2016 IEEE 40th Annual Computer Software and Applications Conference

Centrality and Cluster Analysis of Yelp Mutual Customer Business Graph

Brian McClanahan and Swapna S. Gokhale
Dept. of Computer Science and Engineering
University of Connecticut, Storrs, CT 06269
Email: {brian.mcclanahan, swapna.gokhale}@uconn.edu

***Abstract*—This paper proposes a novel approach to understand customer relationships among businesses and the type of information that can be inferred from these relationships. Our approach is grounded in a unique method of constructing a mutual customer business graph, where businesses are represented by nodes and the weight of the edge connecting two businesses reflects the strength of their mutual customer population, which is estimated based on the reviews from the Yelp academic data set. We construct and analyze these mutual customer business graphs for cities of Las Vegas and Phoenix using centrality and spectral analysis techniques. Centrality analysis computes unweighted and weighted versions of degree and PageRank graph measures; the results reveal that businesses with high graph centralities also tend to be geographically central relative to other businesses. Spectral clustering partitions the graph to group businesses that are frequented by the same set of customers. An analysis of the frequency distribution of words from the reviews within each cluster suggests that businesses aggregate around a theme. Taken together, these findings suggest that customers prefer to visit businesses that are geographically proximate and/or offer similar products and services. We discuss how businesses could strategically position themselves by considering the impact of these two factors in attracting clientele.**

I. Introduction

Many groups of businesses share mutual or common customers who visit them frequently. Such customers shared by a group of businesses can indicate different types of relationships among them, as there may be some underlying factors that can cause businesses to share a few, a majority or almost no customers with each other. As an example, it is likely that customers who visit coffee shops like coffee and consequently have tried many other coffee shops. Thus, in this case the nature of the businesses or the types of products and services they offer leads to a sharing of customers between them. On the other hand, although two stores may offer disparate products and services, they may belong to the same shopping center. As a result, they may be visited by the same customers, so the factor of geographic proximity can also potentially influence the shared clientele.

This paper proposes a novel approach to understand the impact of two factors, namely, the nature of a business and its geographic proximity, on its customer relationships with other businesses. Underlying our approach, is a unique method to construct mutual customer business graphs. These graphs comprise businesses as nodes and the edges connecting the businesses have weights which are proportional to the strengths of their mutual customer bases. The number of mutual customers between pairs of businesses is estimated based on the business reviews from the Yelp academic data set [1] for the cities of Las Vegas, Nevada and Phoenix, Arizona. Separate graphs are constructed for each city, so the results of the following two types of analysis can be interpreted from a local perspective.

- *Centrality analysis* computes unweighted and weighted versions of degree and PageRank measures for each business, to understand the volume and strength of the business's connections. These centrality measures are then concurrently viewed with business locations to determine if any correlation exists between business centrality and business placement. Our findings from centrality analysis suggest that businesses which are geographically central tend to have the strongest customer relationships.
- *Spectral clustering* partitions the graph to group businesses together based on the strength of their customer relationships. Subsequently, an analysis of the frequency of words from the business reviews for each cluster reveals that businesses in most clusters are grouped according to a theme. This theme-based clustering suggests that similarity in products and services offered can also contribute to a strong mutual customer base.

Taken together, centrality and spectral analysis highlight the importance of geographic proximity and similarities between services to the development of mutual customer population. We conclude the paper by discussing how businesses could strategically consider the impact of these two factors to position themselves to attract customers.

The organization of this paper is as follows. Section II presents the procedure used for graph construction. Section III describes how the Yelp data set can used to construct the business graphs. Sections IV and V present centrality and spectral clustering analysis respectively. Section VI summarizes our findings, its applications, and threats to validity. Section VII compares and contrasts related work. Section VIII concludes the paper and provides future directions.

II. Mutual Customer Business Graph

We describe a novel approach to construct a mutual customer business graph that compactly represents the connections between the businesses. Nodes in our graph represent businesses, and the edges represent relationships among these businesses. The links between the businesses are based on the

0730-3157/16 $31.00 © 2016 IEEE
DOI 10.1109/COMPSAC.2016.79

intuition that if two businesses share a large number of mutual customers, then there may be a strong connection between them. An edge between two nodes could be unweighted with a binary interpretation, that is, the presence of an edge indicates that the number of mutual customers between the two businesses exceeds a certain threshold. However, because the number of mutual customers indicates the strength of the connection between the two businesses, we use weighted edges, where the weight of the edge is proportional to the number of mutual customers.

Using the raw count of the mutual customer base between the two businesses to compute the weight of the edge between them, however, could lead to a situation where two businesses which individually have larger customer bases will naturally have a larger edge weight compared to two businesses with smaller individual customer populations. To overcome this bias in favor of businesses with larger individual customer populations, we compute the weight of an edge by first calculating the fraction of mutual customers for each business, and then computing the average of this fraction. To explain the computation of edge weight, we consider two businesses A and B, respectively with 300 and 600 customers. Setting the customers shared between the two businesses to 150, the fraction of the mutual customers for A with B is $150/300 = 1/2$, and for B with A is $150/600 = 1/4$. The weight of the edge connecting business A and B is then computed as the average of these two fractions, which is $3/8$.

To formalize the computation of edge weight, we let $G = (B, E)$ be an undirected graph with vertices B and edges E, $b_i \in B$ and $b_j \in B$ denote businesses i and j respectively, and $c(b_i)$ and $c(b_j)$ denote the sets of customers of businesses b_i and b_j respectively. The edge weight between businesses b_i and b_j, denoted $w_{i,j}$, is defined as:

$$w_{i,j} = \frac{\frac{|c(b_i) \cap c(b_j)|}{c(b_i)} + \frac{|c(b_i) \cap c(b_j)|}{c(b_j)}}{2} \quad (1)$$

With this construction, the weight of the edge linking two businesses is high only if the fraction of their mutual customers out of their total customer population is high. We argue that these fractions of mutual customers are more indicative of the strength of the connections between the two businesses, which can be easily inflated by the scale of the customer populations of individual businesses.

