Analyzing and Predicting Internal Migration Patterns in the USA

Noah Friedman, Sam Gussman, and Jeremy Wood

Introduction

The story of the United States is a story of migration. From the initial colonization of the Americas, to the westward migration with manifest destiny, to the California gold rush, vast movements of people have defined the history of the country for good or ill. In the last century, availability of jobs, social conditions, and a host of other factors have caused massive internal migrations that have reshaped the character of the nation. The great migration (1910-1960, see figure 1.) saw six million African Americans move from the south to the north and midwest in search of better economic opportunities and less acute racism and exclusion. After 1960, migration to the sunbelt has reshaped the country by shifting the demographic center of the United States away from the north and midwest and towards sunny and rapidly growing states like California, Arizona, Florida and Texas. For muchs of the 60s, 70s and 80s, American cities either shrunk or stagnated as wealthier families migrated to the suburbs. Now this trend shows signs of reversing as many young professionals are drawn to city centers.

What trend is next? Are we in the midst of the next migration that will change the fabric of our country? Can we predict where migration flows are headed in the future? Are there interesting undiscovered migration trends that lurk beneath the data? These are the questions that our project seeks to answer. We have taken advantage of the wealth of data that the US census bureau has collected on county to county migration flows over the past three decades. Leveraging this data, we have analyzed the dynamics of past migration and built a dynamic model that we use to predict future migration. By applying and extending the computational techniques previously used to study migration patterns, we can improve our ability to predict future migration patterns, and unearth novel and heretofore unknown features of the network of internal migration in the United States.

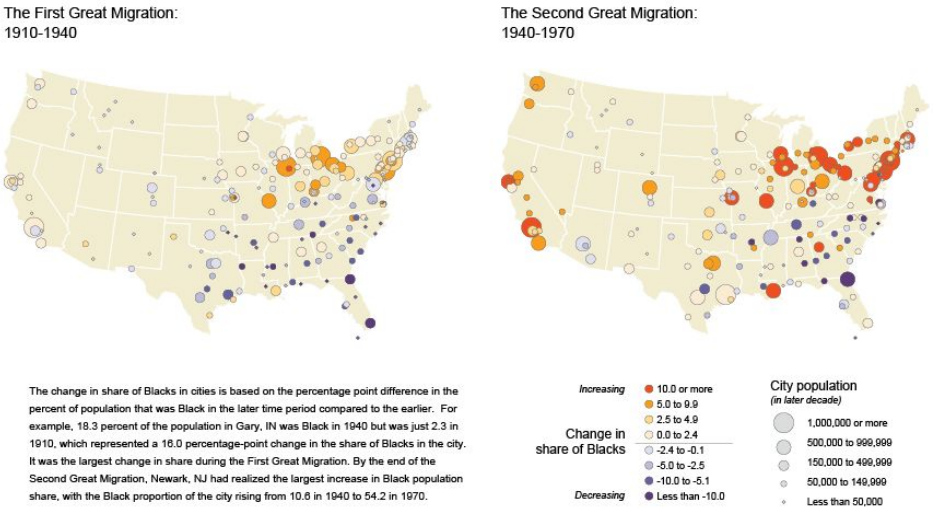


figure 1. The great migration was one of the largest internal migrations in US history

Related Work

Our approach to analyzing migration networks is heavily indebted to the analytic framework presented in *A universal model for mobility and migration patterns* by Filipo Simini, Marta Gonzalez, Amos Maritan, and Albert Laszlo Barbasi. Their paper presents and summarizes two different models for predicting commuting (analogous to migration), the gravity model and a model they created, the radiation model. The authors illustrate the limitations of the gravity model and present solutions in their design of the radiation model. The authors then go on to demonstrate that the radiation model performs markedly better than the gravity model when predicting population movements between counties, allowing the authors to propose it as a new state of the art model for predicting migration flow. However, as figure 2 shows, it is still far from perfect, leaving ample opportunities for a more specialized algorithm to further improve performance.

