Analyzing Inter-Urban Spatio-Temporal Network Patterns
CS 224W Final Project

Tanner Gilligan, Travis Le, Pavitra Rengarajan

1 Introduction

The ease with which mobile phones of our generation can communicate their precise location and GPS coordinates has allowed location-based services (LBS) to produce vast amounts of data regarding the whereabouts of their users. Social networks like Foursquare have taken advantage of this capability by using LBS to aid social encounter and place discovery in cities. Such statistics give way to revealing spatial and temporal analyses of the aggregate activity generated by such networks. In this paper, we study Foursquare user behavior by analyzing checkin transitions between venues, as well as information on particular venues, with the aim of gaining a deeper understanding of 5 US cities: Miami, Boston, San Francisco, Los Angeles, and New York. We aim to answer some key questions -- which places are most "important" in the urban fabric in a city? How does this change over time? How do these trends and important types of places differ between cities?

2 Related Work

Network analyses over the past few years have examined the relationship between online LBS-based social networks and urban landscapes of cities. Much of the research that has been done has addressed core facets of temporal network evolution: how edges are created, how social triangles are created, and how users' mobility affects new edges. [3], for example, found that a gravity model which combines both node degree and geographic distance is most suitable in modeling edge attachment. Another interesting finding based on Gowalla data was that users with higher degree created new ties at a faster rate in accordance with the "forest-fire" copying process proposed by Leskovec et al. [5]. This process says that when a new node joins the network and connects to a first neighbor, it starts recursively creating new links among the neighbors of the first neighbor, effectively copying the connection already in place; this process mixes preferential attachment and transitivity. This analysis uses Gowalla data, however, which contains information on users, whereas our Foursquare dataset provides more details on the place network without giving any information on users or the user ids.

This is not to say that network analyses on the Foursquare place network dataset has not been done. In [1], Noulas et al. gathered Foursquare checkins to find interesting relationships between spatial and temporal differences in checkins. Their analysis was two-fold: in the first part, they analyzed how checkins vary at different hours during the day, and they found that on weekends there were 3 spikes in checkins (around 9 am, 1 pm, and 7 pm), and on weekends there was a continuous plateau from about 2 pm to 10 pm in which checkins remained relatively constant. They also analyzed transitions between checkins by looking at the probability that people who checked in at one location would check in at a second location, by separating transition likelihoods based on time between checkins, and found four different patterns for different time scales. For example, on the 0-10 minute scale, train station to train station had the highest probability, while on the 100-500 minute scale, airport to gate had the highest probability. Another paper

by Bawa et al. [2] uses the scan-based spatial density algorithm DBScan to investigate spatial fragmentation. The DBScan algorithm works by iteratively aggregating geo-located points into clusters based on some threshold distance and a minimum cluster size, so this algorithm "scans" outwards starting with a random point, and iteratively adds points to clusters when they are within the threshold distance of any existing member of the cluster, with unassigned points marked as "noise," until each point has been examined exactly once. Parisian activity stood out as less spatially fragmented than the other cities, looking more like a contiguous blanket of social activity, while London and New York City looked more like "social archipelagos," fragmented sets of islands characterized by high-density social activity. We aim to expand on this analysis by not just considering the structure of the network but also examining the centrality of different types of places and looking at these central places' evolution over time across five different U.S. cities.

3 Data Collection & Preprocessing

In this study, we examined the Stanford Foursquare Place Graph Dataset provided by Professor Jure Leskovec. Every day, thousands of people check in to various locations on Foursquare and, in the process, create vast amounts of data about the connections between places. Our dataset contained metadata about over 160,000 popular public venues, and over 21 million anonymous checkin transitions, which represented trips between the venues. When examining our data, we realized that we had very detailed place information and no data on the people themselves. Therefore, we specifically examined our data with the intention of learning about the places rather than the people. Our resulting weighted place-graph used individual venues as the nodes, checkin transitions as the edges, and weights (number of times this edge appears) as the attribute on each edge. Due to the large size of the files being dealt with (in excess of 1.5GB), we utilized a number of preprocessing techniques in order to reduce their size, including remapping 24-character venue IDs to shorter integer IDs that could be used by snap, partitioning the data by month, and replacing repeat checkins with a weighted edge representing number of times that checkin was seen. Overall, we reduced the total size of our files by a factor of 6 (approximately 200 MB instead of 1.5 GB).