III. Yelp Data

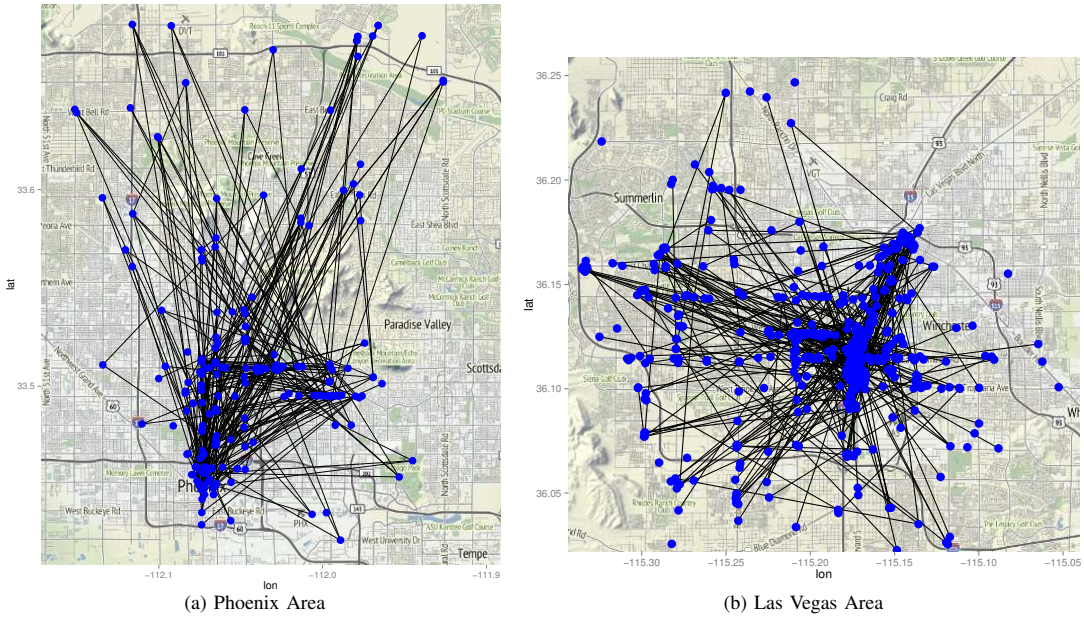
We estimate the customer population $c(b_i)$ for each business b_i based on the reviews in the Yelp academic data set [1]. Yelp is a social media platform designed to help users find and evaluate local businesses of interest. Users rate and review businesses and have their opinions shared with other Yelp users. The data set comprises a wealth of information on businesses and Yelp users across 172 U.S. cities. Additionally, it contains a corpus of business reviews. The data is stored in JSON format and comprises several objects, of which the following three are relevant for our study:

- **Business Objects:** These objects contain information on individual businesses, such as a unique business id, business name, state, city, latitude and longitude coordinates, average star rating, etc.
- **User Objects:** These objects provide information on Yelp users, such as user id, name, friends, review count, average star rating, etc.
- **Review Objects:** These are reviews of businesses by Yelp users. They contain the user id, business id, review text, and star rating.

Because our objective is to understand customer relationships among businesses, our graph only includes businesses from a single city, as customers from one city are unlikely to visit businesses from other cities regularly. In fact, analysis over these city-wide graphs creates groups of local businesses, and allows us to interpret the results from a local perspective. We selected the cities of Las Vegas, Nevada and Phoenix, Arizona for graph construction and analysis, due to the larger number of businesses that were reviewed in those areas. 7449 businesses in the Phoenix area were reviewed, and this number is 12022 for the Las Vegas area. In the Yelp data set, there are quite a few cities with only one review, and the average number of reviews for a city is 234.

Our estimate of $c(b_i)$ will be accurate only if it considers those customers that are satisfied with the business and return, rather than those who use the business occasionally but never return because they were dissatisfied or for other reasons. The data set, however, provides no information about how often a Yelp user uses a business. As a result, we estimate whether a user is satisfied with a business or not based on the user's rating of the business. We assume that if a user has rated a business positively (4 stars or higher), then this user is satisfied and is likely to use the business again. In contrast, a user who has rated a business negatively or is neutral (3 stars or below) may be dissatisfied or unhappy and is unlikely to return. Under this assumption, $c(b_i)$ is estimated as the number of Yelp users who have rated b_i positively. Furthermore, to obtain more reliable estimates of mutual customers and to restrict analysis to more popular businesses, we include in the graph only those businesses with more than a certain threshold number of reviews, which we set to 100 for the sake of illustration. Another reason for only considering businesses with more than 100 reviews is that there are a large number of businesses with a small number reviews, so setting a threshold of 100 prevents the graph from becoming too dense.

The business graphs constructed over the cities of Las Vegas and Phoenix are shown in Figures 1a and 1b respectively. The number of vertices in Phoenix and Las Vegas Graphs is 241 and 749 respectively and the number of edges is 26300 and 217360 respectively. To make Figures 1a and 1b more viewable a random sample of only 400 edges is used for each. Examining these figures it appears that both Phoenix and Las Vegas each contain a sub-region with a denser population of businesses than the rest of the region. It also appears that these dense regions account for much of the edge density in the networks. Businesses closer to the boundaries of the regions



(a) Phoenix Area (b) Las Vegas Area

Fig. 1: Mutual Customer Business Graphs

appear to be less connected than nodes in the center.

IV. Centrality Analysis

In this section, we examine the correlation between node centralities of the businesses and their locations. Specifically, the node centrality of a business is a measure of its importance within the graph. Therefore, the objective of this analysis is to investigate how the geographic location of a business is related to its importance in the business graph. The node centrality of a business is influenced by two factors, namely, the number of connections that a business shares with other businesses, and the strength of those connections. Several centrality metrics measure different aspects of a node's connections to other nodes [2]. We elect to use degree and PageRank centralities, both these measures have weighted and unweighted versions.

Degree centrality measures a node's direct connections with other nodes. The unweighted degree of a node is simply the number of adjacent nodes in the graph. Let $a(b_i, b_j)$ be an indicator function which takes the value 1 if business b_i is adjacent to business b_j. Then, the unweighted degree of a business b_i, denoted by $degree_u(b_i)$ is defined as:

$$degree_u(b_i) = \sum_{j=1}^{n} a(b_i, b_j) \tag{2}$$

where n is the number of businesses in the graph. The weighted degree of business b_i, denoted $degree_w(b_i)$ is the following simple extension of $degree_u(b_i)$.

$$degree_w(b_i) = \sum_{j=1}^{n} a(b_i, b_j) w_{i,j} \tag{3}$$

Note that $w_{i,j}$ takes the value 0 if b_i and b_j have no mutual customers. Weighted degree can be thought of as measuring the strength of the connections of a business. According to the construction of the business graph, a business will have high unweighted degree if it is connected to a large number of businesses by the virtue of sharing customers with them, regardless of how many customers are shared individually with each business. On the other hand, a business will have a high weighted degree if the summation of the weights of all the adjacent edges is large, or in other words if the aggregate strength of a business's connections is high.

