An Analysis of Growth and Connection in Chicago's Neighborhoods using Eigenvector Centrality

Benjamin Rothschild

MACSS Thesis Rough Draft

4/6/2018

Introduction

There have been many approaches to analyze job and neighborhood growth in cities that spans academic disciplines like economics, statistics and sociology. Much research has focused on important questions like what drives jobs growth, home values, and crime and there are many standard statistical tools researchers have in their tool belt to analyze these questions such as regression, causal frameworks, and social analysis. One aspect of cities that is less studied, however, is the connection between employment centers and housing and how this network creates and distributes wealth within a city. Approaches to quantify important links in networks have been studied in other academic areas such as the geographic network such as trade routes, transportation networks, and social networks. In this paper I will use a common tool used to quantify important nodes in a network, eigenvector centrality, to find important neighborhoods for employment and housing in the commuting network of Chicago. The eigenvector centrality method of computing node importance in a network will help us understand how the relationship between employment centers and housing centers in the city are connected 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 the growth of the city. In this paper I will use the Longitudinal Employer-Household Dynamics dataset published by the US Census in order to apply this method.

Literature Review

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 and order in very complex situations. His hope, which many computational social scientists share, was that underlying complex phenomena might be a mathematical idea that provide a meaningful geographic interpretation. To explore this idea, he maps out the road network of Uganda and creates a connectivity matrix of this network on a binary scale, 1 indicating if two cities are connected and 0 if they are not. He calculated the first four eigenvectors of these matrices in 1921 and 1935 and compared the results. The first eigenvalue is centered around the city of Kampala, which on the map is by far the most connected town owning to the number of direct linkages and its central location. The city with the next highest value, Entebbe, was also very connected. He notes that the successive eigenvectors and eigenvalues are a "pull out" of small regional networks within the trading structure. He then examines a new connectivity matrix of the cities in 1935 and describes how several structural characteristics have been strengthened as new cities are added to the network. Gould makes an attempt, though vague, to describe the meaning of this calculation. He says "vectors representing well-connected towns will not only lie in the middle of a large number of dimensions but will tend, in turn, to lie close to the principal eigenvalue. 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"[1]. This interpretation has been named the "Gould Index of Accessibility".

Other researchers have since tried offered analogous definitions of eigenvalues of the connectivity matrix in geographic and other networks. Tinkler described these eigenvalues in the context of a social network. If there is a social network of with a rumor teller at some vertex *i* at time 0, as time progresses, the rumor will be spread throughout the network according to the

[1] Gould page 66

connections between people in the social network. The distribution of the rumor by time *t* is the equilibrium distribution after a large number of time periods will also be given by Gould's index. In other words, after time progresses the eigenvector is the chance that the rumor has spread to a specific node in the network. So far, the connectivity matrices studied have been adjacency matrices however it is also possible to weight the adjacency matrix edges based on other information we have about the network such has how strong the connections between two nodes are. For example, in a trading network, we might weight the connections by how much trade there is between two cities or in a social network we might weight the connections by how well two people know each other.

Another interpretation of eigenvector centrality was given by J.W. Moon who described the eigenvalues in terms of an iterative round-robin competition between teams. In his example, 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 after a certain amount of turns the ranking of the player will be the same as Gould's index.

Another interesting application of eigenvector centrality was developed 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 create a hyperlink database 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. PageRank can be calculated using an iterative algorithm and corresponds 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.

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. Often there remains further information about the network structure that subsequent eigenvectors can explain. For example, where the first eigenvector is likely to reflect 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. Iaocobucci et al demonstrate that the extraction of only the first eigenvector can be, and in even modest-sized networks insufficient for a more comprehensive understanding of the network. 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. In my analysis I will also try to explain the interpretation between the first and second eigenvectors.