4 Initial Statistics

We used a TNEANet graph in SNAP to represent our data. This graph allowed us to provide directed edges, while also providing an attribute for the weight of each edge. Running initial statistics on the graph for Miami in January, we found that there were 13896 nodes with 98369 edges. Of these nodes, there were 787 zero-in-degree nodes and 834 zero-out-degree nodes. Additionally, the largest connected component includes 99.94% of all nodes and the largest strongly connected component contains 87.66% of all nodes.

5 Methodology

In order to answer the questions presented in the introduction, we needed some measure of importance for each venue in the graph. Logically, we could think of 3 main measures for how important a particular venue is: how many people checked in at a venue, how central a venue is (close to everything else in

terms of edge-distance), and how many different types of people checkin there. To measure each of these, we selected degree, closeness, and betweenness respectively. Degree has an obvious connection to total checkins since one additional degree corresponds exactly to one additional checkin. Closeness answers our second question because it finds those venues that are least distant from all other venues, effectively finding the "central" venue by this measure. Due to the fact that we had no user information, answering this final question was non-trivial. We elected to use betweenness as a quantifiable measure of this, since places with high betweenness have many different shortest paths going through them. Each of these shortest paths corresponds to different types venues (e.g. mall -> grocery store, bar -> bus), which likely correspond to people with differing habits. We acknowledge that this is not truly representative of what we are trying to measure, but it is a fair approximation given the lack of user data.

Unfortunately, the traditional degree, closeness, and betweenness measures presented in class do not account for weights between two nodes. While a few network measures have been proposed for weighted networks, these generalizations have solely focused on only edge weights, and not on the number of edges of a particular node [7]. In order to examine the "importance" of different venues, we wanted to generalize these centrality measures (degree, closeness, and betweenness) such that they account for both the edge weights as well as number of edges. After examining several papers, we implemented our algorithms in Python in accordance with Opsahl et al. [6].

5.1 Degree Centrality

To begin, we implemented a generalized degree centrality algorithm as a first indicator of the involvement of a node in the network. Let us say that k_i is the measure of the number of nodes that a focal node is connected to, and s_i is the sum of weights of a node's neighbors. Then, these can be formalized as follows:

$$k_i = C_d(i) = \sum_j^N x_{ij} \text{ and } s_i = C_D^w(i) = \sum_j^N w_{ij}$$

where i is the focal node, j represents all other nodes, N is the total number of nodes, x is the adjacency matrix where x_{ij} is defined as 1 if node i is connected to node j and 0 otherwise, and w is the weighted adjacency matrix where w_{ij} is greater than 0 if the node i is connected to node j, and the value represents the weight of the edge. Since degree and strength can both be indicators of the level of involvement of a node in the surrounding network, it is important to combine both these measures when studying the importance of a node. In an attempt to combine both degree and strength, we use a tuning parameter α which describes the relative importance of the two measures such that the modified degree centrality measure is a product of the number of nodes that the focal node is connected to and the average weight to these nodes adjusted by the tuning parameter. Formally, we use the following measure:

$$C_D^{w\alpha}(i) = k_i * \left(\frac{s_i}{k_i}\right)^{\alpha} = k_i^{(1-\alpha)} s_i^{\alpha}$$

As described in Opsahl et al. [6], we decided to proceed using the tuning parameter of $\alpha = 0.5$. After implementing this modified algorithm in Python and running on the city of Miami for the month of January, the 20 venues with the highest degree are given in Table 1 of the Appendix. The resulting degree distribution appeared to be similar to that of a power law, as shown in Figure 1, although we see more venues with very minimal (or zero) checkin transitions than expected.

