STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

# BIA 658 Final Project Report

# Reviews Network of Restaurants

Team 2

Ling Ao

Ruizhi Hao

Siqi Jiang

Yusi He

## Executive summary

The goal of this project is predicting the restaurants' network of the Champaign, Illinoi, the USA in 2020. Restaurants and customers reviews stand for nodes and edges separately. Dataset was downloaded from Yelp Dataset official website, and three kinds of link prediction methods were considered:  shortest path similarity, inverse log-weighted similarity and Jaccard similarity. Reviews of 2018 are the training dataset and Reviews of 2019 are the testing dataset. In result, Jaccard similarity method provided the highest accuracy 7.87%, while the accuracy of the random method is just 3.31%. We calculated the results again by using reviews of 2016-2017 as training dataset and reviews of 2018-2019 as test testing dataset. Jaccard similarity method calculated a higher accuracy 24.3%. Then we used Jaccard similarity method to predict the network of 2020 and visualized the network on the real map. Based on the centrality algorithm, ten most popular restaurants were selected.

# Project Introduction

### *Background*

Yelp is a popular and biggest merchant review website in the United States. Found in 2004, it includes merchants in restaurants, shopping centers, hotels, and tourism. Most customers will search for the related information from yelp.com before they go to the restaurant, which they plan to go to.

With a large number of reviews and user ratings, Yelp has a steady adoption rate in the United States. For users who frequently browse restaurant consulting sites, they always want to be able to get as much detail as possible. Yelp is so successful in the United States because it can unify the information and further enrich it. As far as restaurants are concerned, this information includes store name, restaurant category, customer rating (up to 5 stars), location, phone number, business hours, price range, availability of credit card, clothing requirements, dinner availability, parking lot, wheelchair access, customer submissions or photos, etc. Besides, the number of user ratings is also quite rich, so far the number of reviews has been nearly 20 million.

### *Motivation & Research question*

Under the era of the rapid development of Internet technology， the publicity of physical business has become more diversified. We hope to realize the rapid growth of the real market with the help of the highly developed Internet technology. Combining with our own life experience and the big data processing and analysis techniques learned in the classes, we put forward the idea of using link prediction to make predictions on restaurant review networks. restaurants represent nodes, and if a user rates two restaurants, then there will be an edge between them.

Our research objects are restaurants in Champaign, Illinois, USA, and user comments on yelp.com. We plan to find out the similarities of some restaurants, such as similar dishes and tastes, by studying the robust co-commenting network of restaurants on yelp.com, to be able to recommend the restaurants they like to customers. This is a research topic that benefits both supply and demand. At the same time, we can provide potential customers who want this kind of dishes based on the research results to help merchants market. The dataset we used come from the Yelp Dataset Challenge.

# Literature Review

### *Yelp Review*

Online reviews play a critical role in decision making; however, it's usefulness may bring some problems, such as: deceptive opinion spam (Mukherjee et al. n.d.). In case of this problem, yelp one of the most popular review platform uses filtering algorithm to filter reviews, which tend to be more extreme than other reviews. The study of Luca and Zervas shows that: around 16% of restaurant reviews, which tend to be more on Yelp are filtered. Thus, the credibility of Yelp reviews is relatively high.

The reviews are not only beneficial for customers but also restaurants. According to these reviews, customers can easily find out the restaurants they like. Sawant proposed a recommendation system for Yelp based on the Yelp business data and reviews by network-based-inference collaborative filtering algorithm, which are proposed by Ming-sheng &Yan & Duan-bin (2013).Restaurants can gain higher profits after analyzing customer reviews. One report from Harvard business school states that a restaurant can earn 5 to 9 percent in revenue when Yelp rating increases one-star (Luca 2011). The four main points customer care about are: service, subsequently value, take out and decor (Huang 2014). If restaurants focus on the four aspects, the rank and revenues of restaurants will increase.