Data

The main dataset I am using 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 a very 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 this 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 into three broad categories:

1. Goods Producing
2. Trade, Transportation, and Utilities
3. Other

Three age brackets:

1. 29 and younger
2. 30-45
3. 55 and older

And three income ranges:

1. $1,250/month or less
2. $1,251 - $3,333 per month
3. $3,333 per month or more

The data is published every year from 2002-2015.[2] The data that is published shows the number of people who live and work between each census block in the United States. Census blocks are currently the smallest geographic units used in the US Census Bureau statistics. The

[2] Data is available for almost all State-Year combinations except around 9 which the Census department notes there are data integrity issues. Illinois was not noted on this list, so my study is unaffected by missing data.

number of census blocks in the 2010 Census was 11,155,486 so the resultant dataset provides a very fine-grained view of the relationship between places of work and employment. Since the data is so fine-grained the Census Bureau employs a few techniques to protect confidentiality of the users such as noise infusion and other synthetic methods 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	
Number of Jobs of workers age 29 or younger	1,027,445	26.1%
Number of jobs for workers age 30 to 54	2,328,093	59.3%
Number of jobs for workers age 55 or older	568,614	14.4%
Number of jobs with earnings $1250month or less	1,090,632	27.7%
Number of jobs with earnings $1251/month to $3333/month	1,467,733	37.3%
Number of jobs with earnings greater than $3333/month	1,365,787	34.7%
Number of jobs in Goods Producing industry sectors	708,324	18.0%
Number of jobs in Trade, Transportation, and Utilities industry sectors	819,502	20.8%
Number of jobs in All Other Services industry sectors	2,396,326	61.6%

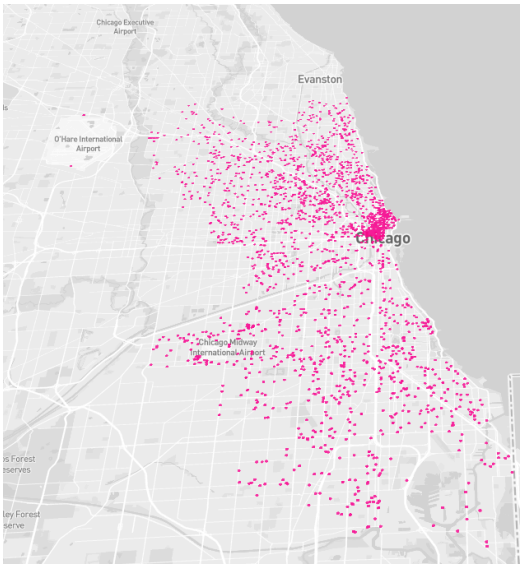
[Double check these numbers, seem high]

I use data from Zillow to measure housing values. Zillow produces data at the neighborhood level for a number of statistics. The datasets I look at are produced monthly from . I look at two measures % of houses increasing in value and median home value per square foot. [Insert additional description of Zillow dataset including a table or a graph…]

I use the building permits and building licenses dataset published by the City of Chicago Department of Buildings in order to measure investments in business properties in the city. This dataset is available from 2006 to the present. Along with each permit application are the

[3] More information about noise infusion and confidentially protection can be found here: https://www2.census.gov/ces/wp/2014/CES-WP-14-30.pdf and differential privacy (https://users.cs.duke.edu/~ashwin/) (http://slideplayer.com/slide/6391900/) http://slideplayer.com/slide/12410369/

following fields Permit type, Date, Estimated Cost, Street Address, Work Description. A visualization of the dataset for data in 2005 is below.



[Also produce visualization of $ spent by census block and table with summary statistics of the data, for example counts on permit type, average amount, etc]

Setup

In order to find the employment centers of a city we will use a setup based on eigenvector centrality that has been used in studying networks such as trade routes, social networks, and webpages. The goal of the method is to take a network of linked nodes and to output the most influential nodes in the network graph. The definition of influential varies depending on the context of each problem for example in the case of social networks, the most important node would be the person who is connected to the most people while in a trade route it might be the city that is connected to the most other cities.

Consider the following 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.

