithm is based on the spectral properties of the network's modularity matrix. By calculating the eigenvector associated with the largest eigenvalue of this matrix, the algorithm partitions the network into communities so as to maximize modularity.

The Label Propagation algorithm works by iteratively propagating node labels through the network. Initially, each node is assigned a unique label, and in each subsequent iteration, nodes adopt the most frequent label among their neighbors. The algorithm converges when the labels no longer change. Lastly, the Walktrap algorithm operates by using random walks to measure the similarity between nodes. It performs short random walks between node pairs and calculates their similarity based on the probability of a random walker transitioning between them. Nodes with high similarity are considered part of the same community, and an agglomerative hierarchical clustering approach is employed to merge communities based on this similarity metric.

3.2.3 Gravity Model

Following the work of Charyyev et al. and Pitoski et al., this paper aims to analyze the "individual impact" of factors by using a gravity model [9, 10]. The gravity model aims to estimate migration using the following relationship:

$$lnM_{st} = \delta + \Sigma_i^n \alpha_i ln(p_i)$$

where M_{st} is the migration flux from s to t (source and target). The sum is over the natural log of the following parameters: (i) source state population; (ii) target state population; (iii) distance from the source to the target state; (iv) number of natural disasters in the source state; (v) number of times the IH Program was declared in the source state; (vi) number of times the IA program was declared in the source state; (vii) number of times the PA program was declared in the source state; (viii) number of times the HM program was declared in the source state; (ix) number of natural disasters in the target state; (x) number of times the IH Program was declared in the target state; (xi) number of times the IA program was declared in the target state; (xii) number of times the PA program was declared in the target state; (xiii) and number of times the HM program was declared in the target state; (xiv) δ is an error term.

To improve results, the natural log of the parameters and migration flux are normalized between 0 and 1. The parameters α_i are estimated by Ordinary Least Squares (OLS) regression. The purpose of the gravity model is not to understand how well OLS fits the data or whether it can accurately predict new migrations; instead,

it aims to qualitatively understand the importance of each parameter in estimating migration flux.

3.3 Graph Neural Networks and Graph Convolutions

Graph Neural Networks (GNNs) are designed to process data represented as graphs. The two variants of GNNs employed in this study are MPNNs and GCNs. GCNs generalize the concept of Convolutional Neural Networks (CNNs) to graph-structured data. The core idea is to perform graph convolutions to learn local features from neighborhood nodes. Formally, given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with nodes $v_i \in \mathcal{V}$ and edges $(v_i, v_j) \in \mathcal{E}$, we define a graph convolution as [**?**]:

$$h_{v_i}^{(l+1)} = \sigma\left(W^{(l)} \cdot \sum_{v_j \in \mathcal{N}(v_i)} \frac{1}{c_{ij}} h_{v_j}^{(l)}\right), \tag{3}$$

where $h_{v_i}^{(l)}$ is the feature vector of node v_i at layer l, $W^{(l)}$ is the weight matrix at layer l, $\mathcal{N}(v_i)$ denotes the neighborhood of node v_i, c_{ij} is a normalization constant, and σ is the activation function.

3.3.1 Data Processing

To preprocess our data, we performed several steps to prepare it for training and evaluation:

3.3.1.1 Train-Test Split To evaluate the performance of our models, we split the data into train and test sets across time. The train test comprises es years 1990 to 2006 while the testing set comprises years 2006 to 2011. The features of the data frame are normalized per year.

3.3.1.2 Subgraph Creation To expand our dataset and allow our models to learn from different overlapping graphs, we created subgraphs using the NetworkX ego graph function. This resulted in batches of time series data, where each time series represents a subgraph centered around a randomly selected central node. This allows our models to learn from different local neighborhoods within the larger graph structure, potentially improving their ability to generalize and adapt to different edge cases.

The size of our dataset can be described as follows: we have a number of subgraphs, with each subgraph containing a certain number of nodes and features at each time step, over a certain number of time steps. Therefore, the input data for our models can be represented as a tensor of shape (number of subgraphs, number of timesteps, number of nodes, number of features).

This approach of creating subgraphs allows us to extract meaningful information from different local neighborhoods within the larger graph structure, potentially improving the robustness and generalizability of our models.

3.3.1.3 Neural Network Training All neural network training was done with WANDB, graphs were created with NetworkX, Spectral, PyTorch Geometric Temporal. Neural networks were built in PyTorch with PyG and TensorFlow with TensorFlow Graphs [].

3.3.2 Graph Neural Network Problem Structure

Our objective is to predict the total migration for a county in the next year. To achieve this, we employ GNNs for use in the training regime as described in Algorithm 1. The input to the algorithm consists of the train set M_{tr} and the test set M_{te}. The output is the learned set of parameters, θ.

Algorithm 1 Graph Neural Network Training Regime

Require: $M_{tr}, M_{te}, \alpha, \alpha_m, n$ epochs
Ensure: θ
Initialize θ randomly ▷ Randomly initialize the model parameters
for $D \in M_{tr}$ **do** ▷ Iterate through training datasets
  for $(T_r, T_e) \in D$ **do** ▷ Iterate through train and test task pairs in each dataset
    for Batch $b \in T_r$ **do** ▷ Iterate through batches of training data
      $\theta_t = \theta - \alpha \nabla_\theta L(f_\theta(b))$ ▷ Update temporary model parameters using gradients
    end for
    $\theta = \theta - \alpha_m \nabla_\theta L(f_{\theta_t}(T_e))/|M_{tr}|$ ▷ Update model parameters using gradients
  end for
end for
for $(T_r, T_e) \in M_{te}$ **do** ▷ Iterate through test dataset task pairs
  for Epoch $e \in n$ epochs **do** ▷ Iterate through epochs for training
    for Batch $b \in T_r$ **do** ▷ Iterate through batches of training data
      $\theta = \theta - \alpha \nabla_\theta L(f_\theta(b))$ ▷ Update model parameters using gradients
    end for
  end for
  $error+ = E(f_\theta(T_e))$ ▷ Compute the error on the test task using the current model parameters
end for
return $error/|M_{te}|$ ▷ Return the average error across all test tasks

The algorithm above, described by [21] is a general graph learning algorithm using gradient descent and batching of nodes. It initializes the parameters, θ, randomly and iterates through to minimize the loss on the task's train set and update the parameters. Next, the algorithm iterates through the test set to train, computing the error function E.

