ERIC KIM | DANIEL KUNIN | JONG HA LEE | DANIEL XIANG

A SOCIAL NETWORK

STRUCTURE WITHIN EDUCATION AND SOCIAL MOBILITY

Abstract

In this report, we investigate the effect of education, demographics, and economics on social mobility within the United States through migration flows. We cluster U.S. counties into 6 "social classes" based on the aforementioned aspects to connect socioeconomic factors to geographic movement. In characterizing movement, we first create a graph network of county-to-county migration outflows. We then apply network based significance testing to filter out migrations with small statistical significance with respect to the dependence between migrating from one cluster to another. After obtaining a cluster of counties with similar social classes and the statistically meaningful migrations between them, we apply a Poisson-distributed LASSO linear model to identify possible bottlenecks in social movement. We find that educational aspects are significant factors to facilitate better mobility between socioeconomic clusters, suggesting possible educational policy improvements.

Introduction

There is no doubt that education plays a major role in an individual's ability to climb the social ladder in the United States. The idea that citizens can lift themselves and their children into a higher social class through hard work and determination has formed the basis for the country's reputation as the land of opportunity. This ability is due in part to the quality of higher education in the United States. The U.S. has by far the largest number of universities in the top 100 global rankings, clocking in an impressive 48 out of the top 100 ranked universities worldwide [1]. These world class universities are often perceived academic bubbles, but they do in fact have a significant and tangible effect on the American lower class student. For instance, a *New York Times* article states that "76% of students who enrolled [at The City College of New York] in the late 1990s and came from families in the bottom fifth of the income distribution have ended up in the top three-fifths of the distribution" [2].

Although the United States leads the world in elite research universities, it has at the same time lagged behind its global peers within the realm of primary and secondary school. In lower socioeconomic areas of the country, public schools struggle with obtaining adequate funding to educate the nation's youth. Standardized test scores have been the main proxy for measuring the effectiveness of a school district's education, due to their ease of measurement and their nationwide adoption. But a factor that is not as frequently considered which also characterizes social mobility is *geographical* mobility. Every year, more than 10% of U.S. residents move according to data produced by the U.S. Census Bureau and IRS [3]. Counties with a higher quality of education tend to produce individuals equipped with the knowledge and skills to move into "hubs" of social and economic growth to pursue their life goals, whereas areas of low quality education do not. Indeed, it can be argued that there exists a relationship between these three factors (Fig. 1), which sets the scene for the following report.

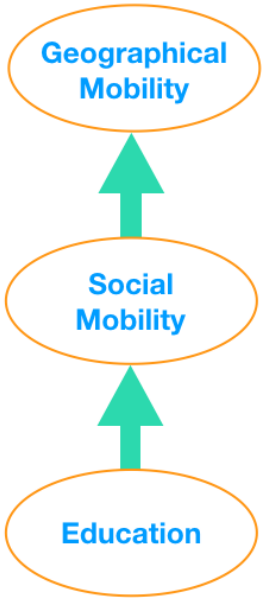

Fig. 1: The flow of the social mobility model. Education facilitates social mobility which enables geographical mobility. More features will be included later in the report as we discover more about the related data.

Nontechnical Summary

Main Questions

Our analysis can be broken down into three main questions:

1. *Is there evidence of a relationship between the geographic location a person is born in the U.S. and their educational and economic opportunity?*
2. *Can we determine the geographic structure of this relationship and model mobility within it?*

3. *Can we use this model to identify bottlenecks within the network of social classes and suggest improvements?*

Key Findings

We found that there was indeed a relationship between the features of interest and geographical mobility within the United States. Using clustering techniques, we were able to group counties into 6 social classes, determined by the K means clustering algorithm. Constructing a graph from these social classes whose edges are geographical migrations of people between them, we found that the lowest social cluster had very little flow back to itself, most of its flow toward the lower middle classes, and extremely small flow to the upper class. On the opposite side of the spectrum, the highest social class had most of its flow back to itself and the middle upper classes, and almost no flow to the first class (see Fig. 11). These skewed distributions of migration help to explain the ever widening gap of inequality in the U.S.

After identifying the migration flow and distributions to different social class clusters, we hypothesized that there may be possible bottlenecks relating specifically to education, among other factors. To test our hypotheses, we implemented a Poisson-distributed LASSO Regression model to identify which educational factors may help facilitate migration flows, remove bottlenecks, and therefore better enable people to move up the social ladder. After running our model on every unique cluster-to-cluster outflow, we find that educational factors such as average number of teachers, enrollment numbers, and pupil per expenditure are proportionally significant and indeed contribute positively to migration flows, after controlling for demographic factors such as population.

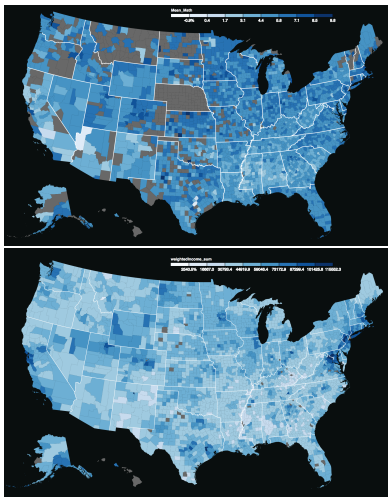

Fig. 2: Heat maps of test scores (top) and median household income (bottom). The grey spots are the counties for which there was no data.

Technical Summary

Initial Data Exploration

Perhaps the most important step in data analysis is to make sure we understand the data. Only then can we discover interesting trends and form informed, answerable hypotheses with the provided data set. Consequently, quality data visualization tools that are easy to use and adaptable to different data sets are crucial to efficient exploration. Keeping this in mind, we developed and adapted several exploratory visualization tools in D3, a JavaScript library used for creating interactive data visualizations in a web browser [7]. These can be accessed at `teamtwo.getforge.io`. These investigatory tools enabled us to explore our data in multiple perspectives efficiently.