### *Link prediction definition & methods*

Link prediction in the network refers to how to predict the possibility of a link between two nodes in the network that have not been connected by the known network nodes and network structure. This prediction includes both predictions for existing yet unknown links and predictions for future links. By using the link prediction method, we hope to develop a recommendation system model with higher accuracy.

Similarity algorithms based on local information, path-based and random walk are the three main link prediction methods based on structural similarity. Common neighbors (Common neighbors, CN) is the simplest similarity indicator based on local information ( Jaccard, 1901). This indicator considers that the similarity of two nodes is proportional to the number of common neighbors. And we use the Jaccard algorithm similar to CN , which is the normalized form of calculating the node similarity in the CN algorithm (Madadhain & Hutchins & Smyth,2005).

### *Link prediction application*

The link prediction method is widely used in various fields, including business, market, healthcare, social, security, education, etc. The following are the three interesting applications.

For the practical application of Link Prediction, researchers believe that a large organization (such as a company) can benefit from the interaction of informal social networks among its members; these connections help to supplement the official hierarchy imposed by the organization itself (Kautz,Selman,&Shah,1997). It can indicate promising interactions or collaborations that have not yet been determined within the organization by analyzing such a social network using the link prediction method. On the other hand, security research has recently begun to emphasize the role of social network analysis, the main motivation of which is to monitor terrorist networks; in this case, link prediction allows people to speculate that specific individuals are working together, even if their interactions are not observed directly (Krebs,2002). Another article mentions that when everyone is concerned about predicting future links, they believe that predicting links that may disappear in the future is equally important. Therefore, a link prediction model is proposed, which can simultaneously predict possible links occurring and links that may disappear in the future and successfully apply it to two different but very related fields, namely healthcare and gene expression networks. The first application focuses on doctors and their interactions, while the second application involves genes and their interactions.

## Data Preprocessing

### *Data collection*

The data was downloaded from the Yelp dataset official website directly. There are two files: one is a JSON file, and another one is a photos file, we chose the JSON file, which was edited on December 13th, 2019. There are five sub Json files in the Json file. After considering, we selected two useful files: yelp_academic_dataset_business. Json and yelp_academic_dataset_review. Json.

### *Data transportation*

Python software and the Json package are adopted to transform json files into csv files. The yelp_academic_dataset_business.csv file lists 12 kinds of information: Business id, Name, Address, City, State, latitude, longitude, Stars, Reviewcount, Is_open, Attribute, Categories. Among these attributes, Business id, name, Address, City, State, latitude, longitude are meanfuling to this research.

The yelp_academic_dataset_review.csv file lists 21 kinds of information: Review_id, User_id, Business_id, Stars_X, Useful, Funny, Cool, Date, Name, Address, City, State, Latitude, longitude,Stars_y, Review-, count, Is_open, Attribute, Categories, Hours. Among these attributes, Review_id, User_id, Business_id,City, State, Latitude, longitude are useful to this research.