By using the Algorithm 1 in combination with graph neural networks and graph convolutions, recent work has been shown to effectively model the non-linear dynamics of spatio-temporal graphs. We employ three different types of graph neural networks in this study. GraphSAGE is a non-temporal message passing GNN, MPNN-LSTM is a message passing recurrent GNN, and DCNN is a graph convolutions network that uses neural network recurrence. These networks were chosen to study the effect of different node embedding mechanisms as reflections of dataset characteristics.

3.3.3 GraphSAGE

GraphSAGE [20] is a non-temporal graph neural network that employs message passing to learn node embeddings. It extends the conventional GNNs by enabling the generation of embeddings for previously unseen nodes through the aggregation of their neighborhood information. This is particularly beneficial for large and dynamic graphs where nodes can be added or removed over time. GraphSAGE operates as follows:

$$h_{v_i}^{(l+1)} = \sigma\left(W^{(l)} \cdot \text{AGGREGATE}\left(h_{v_j}^{(l)} | v_j \in \mathcal{N}(v_i)\right)\right), \tag{4}$$

where $h_{v_i}^{(l)}$ represents the node feature matrix at layer l, $\mathcal{N}(v_i)$ is the set of neighbors of node v_i, $W^{(l)}$ is the weight matrix at layer l, and AGGREGATE is a user-defined aggregation function such as mean, max, or sum.

3.3.4 Diffusion-Convolutional Neural Networks (DCNN)

Diffusion-convolutional neural networks (DCNNs) [22] utilize diffusion-based representations to learn from graph-structured data. We chose to implement a model that uses graph convolutions due to GCN's scalability, computational efficiency and varying node embedding. DCNNs offer several advantages, such as isomorphism-invariant latent representations and polynomial-time prediction and learning. The diffusion-convolution operation is defined as:

$$H^{(l+1)} = \sigma\left(D^{-1}AH^{(l)}W^{(l)}\right), \tag{5}$$

where $H^{(l)}$ denotes the node feature matrix at layer l, A is the adjacency matrix of the graph, D is the diagonal degree matrix, and $W^{(l)}$ is the weight matrix at layer l. Unlike MPNN-LSTM and GraphSAGE, which rely on fixed-size node embeddings, DCNNs can generate varying-sized embeddings based on the structure of the graph. This enables the model to better capture local and global information from the graph, leading to more accurate and informative embeddings.

3.3.5 Message Passing LSTM

The message passing LSTM model combines the idea of message passing in GNNs with the memory and learning capabilities of long short-term memory (LSTM) networks [21]. This approach allows the model to capture the long-range temporal dependencies in the time series data. Given a sequence of graphs $G(1), G(2), \ldots, G(T)$ corresponding to a sequence of dates, we utilize a message passing neural network (MPNN) at each time step, obtaining a sequence of representations $H(1), H(2), \ldots, H(T)$. These representations are then fed into an LSTM network:

$$h_{v_i}^{(t+1)} = \text{LSTM}\left(W^{(t)} \cdot \sum_{v_j \in \mathcal{N}(v_i)} \frac{1}{c_{ij}} h_{v_j}^{(t)}\right). \tag{6}$$

We expect the hidden states of the LSTM to capture the spreading dynamics based on the mobility information encoded into the node representations. A stack of two LSTM layers is used in this architecture. The new representations of the regions correspond to the hidden state of the last time step of the second LSTM layer. These representations are then passed on to an output layer similar to the MPNN, along with the initial features for each time step.

4. Results

4.1 Network Analysis

4.1.1 Centrality Analysis

Many of the states have similar centralities in 2006 and the baseline. The betweenness centrality seems to show about 10 states having different scores between 2006 and the baseline. The hub centrality is much higher for Florida in 2006 than for the median. Also notable is that DC's centrality is much lower in 2006. In plain words, there were more out-migrants for Florida in 2006 than the baseline and there were less out-migrants for DC in 2006 than the baseline. From the eigenvector and PageRank centralities, the migration for Texas, Florida, Georgia, and Washington differ between 2006 and the baseline. From the PageRank centrality, Texas and Florida have higher in-migration in the baseline than in 2006, and vice versa for Washington and Georgia. Existing studies (namely Charyyev et al.) do not establish the significance of these differences, so it is difficult to gauge whether these changes are significant [10]. As such, the description of these results will stay qualitative and limited to the above remarks.