Our interest at large was the relationship between demographic, economic, education, and geographical migration features at the county level. To understand how these features interact, we first tried to understand their distribution throughout the country through a heat map, where different levels of education quality or median household income can be indicated by a gradient of color.

Since darker blue on the heat maps represent higher test scores and higher median income respectively, the above heat maps seem to agree that the northeast has a high density of good test scores and wealth, and that the midwest is scattered in terms of density of high income and test scores.

Other factors which we suspect are also related to education and social mobility are poverty demographics and employment. A plot of interest is then a scatter plot [1] between poverty rate and percent of men employed, shown below.

[1] Click `teamtwo.getforge.io` to explore this visualization.

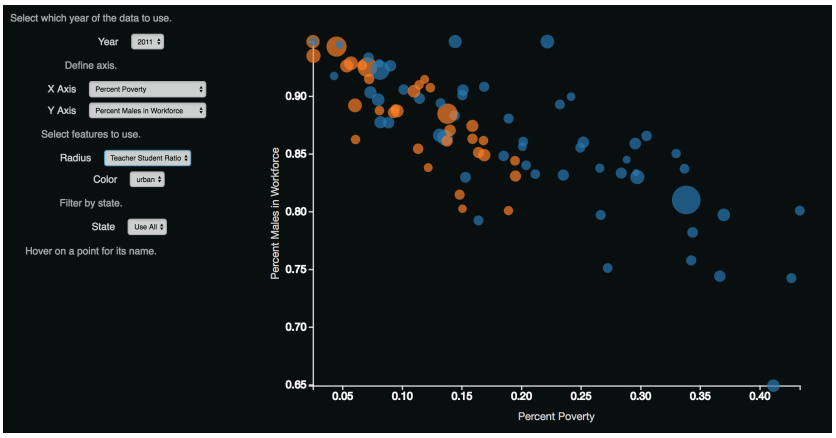

Fig. 3: The first of our three D3 visualizations is an interactive scatter plot of the SEDA databse. Users can filter by year and/or state, choose which features to plot on either axis and change the color and size of the plotted points. We encourage the reader to use these tools to explore the datasets used in this report and formulate their own questions.

It appears that the poverty percentage is negatively correlated with the percent of males in the workforce.

By making some quick plots exploring the possible trends between the factors of interest (Fig. 4), we now have an idea of what kind of feature matrix to clean and construct. Further exploration will be done by clustering with these features later in the report.

Data Wrangling and Engineering

After exploring the data and noticing that geography plays an important role in educational and economic opportunity, we wished to investigate whether geographic mobility is an indicator for social mobility. Acknowledging that one's geographic mobility and decision to move is a complex problem not only attributed to education or economic related reasons, we decided to incorporate various other

factors such as age, racial distributions, and population numbers. Thus, our analyses required extensive data wrangling to generate relevant features pertaining to three domains: education, economics, and demographics.

The greatest challenge we encountered during this process was that our datasets were at varying spatial granularities, ranging from district-level data per year to county and state-level data. The primary geographic level of granularity we wished to analyze was on the county level, requiring aggregations of `districts`-related datasets such as SEDA. We were then able to join the necessary datasets on the county-level through county ids which were standardized across both SEDA, government, and external datasets. After joining our various data sources we cleaned and organized it into a single data source for the rest of our analysis.

Having obtained our cleaned dataset of county features, we then wished to collect data on the relationship between counties in hopes of adding depth to our understanding of geography. After extensive searching, we found Internal Revenue Services Statistics of Income (SOI) program's county-to-county migration estimates based on tax redemptions and returns [4]. From this dataset, we were able to understand the migration flows between counties, which was crucial to the network analysis approaches we describe later.

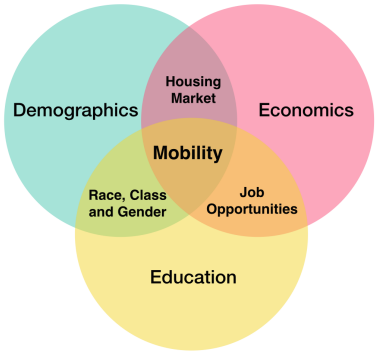

Fig. 4: A conceptual Venn Diagram visualizing the three pillars of our clustering and how they relate to each other.

Further Exploration with Clustering

When trying to uncover trends in unsupervised data, clustering is often a useful exploratory tool. Our first question involved understanding whether or not geographic location of birth had a relationship to someone's educational and economic opportunity. After obtaining aggregated data on education, demographics, and economics on the county level, it was natural to investigate the general topography of these features over the geographic domain. To this end, we turned to clustering.

Since we decided to cluster counties based on features related to social status, we needed to pick a the number of clusters in a way that reflects the current state of social structure in the United States. Sociologist Dennis Gilbert postulates in his book *The American Class Structure* [5] that there are six main social classes. These classes roughly can be described as an upper class, an upper middle class (educated professionals), middle class (white collar), a lower middle class (semi-professional), a working class (blue collar), and lastly a lower class consisting of the poor and unemployed. Applying this domain specific information to direct our investigation, we pick $K = 6$ to be the number of clusters in our K means clustering

approach.

To apply K means, we first need to determine which features of the data set best differentiate between potential social classes, and in order to stay conscious about the curse of dimensionality with K means [2], we need to pick these features wisely. As such, we perform some more exploratory data analysis to find the features that best allow us to create clusters with relevant information about social classes. Further, we look only at the features below for the year 2013.

Since education is vital in our analysis, we decide to look at education, in particular, test scores, and how they relate to other features. We turn to two main metrics of education, which are the Mean ELA and Mean Math scores over a county, and we noticed that there is a relatively strong linear relationship between them. Because of this, we decide to average over the two scores to come up with a generic `Education` feature.