5.2 Closeness and Betweenness Centrality

The closeness and betweenness centralities are calculated exactly as they have been defined previously, with a slight exception. Instead of using an unweighted distance, we now have to utilize a weighted distance, which is defined below. Thus as discussed in Opsahl et al. [6], the formula for calculating the closeness centrality is now: $C_C^{w\alpha}(i) = \left[\sum_{j=1}^{N} d^{w\alpha}(i,j)\right]^{-1}$. Additionally, the formula for betweenness centrality is defined as: $C_B^{w}(i) = \frac{g_{jk}^{w\alpha}(i)}{g_{jk}^{w\alpha}}$. In this equation, $g_{jk}^{w\alpha}$ indicates the total number of weighted shortest paths from an arbitrary node j to an arbitrary node k subject to $j \neq k$. In addition, $g_{jk}^{w\alpha}(i)$ indicates how many of these shortest paths pass through node i, so $\frac{g_{jk}^{w\alpha}(i)}{g_{jk}^{w\alpha}}$ represents the proportion of all shortest paths that travel through node i. It should be noted that if there are multiple shortest paths from j to k, then the value is split across each of the paths, just as in betweenness calculation for an unweighted graph.

5.2.1 Weighted Distance

We define the weighted distance as given by Opsahl et al. [6]: $d^{w\alpha}(i,j) = min\left(\frac{1}{(w_{ih})^{\alpha}} + \ldots + \frac{1}{(w_{hj})^{\alpha}}\right)$, setting the tuning parameter $\alpha = 0.5$, like previously. Notice that if $\alpha = 0$, that we get an unweighted distance and that if $\alpha = 1$ that the inverse of the weights of each edge are taken as the actual distance.

5.2.2 Implementation

Due to the fact that snap does not provide functionality for computing the closeness and betweenness centrality of a weighted graph, we had to implement these ourselves. For both of these measures, we explored the graph in the same manner. For a given node i, we perform Dijkstra's breadth-first search on the graph until we have seen every node. During Dijkstra's search, whenever we pop a node j off the queue, we know that this was the shortest path from i to j. Based on which centrality we are measuring, our next step differs. For closeness, we simply add the computed distance from i to j to a running sum, which represents our $\sum_{j=1}^{N} d^{w\alpha}(i,j)$ term. In the case of betweenness, the steps taken are more complicated. When we pop node j off the queue, we record the path taken from i to j, and store these until we complete Dijkstra's. Once we finish, we have a complete list of shortest paths from node i to every other node in its connected component. For each of these paths, we iterate through every node in the body of the path (i.e. not the start or end node), and increment its betweenness by $\frac{1}{p_{ij}}$ where p_{ij} represents the number of shortest paths from node i to j. By running this BFS for every node i, we will find every shortest path in the graph, and properly compute their betweennesses and closenesses.

5.2.3 Parallelization

With the massive amount of data that we were presented with, we found that it took over three hours to compute a single centrality measure on just one month of the smallest city. In order to address this

problem, we incorporated both multithreading and parallel computing to reduce the time down to about 20 minutes for the smallest dataset and about 4-5 hours for the largest data set. We redesigned our algorithms to handle multithreading and were able to implement them using a threadpool, with different thread counts depending which measure we were computing. However, since this took up a large amount of RAM (about 10 GB-20GB), we weren't able to run this on our own laptops, and thus had to use the Stanford Corn Machines to run these scripts. By manual distributing the workload over 12 different machines (one month of data per machine), we were able to obtain weighted centrality measures for all 5 cities in a reasonable amount of time.

6 Results and Findings

By examining the graphs of these 5 major U.S. cities through visualization, generalized centrality algorithms, repeated checkins, and geographical location, we garnered some interesting results.