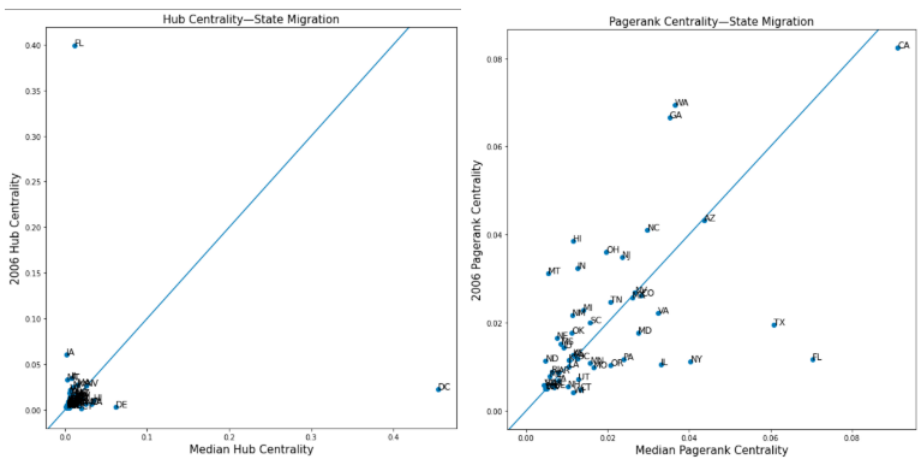

Figure 3. Hub and Pagerank centrality for state level migration flux (2006 migration vs. baseline migration). All other centrality metrics shown in appendix 1.

4.1.2 Community Detection

The Louvain community detection algorithm detected a larger number of well-defined communities than the rest of the algorithms. In particular, when applying InfoMap, the entire state-level map was assigned into one cluster, which could either be due to an implementation error of the algorithm by the package (iGraph) or by the nature of the algorithm itself. In a few of the community detection methods, such as Louvain in Figure 4, we observe the state of Florida being grouped with more northeastern states in 2006 than in the baseline.

The rest of the graphs can be found in Appendix 2.

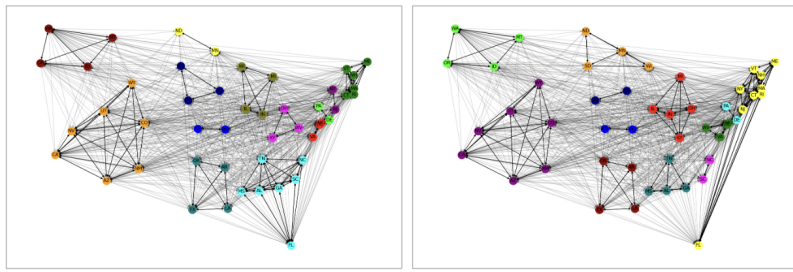

Figure 4. Louvain Community detection for median (left) and 2006 (right) migration

4.1.3 Gravity Model

The OLS results are in the table below. The source population, target population, and distance remain at the same level of importance for the baseline and 2006. However, the coefficients related to natural disasters are different between the median and 2006 migrations. The importance of the IH programs was most striking. For both the source and target states, the importance of the IH programs in 2006 is $\approx$ 15 orders of magnitude greater than the importance of the same factor in the median migration. The number of natural disasters in both the source and target states for 2006 is more than two times as important as for the median migration (note that this is on a logarithmic scale). The rest of the impact seems non-uniform—coefficients change but with no discernible pattern.

4.2 GNN

Despite the additional different data-augmentation and regularization parameters, none of the three GNNs were able to adequately learn patterns in the migration data. When a GNN is trained on data that lacks clear patterns, relationships, or features that can effectively distinguish between different samples, the model may fail to capture any meaningful information. As seen in this study, the GNN may resort to predicting the mean or a constant value as its output. Predicting the mean often minimizes the error in the absence of any other discernable patterns. It is important to note that the loss and validation loss scores represent cumulative sums over the graph over an epoch. Per node, the error is on average around +/- 200 people per county. Due to lack of

	Median Migration Coefficient	2006 Migration Coefficient
Distance	-0.0625	-0.0715
Source Figures		
Population	-0.0143	-0.0185
Number of Disasters	-0.0111	-0.0400
IH Declarations	1.363e-18	0.0014
IA Declarations	0.0061	0.0014
PA Declarations	-0.0085	0.0506
HM Declarations	0.0028	-0.0080
Target Figures		
Population	0.0274	0.0213
Number of Disasters	-0.0146	-0.0336
IH Declarations	6.146ee-18	-0.0002
IA Declarations	0.0005	-0.0002
PA Declarations	0.0028	0.0586
HM Declarations	0.0035	-0.0124
Error	0.0589	0.0539

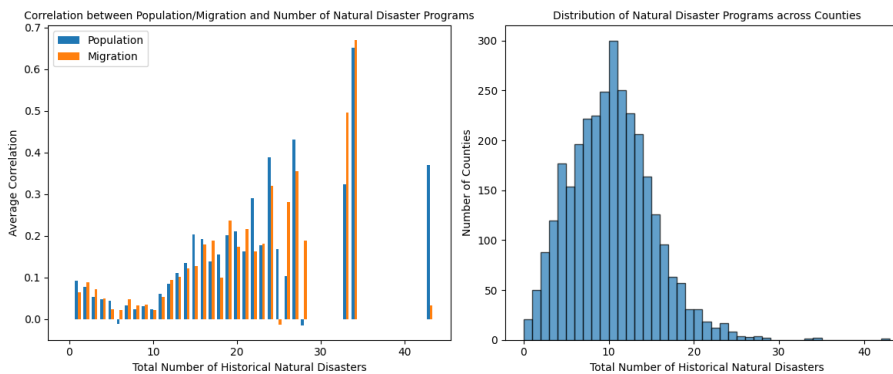

Figure 5. Average correlation between the number of natural disaster programs and both population and migration for each county over the 22 year study period. The x-axis represents the number of natural disasters, while the y-axis displays the average correlation values. The graph consists of two sets of bars: one for population (blue) and the other for migration (orange).