To best capture economics and in particular, economic inequality, we decide to use the Gini Index over each county as one of our features. As for income inequality, we use the weighted average income over each county, which was computed by taking the midpoint within various income ranges and weighting them by the proportion of residents in the county within that bin. [3]

With these three features, we looked at their distributions, correlations, and scatter plot matrices and noticed that the distributions all look rather Gaussian. So we performed pairwise regressions and find that even after controlling the significance levels with a Bonferroni correction, the slope coefficients are highly significant, which confirms our choice to use these features for our cluster analysis.

After performing K means, we plot the counties on a map of the continental US and color by the clusters from K means results with the clusters ordered from lowest within cluster mean weighted income (cluster 1) to the highest (cluster 6). Further, we adjust the size of each point by that county's weighted income. We can start to see some patterns from this map. In particular, the highest income counties are the large cities while lower income counties are in the more rural areas.

We now take a look a closer look at our clusters and the features we clustered with in order to analyze the migration of people between them. We notice first that cluster 6 (the cluster with the highest average weighted county income) had the fewest number of people in the clusters, which may suggest that it is rather difficult to move into this high social class.

We also examine the scatter plot matrix of our features used for clustering. K means successfully split up the clusters by weighted income and education scores nicely, but wasn't as cleanly split for the

[2] In high dimensional Euclidean space, i.e. the feature space of data points, distances are typically inflated.

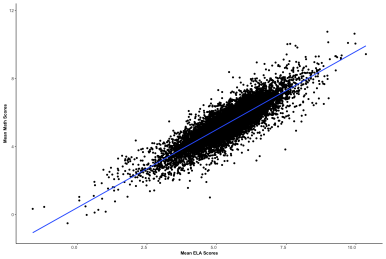

Fig. 5: Mean Math Scores vs Mean ELA Scores

[3] The ranges of these income bins are $[0, 10000)$, $[10000, 15000)$, $[15000, 25000)$, $[25000, 35000)$, $[35000, 50000)$, $[50000, 75000)$, $[75000, 100000)$, $[100000, 150000)$, $[150000, 200000)$, $[200000, \infty)$.

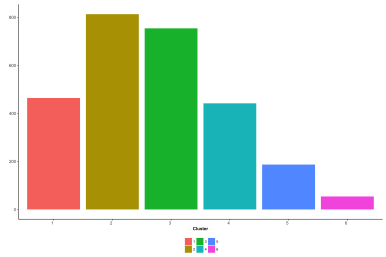

Fig. 6: Relative Size of Clusters

Fig. 7: Scatter plots of the paired features of interest, with corresponding color classifications into the 6 social classes. Each point represents a county.

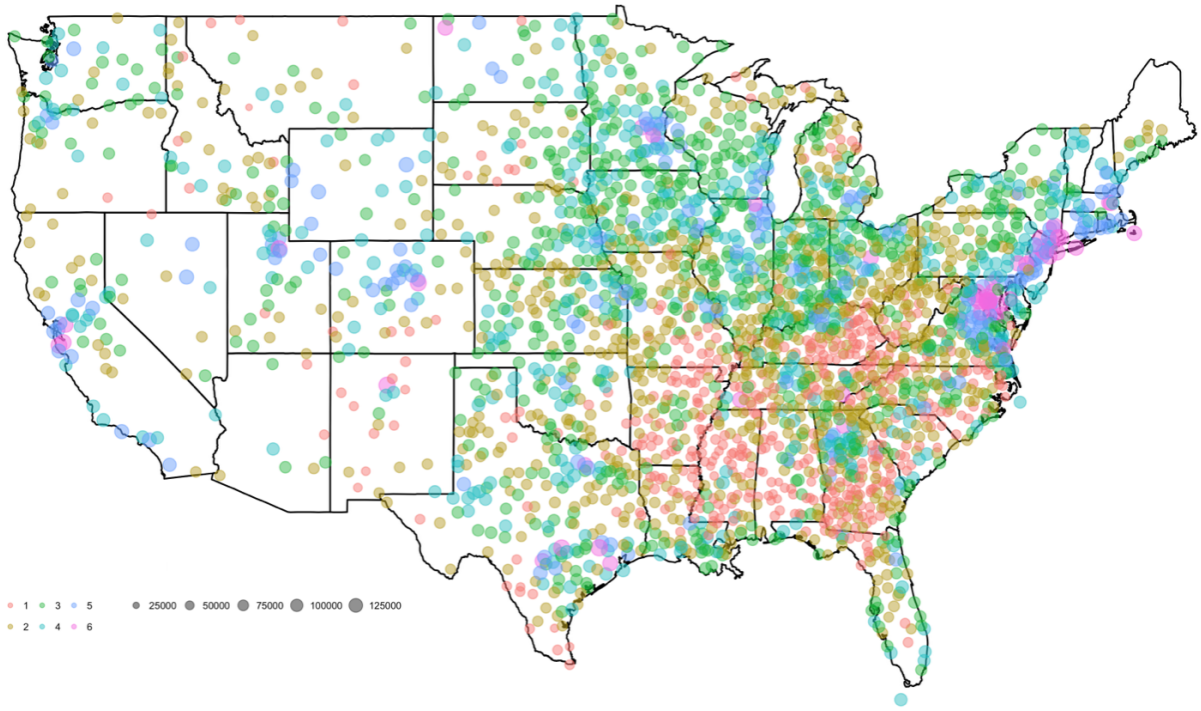

Fig. 8: Clustered Counties, Point Scaled by Weighted Income

Gini Index. This, along with the bar plot (Fig. 6) showing the non-uniform distribution of clusters, confirms that the K means did not simply bucket our data uniformly nor did we fall victim to the curse of dimensionality.

The Model

Notations

We will define some of the notation in this section, and then some more throughout the derivation of the model as it becomes more relevant.

Suppose the output of the K means clustering gives K different clusters, denoted $C_1, \ldots, C_K$. The interpretation to these clusters is that each represents a distinct social class, distinguished by a mixture of education, economic, and demographic features.