PageRank centrality measure gives high centrality values to nodes which are connected to well connected nodes. PageRank centrality defined below is for a directed graph. Our mutual customer business graph is undirected, so to compute PageRank we simply convert the undirected graph to a graph with bi-directional links. Let $PR_u(b_i)$ be the unweighted PageRank of business b_i and $Adj(b_i)$ be the set of businesses connected to b_i. The unweighted PageRank of business b_i is then defined as follows [3]:

$$PR_u(b_i) = \frac{1-d}{N} + d\Big(\sum_{j:b_j \in Adj(b_i)} \frac{PR_u(b_j)}{degree_u(b_j)}\Big) \tag{4}$$

In Equation (4), $d \in [0, 1]$ is a damping factor.

The PageRank scores form a probability distribution over businesses. An iterative algorithm can be used to compute PageRank, and the PageRank scores of nodes correspond to the principle eigenvector of the normalized adjacency matrix [3]. PageRank was developed by Brin and Page [3], and they describe it in terms of a "random surfer" model, in which a Web user starts on a web page and clicks links at random. With probability $1 - d$, the user will get bored and decide to stop clicking links and randomly jump to another web page. Unweighted PageRank is computed based on the assumption

that once a user is on a web page, the links on that web page have an equal probability of being clicked. To eliminate this assumption, the links can be given weights and the probability of clicking a link can be made proportional to its weight. This intuition leads to the following formulation of weighted PageRank:

$$PR_w(b_i) = \frac{1-d}{N} + d(\sum_{j:b_j \in Adj(b_i)} PR_w(b_j) * \frac{w_{i,j}}{degree_w(b_i)}) \quad (5)$$

In the context of the business graph, we analogize the "random surfer" model with the "random shopper" model. In our "random shopper" model a shopper starts at some business and randomly visits businesses which are adjacent. With probability $1-d$, the shopper decides to jump to a random business ignoring adjacency. In our experiments, we set $d = 0.15$. The links between businesses can be unweighted or weighted, leading to a uniform or a non-uniform probability distribution over adjacent businesses. If links are unweighted then each business adjacent to the current has an equal probability of being visited. In contrast if the links are weighted the businesses may have an unequal probability of being visited. Both the unweighted and weighted PageRank scores were computed using the Python igraph library [4].

Figures 2 and 3 show both unweighted and weighted degree centralities of businesses for Las Vegas and Phoenix areas respectively. An inspection of Figure 2a reveals that many businesses have a high degree, with most degrees being over 400. The businesses with the highest degree appear along the Las Vegas Strip, a major tourist attraction in Las Vegas, known to contain some of the largest hotel, casino, and resort properties in the world [5]. Although many businesses along this strip have a large number of connections, most of these connections are weak, as is evidenced by inspection of the edge weight histogram in Figure 4 and the weighted degree centralities in Figure 2. Most edge weights are less than $.04$, indicating that few businesses share more than 4% of their customer base with other businesses. This could be because most of the businesses along the strip are of a similar nature, they are all casinos and gambling places, and allied services. Each casino is likely to have a loyal fan base, and they are unlikely to go elsewhere. The edge weight histogram for the Phoenix area in Figure 4b shows the same effect, although at a slightly lower intensity. As a result, few businesses have a high weighted degree. An interesting observation is that businesses that are geographically central relative to other businesses on the Las Vegas Strip are the ones that have the highest weighted degree. This may suggest that there is an advantage to being geographically central in terms of attracting customers from other businesses. The observations, noted for the unweighted and weighted degree centralities of businesses in Las Vegas, also apply to the business graph for Phoenix. Most businesses in Phoenix have over 200 connections, however, the weighted degrees of most businesses is low. As observed for Las Vegas, businesses with highest weighted degree in Phoenix appear in locations geographically central relative to other businesses.

Figures 5 and 6 show the unweighted and weighted PageRank centralities for businesses in Las Vegas and Phoenix areas respectively. The unweighted and weighted PageRank plots in Figure 5 and 6 are virtually identical to the unweighted and weighted degree centrality plots in Figures 2 and 3 with the only significant difference being the scale of the values. Thus in the "random shopper" model, the probability of a shopper visiting a particular business is highly dependent on the businesses immediately connected to it, similar to a "random surfer" model where the pages a user views strongly depends on the links available on that current page.

To explore the relationship between geographic and graph centralities, we further formalize the former notion. Towards this end, we compute the average distance of each business from the other businesses in that area. We then consider a business to be geographically central if this average distance is low. Figure 7 contains scatter plots showing the average distance vs. weighted degree centrality for each business for both the Las Vegas and Phoenix areas. Observing Figure 7, somewhat of a negative correlation can be seen between average distance and weighted degree for Phoenix. Although this correlation is not as clear for Las Vegas, it can be observed that one of the businesses with the shortest average distance has the highest weighted degree and also that the business with the greatest average distance has one of the lowest degrees. We choose weighted degree here as a representative for graph centrality. Both unweighted and weighted PageRank centralities and also unweighted degree centralities show negative correlation patterns similar to those shown here for both Phoenix and Las Vegas areas. These figures suggest that average distance can impact a business's graph centrality, however, depending on the region the influence of distance may be more or less of a determinant in the customers that the businesses share.