### *Data selection*

From the converted "yelp_academic_dataset_business.csv" and "yelp_academic_dataset_review.csv" files, we use 'pd.merge' function in python3 to combining these two tables by the 'business_id' column. So we can get the restaurants we want and the reviews we want about them(about 20000 reviews information). Using the 'data' column to filter the 'business_reviews' table, we obtain the information from 2016 to 2017 (about 6000 reviews). It was used to 'training data' to verify the accuracy of the model. The data from 2018 to 2019 was used to predict the future link by using the link prediction model.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | review_id | user_id | business_id | stars_x | useful | funny | cool | date | name | address | city | state | latitude | longitude | stars_y | review_co | is_open |
| 415 | 3124542 | gyyw8nEK | W0TobKzk | 0FrYsoVHheQG | 3 | 2 | 1 | 1 | 2019/5/15 20:26 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 416 | 2900838 | aS38d2y8s | xeS2RUkq | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2018/4/1 4:00 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 417 | 3058606 | y8sdRtMV | Y-uZHZivc | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2019/3/20 11:53 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 418 | 2412526 | k39wH30( | CbdHvRQ/ | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2018/2/11 18:06 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 419 | 2647998 | JlByijyY2-C | zRQ4hwXi | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2018/4/8 17:55 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 420 | 3010596 | Kzi0e9NfE | w5xsTd0r | 0FrYsoVHheQG | 4 | 0 | 0 | 0 | 2019/1/21 16:25 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 421 | 2868165 | CYUn1O4l | z1H1BAL-c | 0FrYsoVHheQG | 4 | 0 | 0 | 0 | 2018/8/14 2:21 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 422 | 2996429 | bloGNbPp | HWSSXWJ | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2019/7/1 23:53 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 423 | 2992564 | NbdQypRI | ohYnCSs6 | 0FrYsoVHheQG | 4 | 1 | 0 | 1 | 2019/4/22 15:07 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 424 | 2897388 | 7jV8Lb0is\ | _A_VabDF | 0FrYsoVHheQG | 2 | 0 | 0 | 0 | 2018/9/12 19:52 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 425 | 2531416 | VLeSZgWo | XItANpHK( | 0FrYsoVHheQG | 4 | 0 | 0 | 0 | 2018/8/6 19:11 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 426 | 2900379 | EB1HfxhEI | bCxMWA9 | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2018/10/30 14:18 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 427 | 3072071 | UwD5Zao! | vpkH8W_I | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2019/7/26 17:32 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 428 | 3081610 | AVj1L83T_ | hc2zTm4a | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2019/6/20 22:59 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 429 | 3168969 | hrP2d_39( | linswYQPu | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2019/4/9 17:25 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 430 | 3172577 | 4qe9BtA8( | NkRr4pAJI | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2019/4/10 2:00 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 431 | 2925775 | EGgERH-_ | YVx6wKdD | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2018/11/18 17:05 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 432 | 2464049 | N0YL7H4Ir | wsifeUGVI | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2018/2/12 8:04 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 433 | 2466720 | NXFLNOdT | kSXl89kfP( | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2018/4/17 16:24 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |
| 434 | 3179100 | soYlYWj22 | oDx_sZFc8 | 0FrYsoVHheQG | 5 | 0 | 0 | 0 | 2019/8/16 22:26 | Aroma Ca | 118 N Neil | Champaig | IL | 40.11706 | -88.2432 | 4.5 | 149 | 1 |

Figure 1.reviews information

## Mthodology & Result

After we got the data of reviews in Champion city from 2018 to 2019, the next goal was finding the links between restaurants and recording them in a desired form for the future link prediction. This part mainly contains three sections: create networks based on the reviews, use the data of 2018 as the training set and data of 2019 as testing data to selected method, use the data of 2016-2017 as training set and data of 2018-2019 testing data to verify the selected method again. The following are the specific steps:

### ***Training set and Test set***

In order to verify the effectiveness of the method, we treated the reviews in 2018 as the training set and the reviews in 2019 as the test set. We want to use the restaurant's links in 2018 to predict the links in 2019 and use the real links in 2019 to calculate the prediction accuracy. Thus, the first step is to create the links.

### ***Create network***

Our requirement for this dataset is that it contains at least two columns: X1 and X2, which represent two vertices of each link. Besides, we also recorded the repeated links between the restaurants, which was recorded in column "edges number". For example, if restaurant "A" and restaurant "B" have reviews from the same 2 customers, there will be one row recording "A", "B" and "2". At the same time, we should make sure that there isn't another term "B", "A" and "2" in the data set because the links between restaurants are indirect.

Here are the steps:

- Define a function to calculate the review times of each customer;
- Select the customers who have reviewed more than 1 times (because only these customers can create links between the restaurants);
- Select the reviews made by these particular customers;
- Based on these reviews, figure out the not repeated or omitted business id of all restaurants;
- Define a function whose inputs are two business id of two restaurants and output is the number of the same customers who have done reviews to these two restaurants;
- Use two loops to create links (if the number of restaurants is n, to avoid repeating, we let i equals from 1 to n-1 and j equals from i+1 to n);
- Create the 2018 links and 2019 links and export them to CSV.

