1 Introduction

Human migration is a revealing social phenomenon. Migration may indicate destinations with greater economic potential or cultural growth, or it can also define properties of a country, such as its internal stability. In addition to individual qualities, migration can indicate relationships between regions such as a shared culture or language, demographic and age changes with respect to certain areas, or regional dependence.

In this project, we first explore properties of US domestic migration, looking at patterns of migration over various demographics. We also search for higher-level conclusions by running a number of community detection methods on our dataset. With network analysis techniques, we hope to learn more from our migration data than simply the number of people who migrated from one region to another. More specifically, we will see if migration data can give us new insights about US communities beyond state boundaries, patterns of movement for people in different demographics, and insights about cities and their ties to respective communities.

2 Background

Prior work on graph network analyses of migration is limited. One Elementary exploratory work is is Maier et al’s paper on social network analysis on internal migration within US states [3], which tested many basic techniques to suggest promising directions. The primary purpose of the paper was to test the waters and see if social network analysis could provide insight into migration patterns that matched those derived from more traditional means.

2.1 Evaluating Community Structure

Newman and Girvan [4] propose a set of algorithms for finding in community structure in networks and a corresponding a measure of the strength of the community structure. They primarily use a “divisive” technique where they iteratively remove edges from the network to thereby split it into communities. They choose the edge to remove by looking at betweenness measures, such as shortest-path betweenness. After each edge is removed, the betweenness measure is recalculated.

The article also defines a measure of the quality of the division of the network called modularity. The modularity is calculated as the trace of matrix $e$ minus the sum of the elements of $e^2$ where an entry $e_{ij}$ of $e$ is defined as fraction of all edges in the network that connect community $i$ with community $j$.

$$\text{modularity} = Q = \text{Tr}(e) - ||e^2||$$

So one way of looking at community divisions is constructing a dendrogram, or a tree displaying the structure of communities in the network, and looking at which height in the tree modularity reaches a maximum.
2.2 Graph similarity scoring and matching

Not yet applied to migration, Zager and Verghese introduce similarity measures on separate graphs [5]. The similarity measures are based on node similarity metrics for a single graph. The definition is extended to apply to similarity between nodes of different graphs. This is then used to calculate similarity of edges, which is based on the similarity of the source and destination nodes in the respective graphs. Zager and Verghese derive a single matrix update equation $s_k = Ms_{k-1}$ for the iterative algorithm, where $s$ is a vector with all pairwise node and edge scores between the two graphs and $M$ is a matrix derived from the source-edge matrices and terminus-edge matrices for the two graphs.

This was implemented as two update steps, where $y$ is a vector of edge-edge similarities and $x$ is a vector of node-node similarities. The update steps for each iteration are

$$y_k \leftarrow Gx_{k-1} \quad (1)$$

$$x_k \leftarrow G^Ty_{k-1} \quad (2)$$

where $G = A_S^T \otimes B^T_S + A_T^T \otimes B^T_T$. $N_S$ and $N_T$ are the source-edge and terminus-edge matrices, respectively, of a network $N$, and $\otimes$ represents the Kronecker product between matrices.

3 Data

We draw our data from the County-to-County Migration Flows data from the United States Census Bureau collected by the American Community Survey (ACS) and the Puerto Rico Community Survey. This data is available to the public from the U.S. Census Bureau’s website. ACS estimates data based on a series of monthly samples, with greater numbers of years of data used to estimate geographies of smaller populations.

We used the following years of data:

- 2006-2010 county-to-county data crossed by age, sex, race, and Hispanic or Latino origin
- 2007-2011 county-to-county data crossed by educational attainment, household income, and individual income
- 2008-2012 county-to-county data crossed by employment status, occupation, and work status
- 2009-2013 county-to-county data crossed by ability to speak English, place of birth, and years in the United States

Each of the characteristics (e.g. age, sex, race) were each separately used to divide the data. We treated each county as its own node, assigning it a unique label based on the state and county Federal Information Processing Standards (FIPS) codes. We focused primarily on the major attributes of age, sex, race, education, and employment. For datasets with characteristics, we assigned edges additional attributes corresponding to the value for the characteristic.