We obtained a migration dataset from which we constructed the matrix M whose $(i,j)^{th}$ entry, denoted M_{ij}, is the number of people who migrate from county i to county j. Also keep in mind there is migration data for around 3000 counties in this dataset, and so $M \in \mathbb{R}^{3000\times3000}$ is roughly the size of our migration matrix. We

denote

$$M_i^{\text{leave}} \doteq \left(\sum_{j=1}^{3000} M_{ij}\right) - M_{ii}$$
$$M_j^{\text{enter}} \doteq \left(\sum_{i=1}^{3000} M_{ij}\right) - M_{jj},$$

i.e. M_i^{leave} is the number of people who strictly leave county i. Another way to see this is that M_i^{leave} is the i^{th} row sum of M minus the i^{th} diagonal entry. Similarly, M_j^{enter} is the number of people who strictly enter county j, which is equivalent to the j^{th} column sum of M minus the j^{th} diagonal entry. [4]

The points that were clustered in the K means algorithm had labels, i.e. their county names. So if $\{x_i\}$ are counties that got classified into C_k, then we use the notation

$$|C_k| \doteq \sum_{x_i \in C_k} (M_{x_i}^{\text{leave}} + M_{ii}).$$

In words, $|C_k|$ is the total number of people who "leave" counties within cluster C_k, i.e. the sum of row sums that correspond to the counties within cluster k. Note that we can simply see $|C_k|$ as the total number of people born into the social class C_k, by making the simplification that everyone "leaves" their cluster. In the case where they don't actually leave their cluster, we just say they left that cluster and entered the same cluster.

[4] In case it isn't already clear, by "strictly leave", we are referring to people who leave nontrivially, i.e. don't just leave and immediately return. Same thing for "strictly enter".

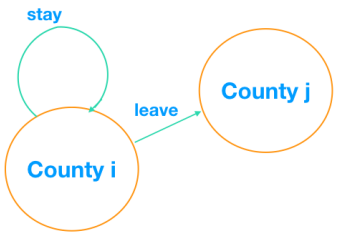

Fig. 9: The migration matrix describes the flow of people between counties. Within this model an agent can either leave or remain in their county.

Defining the Model

We are interested in the relationship between these clusters, specifically whether or not there is any kind of dependence in leaving and entering any two given clusters. It is then natural to define the sample space

$$\Omega \doteq \{(i,j) : i,j = 1,\ldots,K\} = \{1,\ldots,K\} \times \{1,\ldots,K\}.$$

The outcome (i,j) corresponds to a person being born into a class C_i and migrating to a county within class C_j. [5] Within this framework, we are now interested in the relationship between two probability measures. The first is the "true" joint probability measure $\mathbb{P}_{\text{joint}}$ that generated the data. The second is the product of the marginal probability measures, $\mathbb{P}_1\mathbb{P}_2$, i.e. the null probability distribution which assumes that the clusters C_i and C_j are independent for any i,j pair.

The motivation for the above probability space definition is that we expect to see *some* kind of structure between social classes, which would be reflected by the correlation between the above two distributions. For a concrete example, imagine C_i to be lower class and C_j to

[5] The outcome (i,i) represents a person being born into class C_i and never leaving.

be the top 0.5%. Then we'd imagine that its less likely for a migrant from C_i to enter C_j than the reverse, due to various factors such as race, education, and family income that were taken into account by the clustering algorithm.

Comparing the Probability Measures

For ease of interpretation, suppose that there is a person who has not yet been born, who we will refer to as Ken. We now define two indicator random variables defined on Ω that will allow us to compare the distributions $\mathbb{P}_{\text{joint}}$ and $\mathbb{P}_1\mathbb{P}_2$. Let

$$I_{\text{leave } C_k}(i,j) \doteq \begin{cases} 1 & \text{if } k = i \neq j \\ 0 & \text{otherwise} \end{cases}$$

$$I_{\text{enter } C_k}(i,j) \doteq \begin{cases} 1 & \text{if } k = j \neq i \\ 0 & \text{otherwise} \end{cases}.$$

In words, the first indicator random variable is equal to 1 when Ken is born into class C_k and migrates to a different class, and 0 otherwise. The second indicator random variable is equal to 1 when Ken is born into some class that isn't C_k and migrates to C_k, and 0 otherwise. These two indicators will roughly correspond to the marginals $\mathbb{P}_1$ and $\mathbb{P}_2$, while the indicator we define below corresponds to the joint distribution. Let

$$I_{\text{leave } C_i,\, \text{enter } C_j} = I_{\text{leave } C_i} I_{\text{enter } C_j}.$$

i.e. the indicator of the joint event where Ken is born into class C_i and migrates to class C_j. [6]

[6] The indicator function of an intersection of events is equal to the product of the indicators.

Then the relationship between the two clusters C_i and C_j is related to the covariance

$$\begin{aligned} \text{Cov}(I_{\text{leave } C_i}, I_{\text{enter } C_j}) &= E(I_{\text{leave } C_i} I_{\text{enter } C_j}) - E(I_{\text{leave } C_i})E(I_{\text{enter } C_j}) \\ &= E(I_{\text{leave } C_i,\, \text{enter } C_j}) - E(I_{\text{leave } C_i})E(I_{\text{enter } C_j}). \end{aligned}$$

Since the expectation of an indicator is simply the probability of the event it indicates, the above becomes

$$= P(\text{leave } C_i \text{ and enter } C_j) - P(\text{leave } C_i)P(\text{enter } C_j).$$

Before we jump into estimating the above probabilities, we take a step back and reiterate the intuition behind the above measure of dependence. The joint probability is the likelihood that Ken is born into class i and migrates to class j, given that he started out in class i. This conditional statement is what distinguishes the joint from the

product of marginals, since the product of marginal distributions treats this conditional probability as the same as the probability that a person from an unspecified class enters class j.