6.1 Visualization

In order to gain a better understanding of what the these city networks looked like, we decided to visualize the network, and search for clusters. Below is a filtered visualization of Miami in January.

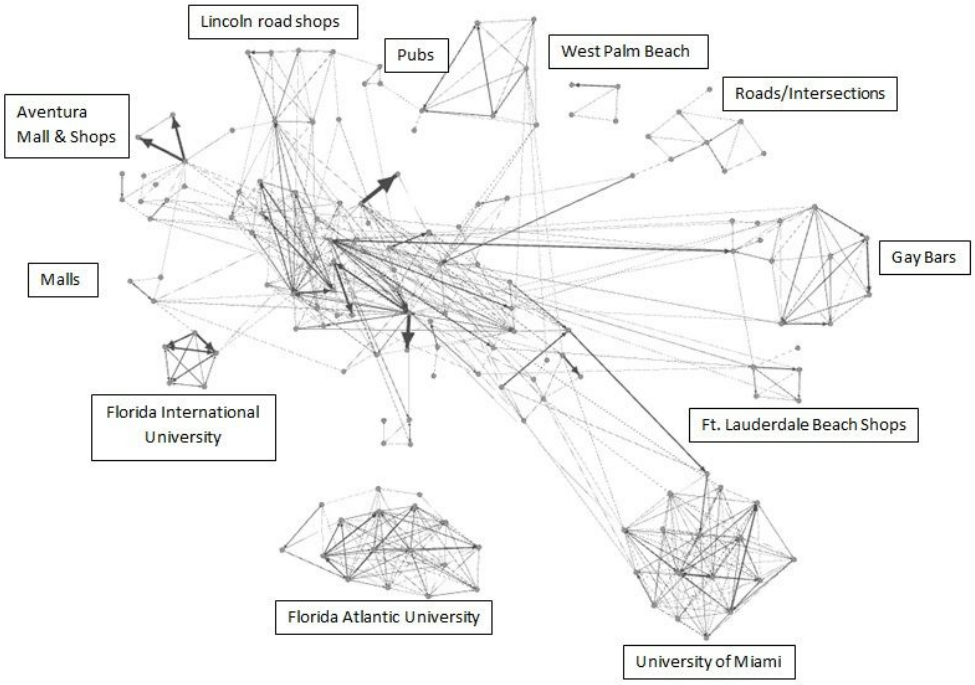

Figure 1. Network visualization of Miami in January

We used Gephi to create this graph. Due to the huge amount of nodes and edges in the original graph though, we had to do some filtering to extract a meaningful visualization. To retain only the most relevant places, we keep nodes with total-degree greater than 20, and edges with weight of at least 2. Unfortunately, none of the built-in layout algorithms produced an informative visualization, so we had to

manually move nodes around to find clusters. We then cross-referenced each cluster's nodes with our original venue data, and tried to identify the overarching theme associated with each of the clusters. In the above graph, each of the clusters is labelled according to the venues contained within it. The unlabelled mass in the center of the graph is mostly comprised of transportation hubs, such as airports and bus stations. In addition, there are also a notable number of restaurants in the center region. It makes sense that these venues would be in the center of the graph given many different types of people visit them so frequently, causing the venues to be connected to many different places. An interesting feature of this network is that so many universities come up as clusters, and in the case of Florida Atlantic University, as a disconnected component all-together. This demonstrates a high usage of Foursquare by college students, and that those college students venture outside the campus relatively rarely. One interesting thing we noted about this graph is the cluster labelled "Roads/Intersections". Unfortunately, our data doesn't provide information about whether these checkins occurred while the individual was driving or walking, but this would have been valuable to know. If it were the case that people were checking in while driving, it could be beneficial for FourSquare to disable to checkins on roads and intersections to deter dangerous driving habits.