4 Methods

As baselines, we extended many of the methods described in the background and modified them to our dataset. We also performed novel analyses. The extensions and innovations are described in this section.

4.1 Weighted Conductance

In Leskovec et al. [2], conductance is defined as a measure of the quality of a community, where a high-quality community is one that has a densely linked set of nodes within itself, but is attached to the rest of the network with very few edges. The conductance of a given community is simply

$$\text{conductance} = \frac{|s_C|}{2|s_E| + |s_C|}$$
Where $s_C$ is the set of edges that connect nodes in the given community to nodes outside of the community, and $s_E$ is the set of edges whose endpoints are both within the community.

Because the US migration network is very much a weighted graph, we want to be able to take into account the weight of each edge instead of just its existence. Therefore, we can define a weighted version of the conductance measure as the following:

$$\text{weighted conductance} = \frac{w(s_C)}{w(s_E) + w(s_C)}$$

Where $w(x)$ is the summation of the weights of the edges $x$.

### 4.2 Diversity Factor

Once we have divided up our counties into different communities, it will be interesting to see two different pieces of information - first, whether the migrants coming into a county tend to be from that county’s community, and second, whether the migrants leaving a county tend to go to another county in that community (or do they go far far away). To measure how diverse a county’s in-migrants and out-migrants are, we define the notion of diversity factor as

$$\text{diversity factor}(n) = \frac{w(\text{edges in community}(n))}{w(\text{edges}(n))}$$

Where $w(x)$ is once again the summation of the weights of the edges $x$, and $\text{edges in community}(n)$ are the edges from $n$ to other nodes in $n$’s community (or vice versa, depending on whether we’re trying to measure the diversity of the out-migrants or the diversity of the in-migrants), and $\text{edges}(n)$ are all of the edges containing $n$ as a starting point (or as an ending point if we’re calculating diversity factor of in-migrants).

### 4.3 Spectral Graph Methods

We applied methods from spectral graph theory. We first applied the graph similarity scoring method from Zager and Vergheese, after reproducing their initial work and examples. In the context of migration, we want to measure the similarity of nodes in one migration network (such as female migration) to the nodes in another migration network (such as male migration). In the standard analysis, the source-edge matrix $S$ is defined as having $S_{ij} = 1$ if the source of edge $j$ is node $i$ and $S_{ij} = 0$ otherwise. We extend this to our weighted network by taking $S_{ij} = w_j$ where $w_j$ is the weight of edge $j$ and $i$ is the source node of edge $j$. We apply a similar extension to the terminus-edge matrix.

We also examined properties of a similarity matrix $W$ where $W_{ij}$ measures the similarity between counties $i$ and $j$. We take $W_{ij} = (M_{ij} + M_{ji})^2 / (P_i P_j)$, where $M_{kl}$ is the number of migrants from county $k$ to county $l$ and $P_k$ is the population of county $k$, as suggested by Cucuringu et al[1]. We took $W_{ii} = P_i$.

### 5 Results

After collecting and cleaning the data, we first looked at some basic measures of centrality and connectedness. Afterwards, we compared the migration patterns across different attributes. Lastly, we focused on our two main analytical tools - community detection and graph similarity.

#### 5.1 Summary Statistics

Each of the graphs had around 3200-3500 nodes and on the order of 150,000 to 450,000 edges. We provide a table below for analysis on the graphs from each of the time frames without any characteristics:
The above information seems to suggest that the data did not change significantly over time. But it does give us a good sense of what the graph looks like overall.

Next we focus on only one graph in particular, the one from 2009 – 2013 and look at various ways of cleaning the data. Many of the edges have extremely low weights. In other words, some edges in our graph only have a couple of people migrating from one county to another. These are fairly insignificant when looking at large-scale migration patterns where thousands of people move to or from a county per year.