V. Clustering Analysis

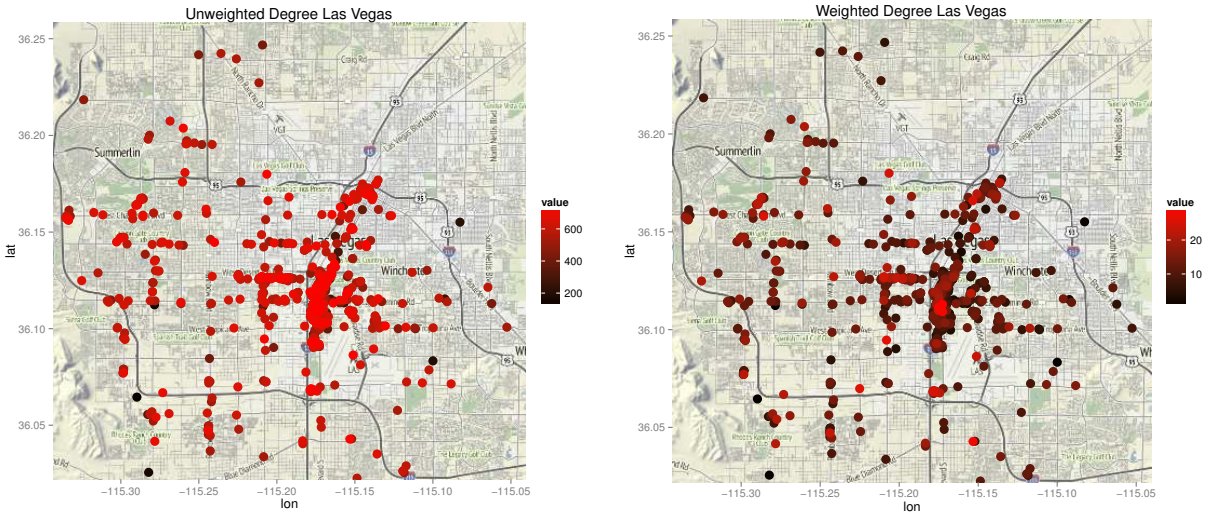
In clustering analysis, we partition the business graph for each area to identify groups of businesses which share a large number of customers. A natural method to identify such groups is to partition the graph into subsets such that businesses within the same subset share strong connections, while businesses between subsets share weaker connections.

We partitioned the weighted business graph of each area using spectral clustering [6]. Spectral clustering partitions a graph in a way which solves a relaxation of the normalized cut problem. Let $A_1, \ldots, A_k$ denote k subsets of vertices in a partition of B. Normalized cut, denoted as $Ncut$, is then defined as follows [6]:

$$Ncut(A_1, \ldots, A_k) = \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{vol(A_i)} \quad (6)$$

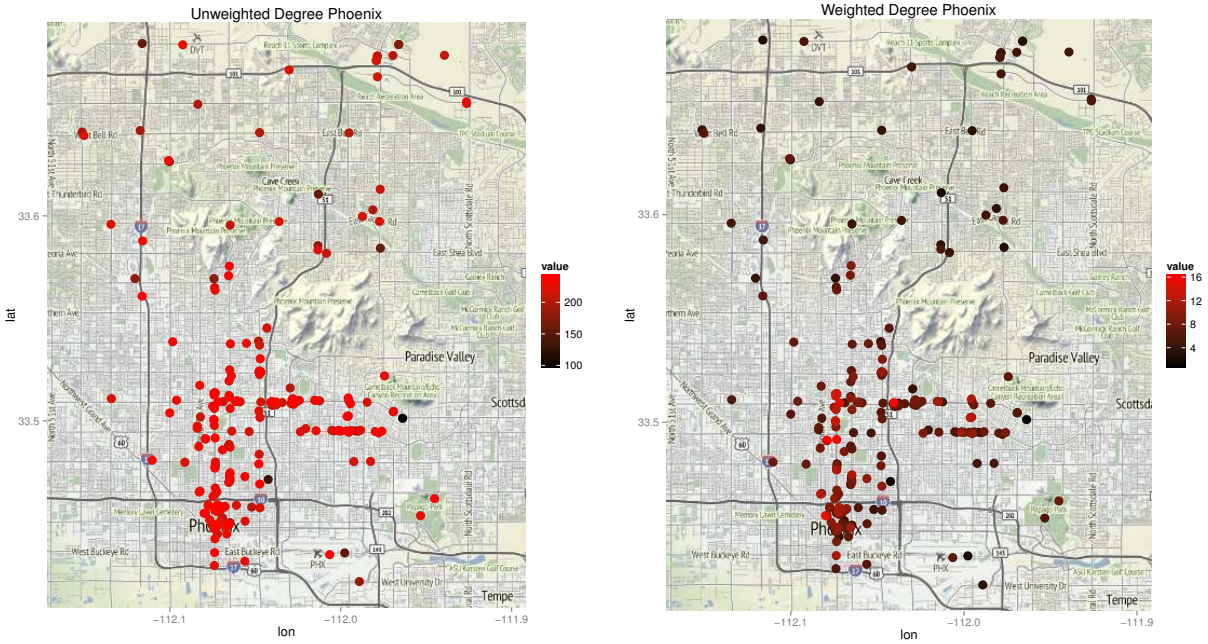
$$W(A_i, \overline{A_i}) = \sum_{\{m,n:b_m \in A_i, b_n \in \overline{A_i}\}} w_{mn} \quad (7)$$