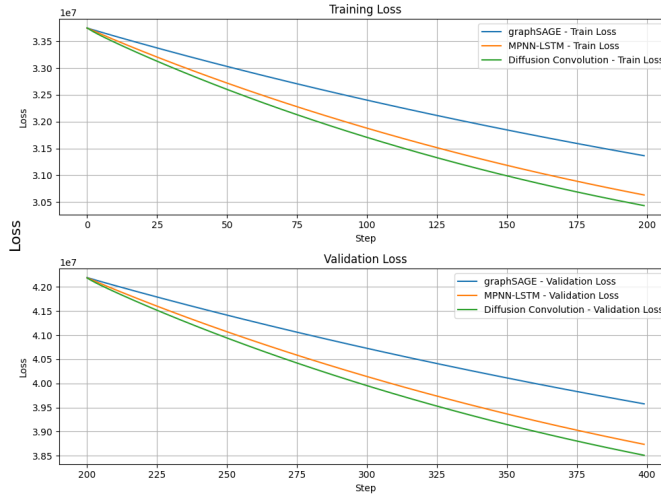

Figure 6. Training and Validation Loss Curves for Graph Neural Networks.

informative results and consistent mean predictions, we do not show the results of the GNN. On-going work will address the shortcomings of this method.

5. Discussion

5.1 Network Analysis

As mentioned in the results, the state of Florida is grouped with more northeastern states in 2006 than in the median. That could potentially point towards a migration flow switch due to Hurricane Katrina which affected the Southeastern US. It would be interesting to see a county-level comparison of the two datasets, baseline and 2006, in those afflicted areas as well, to see what the more fine-level consequences of the natural disaster were. There is also evidence of the US Southeast being affected in the centrality analysis.

Another interesting observation to be made is concerning the gravity model. Specifically in year 2006, the coefficient of IH programs established was orders of magnitude greater than the median. Given that the coefficients are on a log scale, we can also note that the number of disasters was also an important factor in determining migration. Broadly, it seems that natural disasters are more important for predicting migration in 2006 than the baseline.

There are a number of limitations to our methods. Perhaps most salient are the lack of robustness, the inabiliity to establish significance, and an inability to attribute changes to a specific factor. First, as Charyyev et al. mention in their paper, some network analysis techniques are not robust enough to withstand natural yearly variations- for instance, node centrality naturally fluctuating over time. In this regard, the merit of our work is difficult to gauge–did we pick out a true difference or was it simply a yearly fluctuation? In a similar respect, we are unable to establish significance for these metrics. As mentioned, network analysis is often used for qualitative analyses

without establishing significance. Finally, since our data is annual and migration patterns are highly complex and impacted by myriad factors, it's difficult to attribute our changes to natural disasters.

5.2 GNN

An intriguing finding from 5 is the absence of negative correlations in population, which contradicts the hypothesis that people will tend to leave disaster prone areas. The Population Studies Lab hypothesized that the frequency of natural disasters would have a positive correlation with migration and a negative correlation with population. However, our observations indicate that as the cumulative number of historical natural disasters increases in a particular county, the average correlation across all counties with both population and outward migration also increases.

This suggests that natural disasters influence population and migration in regions with exceptionally high frequencies of such events. Nonetheless, only a small portion of the regions falls within this category. In other words, our findings reveal that natural disasters do not have the expected adverse impact on population growth, and they do not necessarily drive people away from affected areas in most cases. Instead, natural disasters appear to have a more significant effect on population and migration dynamics in regions where their occurrence is extremely frequent. This unexpected result invites further exploration and investigation to uncover the underlying factors and reasons for this observed relationship.

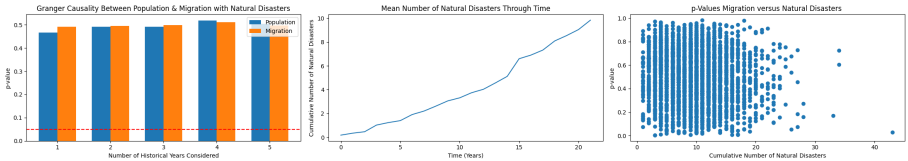

Figure 7. Granger causality test results for the relationship between population change, migration, and natural disasters. The figure includes three subplots: the first subplot shows the Granger causality test results for both population and migration with respect to natural disasters. The red dashed line at y=0.05 indicates the significance threshold for the test. The second subplot depicts the average cumulative natural disaster count per county across our 22 study period. The third subplot shows the Granger causality p-values of migration-natural disaster for each county plotted against the cumulative number of natural disasters.

This suggests that factors other than natural disasters may have a more significant influence on population and migration patterns. The increasing correlation could be attributed to the broader impacts of natural disasters on the economy, infrastructure, or social dynamics in affected areas. Furthermore, it might indicate that people are more inclined to move to or remain in areas with disaster relief programs, as these programs could provide a sense of security or stability. The influx of contractors and disaster relief workers could also offset outflows.

To explore the potential use of natural disasters as a feature in predicting population, we applied the Granger causality technique to analyze historical natural disaster data from up to five years in the past 7. Granger causality is a widely used time series

analysis method for evaluating the predictive power of specific random variables. We chose a maximum lag-time of five years based on our LSTM window.

Our results, averaged over all counties in the United States, indicate that natural disasters alone are not sufficient for prediction. Supporting this finding, the scatter plot between Granger causality p-values and cumulative natural disasters shows no clear correlation. Counterintuitively, this demonstrates that natural disasters may not be useful signals, even in regions with many significant events, as they do not necessarily correlate well. In fact, statistical signficance decreases as the historical number of natural disasters increasing, which reinforces the theory that only recent events will cause population fluxes.