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

$$l = \begin{bmatrix} hydepark \\ downtown \end{bmatrix} = \begin{bmatrix} 10 \\ 100 \end{bmatrix} \quad (2)$$

$$w = \begin{bmatrix} hydepark \\ downtown \end{bmatrix} = \begin{bmatrix} 15 \\ 95 \end{bmatrix} \quad (3)$$

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

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

$$\begin{bmatrix} 0.5 & 0.1 \\ 0.5 & 0.9 \end{bmatrix} \begin{bmatrix} 10 \\ 100 \end{bmatrix} = \begin{bmatrix} 15 \\ 95 \end{bmatrix}$$

The key to understanding how this system behaves is through eigenvalues and eigenvectors. The commuter's flow matrix C is now a column-stochastic matrix since all its entries are nonnegative and the entries in each column sum to one. We can calculate the eigenvalues of this matrix and know that it will know that the largest eigenvalue is 1 (prove this?).

[Insert how these definitions relate to using Perron-Frobenius Theorem this theorem asserts that a real square matrix with positive entries has a unique largest real eigenvalue and that the corresponding eigenvector can be chosen to have strictly positive components.]

[Calculate eigenvector for this example and give an interpretation of the result. What will the Home and live vectors be at equilibrium.]

[Give another example of changing one of the base vectors, for example 20 people move to Hyde Park, how does this effect the eigenvalue and work vector, show how this changes the values from the previous example].

Analysis

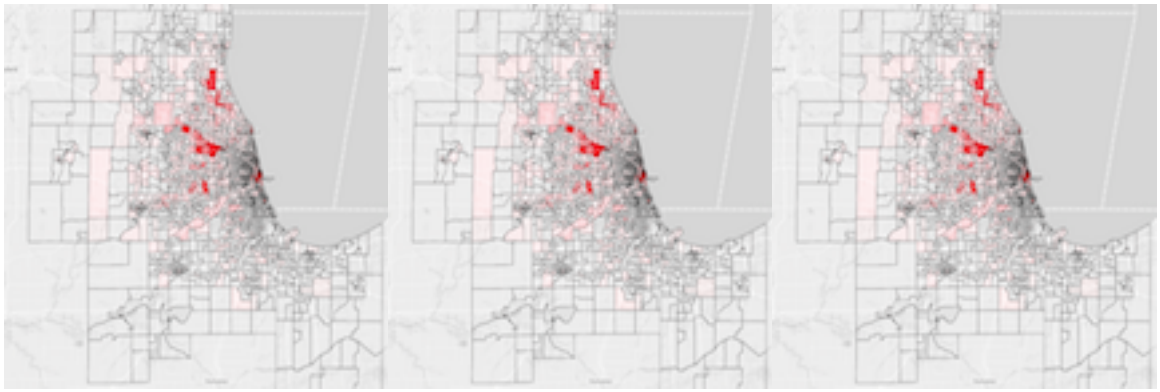
The basis of my analysis is based off of the eigenvector centrality of the live-work network of census tracts in Chicago. Though the LEHD dataset provides census tract level statistics I found that interpretability improves at a slightly larger area so decided to use census blocks, of which there are 2215 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 on the left. The map on the right shows connections between a people living in a census tract in Hyde Park who either commute to the University of Chicago or the Loop downtown.

These maps are also available in an interactive format online at https://chicago.bnroths.com/maps Darker lines indicate more commuters.

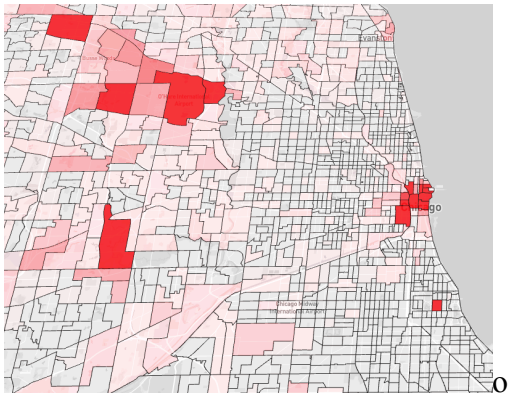
As can be seen, the main employment center in Chicago is downtown in the "Loop" as this is where most of the arcs are directed. Below I will concentrate on three separate questions and show how the eigenvector centrality method can help us understand how neighborhoods in Chicago grow and are connected.

How do the ranks of employment tracts change over time and how does this correlate to business grant investments?

Below I plot the eigenvectors that correspond to the principle eigenvector (Gould's Index) in 2002, 2008, and 2015 for places of employment.