At a higher level, apart from transportation and shops, this network displays a noticeable focus on nightlife in the form of colleges, pubs, and gay bars. It seemed strange to us initially that the "pubs" and "gay bars" clusters were so highly connected, since an edge indicates two consecutive checkins by a single individual. This would indicate that people were repeatedly checking in at only these locations, and did not checkin anywhere else in between. We realized that this was likely a result of people's personal checkin patterns, since they may decide to only check in when they go out to have a good time, as opposed to something mundane like being on the train to work.

6.2 Centrality Measures

We implemented the aforementioned algorithms for the 5 major U.S. cities of San Francisco, New York, Boston, Los Angeles, and Miami. We found the comparison of San Francisco, Boston, and Miami to be the most revealing.

We wanted to gain some knowledge about the types of places that were most popular across different cities. To begin, we considered our weighted degree centrality algorithm and found the 100 most popular places for each city where people would check in. Using the category information of each venue that we were provided with, we considered which categories were most represented in the top 100 places of highest degree. Below is a comparison of central place types for SF, Boston, and Miami.

San Francisco: **Boston:** **Miami:**

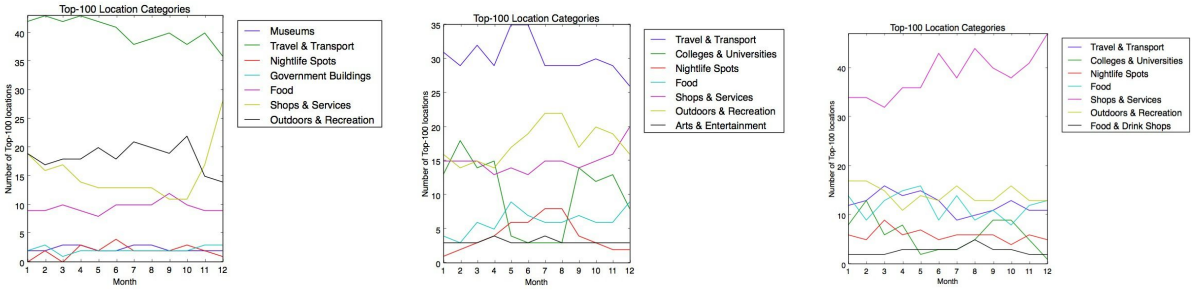

Figure 2. Examination of categories for top 100 places with highest degree centrality for San Francisco, Boston, and Miami, respectively.

This examination of degree gives us information about the volume of checkins through the inflow and outflow; for our graph, it is a combination of the number of people that check into a particular location and the weight of the edges into the location. Several interesting observations arise. On a higher level, it is interesting to note that colleges and universities appear in the top 7 categories for Boston and Miami, albeit to a lesser degree, while the university category does not appear in San Francisco's top 7 categories. This suggests that Boston is more of a college town with higher activity pertaining to the local colleges and universities (which makes sense because of the large number of colleges and universities located in Boston), and Miami actually also has a significant portion of its city life based around local universities. San Francisco, on the other hand, does not seem to be a college town, as colleges and universities do not even make its top 7 most degree-central place categories. Note also that Boston's sharp decline in college activity during the summer months is more notable (Miami's is notable as well but to a lesser degree), as would be expected. We also notice that in the degree graphs, the travel & transport category does not completely dominate over the other categories as it does for betweenness and closeness, which suggests that these travel and transportation places are more a function of the placement as opposed to volume; even though travel & transport places do not have high degree of checkins or a large number of ties, they are more close because they are more centrally-placed in the graph. We also see that Miami's top category is shops & services, whereas San Francisco and Boston's top categories are travel & transport; this suggest that tourism and shopping are likely a more central and important part of Miami's urban fabric, moreso than in San Francisco or Boston. Furthermore, the fact that food & drink shops show up on Miami's top 7 place categories is further indication that Miami's most popular places are usually food and drink shops, while other cities see a more varied distribution of most popular places for socializing. Additionally, for San Francisco, we note that government buildings actually see a high number of checkins, suggesting that San Francisco's landmarks such as the Ferry Building, Coit Tower, and SF City Hall are seen as more noteworthy than government buildings and landmarks in Boston or Miami.