Unaccounted confounding variables, such as economic conditions, political stability, or geographical factors, might play more significant roles than natural disasters. Non-linear relationships, sample size limitations, and insufficient lag-time could also account for the low statistical significance observed. It is important to note that correlation does not necessarily imply causation, and further research would be needed to investigate the underlying reasons for these observed correlations. However, these graph provides valuable insights into the complex relationship between natural disasters and population/migration patterns and can serve as a starting point for future studies aiming to understand these relationships more deeply.

6. Conclusion

In this study, we endeavored to investigate the complex relationship between natural disasters and population migration patterns. To gain a holistic understanding of the problem, we employed two separate methodologies–network analysis and graph neural networks. The network analysis was applied primarily as a proof of concept. In this regard, we were able to recognize qualitative differences between the state migration patterns post Hurricane Katrina and a baseline migration. While our GNN findings may not have yielded statistically significant results, we believe that the pipeline construction, problem-specification, and model architectures proposed in this project have provided valuable insights and established a foundation for future research in this field.

State-level migration network analysis shed light on the relationship between natural disasters and population migration. This exploration could lead to future analyses which could be of interest. In particular, completing a similar network analysis at the county-level could yield more explainable results. Further, expanding our analysis to encompass all years of our data instead of attempting to recognize a difference between a baseline and a specific natural disaster instance is an obvious extension of our work. Finally, since we were able to establish some qualitative difference (especially in the US southeast), it would be prudent to apply quantitative methods to gauge the significance and magnitude of these changes.

Moreover, our study has contributed to the ongoing dialogue surrounding the impact of climate change on human migration patterns. By examining the subject through the lens of natural disasters, we have helped to contextualize the challenges and limitations of current approaches. Our work also serves as a stepping stone for other researchers to build upon, refining and expanding our understanding of

this critical issue. This exploration has led to the identification of potential areas for improvement in our models and data processing techniques, which can be addressed in future research endeavors. Some obvious improvements for future studies include:

- Addressing limitations in our migration data due to processing errors in the IRS. With additional historical data, we would be able to train our models more effectively and potentially achieve more accurate results.
- Expanding our data scope to include specific population characteristics, such as race, gender, and socioeconomic status. Incorporating these factors could provide deeper insights into the underlying drivers of migration patterns and enable a more nuanced understanding of the relationship between natural disasters and population movements.
- Building upon our innovative framework that employs GNNs for population dynamics. As one of the first studies to apply GNNs in the context of migration and natural disasters, our work lays the groundwork for continued research in this area. By refining and expanding upon our approach, future studies can contribute to the development of more effective and comprehensive models for understanding and predicting population migration patterns in relation to natural disasters.

In conclusion, while our results may not have been as definitive as we had hoped, our efforts have still contributed to the broader body of knowledge in this area. We have identified key aspects that warrant further exploration, and we are confident that with continued research and refinement, more conclusive findings can be uncovered. Ultimately, our work has laid the groundwork for future studies to better understand and address the complex relationship between natural disasters and human migration patterns.

6.0.0.1 Special Thanks We would like to thank our professor Karianne Bergen and the Brown Population Studies Center for being crucial to the conception and completion of this paper.

References

[1] Brown, Oli. "Migration and Climate Change." *IOM Migration Research Series*, 2008, https://doi.org/10.18356/26de4416-en.

[2] Lustgarten, Abrahm, and Meridith Kohut. "Climate Change Will Force a New American Migration." *ProPublica*, 15 Sept. 2020, https://www.propublica.org/article/climate-change-will-force-a-new-american-migration.

[3] Climate Migration Report Draft - *White House*. The White House, Oct. 2021, https://www.whitehouse.gov/wp-content/uploads/2021/10/Report-on-the-Impact-of-Climate-Change-on-Migration.pdf.

[4] Robinson, Caleb, et al. "Modeling Migration Patterns in the USA under Sea Level Rise." *PLOS ONE*, vol. 15, no. 1, 2020, https://doi.org/10.1371/journal.pone.0227436.

[5] Wu, Zonghan, et al. "A Comprehensive Survey on Graph Neural Networks." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, 4 Dec. 2019, https://doi.org/10.1109/tnnls.2020.2978386.

[6] Bilecen, Başak, et al. "The Missing Link: Social Network Analysis in Migration and Transnationalism." *Social Networks*, vol. 53, May 2018, pp. 1–3., https://doi.org/10.1016/j.socnet.2017.07.001.

[7] Kemper, Franz-Josef. "Internal Migration in Eastern and Western Germany: Convergence or Divergence of Spatial Trends after Unification?" *Regional Studies*, vol. 38, no. 6, 2004, pp. 659–678., https://doi.org/10.1080/0034340042000240969.

[8] Tranos, Emmanouil, et al. "International Migration: A Global Complex Network." *Environment and Planning B: Planning and Design*, vol. 42, no. 1, 2015, pp. 4–22., https://doi.org/10.1068/b39042.

[9] Pitoski, Dino, et al. "Network Analysis of Internal Migration in Croatia." *Computational Social Networks*, vol. 8, no. 1, 2021, https://doi.org/10.1186/s40649-021-00093-0.

[10] Charyyev, Batyr, and Mehmet Hadi Gunes. "Complex Network of United States Migration." *Computational Social Networks*, vol. 6, no. 1, 2019, https://doi.org/10.1186/s40649-019-0061-6.

[11] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

[12] Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E, 74(3).* https://doi.org/10.1103/physreve.74.036104

[13] Christensen, A. P., Garrido, L. E., Guerra-Peña, K., &; Golino, H. (2020). *Comparing Community Detection Algorithms in Psychological Data: A Monte Carlo Simulation*. https://doi.org/10.31234/osf.io/hz89e

[14] Hauer, Mathew, and James Byars. "IRS County-to-County Migration Data, 1990–2010." *Demographic Research*, vol. 40, 2019, pp. 1153–1166., https://doi.org/10.4054/demres.2019.40.40.

