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

Benjamin Rothschild

MACSS Thesis Rough Draft

Introduction

There have been many approaches to analyze job and neighborhood growth in cities that spans academic disciplines like economics, statistics and sociology. This research has focused on many 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. There has been less research, however, about how neighborhoods are connected between employment centers and housing and how this network creates and distributes wealth within a city. Some approaches to quantify important links in networks have been studied and successfully applied in a 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 in the commuting network in 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 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 theory.

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” (page 53). His hope, which many computational social scientists share, was that underlying such complex situations might be a mathematical idea that provide a meaningful geographic interpretation. To explore this idea, he maps out the road network of Uganda in the years 1921 and 1935 and creates a connectivity matrix of this network with a binary scale (1 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. This gives the city the highest eigenvalue. The city with the next highest value (Entebbe) is also very connected. He notes that the successive eigenvectors and eigenvalues are a “pull out” of a small regional network. 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” (page 66). This description has been named the “Gould Index of Accessibility”

Other researchers have since tried offered analogous definitions of eigenvalues of the connectivity matrix. 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 edges in the social network. The distribution of the rumor by time t the equilibrium distribution after a large number of time periods will also be given by Gould’s index. Put another way, the equilibrium growth rate of a rumor spreading. Another interpretation was given by J.W. Moon who described how we could measure the relative strengths of players in a round-robin tournament. In his example a player gets a ranking by beating another player, however if they beat a stronger player then they will get a higher rating boost than if they beat a weaker player. After the playing has elapsed after a certain amount of turns the ranking of the player will be the same as Gould’s index.

[Insert how these definitions relate to using Perron-Frobenius Theorem]

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 important 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 a more recent paper

k. However, there often 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. The research in this paper is conducted to demonstrate that the extraction of only the first eigenvector can be, and in even modest-sized networks typically will be, insufficient for a more comprehensive understanding of the network.

https://www.cmu.edu/joss/content/articles/volume18/IacobucciMcBridePopovich2017.pdf

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, Unimployment 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. Public datasets used include the census, and shapefile. 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 (primary/dominant job):

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

Three age brackets:

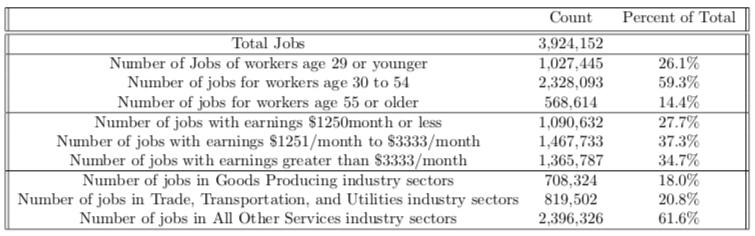
1. 29 and younger
2. 30-45
3. 55 and older

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.[[1]](#footnote-1) 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 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.[[2]](#footnote-2)

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.



Setup

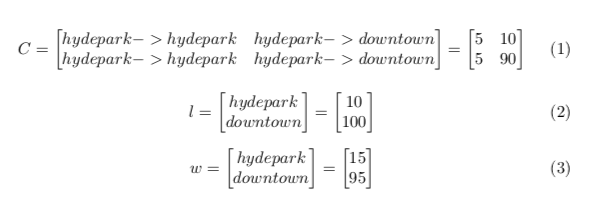
Analysis

Conclusion

Finding Employment Centers with Eigenvectors

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 work downtown and 10 work 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.



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).

Sources

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

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

1. 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. [↑](#footnote-ref-1)
2. 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 privarcy (https://users.cs.duke.edu/~ashwin/) (<http://slideplayer.com/slide/6391900/>) http://slideplayer.com/slide/12410369/ [↑](#footnote-ref-2)