Hence if the above quantity is positive, it is more likely that Ken enters C_j given that he started out in C_i. If it is negative, then it is less likely to end up in C_j, having started out in C_i. A clean way to see this mathematically is to write the above expression as [7]

[7] Recall the definition of conditional probability, $P(A \mid B) = \frac{P(A,B)}{P(B)}$, which implies we can make the factorization of the joint, $P(A, B) = P(A \mid B)P(B)$.

$$\begin{aligned} &= P(\text{leave } C_i)P(\text{enter } C_j \mid \text{left from } C_i) - P(\text{leave } C_i)P(\text{enter } C_j) \\ &= P(\text{leave } C_i)(P(\text{enter } C_j \mid \text{left from } C_i) - P(\text{enter } C_j)). \end{aligned}$$

Then observe that this expression is positive if and only if

$$P(\text{enter } C_j \mid \text{left from } C_i) > P(\text{enter } C_j).$$

Estimating the Joint and Marginals

It remains to estimate these probabilities and normalize by the standard deviations [8] to get the correlation between clusters. The first probability can be estimated using maximum likelihood, which for a Bernoulli random variable (our indicator function) is simply the empirical frequency of people leaving C_i and entering C_j, i.e.

[8] which we also need to estimate!

$$P(\text{leave } C_i \text{ and enter } C_j) \approx \frac{\sum_{x_l \in C_j} \sum_{x_k \in C_i} M_{kl}}{N}$$

where N is the sum of all entries of M, i.e. the total number of migrants, which is equal to the total population (a person who doesn't migrate can be thought of as migrating to their starting point, and this is reflected within the construction of M). For ease of index notation, we will write the double sum above as $\sum_{x_i \in C_i} \sum_{x_j \in C_j} M_{ij}$ to refer to a sum over the counties within clusters i and j, so we don't have to keep track of 4 index variables.

We will estimate the marginals similarly to how we estimated the joint probability, but we will instead derive the estimates in a way that provides more insight into the problem, i.e. conditioning on the class into which Ken is born. Using the definition of conditional probability, we can write

$$\begin{aligned} P(\text{leave } C_i) &= P(\text{born into } C_i)P(\text{leave } C_i \mid \text{born into } C_i) \\ &\approx \frac{|C_i|}{N} \cdot \frac{\sum_{x_i \in C_i} M_i^{\text{leave}}}{|C_i|} \\ &= \frac{\sum_{x_i \in C_i} M_i^{\text{leave}}}{N}, \end{aligned}$$

since $\frac{|C_i|}{N}$ is the proportion of the population within class C_i, and $\frac{\sum_{x_i \in C_i} M_i^{\text{leave}}}{|C_i|}$ is the proportion of those people who leave given they start out in class C_i. Similarly, we have

$$\begin{aligned} P(\text{enter } C_j) &= P(\text{not born into } C_j) P(\text{enter } C_j \mid \text{not born into } C_j) \\ &\approx \frac{N - |C_j|}{N} \cdot \frac{\sum_{x_j \in C_j} M_j^{\text{enter}}}{N - |C_j|} \\ &= \frac{\sum_{x_j \in C_j} M_j^{\text{leave}}}{N}. \end{aligned}$$

The reason we didn't directly say to estimate the marginals with these empirical proportions [9] was because when we factorized the marginal into a conditional probability and a prior on social class, we came up with an estimate for another interesting set of probabilities, namely that

[9] which was what we did with the estimate of the joint probability

$$P(\text{leave } C_i \mid \text{born into } C_i) \approx \frac{\sum_{x_i \in C_i} M_i^{\text{leave}}}{|C_i|} \tag{1}$$

$$P(\text{enter } C_j \mid \text{not born into } C_j) \approx \frac{\sum_{x_j \in C_j} M_j^{\text{enter}}}{N - |C_j|}. \tag{2}$$

These conditional estimates themselves already provide insight into ease of movement between the social clusters, and are directly estimable from the matrix M and the clusters $\{C_k\}_{k=1}^K$. But for now, we've estimated the marginals and can finally give an expression for the correlation between classes. Plugging in our estimates, we have

$$\begin{aligned} \text{Cov}(I_{\text{leave } C_i}, I_{\text{enter } C_j}) &= P(\text{leave } C_i \text{ and enter } C_j) - P(\text{leave } C_i) P(\text{enter } C_j) \\ &\approx \left(\sum_{x_j \in C_j} \sum_{x_i \in C_i} \frac{M_{ij}}{N} \right) - \left(\sum_{x_i \in C_i} \frac{M_i^{\text{leave}}}{N} \right) \left(\sum_{x_j \in C_j} \frac{M_j^{\text{enter}}}{N} \right) \\ &= \sum_{x_j \in C_j} \sum_{x_i \in C_i} \left(\frac{M_{ij}}{N} - \frac{M_i^{\text{leave}}}{N} \cdot \frac{M_j^{\text{enter}}}{N} \right). \end{aligned}$$

Normalizing by the product of standard deviations of $I_{\text{leave } i}$ and $I_{\text{enter } j}$, the estimate for the correlation becomes

$$\widehat{\rho}_{ij} = \frac{\sum_{x_j \in C_j} \sum_{x_i \in C_i} \left(\frac{M_{ij}}{N} - \frac{M_i^{\text{leave}} M_j^{\text{enter}}}{N^2} \right)}{\hat{\sigma}_i \hat{\sigma}_j}$$

Since $E(I_{\text{leave } i}^2) = E(I_{\text{leave } i})$ by nature of indicator functions, we plug in the estimates for the standard deviations to get that the above is

equal to

$$\widehat{\rho}_{ij} = \frac{\sum_{x_j \in C_j} \sum_{x_i \in C_i} \left(\frac{M_{ij}}{N} - \frac{M_i^{\text{leave}} M_j^{\text{enter}}}{N^2} \right)}{\sqrt{\left(\frac{\sum_{x_i \in C_i} M_i^{\text{leave}}}{N} - \left(\frac{\sum_{x_i \in C_i} M_i^{\text{leave}}}{N} \right)^2 \right) \left(\frac{\sum_{x_j \in C_j} M_j^{\text{enter}}}{N} - \left(\frac{\sum_{x_j \in C_j} M_j^{\text{enter}}}{N} \right)^2 \right)}}. \tag{3}$$