It is worth examining the basic intuition of the gravity and radiation models as they form the basis of our project. The gravity model predicts migration based on a mechanism analogous to the Newtonian

$$T_{ij} = \frac{m_i^{\alpha} n_j^{\beta}}{f(r_{ij})}$$

law of gravity. The gravity model represents the propensity to move between two cities as proportional to their respective populations over the distance between the two of them. Formally, it is defined as follows:

Tij is the number of individuals that we predict will migrate from city *i* (the source) to city *j* (the destination).
m represents the population at the source, *n* is the population at the destination, alpha and beta are adjustable exponents and *f()* is a function to model the "distance" between two locations.

The authors extend the gravity model in two ways to create their new and improved radiation model. Firstly, they take into account the number of people who actually move from place to place in a given year relative to the entire population. Secondly, they take into account the total population that lives in the area between the two cities. This is a quite insightful modification--intuitively it makes sense

$$\langle T_{ij} \rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}$$

that when deciding to move from one city to another, an individual is probably considering a wide range of cities near the city they actually decide to move to. Formally, the authors define the radiation model equation as follows:

Here *Ti* represents the number of people who migrate from city *i* in a given year. *Sij* represents the number of people living near cities *i* and *j,* specifically the number of people who live inside the circle

centered on city *i* with its radius equal to the distance between city i and j, excluding the people who live in cities *i* and *j* themselves.

Also of note is Thomas Schelling's chapter, *Micromotives and Macrobehavior,* from *Norton 1978, Chapter 4: "Sorting and Mixing."* Schelling's article looks at how different small motives can sum up to create macro-phenomena such as segregation. He takes into account how even in the absence of top down directed efforts at segregation such as discriminatory policies, segregation is still often an inevitable product because of the dynamics that drive individual decisions. He explains a model that he calls the self-forming neighborhood model which can explain the way in which neighborhoods become segregated. Schelling provides an interesting perspective on the sorts of micro level preferences or structures in the graph that can lead to large structural properties. His frame of analysis and research heavily influences the way in which we featurize our dataset. We selectively include some network measures and not others and design augmented network based machine learning features with Schelling's frame of analysis in mind.

Model and Algorithmic Method

Our project implements the gravity and migration models presented in Simini, Gonzalez, Maritan, and Laszlo with the appropriate modifications necessary to apply the models to the US census dataset. We then build a competing machine learning based model and compare its efficacy to the radiation and gravity models. We featurize our dataset, incorporating node based features (characteristics like population) and network features. We then train a model using those features using SGD and linear regression. By applying machine learning side by side with the gravity and radiation models we hope to be able to discover the features that make each algorithm efficacious and the features that limit them. We then hope to be able to propose changes to the gravity and radiation models based on our machine learning results that can create a maximally effective combined algorithm for predicting migration.

i. Adapting the Gravity and Radiation Models

Adapting the Simini, Gonzalez, Maritan and Laszlo model to the US Census migration dataset requires myriad changes and optimizations. Below are the mathematical definitions of the gravity and radiation model:

In the paper, *Tij* represents the total number of people who are expected to *commute* between cities *i* and *j*. We make it applicable to migration by redefining *Tij* as the number of people we expect to *migrate* from

$$T_{ij} = \frac{m_i^{\alpha} n_j^{\beta}}{f(r_{ij})} \qquad \langle T_{ij} \rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}$$

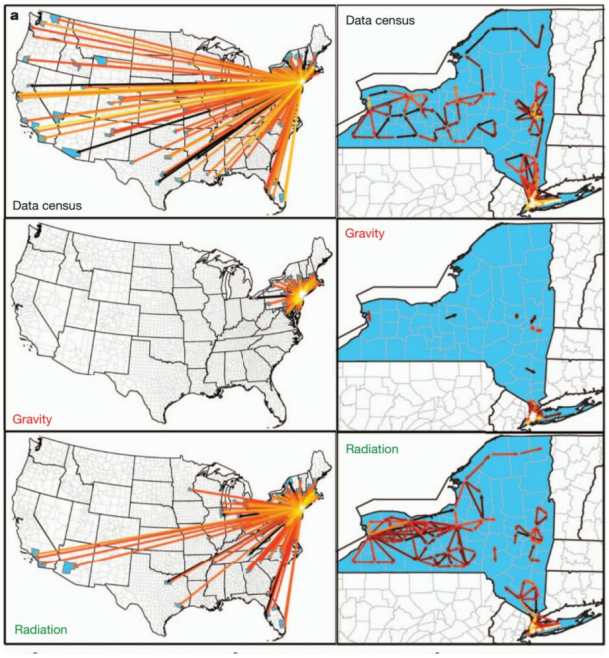
cities *i* to city *j* in a single year (as our dataset shows year to year migration). *m* and *n,* which represent

the total populations of cities *i* and *j* remain unchanged in our adaptation of the formula. Simini et al set alpha + beta = 1**.** Simini et al. define *f(rij)* as *r^4.* We use this same function for our baseline re-implementation of the gravity model.