[15] Molloy, Raven, et al. "Internal Migration in the United States." Journal of Economic Perspectives, vol. 25, no. 3, 2011, pp. 173–196., https://doi.org/10.1257/jep.25.3.173.

[16] Gross, E. "Internal Revenue Service area-to-area migration data: Strengths, limitations, and current trends." *Washington, D.C.: Internal Revenue Service*, 2005 (SOI Working Paper).

[17] DeWaard, Jack, et al. "User Beware: Concerning Findings from the Post 2011–2012 U.S. Internal Revenue Service Migration Data." *Population Research and Policy Review*, vol. 41, no. 2, 2021, pp. 437–448., https://doi.org/10.1007/s11113-021-09663-6.

[18] United States, Department of Homeland Security, FEMA. *OpenFEMA*, Disaster Declaration Summaries, FEMA, 2023. Accessed 4/01/2023. This product uses the Federal Emergency Management Agency's API but is not endorsed by FEMA.

[19] OpenFEMA. "Data Visualizations." *FEMA.gov*, FEMA, 9 Mar. 2023, https://www.fema.gov/about/reports-and-data/data-visualizations.

[20] Hamilton, William L., Ying, Rex, and Leskovec, Jure. "Inductive Representation Learning on Large Graphs." *No date.*

[21] Panagopoulos, George, Nikolentzos, Giannis, and Vazirgiannis, Michalis. "Transfer Graph Neural Networks for Pandemic Forecasting." *arXiv*, 2021, http://arxiv.org/abs/2009.08388. doi:10.48550/arXiv.2009.08388.

[22] Li, Yaguang, Yu, Rose, Shahabi, Cyrus, and Liu, Yan. "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting." *arXiv*, 2018, http://arxiv.org/abs/1707.01926. doi:10.48550/arXiv.1707.01926.

Appendix 1. Centrality Analysis Graphs

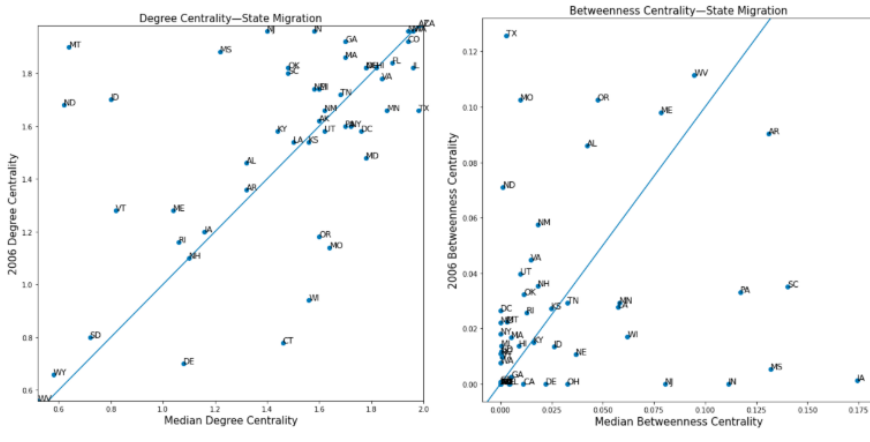

Figure 8. Degree Centrality on the left and Betweenness Centrality on the right. Median centrality on x-axis, 2006 centrality on y-axis.

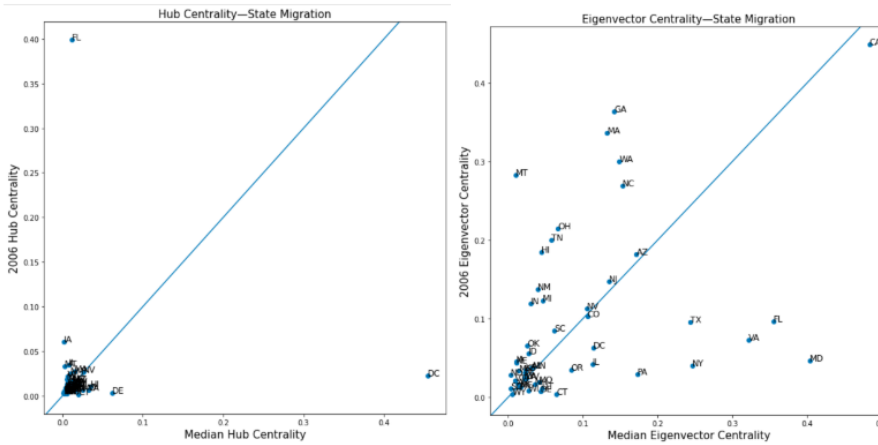

Figure 9. Hub Centrality on the left and Eigenvector Centrality on the right. Median centrality on x-axis, 2006 centrality on y-axis.

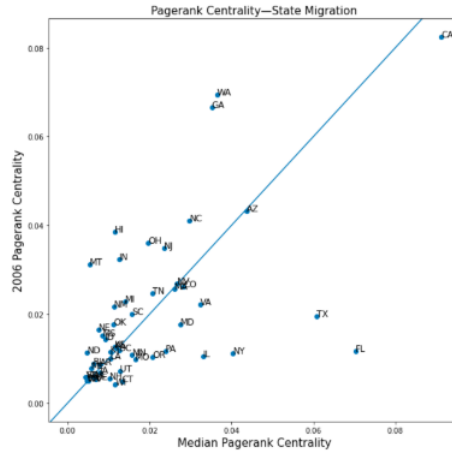

Figure 10. PageRank centrality, median centrality on x-axis, 2006 centrality on y-axis.