Note that we can recover the correlation between two specific counties (not clusters) by setting K equal to the total number of counties. In this particular case, the above expression can be coaxed into the following form by multiplying by $1 = \frac{N^2}{N^2}$,

$$\widehat{\rho}_{ij} = \frac{M_{ij} N - M_i^{\text{leave}} M_j^{\text{enter}}}{\sqrt{M_i^{\text{leave}} M_j^{\text{enter}} (N - M_i^{\text{leave}})(N - M_j^{\text{enter}})}}, \tag{4}$$

which is nearly the form for the correlation coefficient of an undirected network [4] whose ij^{th} edge is M_{ij}, and whose nodes are the counties $\{x_k\}$. [10] The only difference is that M_j^{enter} represents a column sum rather than a row sum (expression 1 in [4]), which reflects the fact that M is not symmetric, due to the directed nature of migration.

[10] Note that our expression should be slightly different from the one in this paper, since our graph is directed.

Testing for Significance

The t statistics corresponding to the correlation estimates in (3) or (4) depending on the choice of K are given by the following expression,

$$t_{ij} = \frac{\widehat{\rho}_{ij} \sqrt{N - 2}}{\sqrt{1 - \widehat{\rho}_{ij}^2}}.$$

We can now use a t test to filter out the insignificant edges (migrations between clusters) by discarding those corresponding to the correlations with low p values. [11]

We now have a directed network model of the social structure in the U.S. whose nodes $\{C_k\}_{k=1}^{K}$ are connected by edges which are the statistically significant geographical movements between social clusters. Also note that the generality of the model allows us to set K equal to the number of counties in order to recover the county to county migration network [12].

When we do set K equal to the number of counties, we recover the migration network of people flowing through counties, pictured in Fig. 10. This figure shows the network of county migrations before and after applying the above significance filtering scheme. From this figure we see that the majority of migrations are statistically insignificant according to the notion of cluster dependence developed in the Modeling section.

[11] Just to reiterate, the null hypothesis is the one corresponding to the null distribution, which is the product of marginals. We want to filter out the migrations that don't say anything statistically meaningful about the relationship between leaving one social class and entering another.

[12] If we set K to be the number of counties, then the K nearest neighbor algorithm will classify every county to its own cluster, yielding the original migration network between counties.

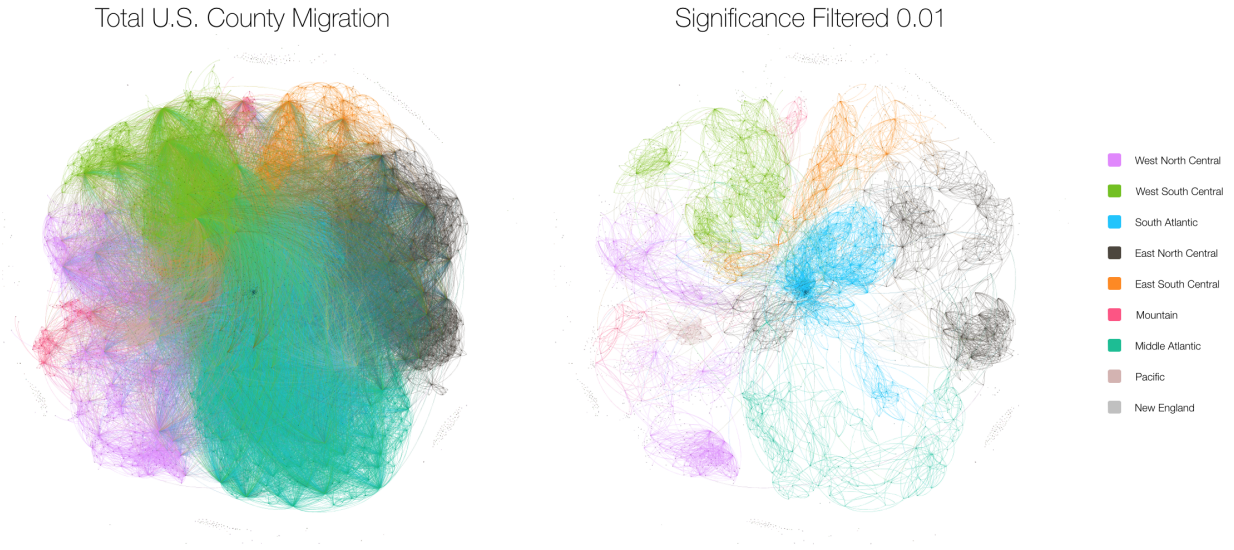

Fig. 10: A stunning visualization of county to county migration from 2012 to 2013 in the U.S. using the *Gephi* software [12]. A node represents a county and an edge represents the migration of residents from source county to target county. The left network has 3,142 nodes and 94,601 edges. The right network has 2,003 nodes and 8,962 edges.

Results

Geographic Mobility throughout K-Means Clusters

To re-emphasize, we assumed the notion that geographic mobility is an indicator of one's social mobility, where one moves into better-education, high income, better housing areas due to one's increase in wealth and hence moving up the social ladder. Through unsupervised K-Means clustering, we generated 6 clusters of counties in the dimension of economic, demographic, and educational factors, as well as identifying which migration flows are significant and not simple noise at the county level.

We now connect the significant migration flows with the unsupervised clustering of counties to determine geographic - and hence social - mobility across clusters. For each source cluster, we determined the proportion of total migration flow in that source cluster, for each destination cluster. In other words, we utilized the clustering of source and destination county, to aggregate migration flows on a cluster level. Doing so allows us to explore the overall movements into different socioeconomic clusters, as well as determine where bottlenecks may be in social mobility. A summarizing figure (Fig. 11) is shown below.