One way in which we can deal with these insignificant edges is to eliminate all edges with a weight below a certain threshold. We experimented with various thresholds in order to find a balance between eliminating useless edges and preserving as much information as possible. A plot comparing the threshold to the number of edges is shown below:

![Graph of edge count vs threshold](image)

Just as some additional insight, we see that all graphs are weakly connected. The location with the highest in-degree is tied between several counties. The location with the highest out-degree however is Asia. When we raise the threshold to be above 500, the location with the highest in-degree becomes Los Angeles, California. This aligns with common sense: Los Angeles has the most places from which it attracts a substantial (> 500) number of people.

### 5.2 Exploratory Analysis

Next, we look at migration patterns splitting by gender, race, education, and employment by looking at visualizations of our data on a map of the United States.
We graph the counties in their respective regions in the US using Gephi. We weight the size of each node by the county’s population and we weight the thickness of the edges by the number of people migrating along that route.

For all of these visualizations, we trim the edges of our graph to only show those edges that have a size at least 80% as large as the largest edge in the graph rather than using the absolute threshold described previously. This deals with the issue of an overabundance of edges for visualization.

We begin by looking at gender. We see that since both genders have roughly the same number of people migrating (50823 females and 54055 males), if we believe the migration patterns for both genders to be roughly equal, we will have very similar graphics. We see that we do have very similar patterns for male and female migration.

For race, we compare the migration patterns, the census data is divided into white, black, and Asian populations.

If we look at the migration patterns of white people, we see that there’s a considerable amount of migration from areas surrounding a major city into the city itself. The few long-distance migration edges that exist are between Los Angeles and San Francisco, Los Angeles and Phoenix, Los Angeles and Houston, Los Angeles and New York, and Los Angeles and Chicago.

However, looking at the migration patterns of black people, we see that there is migration out of Los Angeles into the nearby San Bernadino and Kern counties, but elsewhere there is the same pattern of moving from surrounding areas towards urban centers of large cities. Interestingly, there is a large amount of migration between Houston and New Orleans and between Chicago and Minneapolis.

Note that for the migration patterns of Asian people, the pattern of moving from surrounding areas into major cities is much less pronounced. However, note that there is a large amount of migration within the Bay Area and between the Bay Area and Los Angeles. Furthermore, Hawaii is finally on the map and reasonably so given the proportion of Asians living in Hawaii.

For education, we see some interesting results when we compare the migration patterns of those with higher degrees (bachelor’s and graduate degrees) to those with little to no education. Those that dropped out before high school or only have a high school diploma seem to travel much less and usually migrate between Los Angeles, Las Vegas, Phoenix, Chicago, Houston, and New Orleans.

Compare this to the migration patterns of people who have bachelor’s or graduate degrees. There is a large concentration of migration to Southern California (Los Angeles, San Diego, Riverside), the Bay Area (Santa Clara, Alameda, San Francisco, San Mateo counties), Seattle, New York, Philadelphia, Washington D.C., and Boston.

Lastly, we look at migration with respect to employment. Again, we see that those who are employed tend towards the larger cities that often have more business, such as San Francisco, Chicago, Austin, Houston, Detroit, Dallas, Miami, New York, and Los Angeles.
The map showing the migration of unemployed people is much less enlightening as it seems to be primarily between the largest cities in the US.

Particularly interesting is if we look at the people serving in the Armed Forces. The major destinations of people in the armed forces seem to be San Diego, Kitsap County (Washington), and Virginia Beach. On further research, San Diego is home to the largest naval base on the west coast, Virginia Beach is the location of three military bases, and Kitsap County has a naval base as well. Additionally, Escambia County (Florida) and Lake County (Illinois) seem to be the source of migration for armed forces members. We note
that Naval Air Station Pensacola is located in Escambia County.

5.3 Community Detection

We ran a variety of community detection methods (both the original and their weighted counterparts) on our data. The two methods that led to the best results, based on weighted conductance scores, were actually two of the most well known - Clauset-Newman-Moore and Louvain’s algorithms. Visualizations of the
communities for each of those methods are presented below.