The radiation model includes the term *Ti.* Simini et al. define that *Ti* as the number of people from city *i* who commute each day. They define *Ti* as the population of city *i* times the fraction of the population of the country as a whole that commutes each day. We define our *Ti* to be equal to the fraction of people in a given city that migrate each year (*p*) times the population of the city. We have empirically defined *p* to be .0376 based on the following calculation:

> There were 312,295,000 people in the most recent census data. According to the data, 36,324,000 migrated. Subtracting out those who moved within the same county (because our dataset only takes into account intra-county migration), and those who moved to or from a foreign country, we are left with 11,746,000 internal migrants, who comprise 3.76% of the US population. Thus we define *Ti* to be equal to .0376**mi* (the population of city *i*).

As mentioned previously, the radiation model defines *sij* to be the total number of people who live within a circle of radius *rij* centered at city *i* (see figure 3). We define the circle by treating cities *i* and *j* as two points in two dimensional x, y space based on their latitude and longitude. We discount the effect of the curvature of the earth on the distances and geometry for simplicity and because the distances within the United States are small enough not to be dramatically affected by the curvature of the earth.



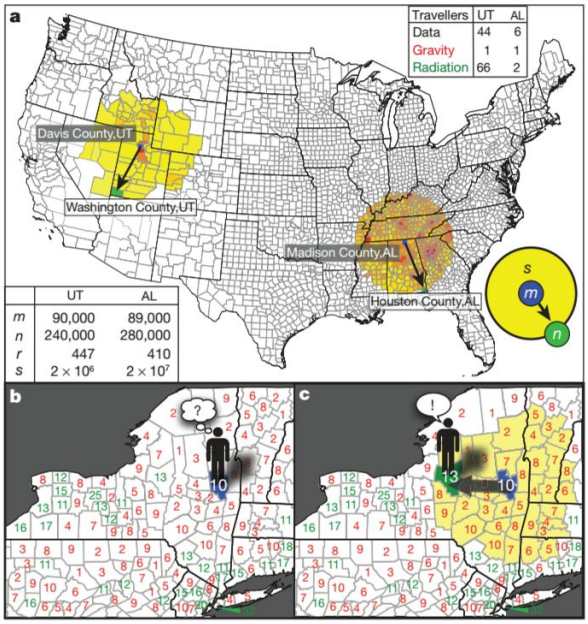


figure 2 (top left). Simini et al. demonstrate graphically how their model significantly outperforms the gravity model in predicting migration

figure 3(top right). Simini et al. show how the circle of surrounding counties works in their model

ii. Featurizing the Dataset and Designing our Linear Classifier

After implementing the gravity and radiation models from Simini et al, we worked to design a competing machine learning application. Our hope was to use the results from the machine learning migration prediction algorithm to contextualize and suggest improvements in the design of the gravity and radiation models.

Both the gravity and the radiation model are fundamentally centered on geographic distance. The gravity model has the parameter *rij,* representing the geographic distance between two cities. The radiation model uses *rij* to define *sij,* a circle that encompasses all the cities. The algorithms are premised on the intuition that all else being equal, people prefer to move to places that are *closer* to where the live. But *closeness* is not the same as geographic distance. To motivate this understanding, consider the following example: Palo Alto, East Palo Alto, and Tiburon are cities in the Bay Area. Palo Alto and East Palo Alto sit adjacent to each other in the south bay area, while Tiburon is about 50 miles away, north of San Francisco in Marin. Palo Alto and Tiburon are both extremely affluent communities with median home prices well over two million dollars. East Palo Alto on the other hand, has a significantly lower median house value (around 600,000). The gravity and radiation model only factor in geographic distance. As such, they would believe that someone is significantly more likely to move from Palo Alto to East Palo Alto than from Palo Alto to Tiburon. We however know that this is not the case, someone in Palo Alto may feel that the community in Tiburon is *closer* socioeconomically and demographically and prefer to migrate there even though it is significantly further away geographically. By analyzing the features we use for our linear classifier and the efficacy they have in improving our results, we hope to be able to see concrete ways to propose improvements to the gravity and radiation models that can overcome these limitations.