First, we note that the clusters are in order of average weighted median income, for easier interpretation. We see that for clusters 1 through 5, the distribution of proportion of migrations into different destination clusters is roughly the same; there is little movement into

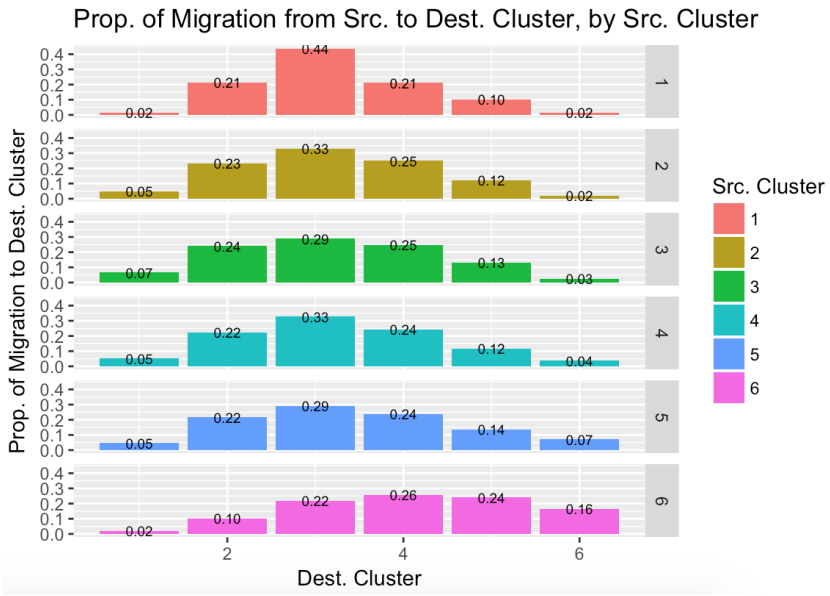

Fig. 11: Proportion of Migration from each source cluster to each destination cluster. Each facet of bar charts represent the source cluster, with its x-axis of categories being the destination clusters.

the either end of the tails - the extreme ends of the socioeconomic clusters. Furthermore, we see a small, but growing migration trend into higher socioeconomic destination clusters as the socioeconomic level of source clusters also move up, especially into clusters 5 and 6, which supports our intuition.

On the other hand, however, we also detect interesting trends and possibilities. For example, we see that the middle-class source clusters 2 through 5 possess very similar distributions and specifically, proportions of migrations in each of the destination clusters. This indicates that in the middle-class, there is no significant change in migration movements into higher destination clusters, even when the source cluster socioeconomic level is increasing. Migration trends remain constant across the five clusters, and may indicate possible bottlenecks in climbing the ladder into upper socioeconomic classes.

Another interesting quality we see is in source cluster 6, which share a more uniform distribution in migration into destination clusters. Specifically, people in high-income counties tend to stay in that socioeconomic level (or slightly below in cluster 5), and seems to circulate within medium-high to high socioeconomic clusters. This demonstrates a hint that the rich stays as the rich in comparison to other economic classes.

Overall, the distribution bar plots show that while the poorest are able to easily move into the medium classes, there is a possible bottleneck in moving to the higher-level socioeconomic clusters from middle income clusters. Based on this initial result, we analyze more

Fig. 12: Network of social clusters with directed edges representing likelihood of migration in that direction. These probabilities of entering one class conditional on starting in another are estimated using (1) and (2), which are derived above in the Modeling section. Thicker arrows correspond to higher likelihood of migration.

deeply in identifying largest bottleneck areas, possible causes, and recommendations related to educational aspects.

Estimating Effects of Education on Solving Bottlenecks through Regression

After discovering the potential of bottlenecks in mobility across different clusters, we narrowed our focus to analyzing whether educational aspects specifically impacted the likelihood and propensity for people to move across clusters. We hypothesized that educational opportunity is a potential cause of migration bottlenecks, whether it be due to the quality of education in one's current residence (source cluster), or the number of schools in the destination cluster, etc.

To estimate the impacts of education on migration, we utilized a LASSO Poisson-distributed generalized linear regression model, since our dependent variable focuses on the *count* of people migrating. We first specifically chose a source and a destination cluster to see the potential migration bottleneck in the migration outflow. Our observations (rows) in the cluster-to-cluster outflow were the county-to-county outflow between the two clusters, with predictors being various educational, as well as demographic, social, and economic features of the two counties in question. Our dependent variable was the number of people in this county-to-county outflow. The regression equation can be written as

$$P(\mathbf{h}) = \mathbf{w}^T\mathbf{h} + \varepsilon$$

where $\mathbf{w}$ is the coefficients vector, $\mathbf{h}$ is the column vector of features, ε is the noise, and $P(\mathbf{h}) \sim$ Poisson [13]. We used the L1-penalized squared error loss function,

$$L(\mathbf{w}) = \frac{1}{2}\|P(\mathbf{h}) - \mathbf{w}^T\mathbf{h}\|_2^2 + \lambda\|\mathbf{w}\|_1.$$

As an illustrating example, we show below the estimated coefficients for the county-to-county outflow between cluster 2 and cluster 3, as well as the proportion of times each coefficient was included as "significant" - meaning nonzero - out of the 36 possible cluster-to-cluster outflows. Note that we omit the standard errors in this cross-validated LASSO model due to the fact that heavily penalized methods give strongly biased estimates [11], and so the variance of the estimate plays the much smaller role in the error. [14]

Analyzing the coefficients of cluster 2 to cluster 3 outflow model, we see that multiple factors relating to education have an impact on the number of migrating people across the two clusters. For example, we see that for a unit increase in the average number of teachers in

