Evolving Community Structures in a Geographic Commuting Graph

Stanford 224W Fall 2018

Daniel Gardner (dangard@stanford.edu)

December 10, 2018

Abstract
Where people choose to live and work is central to the makeup of a metropolitan area. Some cities have large employers scattered throughout, while others have a few central employment hubs that draw workers from all corners of the region. Determining the commute dynamics of an urban area is vitally important to planners and local leaders as they develop ideas for future infrastructure projects. Identifying sub-regions that segment the area into self-contained smaller networks is helpful in this process. We perform this segmentation by running the Louvain community detection algorithm on the commuter network of the San Francisco Bay Area, utilizing the LODES dataset from the US Census Office. Running the original Louvain algorithm yields four distinct communities with modularity of 0.35. Adapting the community detection algorithm to take into account weighted and directed edges increases the modularity by 0.05 for the entire graph, and more than 0.1 for commute networks of specific socioeconomic and demographic groups. We find that over 14 years commute distance increases, while the maximized modularity decreases.

1 Introduction

This project examines how metropolitan areas grow over time and how new and existing workers change their location preferences for where they work and live. We aim to quantify this growth using network analysis techniques on novel data. In particular, we will evaluate the San Francisco Bay Area over a number of years and examine where in the network jobs are created, paying special attention to centralized network hubs. We will then use commute data to formalize networks showing worker flow to these large employment areas. We will perform community detection techniques to identify commute sub-networks, and then assess how these communities evolve over time.

We are interested to see if new workers in the Bay Area undertake longer commutes to reach the same jobs as established employees living closer to their work site. The commute networks may react to worker demand by densifying or by extending the size of the outermost segments (or both). We begin by identifying the key regions of the Bay Area commute network using measures of centrality. We then perform community detection using the Louvain algorithm and calculate the modularity score of the graph.

We repeat the community detection step over multiple years, assessing how the selected communities change over time. For each year, we calculate the number of commuters in each community, as well as the average commute distance within the community. Finally, we analyze sub-networks of commuters from various demographic backgrounds and see if their commuter communities are quantifiably different in any of these metrics.

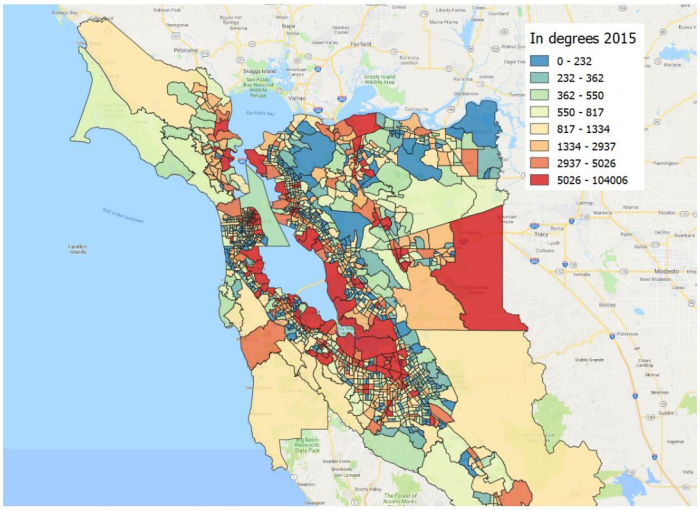

Figure 1: In degree heat map of Bay Area 2015

For our initial community detection we use the original Louvain algorithm introduced in the paper by Blondel et al. (2008). The optimized modularity for the vanilla Louvain treats the graph as undirected and unweighted, which likely produces less than ideal community groupings. To mitigate this, we introduce and utilize a modified Louvain algorithm that incorporates directed and weighted edges into its modularity maximization (Dugué & Perez, 2015). We then quantitatively and qualitatively compare the results of the two Louvain algorithms to determine if the more advanced method produces better community partitions.

2 Literature Review

Most of the work in this area has relied on travel surveys and most recently, smart card data from mass transit systems. The aim of this literature is to identify city centers and partition metro regions to inform transportation and infrastructure planning. Additionally, identifying city centers informs city planners whether the flows of people within the city follows their intended design.