We decided to implement linear regression using a stochastic gradient descent model trained on hand selected features. We extracted multiple features from our dataset. We are predicting an edge feature, migrant flow, with a combination of node and edge features. As such, we sought to design our features so that they would be relevant for predicting edge characteristics rather than node characteristics.

We reuse distance and population as features even though they show up as parameters in the gravity and radiation model. We do this because we believe that there may be other more subtle effects of population besides what the gravity and radiation model capture. Simini et al. view population in analogy to mass-the more of it there is the more attraction a city exerts on potential migrants. Intuitively there are likely also other more subtle effects of population at play. For example, maybe people from small towns are more predisposed to relocate to other small towns or people in cities are more prone to relocate to midsize suburbs. We try to prime our linear regression model to capture these subtleties by giving it population as a feature along with auxiliary features such as the product of the population of two cities or the inverse of its population.

Our model also integrates network features and leverages them as components of our linear regression model. We use the following features:

- Closeness Centrality
- Degree Centrality
- Node Cluster Coefficient
- Graph Cluster Coefficient
- Betweenness Centrality

- In Degree
- Out Degree
- PageRank
- ID corresponding to its connected component

Each one of these features offers a unique lens for interpreting the migration data. Closeness centrality captures cities that are hubs, where connections to people who have migrated there in the past are plentiful. Degree centrality similarly will capture hubs of the migration graph, but in a subtly different way. Cities with high degree centrality will be cities with dynamic populations with people freely and rapidly flowing in and out. The clustering coefficients can capture areas of the graph where there is a dense network centered around a particular region. This can capture networks of areas where people move freely and oftenly because of the connections that link the areas. In degree and out degree can capture cities that are attracting disproportionate numbers of migrants or who are shedding large numbers of migrants. This can actually serve as an interesting proxy of city ascension or decline as validated by changes in population. PageRank offers insights into the dynamics of the network-whether there are cities that are particularly likely to be visited by migrants, namely cities that are likely to draw people at some portion of their life. Finally, the ID of the connected component feature could help pinpoint the regionality of certain migration networks or the ways in which

Our linear regression model takes all of these features into account and uses them to predict volumes of migration flow. We compare this to the results that the gravity and radiation models give us.

iii. Technical Implementation Challenges

Working with a massive dataset has presented us with numerous unforeseen technical challenges. The US Census data on migration represents a graph with thousands of nodes and millions of edges. Furthermore, the edges all change for each year of migration represented in the dataset. To mitigate this challenge, we compile all the data from multiple years (which the census bureau gives us in myriad separate files) into a single combined file. We then preprocess the data file so, which is written in as a JSON before we run our main linear regression algorithm. This gives us tremendous runtime performance improvement. Furthermore, we face challenges with calculating some of the network features, because they require $O(n^2)$ time. To solve this we calculate our network measures ahead of time because we are analyzing a static graph.

Results and Findings

We found that our linear classifier significantly outperformed both the gravity and radiation models in predicting the migration flows between counties. We evaluated all models by comparing their predicted migration flows to the actual observed migration flows. This gave us a grasp on the different magnitudes by which the algorithms succeeded or failed in predicting migration. We then looked at our linear regression model and quantified the degree to which different features affected our predictive power. By doing this we pinpointed the features that were most useful for the linear regression model, and thus what we believe to be the most important features impacting migration.

Our model managed to predict migrant flows with 7% standard error, whereas the gravity model performed atrociously with 19,380 standard error, and the radiation model performed well, but still worse

than our algorithm, with -0.603 standard error. We were surprised by this result. We suspected that our model would be vulnerable to overfitting, or that it would overemphasize migration to large cities. Instead, it seems that a simple linear classifier was able to discover and adequately understand the majority of migration in the United States. Our results demonstrate that in this context the brute power of machine learning can still outstrip our best dynamical based explanations.