San Francisco: **Boston:** **Miami:**

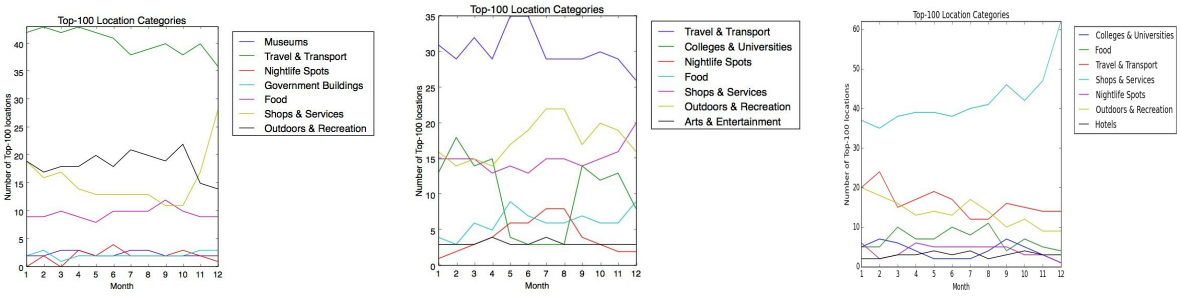

Figure 3. Examination of categories for top 100 places with highest betweenness centrality for San Francisco, Boston, and Miami, respectively.

An inspection of places with highest betweenness for San Francisco, Boston, and Miami also yields interesting results. People stop at high betweenness places on their way between different cities. First, note that Miami sees high degree but low betweenness for colleges. This observation is in line with the visualization seen in Figure 1, where Florida Atlantic University is in its own disconnected component, and the University of Miami is in its own nearly disconnected component. This suggests that colleges and universities are frequented, but are often in their own cluster and do not facilitate much movement outside of the college/university realm. It is also notable that in Miami, nightlife and hotels make the list of the top 7 categories, indicating that nightlife and tourism are a bigger part of the culture. The fact that Miami's top category is shops & services, which sees a rise during the months leading up to Christmas, also fits in with the idea that Miami's core industries are less varied than that of San Francisco or Boston. Boston and San Francisco, on the other hand, see travel and transport as their most popular category, seeing more movement of people between places in their respective cities. San Francisco and Boston have a similar involvement proportionally of outdoors and recreation, although outdoors & recreation decreases more dramatically during the winter in Boston than it does in San Francisco (likely because of the weather). All cities have certain categories that inevitably serve as hubs, including college and universities since college students and younger people are more likely to use an app like Foursquare, food because it is often a hub of social activity, travel and transport to facilitate movement of people, shops and services, and outdoors and recreation.

San Francisco: **New York:** **Miami:**

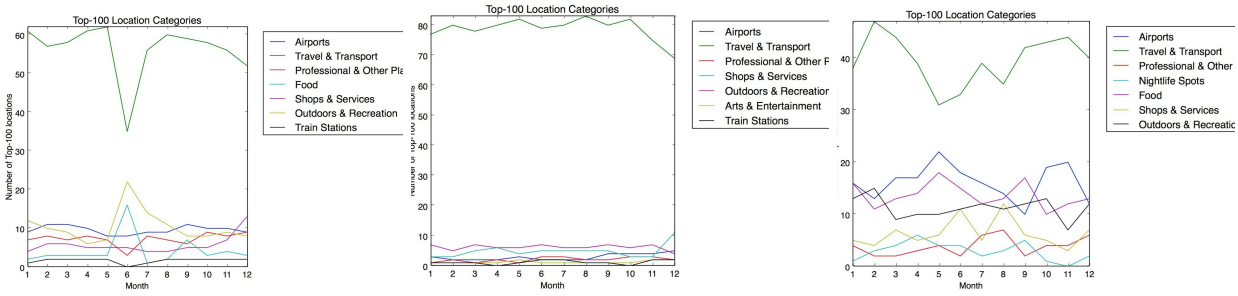

Figure 4. Examination of categories for top 100 places with highest closeness centrality for San Francisco, New York, and Miami, respectively.