Urban spatial structure and statistical analysis: Before the recent emergence of graph analysis for detecting urban spatial structures, cutting edge research in urban planning relied on statistical analysis. A prominent paper featuring mass transit smart card data is Roth et al. (2011). Here the authors use the rider movements of London's transit system to characterize London's polycentric structure. Ranking stations using the total number of in-commuters, the authors reveal London has three levels of commuting centers: massive transit stations, middling transit stations, and less-frequented transit stations. Using histograms they reveal that for most stations, the number one destination of travelers lies within the first tier of commuting centers, while the tenth most popular destination lies within the third tier of commuting centers. Using statistical methods such as histograms and summation of node degrees, the authors encapsulate London's polycentric nature, departing from traditional theoretical models of monocentric

cities. Similarly, Jiang et al. (2012) use a large time survey data for the city of Chicago to determine different mobility patterns for different types of travelers, such as students and workers.

Urban spatial structures and graph analysis: Zhong et al. (2014) also use mass transit data from smart cards, but in Singapore instead of London. They also have the data for three years, allowing them to detect dynamics of urban spatial structure. In their paper they outline three central concepts to urban planners and innovate by pairing these concepts with common graph analysis measures.

The first concept they define is that of city hubs, which are areas in a city which have many travelers pass through them. To quantify this concept, the authors use betweenness centrality, which indicates region of a graph that has significant traffic. Second, they discuss centers, which are hubs ranked using the pagerank algorithm. Third, they discuss borders, which exist as the result of community detection in a graph setting.

Our key extension of this nascent literature is to analyze hubs and communities over the span of more than a decade to provide a fuller description of city dynamics. Additionally, our data provides three unique advantages over mass transit data as it relates to describing urban structure. The first key advantage is that we have commuting paths for all workers in the Bay Area, not only those who take public transit. Secondly, because our data isolates commutes only due to travel to and from work, we know the communities we classify are commuting communities, unobscured by the noise of other types of travel (leisure, entertainment, etc.). Finally, we are able to examine sub-networks of different demographic groups and see how they differ.

Community Detection for weighted/directed networks: The classic Louvain algorithm introduced by Blondel et al. (2008) is an excellent method for community detection, but is limited to working on undirected and unweighted networks. We utilize a directed modularity optimization function from Leicht & Newman (2008), along with a directed Louvain from Dugué & Perez (2015) to run community detection on the weighted and directed SF Bay Area commuter network.

3 Methods

3.1 Data

For this project I will use the Longitudinal Origin Destination Employment Statistics (LODES) data set. This data set is the result of a collaboration between the United States Census Office and each state's employment statistics office. The data set contains employment statistics for all 11 million census blocks in the US. A census block can roughly be pictured as the equivalent of a neighborhood block. The key employment statistic for my analysis is the number of workers who reside in one census block and work in another for each census block pair. Therefore, in my network, census blocks are nodes and edges are created when at least one worker commutes from one census block to another. Edges contain a feature of magnitude; the more workers who commute from one census block to another, the more heavily weighted the edge.

I decided to focus on the Bay Area, which allows me to gauge my results with known-to-me ground truth information. I define the Bay Area as the following counties: San Francisco, San Mateo, Santa Clara, Marin, Alameda, and Contra Costa. I filtered the CA LODES data to only include commutes with destinations in these counties. I further restricted the data to commutes that originated in these counties, as well as five additional surrounding counties. This last step is crucial to identifying lengthy commutes

from outlying regions into the Bay Area. I further aggregated the census blocks into larger census tracts, which produces just under 1800 location nodes for the Bay Area.

An interesting feature of my data is that worker flows stratified by income groups, industry, and age are also included. Therefore I can see how commuter communities differ for various populations.

3.1.1 Summary Statistics

Table 1: Summary Statistics for Origin Destination Matrix

	2010	2015
Nodes	90,485	82,434
Edges	2,121,308	2,486,755
Self-Edges	1537	1755
Total Workers	2,284,949	2,704,262

We show some basic summary statistics from the Bay Area LODES graph for 2010 and 2015 in table 1. It is interesting to note that while the number of workers and edges increases between the two years, the number of nodes goes down. This might suggest that the network is densifying rather than growing outwards. I will repeat this process for all years and see if any trends develop. Below in figure 1 and 2, we show in and out-degrees log-log plots for 2015 respectively. In both cases, there are many nodes with zero, given that most census tracts are entirely business or residential zoned. Most nodes have out-degrees between 10 and 500, while some nodes have in-degrees greater than 10,000. It makes sense that the in-degrees plot has a long tail, since some nodes have many thousands of workers in a small area.