### ***Prediction Accuracy***

Our idea is using 3 different methods to generate a recommended links list as "my answer" and using the real links to generate a real list as the "real answer". Then using the apk() function in R to calculate the accuracy. Thus, the first step is creating the real answer of links in 2019.

#### **Real Answer**

Here are the steps:

- Choose the business id which appears both in 2018 links and 2019 links;
- Define a function whose input is the business id of one restaurant and output is a vector which contains all business id of restaurants linking to this restaurant in 2019;
- Use a loop and the function to create the correct answer list for the selected business id in the first step.

**Methods Implement**

Firstly, we want to use random recommended links to get the accuracy. If the accuracy based on the methods is better than the random recommendation, it means that the relationship between hotels is regular, and our project is meaningful. Here are three methods based on three different ways to measure similarity(Golbeck,2013).

- Shortest paths similarity;
- Inverse log-weighted similarity;
- Jaccard similarity.

And here are steps to get recommendation links as “my answer”:

- Assign the 2018 links to training edges;
- Transfer the type of training edges to igraph;
- Use different methods to generate a similarity matrix based on training edges;
- Define a function whose inputs are the business id, similarity matrix, the number of recommendation links, and output is a vector which contains the recommend links;
- Use lapply function to get recommendation links for all business id and make a list;
- Use mapk function to calculate the accuracy.

Here is the result:

| **Table 1. One-year Training and One-year Test Accuracy** | |
|---|---|
| Methods | Accuracy (%) |
| Random recommend | 3.31 |
| Shortest path similarity | 2.21 |
| Inverse log-weighted similarity | 5.15 |
| Jaccard similarity | 7.87 |

Table 1. One-year Training and One-year Test Accuracy

**Analysis**

Although the accuracy of each method is very low, at least the Jaccard similarity method is much better than the random recommendation, which means there is a basis for creating a network graph based on customer reviews. Besides, one reason for low accuracy is because of the short time horizon. Some correct answers just contain one or two elements. Thus, we infer that if we extend the training set and test set time horizon from 1 year to 2 years, the result will be much better. The total links of 2 years are not just the sum of two 1-year links.

**Extend Time Horizon**

We still have the methods we used before, except for the training set and the test set. This time, we use the links in 2016-2017 as training sets and the links in 2018-2019 as test sets. Moreover, here is the result:

| **Table 2. Two-year Training and Two-year Test Accuracy** | |
|---|---|
| Methods | Accuracy (%) |
| Random recommend | 4.70 |
| Shortest path similarity | 6.54 |
| Inverse log-weighted similarity | 15.18 |
| Jaccard similarity | 24.30 |

Table 2. Two-year Training and Two-year Test Accuracy

Due to the better performance of the methods after we extend the time horizon, we decide to predict the 2020 links based on the links in 2018 and 2019.

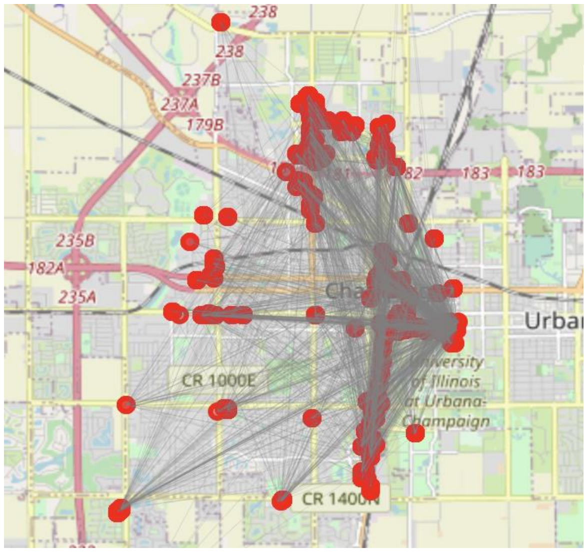
### *Prediction of 2020*

We use the 2018 and 2019 network as the train edges. After comparing the accuracy of each method, we choose the method with the highest accuracy. Therefore, we calculate the jaccard similarity as the prediction feature. Then we use the model above, and we can get the matrix of the 2020 network. For each node, we select ten links and write the outcome into a csv file.

