Abstract

Eigenvector centrality is a common measure of actor power in networks and has proven to be especially useful in analyzing networks because it takes into account the both the connections of an actors as well as the power of the actors it is connected with. This has proven especially useful in analyzing networks where more analog measures like degree centrality fail. In this paper, I calculate the eigenvector centrality of neighborhoods in Chicago by using the Longitudinal Employer-Household Dynamics (LEHD) dataset published by the US Census Bureau which is used to create a commuting (live-work) network. I then explain the meaning of a neighborhood's centrality as well as use the centrality ranking to further analyze attributes of neighborhoods such as neighborhood investment, crime, and housing valuation.

1 Introduction

Network theory has been used to study how actors interact in a diverse set of disciplines from social network analysis to and trading networks to webpages and molecular networks. In many fields it has lead to important new insights and created entire industries. Many mathematical tools have been developed to help understand how networks are structured and rank actor importance. I will use a common measure used to quantify actor importance, eigenvector centrality, to find important neighborhoods for employment and housing in the network of neighborhoods in Chicago. The eigenvector centrality method for computing actor importance will show the relationship between employment and housing centers in Chicago and allow us to ask additional research questions such as how wealth and employment is distributed across the city and how connections between neighborhoods influence other local socioeconomic attributes. of a neighborhood. While there are many approaches to analyze job and neighborhood growth in cities that spans academic disciplines from economics to statistics and sociology and use traditional statistical frameworks such as regression, causal frameworks, and policy evaluation, I hope that this method will provide an additional lens for researchers to analyze the structure of a city and allow them to ask new and different questions. In this paper I will use the Longitudinal Employer-Household Dynamics dataset published by the US Census.

2 Literature Review

Network centrality measures have been developed as a way to analyze how actors are ordered, evolve, and interact. Several measures of centrality have been studied in order to address a central question: *Who is the most important actor in a network*? While there is no agreement on one measure of centrality, many researchers suggested centrality measurements based on how a specific network is structured, how actors interact, and what the researchers mean by “importance”. For example, “degree centrality” is a classic measure of centrality that counts how many connections an actor has. A node is more important if it has many neighbors and less important if it has fewer neighbors. “Closeness centrality” measures the degree to which an individual is near all other individuals in the network and is the inverse of the sum of the shortest distances between each node and every other node in the network. It has been used to study how long it would take to spread information from one actor to all others sequentially. “Betweenness centrality” quantifies the number of times a node acts as a bridge along the shortest path between two nodes and has been used to quantify the control of information in a social network.

In this paper I will focus on “eigenvector centrality” which is a measure of influence of a node in a network and corresponds to the first eigenvector in the connectivity matrix of a network. The first researcher to apply the mathematics of eigenvectors to geography was P.R. Gould. In his paper, On the Geographical Interpretation of Eigenvalues, his goal was less motivated by a specific hypothesis but more of a curiosity to determine if this mathematical structure could uncover a pattern in very complex situations.[1] His hope, which many computational social scientists share, was that underlying complex phenomena might be a mathematical idea that could provide a meaningful geographic interpretation. To explore this idea, he mapped the road network of Uganda and created a connectivity matrix of this network on a binary scale, 1 indicating if two cities were connected and 0 if they were not. He calculated the first four eigenvectors of these matrices in 1921 and 1935 and compared the results between the two years. He found the first eigenvalue, which is centered around the city of Kampala, was by far the most connected town owning to the number of direct linkages and its central location while the city with the next highest value, Entebbe, was also very connected. He then examined a new connectivity matrix of the cities in 1935 and described how several cities have become more important as new roads and cities have been built. He notes that the successive eigenvectors and eigenvalues are a “pull out” of small regional networks within the trading structure of the region. Gould makes an initial attempt, though vague, to describe the meaning of his eigenvector derivation. He explains that vectors representing well-connected towns will not only lie in the middle of a large number of dimensions but will tend to lie close to the principal eigenvalue. On the other hand, towns that are moderately well connected will not lie in the middle of so many dimensions as the well-connected towns and will tend to form small structural clusters on their own. This interpretation has been named the “Gould Index of Accessibility”.