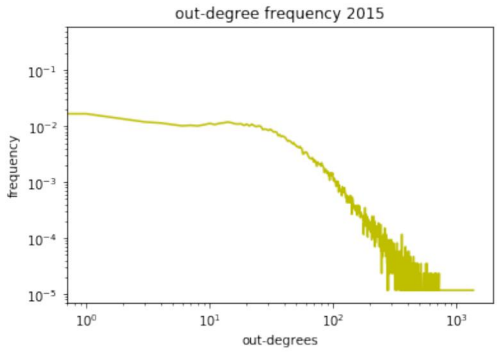

Figure 2: Out degrees

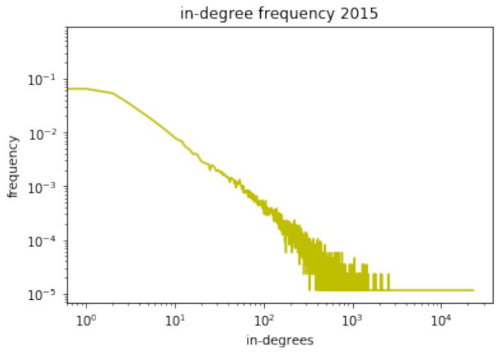

Figure 3: In degrees

3.2 Community Detection

The process of taking the network formed from the commuter edges and segmenting the Bay Area into meaningful communities requires selecting an optimization function and a community detection algorithm, both described below.

3.2.1 Modularity

Community detection involves segmenting a network's nodes into n partitions that maximizes some optimization function. A popular optimization function used for this task is modularity, which calculates

the fraction of edges that exist within the partitioned communities minus the expected fraction if the edges were randomly distributed. Modularity is a number between -1 and 1, with positive modularity indicating that edges are more likely to exist within communities than in-between.

Formally, we define the modularity Q of a community partition C on $G = (V, E)$ as follows:

$$Q = \frac{1}{2m} \sum_{1 \le i,j \le n} \left(\left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j) \right)$$

Where $2m$ is the sum of the adjacency matrix A, A_{ij} is the edge weight between nodes i and j, d_i and d_j are the degrees of nodes i and j, and $\delta(c_i, c_j)$ is the indicator function that is 1 if nodes i and j are in the same community, and 0 if not.

One drawback to the classic modularity metric is that it treats the edges in the network as undirected. For our commuter network, with each edge connecting a residential node to a workplace node, having a sense of direction is quite important for correctly identifying hubs. For this reason, we also utilize a directed modularity function, introduced by Leicht & Newman (2008).

$$Q_d = \frac{1}{m} \sum_{1 \le i,j \le n} \left(\left[A_{ij} - \frac{d_i^{in} d_j^{out}}{m} \right] \delta(c_i, c_j) \right)$$

In this directed modularity, A_{ij} represents the existence of an edge between i and j, d_i^{in} is the in-degree of node i, and d_j^{out} is the out-degree of node j.

3.2.2 Louvain Algorithm

The Louvain community detection algorithm (Blondel et al., 2008) starts with all nodes in their own community, and then performs two steps repeatedly until the maximum modularity is reached. First, we merge nodes into one of their neighbor node's community based on which merge increases modularity the most. When we have reached a steady state in the graph, we aggregate all nodes in each community into one node, and repeat. We attempt to maximize Δ_Q as we make community assignments.

$$\Delta_Q = \left[\frac{\sum_{in} + d_i^C}{2m} - \left(\frac{\sum_{tot} + d_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{d_i}{2m} \right)^2 \right]$$

where d_i^C represents the degree of node i in community C, $\sum_{in}$ is the number of edges in community C and $\sum_{tot}$ is the total degree of community C.
For the directed case the equation is below, where $\sum_{tot}^{in}$ is the total in degrees of community C and $\sum_{out}^{tot}$ is the total out degrees of community C (Dugué & Perez, 2015).

$$\Delta_{Q_d} = \frac{d_i^C}{m} - \left[\frac{d_i^{out} \sum_{tot}^{in} + d_i^{in} \cdot \sum_{tot}^{out}}{m^2} \right]$$

3.3 Experiments

3.3.1 Basic Network

First, we run Louvain community detection on the entire data set, treating the edges as undirected and unweighted. We do this for all years 2002 - 2015, and report the number of community partitions and

modularity achieved. This simplified network is still quite representative of the actual one, since most edges only have one or a few commuters on them. The actual edge count of a node can be a good substitute for total edge weight. It can be difficult, though, to determine if a node is a workplace or residential area however, as both can have high degree.