## Discussion of Networks

### *Network of 2018-2019*

We visualize the network using two methods. The one is visualizing on the map, another is using the circular layout.



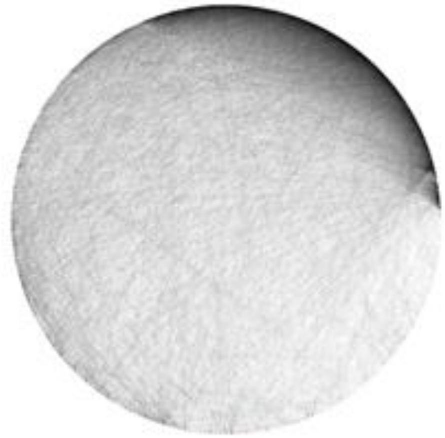



Figure 2.Real map visualization

Figure 3. circular layout

There are 256 nodes and 5822 edges in the original network. It is an unweighted and undirected network. The density which is a measure of how many edges between nodes exist compared to how many edges between nodes are possible of this network is 0.18.

The average degree of the network is 45.48. The histogram shows the degree distribution of nodes. The degree of most nodes is less than 50, while the maximum degree of a node is nearly 175. We can infer that this restaurant is very popular.

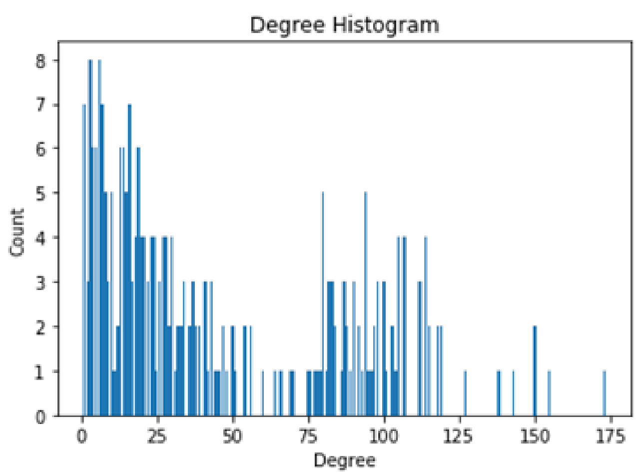



Figure 4. degree histogram

In the research of complex networks, if the nodes in the network can be easily grouped into node sets which are more densely connected internally than with the rest of the network, then the network has a community structure. An important application of community detection is the prediction of missing links. It shows that there are four communities in our network. Different colors of nodes represent different communities. And we roughly think that we can recommend restaurants in the same communities to users.

Figure 5. community detection

***Predicted Network of 2020***

We used the review in 2018 -2019 to predict the recommended system network in 2020. The network structure in 2020 has shown below on the real map.