Examining the closeness centrality measure for the cities of San Francisco, New York, and Miami reveals several surprising results. A high closeness measure of a location suggests that people will often check-in somewhere else after checking in at that location. Additionally, these closeness centrality measures were weighted by count, meaning that the more check-ins between two places, the "closer" they are to each other. Firstly, it was found that the Travel & Transport category dominated all three cities for all months. However, the one interesting feature here is that in San Francisco in the month of June 2012, we noticed a sharp decrease in the number of Travel & Transport places and an increase in the number of Food category places and Shops & Services category places. After doing some investigating, it was found that many of the BART locations and the SFO Airport had dropped out of the top 100. Researching further, we found that the BART closed for about 14 hours in June 2012 due to a fire, but we don't believe that that would have had a significant impact over the number of checkins over a month. Since we couldn't find any other major events that occurred in June 2012 dealing with either the BART or SFO Airport, we are unsure as to why there is such a huge drop. In looking at the percent of places that are Travel & Transport, it was found that Miami had about 40%, while San Francisco and New York had about 60-80%. What this suggests is that Miami is not as travel-centric, whereas San Francisco and New York attract more travellers. Finally, in examining specific categories, only New York and Boston had a few Arts & Entertainment places reach the top 100. Additionally, only Miami had a few Nightlife Spots places reach the top 100 as well. We believe that this has to do with the culture of these cities as New York and Boston are considered arts hubs and Miami is also known for having a thriving night scene.

6.3 Popularity

From our centrality measure calculations, we garnered information about the most "central" places for various different cities.

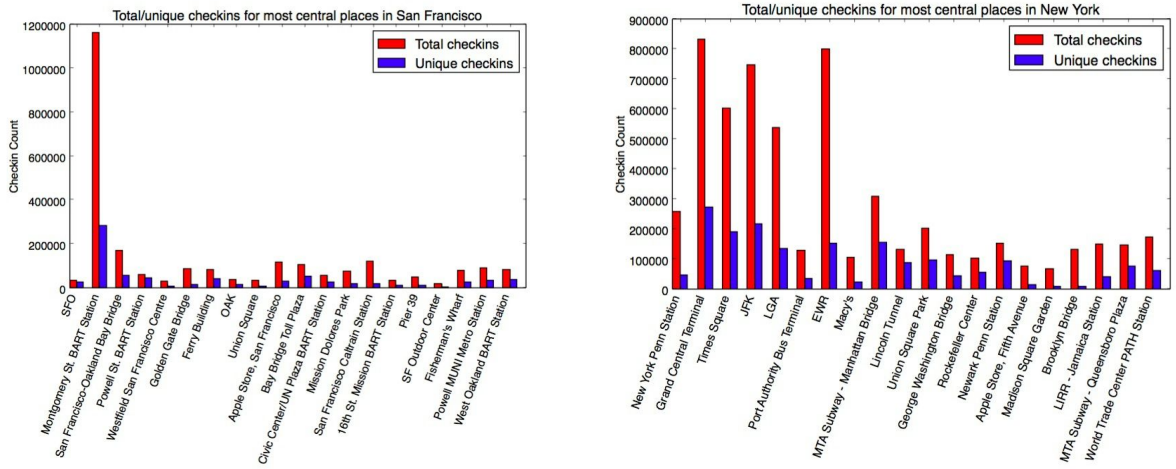

Figure 5. Total checkins and unique checkins as of 2012 for top 20 places in San Francisco and New York as seen across our degree, betweenness, and closeness centrality measure calculations.