3.3.2 Directed and Weighted Network

Next, we run Louvain community detection on the entire data set, utilizing all information about direction and weight of edges. We examine the same years as before, and again report the partitions and modularity. We compare these results to the undirected/unweighted case.

3.3.3 Comparison of commuters by age

We filter the network to only include workers under age 29 and then over age 55. For each age group, we run the directed/weighted Louvain and partition it into communities.

3.3.4 Comparison of commuters by income

We filter the network to only include workers making less than $3333 per month, and run the directed/weighted Louvain and partition it into communities.

3.4 Analysis

Performing these experiments on all 14 years of LODES data involves a total of 70 runs of community detection. Some community partitions are similar across multiple experiments, while others vary in community count and modularity. We compare these values, and also compute the clustering coefficients and average commute distances to help summarize the results.

3.5 Distance

We use the Euclidean distance between two census tract nodes as a proxy for commute distance. The Google Maps API would be an ideal measure, but ours is sufficient for making rough comparisons. To calculate the average commute distance of a community c_i, we sum the total distances of commutes that end in c_i, and divide by number of workers in c_i.

4 Results

4.1 Entire population

Running community detection on the unweighted, undirected network yields good results. For each year 2002 - 2015, either four or five communities were partitioned, and the modularity ranged between 0.3511 and 0.3875 as seen in table 2. Examining the map in figure 4 panel b, we could broadly refer to partitioned communities by county names: SF/Marin, San Mateo, Alameda/Contra Costa, and Santa Clara. We also see a couple census tracts in Oakland/Piedmont being partitioned into the SF community, which makes sense given the wealth of those neighborhoods (bankers, attorneys, etc.).

When we add direction and weight to the edges, the number of communities detected increases to five or six, and modularity increases by about 0.04. The primary difference we see in figure 4 panel a is the East Bay community being broken up into coastal and inland communities. This is because in the

basic unweighted network community detection, there are a large number of low-weighted edges of commuters (going from Walnut Creek to Oakland perhaps) that get equal weight to other high-weighted shorter commutes. This forces a larger East Bay community in the Basic Network. When we shift to the weighted/directed community detection, the modularity can be increased by ignoring these cross-community commuters and focusing on the higher number of short commuters.