Figure 6: Clauset-Newman-Moore’s algorithm for community detection

Figure 7: Louvain’s algorithm for community detection
5.3.1 Quality of Community Detection Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Modularity</th>
<th>Weighted Modularity</th>
<th>Conductance</th>
<th>Weighted Conductance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clauset-Newman-Moore</td>
<td>0.310</td>
<td>0.538</td>
<td>0.820</td>
<td>0.422</td>
</tr>
<tr>
<td>Clauset-Newman-Moore (unweighted)</td>
<td>0.260</td>
<td>0.363</td>
<td>0.597</td>
<td>0.431</td>
</tr>
<tr>
<td>Louvain - 1st iteration</td>
<td>0.170</td>
<td>0.449</td>
<td>0.916</td>
<td>0.657</td>
</tr>
<tr>
<td>Louvain - 2nd iteration</td>
<td>0.277</td>
<td>0.533</td>
<td>0.864</td>
<td>0.486</td>
</tr>
<tr>
<td>Louvain - 3rd iteration</td>
<td>0.303</td>
<td>0.541</td>
<td>0.838</td>
<td>0.439</td>
</tr>
<tr>
<td>Louvain - final / best (4th) iteration</td>
<td>0.303</td>
<td>0.541</td>
<td>0.836</td>
<td>0.437</td>
</tr>
<tr>
<td>Louvain (unweighted) - 1st iteration</td>
<td>0.308</td>
<td>0.454</td>
<td>0.687</td>
<td>0.518</td>
</tr>
<tr>
<td>Louvain (unweighted) - final (2nd) iteration</td>
<td>0.308</td>
<td>0.452</td>
<td>0.683</td>
<td>0.513</td>
</tr>
</tbody>
</table>

In the above table, by looking at weighted conductance values (our way of evaluating the quality of the communities created), we can see that weighted community detection methods tend to do better than normal community detection methods. Furthermore, we can see that Clauset-Newman-Moore divides the network into higher quality communities than Louvain does.

5.3.2 Interesting Observations from the Communities Detected

Because Clauset-Newman-Moore’s method on weighted networks has the best weighted conductance value, we will take a further look at some of the county-community mappings that look a little anomalous. From the figure, we can spot a few interesting things: St Louis’s metropolitan area stretches into one of the counties in northwest Illinois, El Paso tends to be more connected with Arizona / New Mexico than the rest of Texas, Green Bay tends to have migrants coming in/from a county in Michigan than that county does to the rest of Michigan, Pennsylvania and West Virginia are cut in half (possibly due to the Appalachian Mountains being located there), there is lots of movement connecting New England and Florida (probably New England seniors going to Florida to retire), and not a lot of migrants move in / out of Yellowstone from the rest of Wyoming (it’s unclear why / whether you would move to a part of your state that’s a national park)

5.4 County Diversity

After we’ve established the communities formed by Clauset-Newman-Moore on weighted networks as the ones we want to examine, we can look at the diversity factor for major counties with respect to those communities, which is presented in the tables below.

<table>
<thead>
<tr>
<th>County (cities represented in county)</th>
<th>In-Migrant Diversity Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Angeles County (Los Angeles)</td>
<td>0.717</td>
</tr>
<tr>
<td>Harris County (Houston)</td>
<td>0.483</td>
</tr>
<tr>
<td>Maricopa County (Phoenix)</td>
<td>0.561</td>
</tr>
<tr>
<td>Cook County (Chicago)</td>
<td>0.392</td>
</tr>
<tr>
<td>San Diego County (San Diego)</td>
<td>0.641</td>
</tr>
<tr>
<td>New York County (New York)</td>
<td>0.668</td>
</tr>
<tr>
<td>Orange County (Anaheim / Irvine)</td>
<td>0.814</td>
</tr>
<tr>
<td>Riverside County (Riverside)</td>
<td>0.854</td>
</tr>
<tr>
<td>Dallas County (Dallas)</td>
<td>0.590</td>
</tr>
<tr>
<td>King County (Seattle)</td>
<td>0.683</td>
</tr>
</tbody>
</table>