Eigenvector centrality has been studied to explain many of phenomena in the social sciences and some results have given new insights when previous theories and measures of centrality failed. For example, Cook et al. argued that typical centrality measures like degree centrality have failed to predict power distributions in exchange networks (networks where actors bargain or trade which each other).[2] Through theoretical and simulated results they showed that in negotiations it is advantageous to be connected to those who have few options and being in a central position does not make an actor more powerful. This is because if an actor is connected to someone who is powerless they will have more negotiating power because if the powerless actor was connected to other powerful actors, their bargaining power would increase. Thus, they suggest that a more general conception of point centrality needs to be developed that takes into account power dependency as well as closeness. In situations where degree centrality fails, social scientists have proposed other measures that fuse power-dependency and closeness. Bonacich suggested that eigenvector centrality makes a good centrality measure in these networks because it takes into account the number of connections an actor has as well as the centrality of the actor it is negotiating with.[3] Thus, if an actor is connected to a more important node its eigenvector centrality will be higher than if it connected to a less important actor.[1] Eigenvector centrality also has benefits over other measures of centrality as it can be used for signed graphs, adjacency graphs, or value based graphs. For example, networks graphs with negative connections include dating and friendship networks where reciprocation is not necessary or trade where one actor sells a product to another.

Researchers have since applied eigenvector centrality across a number of fields and applications. For example, Tinkler described these eigenvalues in the context of a rumor spreading through a social network. He described a social network where actors were connected through social ties (1 indicating two actors knew each other and 0 indicating they do not) and a rumor starting at some vertex i at time 0. As time progresses, the rumor will be spread throughout the network according to the connections between actors in the social network. If someone knows the rumor they tell it as many times as they heard it to all the people they are in direct contact with. In his example, an actor can also start an "anti-rumor" which is denoted by a negative value which can also propagate through the network. As this process repeats after a large number of periods, the distribution of the rumor will also be given by the principal eigenvector. In other words, after time progresses the eigenvector is the chance that the rumor has spread to a specific actor in the network.[4]

Another interpretation of eigenvector centrality was given by J.W. Moon who described the eigenvalues as the ranking of players after an iterative round-robin competition. In his example, there is a tournament between n players and the win-loss outcomes create a square matrix with 1 if a player beat their opponent and 0 if they lost to their opponent. A player gets a ranking by beating another player, however if they beat a stronger player they will get a higher rating boost than if they beat a weaker player. After the tournament has elapsed into an equilibrium where rankings are consistent, the player's rankings will correspond to the ranking of the principal eigenvector of the win-loss matrix.[5]

[1] bonacich2007some

Eigenvectors are also commonly used to analyze population flows over time. For example, in a scenario with two cities and a constant flow of the population moving between them the equilibrium state can be represented through its eigenvalues and eigenvectors. This is because the information in the eigenvectors and eigenvalues contain all information needed to study how a system moves from one state to the next.

One last interesting application of eigenvector centrality was implemented by Sergey Brin and Lawrence Page in creating a web search engine which served as the basis of the first version of Google's search engine. They created a database with the hyperlink network of over 24 million pages and a PageRank algorithm to order results of a query. Pages are arranged in a network based on their hyperlinks to other pages. PageRank does not count all links equally though, but normalized the weighting by the number of links on the page. The PageRank of a webpage was calculated using an iterative algorithm that corresponded to the principal eigenvector of the normalized link matrix of the web. They give a few intuitive justifications why this ranking works. One was imagining a "random surfer" who is given a web page at random and keeps clicking links. The PageRank of a page is the probability the random surfer will land on a page. Another interpretation was that a page will have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and also have a high PageRank so it was a way to combine a ranking that measured reputation and ubiquity.[6]

Another important and related area of research is studying the secondary eigenvector in the network. For example, in Gould's analysis, the second eigenvector was able to pick out significant geographic subsystems in the transportation network of Uganda. Often there remains further information about the network structure that subsequent eigenvectors can explain. For example, while the first eigenvector reflects volumes and strengths of connections among the actors, a second or third eigenvector can delineate those in separate groups within the network who behave in somewhat equivalent manners, or other elements of network structure that can be informative in understanding the actors and the patterns that link them. Iacobucci et al demonstrate that the extraction of only the first eigenvector can be insufficient in gaining comprehensive understanding of the network.[7] The example they give is from a communication network between researchers. While the first eigenvector retrieves the principal structure of the social network it is often similar to common measures of centrality. By extracting the second and third eigenvector several classes of network structures and actor attributes were clearly pulled out and interpreted. This is because if the first eigenvector the second eigenvector by necessity will be uncorrelated with the previous eigenvectors and therefore uncorrelated with the traditional degree of centrality. This lack of redundancy indicates the supplemental information that the eigenvector can bring to the network modeler. Another example is given by Bonacich [8] in analyzing cliques in a social network. Consider a social network that is made up of many cliques where each clique has zero communalities between another clique and all individuals within the clique are connected to each other. In this case each clique will be represented as an eigenvalue with the largest clique being the principal (largest eigenvalue) and the magnitude of the eigenvalue will be a measure of how well the eigenvector is at summarizing the relationships within the clique. The eigenvector for each eigenvalue will be the popularity score of individuals within the clique. In my analysis, I will also try to explain the interpretation of the first and second eigenvectors.