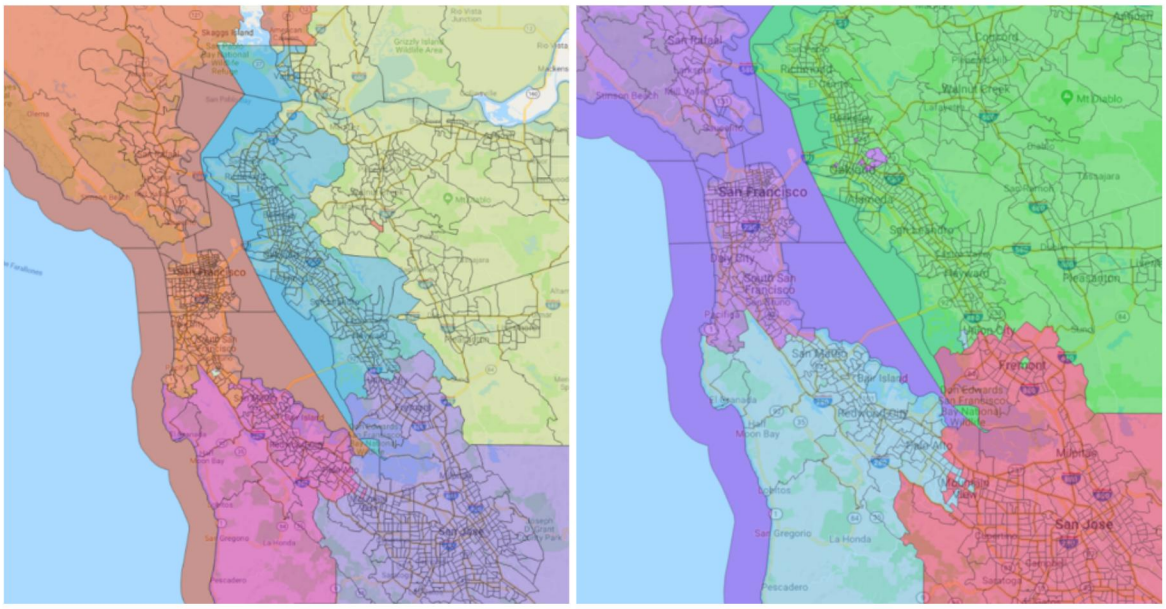

Figure 4: a) Directed and weighted overall b) Basic network

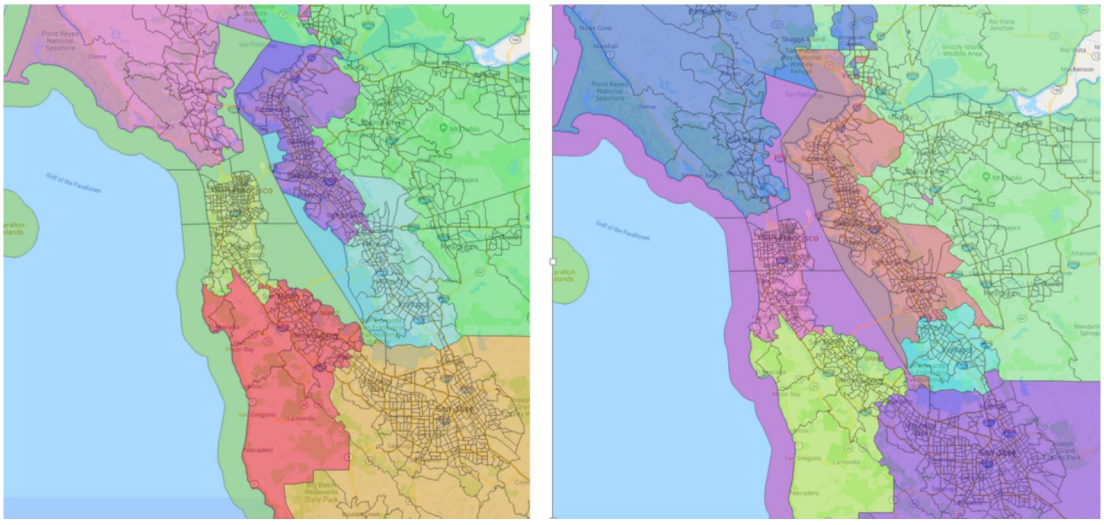

Figure 5: a) DW-Poor, b) DW-Young

4.2 Demographic sub-networks

The networks formed by filtering for low-income, young, and old commuters all generally have higher modularity scores when we perform directed/weighted community detection, and are split into more community partitions. Compared to the entire population network, Marin County is now made into its

own community, and the coastal East Bay is divided north-south as seen in figure 5.

Comparing the community partitions for young and low-income workers is interesting, given that the partitions are quite similar yet the modularity is 0.06 higher for the low-income partition. It is intuitive that the low-income commuters split into more communities and achieve higher modularity, since they have among the shortest commutes on average of any demographic group. They would also be especially averse to commuting across a bridge, given rising toll costs. On the other hand, young people have longer commutes on average, and yet we see smaller communities (especially the one in Fremont). One possible explanation might be a delineation between young workers who have moved to the Bay Area for employment versus native young people. Having grown up in the South Bay myself, I know many people still live at home with their parents and have relatively short commutes. At the same time, young transplants might live in Dublin and take BART to San Francisco for work, increasing the average commute distance dramatically and lowering the modularity.

4.3 Changes 2002 - 2015

Figure 6 shows that the average commute distance in the Bay Area is only about 14 miles (figure 6). This number has been on the rise for all demographic groups from 2002 - 2015, increasing about 1.5 miles. This is the primary cause of decreasing modularity across all community detection over that time span. Rather than new commuter from outlying areas entering, it seems that workers already in the region are traveling to the opposite corner for work. This is manifest on increasingly crowded freeways and BART and CalTrain cars. I also computed the clustering coefficient on the entire network, and it steadily increased from 0.47 to 0.50 in the 14-year span. This indicates the network is also densifying.