$$vol(A_i) = \sum_{\{m,n:b_m \in A_i, b_n \in B\}} w_{mn} \quad (8)$$



(a) Unweighted Degree (b) Weighted Degree

Fig. 2: Degree Centralities of Businesses – Las Vegas

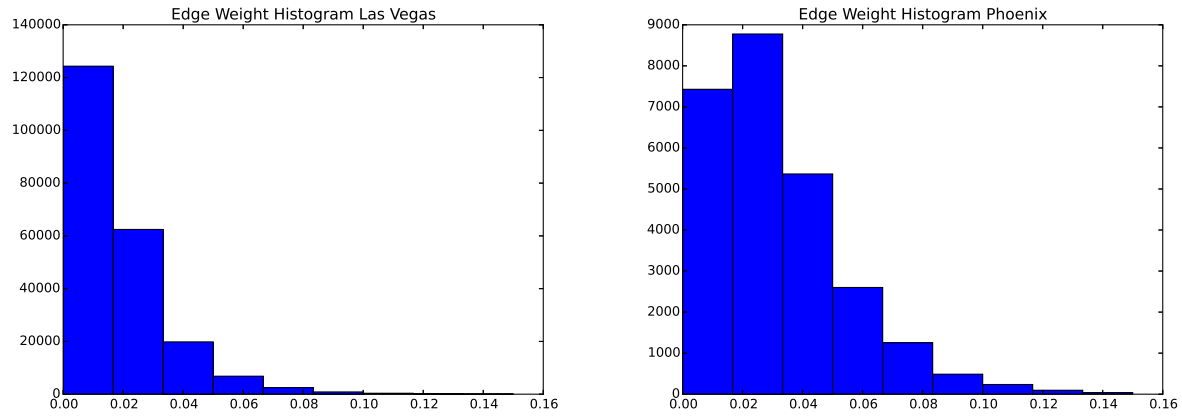


(a) Unweighted Degree (b) Weighted Degree

Fig. 3: Degree Centralities of Businesses – Phoenix

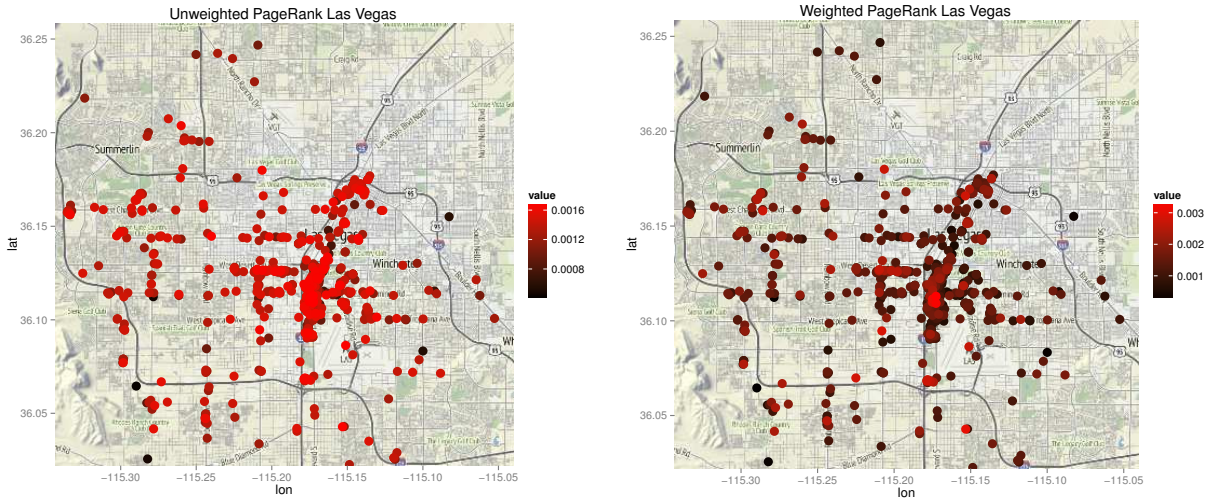
Here $\overline{A_i} = B \setminus A_i$ is the compliment of A_i, $W(A_i, \overline{A_i})$ is the summation over the weights of all edges between A_i and $\overline{A_i}$, and $vol(A_i)$ is the sum over the weights of all edges incident to a node in A_i. Spectral clustering was performed using the Scikit-learn Python machine learning package [7].

The business clusters produced from spectral clustering are shown in Figure 8 for the cities of Phoenix and Las Vegas, with the number of clusters set to 15. The quality of the clusterings is assessed manually; thus no automated procedure could be implemented for identifying the optimal number of clusters. Upon inspecting the results of 5, 10, and 15 clusters, we observed that setting the number to 15 seems to produce the most distinct clusters in terms of business reviews. The color of a business in Figure 8 indicates its membership in a cluster. Businesses which belong to the same cluster often appear close to one another, indicating that the proximity of businesses to one another must have an impact on the number of mutual customers. This confirms our observations