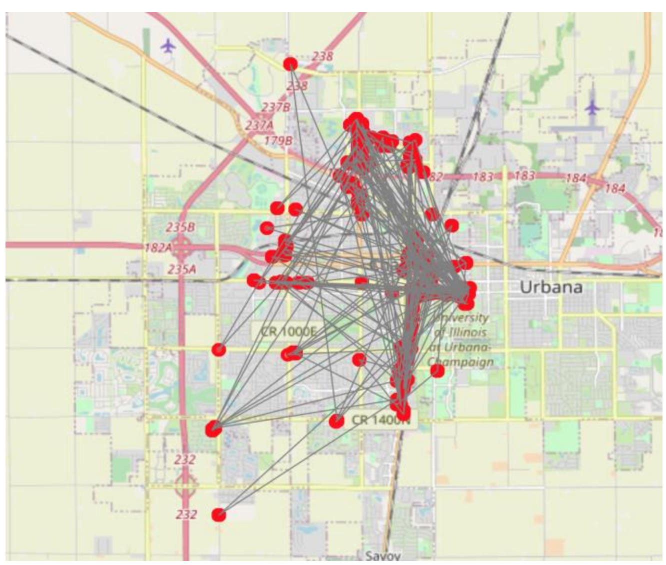



Figure 6. community detection

We can find that there are 256 nodes and 2890 edges. The average degree is 11.733. Moreover, the density of the graph is 0.0397 which is a very normal social network in the real world. Some points have very high degrees and some don't. To analyze the popularity of each restaurant, we compare restaurants with rankings provided by various well-known measures of centrality. Those algorithms are:

**Centrality of Betweenness:**

Betweenness explores the degree to which a node is positioned in a network between others. This is a measure of how much harm there would have been to communication if a given node were removed from the network.

**Closeness centrality:**

Closeness calculates the shortest distance between each person in the network and each other based on the connections between all network members. Here central nodes are the ones that are closest to all other nodes.

**Centrality of the Eigenvector:**

The centrality of each vertex is proportional to the sum of its neighboring centralities.

Although the centrality of the own vector and the centrality of betweenness do not have much significance, and their values are comparatively lower than the centrality of closeness, the rating for different centrality metrics cannot be easily observed. Take the rating of closeness centrality as an example, and we can also find matched restaurants in the top 10 restaurants in the official Yelp. A significant observation we make about the rating of the proximity centrality, the average cost for the users in those restaurants is fairly low. We generate our results below:

| Restaurant | Closeness | Eigenvector | Betweenness |
|---|---|---|---|
| BIGGBY COFFEE | 0.391 | 0.014 | 0.04 |
| Salad Meister | 0.451 | 0.099 | 0.02 |
| Prairie Fire Restaurant | 0.439 | 0.075 | 0.02 |
| Denny's | 0.378 | 0.012 | 0.02 |
| Wingin Out | 0.406 | 0.039 | 0.02 |

Table3. Centrality

Generally, when a restaurant node has a high degree of near centrality in the network, the values of point degree centrality and intermediate centrality are also high.

These restaurants, as the “distance” between the network node and all other points in the network, are very short, and many other points are “close”. Because of this, the point in the transmission of information is more convenient, the more likely to be in the center of the network, become an important node of the network, and therefore become the most priority recommended restaurant.

To make our model more accurate, we consider the influence of three centralities. We normalized the data by using the maximum value in the column to divide each value. After that, we add all the three normalized centralities.

$$General(i) = \frac{closeness(i)}{max\ (closeness)} + \frac{betweenness(i)}{max\ (betweenness)} + \frac{eigenvector(i)}{max\ (eigenvector)}$$

| Restaurant | General | Restaurant | General |
|---|---|---|---|
| That Burger Joint | 8.496 | Minneci's at the Crossing | 6.924 |
| Subway Restaurants | 8.316 | Spoon House Korean Kitchen | 6.842 |
| Red Cape Hot Pot | 7.799 | The Hub Champaign | 6.764 |
| Golden Harbor Authentic Chinese Cuisine | 7.109 | Texas Roadhouse | 6.581 |
| Let's Take a Seat: Thai Cuisine | 6.939 | Sun Singer Wine and Spirits | 6.011 |

Table4. General result

We can compare our results with the ranking of Betweenness centrality. We found that the recommendation of restaurants mainly depends on betweenness centrality. The top five variables are basically the same as those generated by the general variable and even different variables are also in the top ten. So, betweenness centrality holds all the chips.