Appendix 2. Community Detection Graphs

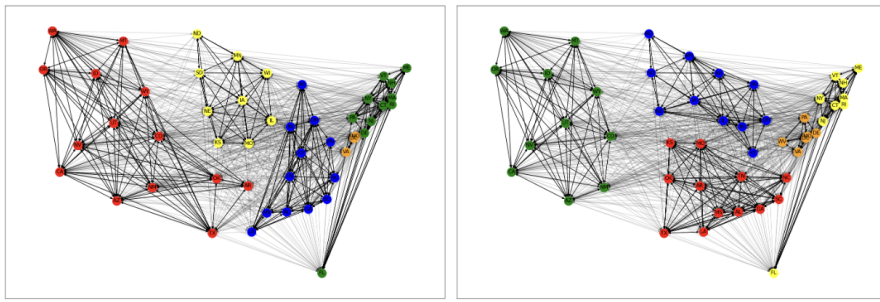

Figure 11. Fast Greedy, median (left), 2006 (right)

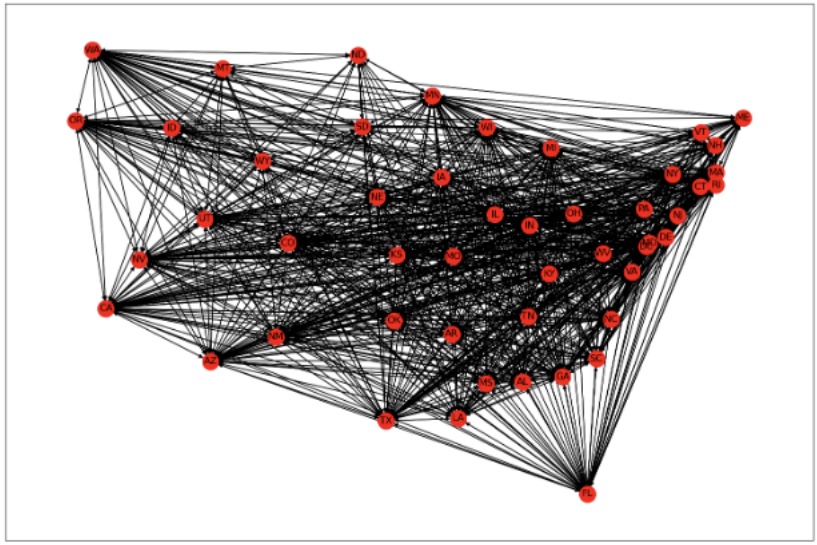

Figure 12. Infomap for median

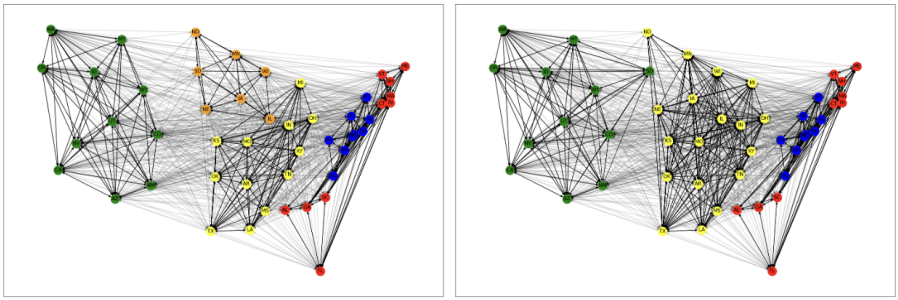

Figure 13. Leading Eigenvector, median (left), 2006 (right)

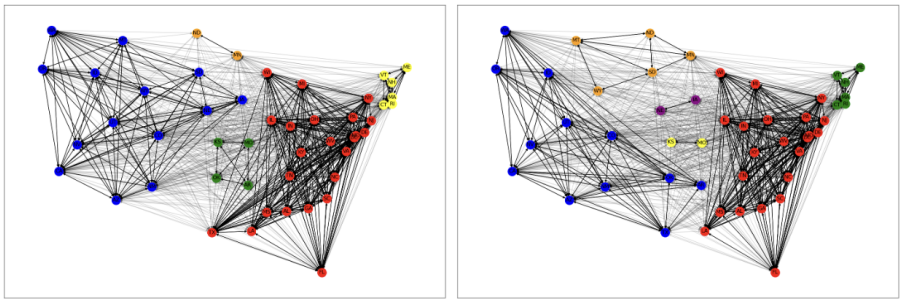

Figure 14. Label Propagation, median (left), 2006 (right)

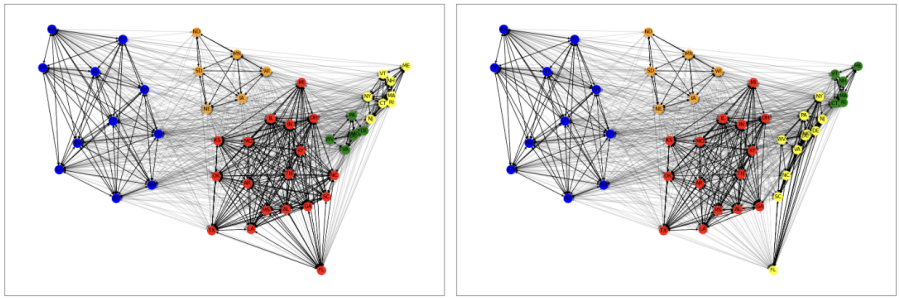

Figure 15. Walktrap, median (left), 2006 (right)