(a) Las Vegas (b) Las Vegas

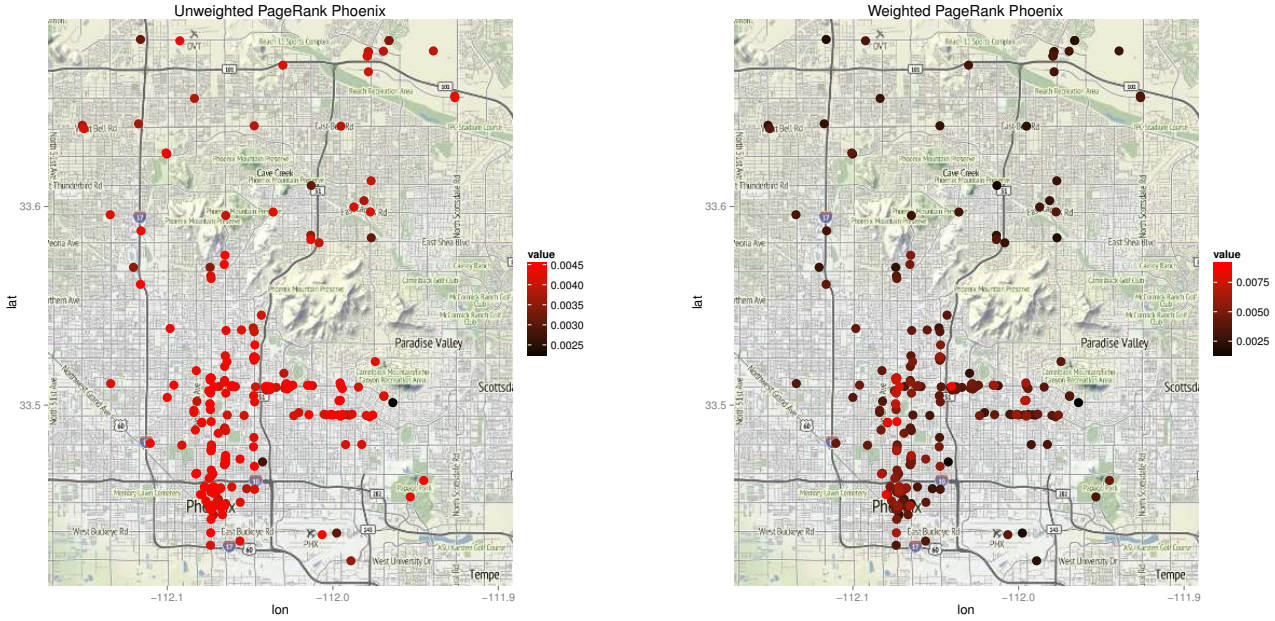
Fig. 4: Edge Weight Histograms



(a) Unweighted PageRank (b) Weighted PageRank

Fig. 5: PageRank Centralities of Businesses – Las Vegas

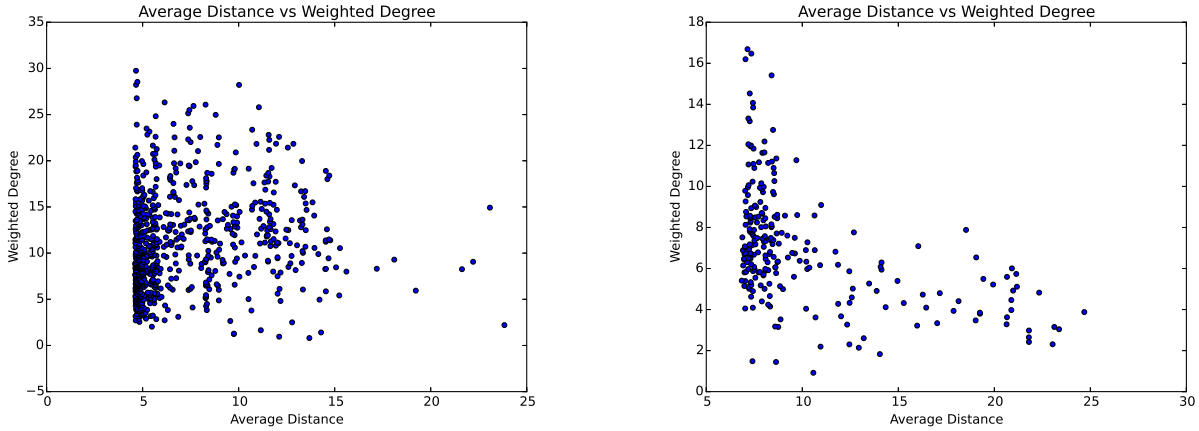
from centrality analysis in Section IV, which indicates that proximity can influence a businesses' graph centrality, and hence, the strength of its connections to other businesses. However, Figure 8 also shows that businesses within the same cluster can also be geographically distant, suggesting that proximity is not the only factor affecting how many customers businesses share. As the geographic location of a business is not the only factor in attracting customers, we postulated that for some businesses the types of services and/or products it offers plays a role its connections with other businesses. To test this hypothesis, we analyzed the reviews of businesses within the same cluster. Although the Yelp dataset provides a "categories" field which can be used to identify business type, we elect to analyze the word distributions of reviews instead of using these categories. Using reviews allows us to obtain more information about businesses within a cluster than what could be inferred from viewing simple categories such as "restaurants" or "nightlife". For example, reviews reveal not only the types of products and/or services that businesses offer but also specifically what customers like or dislike about business offerings. For this analysis, we pre-processed the reviews in two ways. First, we removed the stop words from the reviews; these are the words which commonly appear but are not really associated with any type of theme. Additionally, we also stemmed words to eliminate their different morphological manifestations. Word stemming was performed using the nltk Python package [8]. After pre-processing, we compile all the unique words that occur in the



(a) Unweighted PageRank

(b) Weighted PageRank

Fig. 6: PageRank Centralities of Businesses – Phoenix



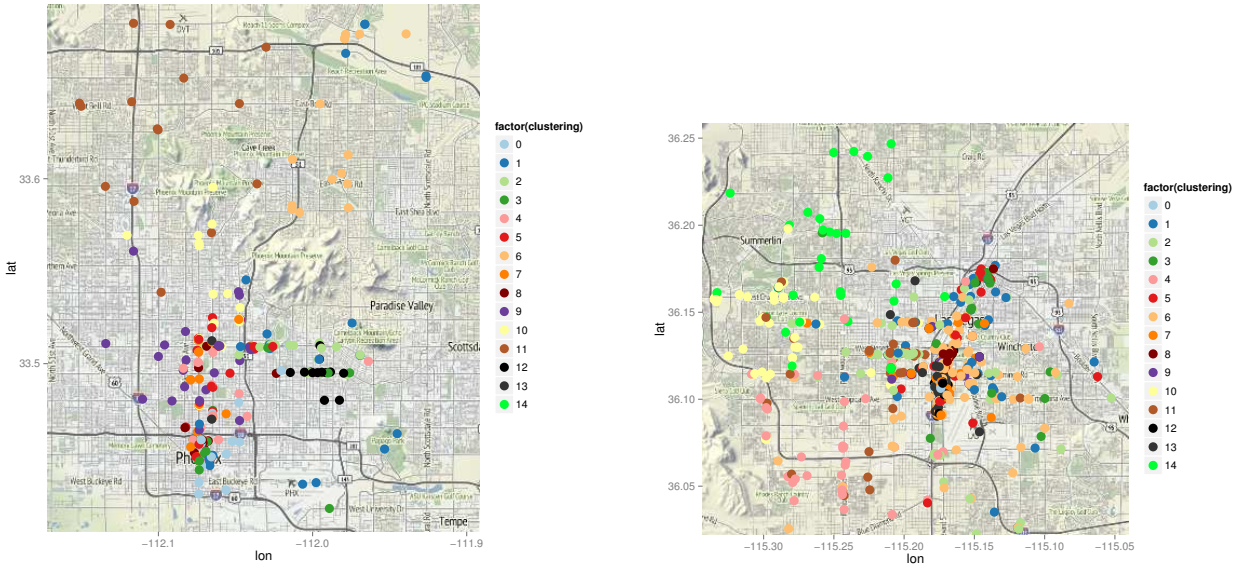
(a) Las Vegas

(b) Phoenix

Fig. 7: Correlation Between Average Distance and Weighted Degree Centrality

reviews of all the businesses in each cluster. For each unique word and for each business, we then compute a normalized count of the number of times the word occurs collectively in all the reviews of that business. The normalization factor is the total number of reviews for that business. For example, if "dog" is one of the unique words that appears in all the reviews in a cluster, and if a pet shop has a total of 30 reviews in which the word "dog" appears 15 times, then the normalized count of the word "dog" in all the reviews is $15/30$.