San Francisco sees a lower number of unique checkins across the board; this can most likely be attributed to the activity of the users, since more people would likely use the Foursquare app being that San Francisco is at the heart of the Silicon Valley. New York also sees more unique checkins at major tourist attractions such as Times Square and Madison Square Garden than SF, indicating that New York likely sees a more varied set of users than San Francisco does. It is surprising to note that tourist attractions such as Union Square, Fisherman's Wharf, Pier 39, and Westfield SF Mall see a lower proportion of unique checkins whereas travel and transportation locations see a higher proportion of unique checkins. This unexpected result might suggest something about the types of people that check in at different places; it is fathomable that tourists would be more likely to check in at travel or transportation stations whereas commuters wouldn't, thus leading to a higher number of unique checkins, whereas everyone (both tourists and locals) would check in at major tourist attractions if they are large enough to be noteworthy, thus leading to a lower number of unique checkins.

6.4 Geographical Maps

We plotted the top 10 non-airport/travel-and-transport (otherwise it would have just been airports and public transit places) based off of their closeness score on a geographical map.

Miami: **New York:** **San Francisco:**

Figure 6. Geographical map containing the top 10 non-airport/travel-and-transport locations.

Geographical maps were created for Miami, New York, and San Francisco. Taking a look at the physical separation between these places gives us an insight into where popular places in a city may exist and how a city is laid out. In looking at Miami, we notice that there are two clusters, one around Fort Lauderdale and one around Miami Beach, suggesting that this is where most people tend to go. Additionally, the types of places that are in the top 10 are mostly food places and beaches, which makes sense given that Miami is located so close to the water. As for New York, we notice that the main cluster is in the heart of Manhattan. Furthermore, the places that appear in the top 10 are typical tourist attractions like Times Square, Rockefeller Center, Herald Square, Macy's, and Madison Square Garden. These make sense as people will often check-in to these popular places and then move on to other places, giving them high closeness centrality scores. Finally, in examining San Francisco, it is mostly spread out near the top with a small cluster of points at SFO. The upper distribution of places is partly indicative of the fact that the San Francisco area is pretty spread out. Additionally, while the small cluster at SFO is comprised mostly of parking structures, the places in the upper spread represent some of the more famous attractions of San Francisco. These popular places include the Golden Gate Bridge, Union Square, the Ferry Building, and

the Bay Bridge. By looking at these geographical maps, we can get a sense of the structure of these cities through their physical separation that would not have been as apparent in a weighted graph.

7 Future Work

The Foursquare Dataset we were provided had very detailed place information, but nearly no data on the people themselves (not even a userid). Given this constraint, we specifically examined our data with the intention of learning about the places rather than the people. An entirely new set of questions would arise regarding users and temporal features if we were able to gain access to a dataset (either of Foursquare or a different LBS) that had more detailed information on the users; we could extrapolate more information on the types of users who check in to particular types of places. If we had user ids for different users, for example, we could analyze people who check in at the same place at the same time, and whether particular individuals are more influential than others. It could also be interesting to infer insights about sequences of checkins, if we could gain basic information about users such as a userid. Additionally, while our dataset spanned the course of a year, it would be interesting to accumulate more data over more time to examine if any places see a "rich gets richer" effect over the years that goes beyond just being an annual trend.

8 Citations

[1] A. Noulas, S. Scellato, C. Mascolo, M. Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. AAAI, 2011.

[2] A. Bawa-Cavia. Sensing the Urban: Using location-based social network data in urban analysis. UCL, 2010.

[3] M. Allamanis, S. Scellato, C. Mascolo. Evolution of a Location-based Online Social Network: Analysis and Models. IMC'12, 2012.

[4] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. Statistical Properties of Community Structure in Large Social and Information Networks. WWW 2008, 695-704.

[5] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters, and possible explanations. In Proceedings of KDD'05, 2005.

[6] T. Opsahl, F. Agneessens, J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. In Social Networks 32 (2010), 245-251.

[7] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani. The architecture of complex weighted networks. In Proceedings of the National Academy of Sciences 101 (11), 3747-3752.