Appendix 3. Causality Analysis

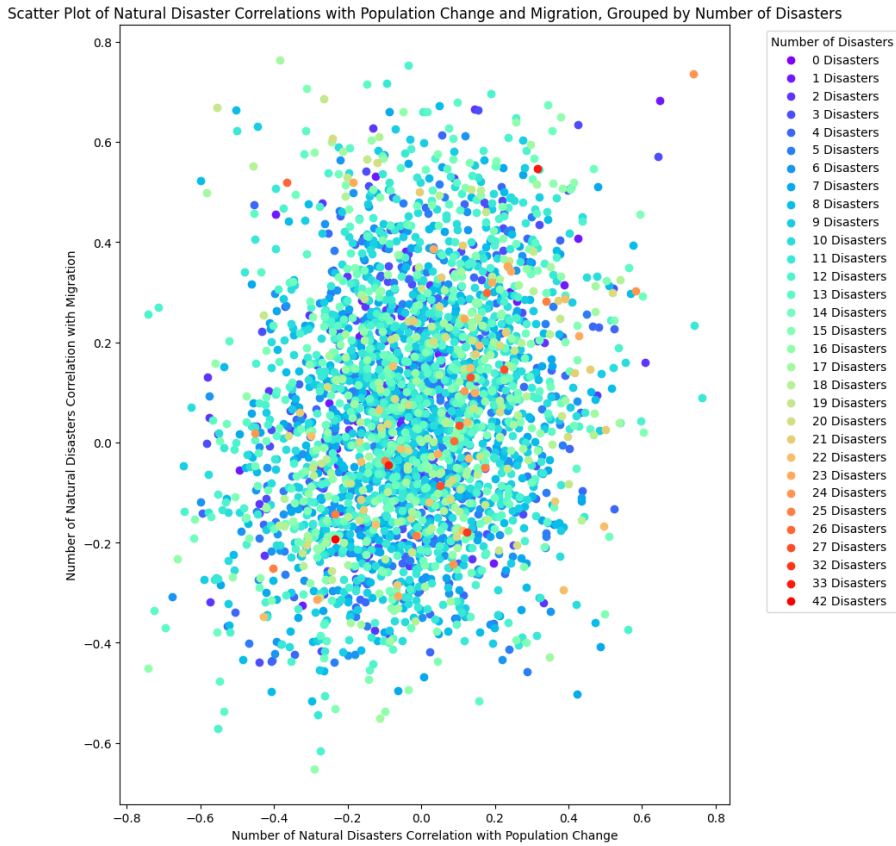

Figure 16. Scatter plot of natural disaster correlations with population change and migration, grouped by number of disasters. Each point represents a county, colored by the number of natural disasters experienced by that county. The x and y axes show the correlation coefficients for each county. The legend indicates the color codes and corresponding number of natural disasters.

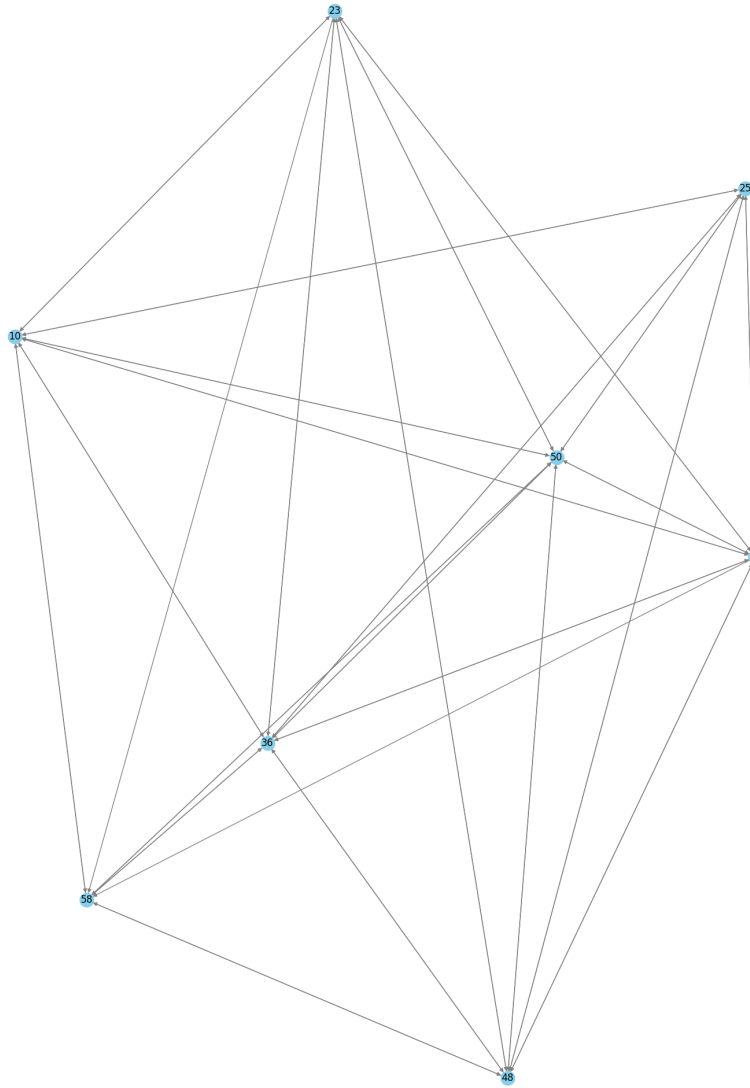

Figure 17. Illustration of the use of subgraphs in graph neural networks for mini-batch sampling. The figure shows an example of an ego-graph (subgraph) created around a central node, which allows the model to extract more information from a larger graph while efficiently batching multiple graphs. By sampling subgraphs of a larger graph, the model can effectively learn from the local structure of the graph, which can lead to improved performance in tasks such as node classification, graph classification, and link prediction.