[13] Note that since ε is zero mean, this implies that the i^{th} response value is distributed Poisson($\mathbf{w}^T\mathbf{h}_i$)

[14] The error can be decomposed into a bias and variance term, and since bias dominates in the case of this LASSO regression, the variance term is insignificant. More details can be found on page 18 of Goeman's paper (citation 11).

the source cluster counties, there was on average an 0.1852 increase in the number of migrating people, suggesting that adding more teachers in the source cluster counties may enable more migration flows, less bottlenecks, and hence more opportunity for upward social migration. A similar statement can be said for destination cluster average enrollment numbers in schools. We also see a relatively larger coefficient on `weight`, which is the proportion of a specific county-to-county outflow over all migration outflows from the source county. Intuitively, this matches our common knowledge of focusing on migrations with heavier weight and importance to improve migration flows. Lastly, it is important to note that we included other features that we may need to control for such as population, age variability, etc to minimize omitted variable bias as much as possible. Other controls and features were not included due to multicollinearity issues.

The next column we see is the proportion of times the specified coefficient was considered nonzero, or significant out of the 36 unique cluster-to-cluster outflow models we could create. Here, besides intuitive and control-related variables such as `weight`, `src_population`, `dest_AgeSE`, etc., we see that source cluster's average number of teachers was considered a significant effect in 25 percent of the models. Other relatively important variables include source cluster's pupil expenditures, destination enrollment numbers, and destination education agencies, all of which are education-related aspects. The fact that these education-related factors were considered important in a higher proportion of cluster-to-cluster outflow models than the other variables imply that these should be the areas of focus in policies aimed at improving social conditions. By impacting these education-related aspects, we can help better facilitate migration flows, and hence provide better opportunities for people to climb the social ladder and reduce the inequality gap.

Table 1: Coefficient Estimates and Proportion of Times of Significance per Coefficient.

	Coef	Prop. Sig. Coef
(Intercept)	3.1541	1.0000
weight	0.2584	0.7222
src_EducAgency	0.0000	0.1111
src_State	0.0000	0.0278
src_numTeachers	0.1852	0.2500
src_PupilExp	0.0000	0.1667
src_Gini	0.0000	0.0556
src_population	0.1964	0.6667
src_Income	0.0000	0.1389
src_AgeSE	0.0748	0.3611
src_AvgScore	0.0000	0.1111
dest_EducAgency	0.0000	0.1389
dest_State	0.0000	0.0556
dest_Enrollment	0.1404	0.2500
dest_PupilExp	0.0000	0.1111
dest_Gini	0.0000	0.0278
dest_Income	0.0000	0.1667
dest_AgeSE	0.0045	0.2778
dest_AvgScore	0.0000	0.1111

Conclusion

Growing socioeconomic disparities throughout the United States have motivated fiery discussions among its citizens regarding how to use our resources in order to reduce the gap. It is widely agreed upon that education and social mobility are nearly impossible to analyze without careful consideration of the other. In fact, many other factors of American life besides education are also intimately intertwined with social class, making the analysis of the concept an incredibly complex task. In this report, we approached the elusive concept of social mobility with a collection of data driven techniques, rigorous reasoning, and creative modeling in order to obtain quantitative results.

The LASSO regression yielded results to support the claim that education related factors are significant and contribute positively to movement within the social mobility network. In addition, our network model of social mobility culminated in a distribution of migration between social classes summarized in Fig. 11. The existence of a right and and more noticeably left skew in the top and bottom rows reinforce intuitive results with a quantitative backing. The interpretation of these skews is that mobility between classes is least observed in the most upper and lower classes, which helps to explain why the disparities seem to become more extreme.

References

[1] USNews. "US News Best Global Universities | US News Education." U.S. News and World Report, U.S. News and World Report, www.usnews.com/education/best-global-universities/rankings.

[2] Leonhardt, David. "America's Great Working-Class Colleges." The New York Times, The New York Times, 18 Jan. 2017, www.nytimes.com/2017/01/18/opinion/sunday/americas-great-working-class-colleges.html.

[3] US Census Bureau. "Americans Moving at Historically Low Rates, Census Bureau Reports." The United States Census Bureau, 16 Nov. 2016, www.census.gov/newsroom/press-releases/2016/cb16-189.html.

[4] Ronen, Shahar, et al. "Links that speak: The global language network and its association with global fame." Proceedings of the National Academy of Sciences, vol. 111, no. 52, 2014, doi:10.1073/pnas.1410931111.

[5] Gilbert, Dennis (1998). The American Class Structure. New York, NY: Wadsworth Publishing. ISBN 0-534-50520-1.

[6] James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. New York. Springer, 2013.

[7] Mike Bostock's Blocks, bl.ocks.org/mbostock.

[8] Tibshirani, Ryan J., and Jonathan Taylor. "The solution path of the generalized lasso." The Annals of Statistics, vol. 39, no. 3, 2011, pp. 1335-1371., doi:10.1214/11-aos878.

[9] Casella, George, and Roger L. Berger. Statistical inference. Brooks/Cole Cengage Learning, 2013.

[10] Wickham, Hadley. "Tidy Data." *The Journal of Statistical Software*, 59, 10 (2014).

[11] Goeman, George, Meijer, and Chaturvedi, Nimisha. "L1 and L2 Penalized Regression Models." https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf.

[12] "The Open Graph Viz Platform." Graph exploration and manipulation, gephi.org/.

Appendix

In terms of future research, we'd like to investigate other network analysis techniques in order to identify counties most "central" to the network. A technique that we'd have liked to try out but didn't have time was eigenvector centrality, which characterizes the connectedness of any particular node to the rest of the graph.

In addition, we didn't have time to implement the computation of correlations [15] between the six social clusters, although we did implement and visualize the case for K equal to the number of counties (see Fig. 10). Given more time, we would have liked to obtain the correlations corresponding to $K = 6$ and see if they were consistent with the rest of our analyses.

[15] given by expression (3)