Zoomed in view focusing on downtown Chicago/Loop area. You can see how certain tracts are pulled out such as the Loop, Hyde Park and O'Hare airport.


o

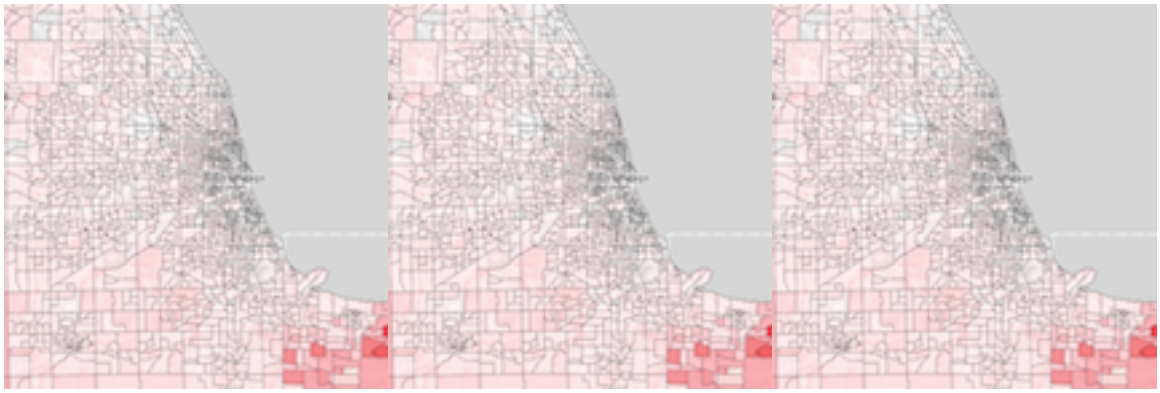
[Produce graph of how neighborhoods change rank over time, for this I think it would make more sense to do at the neighborhood level instead of census tract because this will improve interpretability]

[Correlate change in neighborhoods to business permits or licenses. Explain how this PageRank offers a different and new interpretation of neighborhood development/connection. What additional questions can be asked?]

[Look at the 2nd eigenvalue in the same context. What additional information does it provide?]

How does the ranks of housing neighborhoods change over time?

Below I plot the eigenvectors that correspond to the principle eigenvector (Gould's Index) in 2002, 2008, and 2015 for places of residence. As can be seen compared to the Employment Eigenvector the Housing eigenvector is much more spread out [provide a histogram to compare the differences]



Below is a graph of home prices and eigenvalues for four neighborhoods in Chicago.

[this doesn't show a lot at this point. And there are some other methods I want to try. For example, instead of median home value I want to look at % houses increasing in value and see if this is correlated with eigenvalues. Also, I plan to run a regression and control for some factors like # of houses]

Conclusion

[The general gist is that by using the eigenvector centrality ranking of neighborhoods we can get insight into the structure of the city that is not possible otherwise. Certain interesting things I found looking at the questions above are 1/2/3. Looking at the second eigenvector is also interesting, and I find that. I find…. In addition to what I found I think it would be also

interesting to look at… This work can also be easily replicated in other cities and interesting questions that could then be asked are…]

Sources

Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." *Computer networks and ISDN systems* 30.1-7 (1998): 107-117.

Gould, Peter R. "On the geographical interpretation of eigenvalues." *Transactions of the Institute of British Geographers* (1967): 53-86.

Iacobucci, Dawn, Rebecca S. McBride, and Deidre Popovich. "Eigenvector Centrality: Illustrations Supporting the Utility of Extracting More Than One Eigenvector to Obtain Additional Insights into Networks and Interdependent Structures." (2017).

Moon, John W. *Topics on tournaments in graph theory*. Courier Dover Publications, 2015.

Spizzirri, Leo. "Justification and application of eigenvector centrality." *Algebra in Geography: Eigenvectors of Network*(2011).

Straffin, Philip D. "Linear algebra in geography: Eigenvectors of networks." Mathematics Magazine 53.5 (1980): 269-276.

Tinkler, Keith J. "The physical interpretation of eigenfunctions of dichotomous matrices." *Transactions of the Institute of British Geographers* (1972): 17-46.