While there have been several explanations of eigenvector centrality in networks across many fields, it is often difficult to interpret the meaning of exactly what an eigenvector corresponds to in the real world. For example, many of the examples explain how an eigenvector as a measurement of an exchange or ranking in equilibrium. In this paper I will also attempt to understand the meaning of the principal eigenvector by comparing its value to other neighborhood attributes over time.

3 Data

The main dataset in my analysis is the Longitudinal Employer-Household Dynamics Dataset (LEHD) that is published by the United States Census Bureau. This is a synthetic dataset that joins firm employment data and census demographic data on the census block level and provides a fine-grained view of the connections between where people live and work. This is an innovative way for a government to release data and has many benefits as it creates an interesting dataset at a low cost since it leverages existing datasets and there is no additional burden on respondents such filling out additional surveys. The datasets that are used to produce the LEHD dataset include, Unemployment Insurance wage records, the Quarterly Census of Employment and Wages, and the Statistical Administrative Records System. Some of the data sources that are used to produce the LEHD dataset are confidential and not themselves made public. The current coverage of employment data is limited to jobs covered by the Unemployment Insurance Program which is approximately 95% of jobs in the United States.

Jobs are broken down among job categories, employee age brackets, and employee monthly salary as follows:

1. Job Category:
   - (a) Goods Producing
   - (b) Trade, Transportation, and Utilities
   - (c) Other
2. Age:

    (a) 29 and younger
    (b) 30 - 45
    (c) 55 and older

3. Monthly Salary:
    (a) under $1,250
    (b) $1,251 - $3,333
    (c) over $3,333

The data is published every year from 2002 to 2015 and shows the number of people who live and work between each census block in the United States.[2] Census blocks are currently the smallest geographic units used in the US Census Bureau statistics. The number of census blocks in the 2010 Census was 11,155,486 with an average size of 27 people so the resultant dataset provides a very fine-grained view of the relationship between places of work and employment. Since the data is highly specified, the Census Bureau employs a few techniques to protect confidentiality of citizens such as noise infusion and synthetic data creation using probabilistic differential privacy. [3].

In my analysis I focus on data within the Chicago Metropolitan Statistical Area and a summary of the employment data for the 2002 is below.

	Count	Percent of Total
Total Jobs	3,924,152	100%
Age: 29 or younger	1,027,445	26.1%
Age: 30 to 54	2,328,093	59.3%
Age: 55 or older	568,614	14.4%
Earnings: $1250month or less	1,090,632	27.7%
Earnings: $1251/month to $3333/month	1,467,733	37.3%
Earnings: greater than $3333/month	1,365,787	34.7%
Goods Producing	708,324	18.0%
Trade, Transportation, and Utilities	819,502	20.8%
Other	2,396,326	61.6%

Much of my analysis is based off of the eigenvector centrality of the live-work commuting matrix between census tracts in Chicago broken down by different job and demographic characteristics. Though the LEHD dataset provides census tract level statistics I found that interpretability improves at a slightly larger area so decided to use census blocks and neighborhoods of which there are 2,215

[2] Data is omitted for 9 state-year combinations where the Census Bureau notes there are data integrity issues. Illinois was not noted on this list, so my study is unaffected by missing data.

[3] More information about noise infusion and confidentially protection can be found on the census website

and 190 respectively in the Chicago metropolitan statistical area. To visualize the dataset, I first created a connectivity graph between all census tracts with more than 25 people commuting between them shown below.