Delving further into the results, we were further surprised to learn that the most effective predictive features for our linear classifier were not the characteristics the gravity and radiation model take into account (population, distance etc). We discovered that pagerank and betweeness centrality stood out far and away as the most informative predictive features for our graph. This means that these two network measures are capturing profound and relevant insights about the US internal migration network. This makes intuitive sense for betweeness centrality--a city that is on multiple shortest paths is a city that commonly attracts migrants from a diverse set of areas and that is deeply embedded in the economic networks of the United States. PageRank is central for a similar reason. PageRank captures nodes in a graph that are likely to be visited. In the context of migration, a node that is likely to be visited is a city that will likely draw a constant and wide ranging stream of migrants over time. Being able to pick up on these dynamically central cities turns out to be more informative and powerful that pinpointing cities that simply have high degree centrality.

It is also instructive to look at the network features that did little to improve the performance of our linear classifier. For one, features like population and distance actually had only a marginal impact on the performance of the linear classifier. In the end, we actually threw away many of our augmented features related to distance and population because they significantly slowed down the program without measurably improving performance. This observation is particularly interesting to us in light of the fact that both the gravity and the radiation models use population and distance as the central parameters for their equations. Do we not need these features because the network centrality measures manage to implicitly encode this information from analyzing the structure of the network? Or is it because population and distance are actually fundamentally uninformative for predicting population movements? The answer has eluded us, but our intuition is that it is a little bit of both.

Distance and population must have some bearing on migration. Obviously people are more prone to move to places that are closer to them, and more likely to move to a big city than a small one. But we realized with our linear classifier that much of this information is actually already embedded in the graph itself. The past history of migration--the fact that in the past people have moved from city A to a nearby city, city B, but not to a far away city, city C, is encoded in the graph. We will have edges from city A to city B but not from city A to C. By using a dataset as big and expansive as the US Census dataset, we can largely empirically bootstrap the entirety of the assumptions present in the gravity and radiation model from empirical data. But because we have observed that removing the specific features added to our model to account for population and distance do not measurably affect the accuracy of the results our linear classifier produces, we can deduce that the structure of the migration network exhaustively contains all the relevant effects of population and distance on migration and including these variables is simply redundant. This is an intriguing finding--distance and population may have been the driving force or many of the dynamics and structures of migrations present in the United States, but now today, the past history of the network and its structure tell us all we need to know without looking at the dynamical forces at work.

In the end, our project demonstrated that training a machine learning model on the historical patterns of US migration can create a predictive system that outperforms even the state of the art dynamical prediction models. This is a nice achievement for networks where we have extensive data such as the census network--we conclusively show that building a machine learning model with smart features can lead to quite accurate (7% error) performance. But there are many networks with similar dynamics to the migration network that don't come with such an extensive suite of useful training data. For these networks, we still need to use dynamical models such as gravity and radiation rather than machine learning based on past network history. But we believe that the results demonstrated by our linear classifier can offer fertile ideas for extending and improving the gravity and radiation models. We discovered that a complex set of interdependent network centrality measures presented as features to a linear regression model offer remarkable predictive power. Going forward, we would hope that further research could disentangle the interdependencies of these centrality measures and discover a single parameter or network feature that gives our model it's extensive predictive power. By finding this model we could propose a way forward to a better and more expressive form of the gravity and radiation models and better way to predict migrations in many dynamic contexts.

Works Cited

"City Mayors: Largest 100 US Cities." *City Mayors: Largest 100 US Cities*. N.p., n.d. Web. 09 Dec. 2015.

Schelling, Thomas. *Micromotives and Macrobehavior*. N.p.: n.p., n.d. Print.

Simini, Filipo, Marta C. Gonzalez, Amos Maritan, and Albert Laszlo-Barbasi. "A Universal Model For Migration and Mobility Patterns." *Nature* (n.d.): n. pag. Web.

Wikipedia. Wikimedia Foundation, n.d. Web. 09 Dec. 2015.

Zillow. N.p., n.d. Web.