Let $V = \{v_i : i \in [1, N]\}$ be the set of all unique words observed in the reviews in the cluster after stopword removal, where N is the total number of observed words. For each business b_i within the cluster k and each word $v_i \in V$, the normalized count of word v_i across all reviews of b_i is computed. After computing the normalized counts for each unique word for each business, for each word we then aggregate its normalized counts across all businesses within the cluster. We then rank these aggregated normal-



(a) Phoenx Area

(b) Las Vegas Area

Fig. 8: Spectral Clustering of Businesses

ized counts, and subsequently rank the words in the cluster according to their aggregated normalized counts. We define $clustNormCount_k(v_i)$ to be the sum of all the normalized counts of word v_i for cluster k.

$$normCount_{b_j}(v_i) = \frac{count_{b_j}(v_i)}{numReviews(b_j)} \quad (9)$$

$$clustNormCount_k(v_i) = \sum_{b_j \in C_k} normCount_{b_j}(v_i) \quad (10)$$

In Equation (9), $normCount_{b_j}(v_i)$ is the normalized count of word v_i for business b_j. Alternatively $normCount_{b_j}(v_i)$ can also be viewed as the average number of times word v_i occurs in a review of b_j. $count_{b_j}(v_i)$ is the total number of times word v_i occurred across all reviews of b_j, and $numReviews(b_j)$ is the number of reviews available for business b_j. The words in the cluster k are then ranked, with the highest ranked words being those with the highest values of the metric $clustNormCount_k(v_i)$. The highest ranked words suggest the types of businesses that belong to the cluster.

Instead of using the normalized word counts for ranking, an alternative would have been to simply rank each word based on the total number of times it occurred in all the reviews of all the businesses within a cluster. However, the ranking produced by this approach would be biased, where businesses with most reviews in the cluster would dominate the ranks. It would be desirable for words with the highest rank to be representative of a majority of the businesses within the cluster, not just a few businesses with a large number of reviews.

Tables I and II show the words ranked in top 10 positions for five clusters each from both Las Vegas and Phoenix respectively. These five clusters for each area are chosen solely for the sake of illustration. The first row in the tables gives the theme of the businesses within the cluster as inferred from the top 10 words. Yelp data set has a disproportionate number of restaurant reviews, so most clusters seem to have restaurant themes. Also, we noted that the top 10 words of a cluster do not always appear to correspond to coherent themes. This is because in many clusters there may be other dominant factors, including geographic proximity, that influence the number of customers shared, other than the similarities in products and/or services. Thus, for some clusters a coherent theme does not emerge based on the top 10 most frequently used words. In both the tables, the columns labeled "Misc" for miscellaneous give examples of a cluster with a non-coherent topic.

The top ten words for the four clusters shown in Tables I and II, however, constitute pretty clear themes. For example, in the clubs cluster in Table I we see words such as club, dance, music and crowd, which are all words describing a club scene. In the live shows column we see cirqu as the top ranked word and also words such as ticket, fun, and performance. Looking at the businesses within this cluster, we observe several Cirque Du Soleil shows, The Lion King, Jabbawockeez, and The Blue Man Group, which are all live performances. In the Family Events column of Table II, there are several words which can be associated with family activities, such as zoo, museum, and park. The businesses of the Family Events cluster include Phoenix Zoo, Musical Instrument Museum, South Mountain Park, and Children's Museum Of Phoenix, which are all family oriented places. All examples above offer strong evidence that business similarity can lead to a large number of customers being shared between businesses. These clusters can help business owners or managers understand the various interests

of their customers. This information can be used strategically. For example, a knowledge of what other businesses belong to the cluster may encourage business owners to add products to their inventory which are available at those other businesses.

TABLE I: Example Themes for Las Vegas Clusters

Hotels	Clubs	Live Shows	Japanese Food	Misc
pool	club	cirqu	ramen	chip
clean	danc	shop	sushi	patio
bed	music	ticket	rice	dog
free	girl	store	noodl	downtown
locat	crowd	fun	pork	select
old	floor	seen	roll	cool
floor	guy	end	soup	music
cheap	free	mani	cream	seem
bad	pool	music	fresh	check
use	parti	perform	tea	hot

TABLE II: Example Themes for Phoenix Clusters

Coffee/ Bakeries	Family Events	Mexican Food	Nightlife	Misc
coffe	exhibit	mexican	select	pizza
ice	park	salsa	music	sushi
work	airport	hot	cool	salad
shop	museum	dog	sushi	roll
cooki	kid	big	pub	fresh
cream	hike	spici	downtown	friendli
clean	trail	bean	bartend	dog
cupcak	walk	worth	walk	wine
use	zoo	green	seem	lunch
hi	year	find	live	bread

VI. Summary & Discussion

Based on the results of centrality and spectral analysis of the business graph, a natural question then arises is which factor, namely, location or business type is dominant in determining the number of mutual customers shared by businesses. Our results suggest that neither factor dominates, and indeed both may be equally important. However, depending on the type of business under consideration one factor may be more important than the other. For example, looking at the top left of the Phoenix clusters in Figure 8 it can be seen that all businesses in this sub-area belong to the same group. Google maps suggests that this may be the Deer Valley sub-area within the city of Phoenix, although we note that some businesses in the cluster may be outside of Deer Valley [9]. These businesses are rather distant from the other businesses in Phoenix, so proximity in this case appears to be the factor with most influence. In the Las Vegas clusters, however, businesses which are distant tend to be grouped with one another. Also, towards the bottom of the Phoenix clusters, significant overlap can be seen in the clusters. In this type of situation, the services provided may play a greater role in determining which businesses are grouped together, because businesses are closer to one another in that sub-area.

The insights produced from analyzing the mutual customer business graphs can be used by businesses strategically in several ways. Centrality analysis may be used to identify sub-regions where businesses have high degree and PageRank measures. It is likely that the connections among the businesses in these sub-regions are influenced by proximity. A new business could strategically position itself in such a sub-region in order to improve its chances of quickly gaining customers from the traffic of nearby businesses. Also, an examination of the products and services offered by businesses with large centrality values can provide insights into what type of factors can draw customers in a particular sub-area. For example, some sub-areas may be family friendly, whereas some others may cater towards younger crowds.