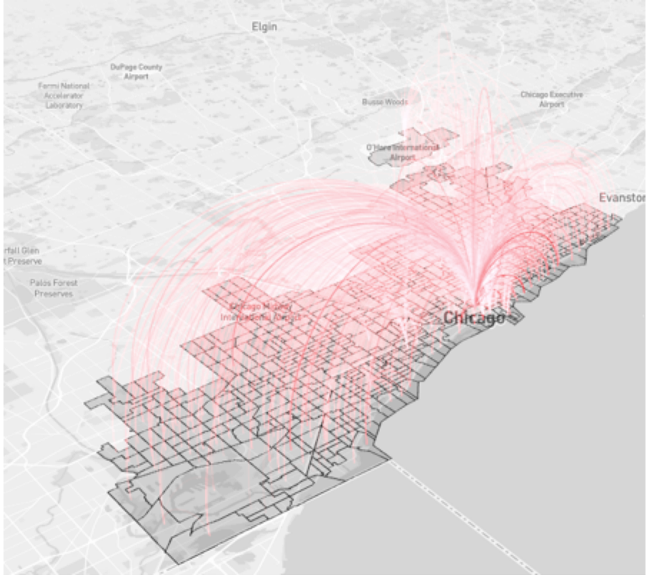
Elgin
DuPage County Airport
Fermi National Accelerator Laboratory
Busse Woods
Chicago Executive Airport
O'Hare International Airport
Evanston
rfall Glen
t Preserve
Palos Forest Preserves
Chicago Midway International Airport
Chicago

Figure 1: This map shows all the connections between where people live and work in Chicago. Note the dominance of The Loop as the main center of employment.

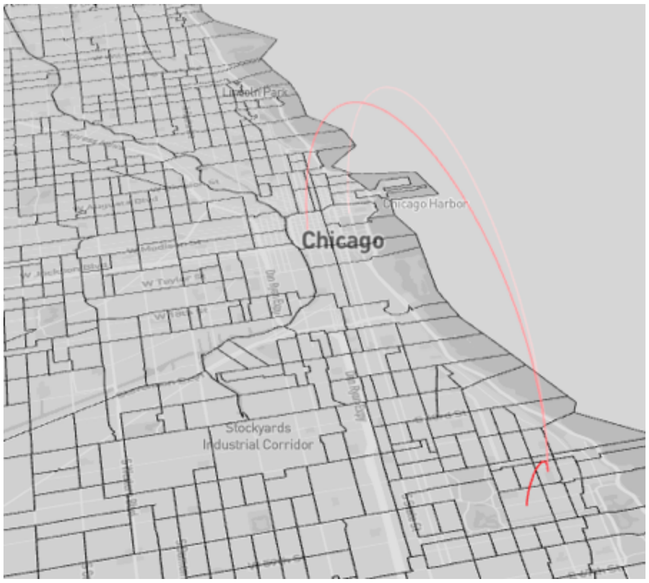

Figure 2: This map shows the connections between a people living Hyde Park who either commute to the University of Chicago or the The Loop. Note: this map only shows combinations on tracts with 25 more people commuting between them

In order to correlate the LEHD calculated centrality measures to other datasets I make the following considerations. When comparing eigenvector centrality to home values, I calculate centrality on the neighborhood level as this is the most fine-grained data level that is available through Zillow. When comparing centrality with data that is only available from the City of Chicago such as Business Licenses and Crime, I calculate centrality on the census block level for the entire Chicago MSA which includes many tracts outside of the City of Chicago Boundary and counties in Wisconsin and Indiana but restrict my analysis to census blocks within the City of Chicago boundary.

Real estate listing data is provided by Zillow, an online real estate database company that tracks detailed listing data of home sales throughout the US starting in 1994. They publish data of home sales for the top 50 markets on the neighborhood level. In my paper I use the following statistics:

1. Median List Price Per Square Foot ($)
2. Monthly Home Sales (Number, Seasonably Adjusted)
3. Zillow Home Value Index: All Homes (SFR, Condo, Co-Op)[4]
4. Zillow Rent Index: All Homes (Multifamily, SFR, Condo, Co-Op)footnotehttps://www.zillow.com/researc
rent-index-methodology-2393/

[4] https://www.zillow.com/research/zhvi-methodology-6032/

Since I am comparing real estate prices to neighborhood rankings, I make the following adjustments to be able to compare the data. First I normalize each neighborhood to its initial value to get an index of the statistic over time. Next I divide by the median value for all neighborhoods in Chicago for the given time period. The result is a statistic that shows how the neighborhood compares to other neighborhoods in Chicago over time. If the value is less than 1, then it is below average whereas if it is greater than 1 it is above average.
For a complete overview of how data was collected, cleaned, visualized, and calculated, I provide the scripts and programs I used on github at https://github.com/b-nroths/chi-data.