From the In-Migrant Diversity Factor table, we can see that among the counties with the greatest number of incoming migrants, the counties corresponding to Chicago and Houston are arguably the ones that have migrants coming from the most diverse locations. In other words, they have the highest percentage of migrants that are not from the same community that they’re located in.
<table>
<thead>
<tr>
<th>County (cities represented in county)</th>
<th>Out-Migrant Diversity Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Angeles County (Los Angeles)</td>
<td>0.820</td>
</tr>
<tr>
<td>Cook County (Chicago)</td>
<td>0.619</td>
</tr>
<tr>
<td>Maricopa County (Phoenix)</td>
<td>0.477</td>
</tr>
<tr>
<td>Harris County (Houston)</td>
<td>0.567</td>
</tr>
<tr>
<td>San Diego County (San Diego)</td>
<td>0.550</td>
</tr>
<tr>
<td>Dallas County (Dallas)</td>
<td>0.806</td>
</tr>
<tr>
<td>Orange County (Anaheim / Irvine)</td>
<td>0.696</td>
</tr>
<tr>
<td>Kings County (San Joaquin Valley)</td>
<td>0.963</td>
</tr>
<tr>
<td>New York County (New York)</td>
<td>0.626</td>
</tr>
<tr>
<td>San Bernadino County (San Bernadino)</td>
<td>0.689</td>
</tr>
</tbody>
</table>

From the above table, we can see that among the counties with the greatest number of outgoing migrants, the counties corresponding to Phoenix and San Diego are arguably the ones whose migrants are going to the most diverse locations. In other words, they have the highest percentage of migrants leaving to communities that are not the ones they’re located in.

5.5 Graph Similarity

In reproducing Zager and Verghese’s work, we found very interpretable similarity scores on their example networks. The resulting similarity score matrix $S$ was very sparse. The nonzero values occurred in the pattern $S_{ij} > 0$ where $i \in C_1, j \in C_2$, where $C_1$ and $C_2$ are sets of nodes in the first and second networks, respectively, that are “similar” to one another. The portion of the similarity score matrix corresponding to the cross product of the two sets generally had positive values, which indicated that these two sets were of interest.

When applied to our networks, however, no such sparsity pattern emerged. The similarity score matrices were nearly fully dense, even after some transformations on the values, so we could not extract useful clusters of corresponding nodes. This is likely caused by the sparsity of our networks after thresholding, which was necessary for computational feasibility. The kind of sparsity in the migration networks involved many nodes having similar patterns of incoming and outgoing edges.

5.6 Low-rank eigenvector visualizations

Our second spectral graph analysis method did not rely on thresholding and did not run into the problems of having a sparse network. We successfully visualized the eigenvectors of the similarity matrix $W$ defined in our methods. $W_{ij} = (M_{ij} + M_{ji})^2 / (P_i P_j)$, where $M_{kl}$ is the number of migrants from county $k$ to county $l$ and $P_k$ is the population of county $k$, and $W_{ii} = P_i$. As suggested by Cucuringu et al.[1], we examined the eigendecomposition of this matrix.

Each eigenvector has a natural interpretation as a scoring over the counties. The scoring is usually organized in some localized way. For example, a heatmap of the 31st eigenvector is shown below.
Here, pink counties correspond to greater values, gray counties correspond to values near zero, and blue counties correspond to smaller values. We see some geographical localization: The 31st eigenvector clearly localizes Alabama, and it also highlights a loose communities along similar longitudes in the Midwest.

6 Future Work

We hope to extend this project both in terms of computational techniques and in terms of types of questions explored. Computationally, we used thresholding, an established and key idea in migration network analysis. However, as we removed edges, some nodes became zero-degree nodes. In particular, this caused issues with our community detection algorithm. As a result, we removed such zero-degree nodes. Unfortunately, removing data is nearly always undesirable. We hope to explore ways to better include these nodes.

We also hope to perform other analysis on this dataset. For example, we might ask, based on the data available, can we predict how many people of a certain demographic are going to move into a given county during a certain time period? Arguably the strongest predictor is the number of people who moved there during a previous time period, but it will be interesting to see what other variables also play a role. Although it may be straightforward to simply apply link prediction techniques to the migration network we are given, now that we have also performed community detection techniques, it will be interesting to predict intra- and inter-community migration.
References