Year	Number of Communities					Modularity score				
	Basic	DW	DW-Poor	DW-Young	DW-Old	Basic	DW	DW-Poor	DW-Young	DW-Old
2002	5	5	7	7	6	0.38	0.42	0.46	0.42	0.47
2003	5	5	7	6	6	0.39	0.42	0.46	0.42	0.47
2004	5	6	7	6	7	0.38	0.42	0.46	0.41	0.47
2005	5	7	7	6	5	0.38	0.42	0.45	0.41	0.46
2006	5	5	7	6	7	0.38	0.42	0.45	0.41	0.46
2007	4	6	7	6	6	0.37	0.41	0.45	0.4	0.45
2008	5	5	7	5	7	0.36	0.41	0.44	0.39	0.45
2009	5	6	6	7	7	0.36	0.41	0.44	0.4	0.44
2010	4	6	7	6	7	0.37	0.4	0.44	0.39	0.44
2011	5	5	7	7	6	0.36	0.4	0.44	0.39	0.43
2012	4	5	7	7	5	0.36	0.39	0.43	0.39	0.43
2013	4	5	7	7	7	0.36	0.39	0.43	0.38	0.42
2014	4	5	7	5	6	0.35	0.39	0.43	0.37	0.42
2015	4	5	7	6	6	0.35	0.39	0.44	0.38	0.42

Table 2: Community partitions and modularity for Basic Entire Network, Directed/Weighted Entire Network, and D/W Poor, Young, and Old commuters

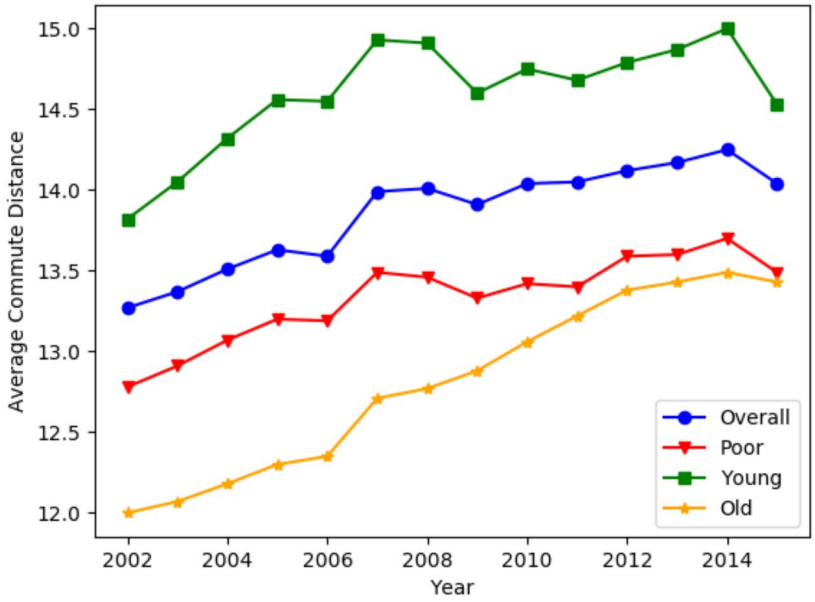

Figure 6: Average commute distance across communities

5 Conclusion and Future Work

The weighted and directed Louvain community detection algorithm produced superior community partitions to the original algorithm. It was particularly impressive to see the spatially contiguous census tracts within communities, and the partitions made sense geographically (bridges, etc.). The extra community partitions provide insights on where the big demographic changes are happening in the Bay Area. It is clear that the most upheaval is happening in the East Bay, as commuters, particularly young people, are working further from home. Since it also remains one of the few affordable places to buy a home, this upheaval in the East Bay is unlikely to slow down. This could in part be caused by tech companies like Facebook and Google sponsoring shuttle buses to carry their employees long distances, but the increasing commute distances is felt on all modes of transportation. Policy makers would be advised to rapidly build more high density housing near high in-degrees workplaces, but efforts to do this have been futile.

The sheer amount of results from the 70 different community partitions was difficult to fully analyze and convey. It would be interesting to calculate commute distance and clustering metrics on specific partitions of the graph that do not change much from year to year. It would also be interesting to do an in-depth analysis of where workers come from to my employer in Livermore, since that lies right on the eastern edge of the Bay Area.

Github Link: https://github.com/gardnerds/commuter

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Dugué, N. and Perez, A. *Directed Louvain: maximizing modularity in directed networks*. PhD thesis, Université d'Orléans, 2015.

Jiang, S., Ferreira Jr, J., and Gonzalez, M. C. Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, pp. 95–102. ACM, 2012.

Leicht, E. A. and Newman, M. E. Community structure in directed networks. *Physical review letters*, 100 (11):118703, 2008.

Roth, C., Kang, S. M., Batty, M., and Barthélemy, M. Structure of urban movements: polycentric activity and entangled hierarchical flows. *PloS one*, 6(1):e15923, 2011.

Zhong, C., Arisona, S. M., Huang, X., Batty, M., and Schmitt, G. Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28(11): 2178–2199, 2014.