4 Methods

In order to find the employment centers of a city I will represent the city's commuting network as a matrix and calculate the important neighborhoods using eigenvector centrality, a method that has been used in studying networks such as trade routes, social networks, and webpages. The goal of the method is to take the network matrix and to output the most influential actors in the network. The definition of influential varies depending on the context and will be explore later in this paper. The matrix representing the connections between actors can be represented a number of way. In social networks, edges represent a connection, in which case the matrix is called an adjacency matrix and is filled with 1 and 0's depending on if two nodes are connected to each other. In a round-robin tournament the matrix could be filled in with 1's and 0's depending on if a team beat the other or by a percentage which would represent a team's win percentage against another team. In a trading network there are multiple ways to describe the network matrix. One simple way is to to use an adjacency matrix (as Gould did) of 1's and 0's if an actor trades with another. Another way would be to represent the size of the actor compared to the country they are trading with or the percent of their total trade an actor makes with a specific partner. Further it could also be represented as a distance of a trading route between two actors. In this case it is common to normalize the distances to the sum of each column is equal to one.

In this paper I will call the network matrix a "commuting matrix". The actor is a neighborhood and the value in the matrix represents the percent of people who live in one neighborhood who commute to the other. This will have the benefit of producing an easy to interpret eigenvector and eigenvalue. Before I demonstrate how the network matrix is built in the context of this paper, it is important to understand why we are guaranteed a positive eigenvalue from the following theorem proven by Oskar Perron and Georg Forbenius.

Theorem 1 (Perron-Frobenius Theorem). Let $C \in \mathbb{R}^{nxn}$ represent a nonnegative primitive matrix (i.e. C: $C_{i,j} > 0$). There exists a positive real number λ_{max}, such that:

1. $\lambda_{max} > 0$
2. λ_{max} has a unique (up to a constant) eigenvector v which has all positive entries
3. $\lambda_{max} > |\lambda|$ for any eigvenvalue $\lambda \neq \lambda_{max}$

Theorem 2. A column stochastic matrix will always have an eigenvalue 1. All other eigenvalues are in absolute value smaller or equal to 1.

To illustrate the network I will explore in my paper, consider a simplified example of a city of 110 people, 100 of whom live Downtown and 10 of whom live in Hyde Park. Of the 100 people who live downtown, 90 works downtown and 10 works in Hyde Park while of the 10 people, 5 work downtown and 5 work in Hyde Park. This network can be represented by the following matrices.

$$C = \begin{bmatrix} hydepark->hydepark & hydepark->downtown \\ downtown->hydepark & downtown->downtown \end{bmatrix} = \begin{bmatrix} 5 & 10 \\ 5 & 90 \end{bmatrix} \tag{1}$$

$$l = \begin{bmatrix} hydepark \\ downtown \end{bmatrix} = \begin{bmatrix} 10 \\ 100 \end{bmatrix} \tag{2}$$

$$w = \begin{bmatrix} hydepark \\ downtown \end{bmatrix} = \begin{bmatrix} 15 \\ 95 \end{bmatrix} \tag{3}$$

From this information we can create a commuting matrix which normalizes the flow between regions of the city and transforms the “live” matrix (2) into the “work” matrix (3). This transforms the above matrices into the following equation.

$$Cw = l \tag{4}$$

$$\begin{bmatrix} 5/15 & 5/95 \\ 10/15 & 90/95 \end{bmatrix} \begin{bmatrix} 15 \\ 95 \end{bmatrix} = \begin{bmatrix} 10 \\ 100 \end{bmatrix} \tag{5}$$

We can also calculate write out the commuting flow from work to home as

$$\text{C}\ l = w$$

$$\begin{bmatrix} 5/10 & 10/100 \\ 5/10 & 90/100 \end{bmatrix} \begin{bmatrix} 10 \\ 100 \end{bmatrix} = \begin{bmatrix} 15 \\ 95 \end{bmatrix} \tag{6}$$

The last connectivity matrix I will study looks at the flow of money between regions modeled by the salary of workers that commute between regions. The first model will be the total flow of money between regions represented by the sum of the salaries of all the workers who commute between regions. For example, if 10 people commute from Hyde Park to Downtown and they each make an average of \$2,500 per month I will consider the money flow between Hyde Park and downtown to be \$25,000. In my dataset, salaries are broken down into three ranges less than \$1,250, \$1,250-\$3,333 and over \$3,333. To simplify the