We noted that some of the clusters retrieved from spectral clustering corresponded to coherent themes, which implies that this type of analysis can be used to identify different business genres. More interestingly, these genres may not solely contain homogeneous business types such as restaurants or bars, but may include different types of businesses, which are all related to some overarching, central theme. For example, the "Family Events" theme shown in Table II includes parks, museums, and a zoo, which are very different types of places but all of them offer opportunities for family activities. Identification of these genres and the businesses within them can be used to recommend businesses to users. A user's interest in a particular genre may be identified by the number of businesses that a user has reviewed under that genre, and also the ratings of all such reviews. After identifying a user's interest in a genre, businesses that belong to that genre can be recommended. Furthermore, any candidate business recommendation within a genre can be weighted by the weights of its connections to other businesses that the user has visited.

Finally, we note some threats to the validity of the analysis performed on the business graphs. One such threat arises from the assumption that if two businesses are rated positively by the same reviewer, then the businesses share the reviewer as a customer. This may not always be the case. For example, the businesses may be competitors and attract customers away from each other. Another threat is from ignoring the time dimension when examining reviews to determine business connections. It is possible for a reviewer to rate a business positively at one point in time and then to opine negatively (or vice versa) later due to declining (increasing) satisfaction. This issue, however, is not very significant for Phoenix and Las Vegas respectively because only 2% and 1.8% of their reviewers have both a review below and above or equal to 4 star threshold for a single business.

VII. Related Research

To the extent of our knowledge, the construction of the mutual customer business graph described in this paper is unique, and thus so is all the subsequent centrality and clustering analysis performed on it. However, we note that the word distributions of the clusters retrieved from spectral clustering appear to correspond to coherent topics. This suggests that our clustering approach can be used as an alternative to more traditional topic modeling approaches such as Latent Dirichlet

Allocation (LDA) [10], when performing topic modeling on business reviews with ratings. When LDA is employed, each review comprises a mixture of topics in different proportions. McAuley *et. al.* [11] combine a latent factor model for predicting ratings with LDA for topic modeling. They make each dimension of the latent factor rating matrix correspond to a topic and transform each individual's latent rating factors to a distribution over topics. The topics retrieved were found to correspond to genres or categories and are similar to the clusters found in this paper. Linshi [12] uses LDA and codewords to identify topics which are more semantically aligned with review ratings. For each review a positive or negative codeword is added after each occurrence of a corresponding adjective. The addition of these codewords makes it so that all of the various adjectives people may use to describe some entity register as one unique word. Thus, LDA can more easily identify topics such as, "Good Food", "Bad Food", "Good Service", or "Bad Service".

Our work differs from the contemporary topic modeling approaches along several dimensions. First, contemporary approaches seek to reveal topics that are semantically related to user ratings. On the other hand, the topics retrieved from spectral clustering correspond to businesses with strong customer relationships. Thus, the semantic interpretation of these topics is certainly different, if not richer because it is not just determined by the co-occurrence of words in documents as in the case of regular LDA. Second, in contemporary works, each review could be associated with multiple topics. In our approach, each business and all of its reviews are associated with a single cluster, so analogizing the topics from LDA with clusters from spectral clustering, it is clear that each review would be associated with only one topic.

Work related to the centrality analysis is that of Tiroshi et al. [13], who also use data from the Yelp data set. They construct a bipartite graph containing users and businesses as vertices, where a user is connected to a business by the virtue of reviewing the business. A second graph is tripartite with users, businesses, and metadata such as business category and business location as vertices. Users are connected to businesses based on reviews and businesses are connected to metadata by relationships. For example, if a business belongs to category "food" then it will be connected to the food metadata vertex. Centrality measures such as degree and PageRank were used to augment the features of a business recommender system constructed using Random Forest Regression model. Other graph measures used by the model include clustering coefficients and node redundancy. The business centrality in this work depends upon the number of reviewers in the case of a bipartite graph, additionally on the association to metadata vertices in the case of tripartite graph. In our analysis, however, business centrality in the mutual customer business graph is solely determined by the customers that the business shares with other businesses.

VIII. Conclusions and Future Research

This paper examines the customer relationships among businesses in the cities of Las Vegas and Phoenix through the use of mutual customer business graphs, constructed based on reviews from the Yelp academic data set. One mutual customer business graph is constructed for each city and both the unweighted and weighted versions of degree and Pagerank centralities are computed for each business in each graph. Our results reveal that businesses with the highest centrality (the most well connected) tend to be geographically central. By visually examining the locations of businesses within a cluster, we found that location can have a significant effect on the customers businesses share. We also identify groups of tightly connected businesses in each graph through the use of spectral clustering. Analyzing the frequency of the words in the reviews within clusters revealed that businesses in most clusters aggregate around a theme. We conclude that both proximity and business type significantly impact the customers shared by businesses, although neither factor dominates.

In this work, spectral clustering identifies genres among businesses from the mutual customer business graph. We plan to extend these genres to comprise not just the businesses but people as well. To achieve this, we propose to construct a bipartite graph containing businesses and users as nodes, with a weighted connection between a user and a business, which reflects a user's sentiment towards the business. We will then partition this graph using a spectral algorithm [14] into clusters containing both businesses and users.

References

[1] "Yelp dataset challenge — yelp," id: 1. [Online]. Available: http://www.yelp.com/dataset_challenge

[2] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," in *Proc. of the 7th Intl. World Wide Web Conference*, 1998, pp. 161–172.

[4] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[5] "Las vegas strip," *Wikipedia*, vol. 2015, no. 11/24, 11/08/2015 2015. [Online]. Available: https://en.wikipedia.org/wiki/Las_Vegas_Strip

[6] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] S. Bird, "Nltk: the natural language toolkit," in *Proc. of the COLING/ACL on Interactive Presentation Sessions*, 2006, pp. 69–72.

[9] "Google maps." [Online]. Available: https://www.google.com/maps/place/Phoenix,+AZ/@33.6056711,-112.4052403,10z/data=!3m1!4b1!4m2!3m1!1s0x872b12ed50a179cb:0x8c69c7f8354a1bac

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[11] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proc. of the 7th ACM conference on Recommender systems*, 2013, pp. 165–172.

[12] J. Linshi, "Personalizing Yelp star ratings: A semantic topic modeling approach," *Yale University*, 2014.

[13] A. Tiroshi, S. Berkovsky, M. A. Kaafar, D. Vallet, T. Chen, and T. Kuflik, "Improving business rating predictions using graph based features," in *Proc. of the 19th Intl. Conf. on Intelligent User Interfaces*, 2014, pp. 17–26.

[14] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. of the 7th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2001, pp. 269–274.