| Restaurant | General | Restaurant | Betweenness |
|---|---|---|---|
| That Burger Joint | 8.496 | Subway Restaurants | 0.033 |
| Subway Restaurants | 8.316 | That Burger Joint | 0.032 |
| Red Cape Hot Pot | 7.799 | Red Cape Hot Pot | 0.03 |
| Golden Harbor Authentic Chinese Cuisine | 7.109 | Golden Harbor Authentic Chinese Cuisine | 0.028 |
| Let's Take a Seat: Thai Cuisine | 6.939 | Minneci's at the Crossing | 0.027 |

Table5. Comparison

# Conclusion

After visualization and analyzing the '2018-2019 reviews network', we can easily find these existing links between restaurants and their community characteristics in distribution. Using the link prediction model, we predict the neighbors that do not appear to be connected now and find out if there will be a connection between them in the future. For a restaurant, we predicted ten neighbors that might link in the future. In this model, the higher the variable value, the higher the probability that the restaurant is recommended. Finally, we visualize this prediction network and analysis the closeness centrality, betweenness centrality, and eigenvector in this picture.

From a practical perspective, this research topic is of practical significance. We can use predictive models to recommend restaurants on yelp websites that customers might like because these recommended restaurants are in line with their preferences to a certain extent. This can provide excellent convenience for customers. At the same time, merchants can also learn about their potential customers through predictive models. They can publicize restaurants in a targeted manner to attract customers and obtain higher profits. Therefore, this is a research topic that both supply and demand are profitable and can promote the development of data and networking in the catering industry.

## Impact

From a business perspective, customers can quickly find their target products and save time. Merchants will know more about their potential customers and make better advertisements to earn higher profits. For example, in this project, two impacts are below: for customers, they can find the restaurants they like and avoid the restaurants they do not like based on the previous reviews; for restaurants, they can find their potential customers fast and know the customers that they may lose in the future. In society angle, link prediction speeds the match process between demander and supplier up. The economic market will be more efficient, which can improve the efficiency of the whole society.

## limitation

The accuracy of the prediction model we use is 24.3%. This is not high enough, and there is room for further improvement. This may be due to two reasons. One is data. We only used the 2016-2017 data as the training set. Perhaps the model accuracy is not high enough due to the small amount of data. On the other hand, it is the method. Probably the Jaccard similarity algorithm we use is not reasonable enough, and there are algorithms more suitable for this model.

The second limitation is that our model focused on a city in Illinois, so it may be geographically specific and not very reasonable for widespread use.

## Reference

Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. (2006, April). Link prediction using supervised learning. In SDM06: workshop on link analysis, counter-terrorism and security (Vol. 30, pp. 798-805).

Almansoori, W., Gao, S., Jarada, T. N., Elsheikh, A. M., Murshed, A. N., Jida, J., ... & Rokne, J. (2012). Link prediction and classification in social networks and its application in healthcare and systems biology. Network Modeling Analysis in Health Informatics and Bioinformatics, 1(1-2), 27-36.

Fu, C., Zhao, M., Fan, L., Chen, X., Chen, J., Wu, Z., ... & Xuan, Q. (2018). Link weight prediction using supervised learning methods and its application to yelp layered network. IEEE Transactions on Knowledge and Data Engineering, 30(8), 1507-1518.

Finn, L. Link Prediction in the Yelp Social and Review Networks.

Golbeck, J. (2013). *Analyzing the social web*. Newnes.

Huang, J., Rogers, S., & Joo, E. (2014). Improving restaurants by extracting subtopics from yelp reviews. iConference 2014 (Social Media Expo).

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat, 37, 547-579.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. Journal of the American society for information science and technology, 58(7), 1019-1031.

Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp. com. Com (March 15, 2016). Harvard Business School NOM Unit Working Paper, (12-016).

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. Management Science, 62(12), 3412-3427.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013, June). What yelp fake review filter might be doing?. In Seventh international AAAI conference on weblogs and social media.

O'Madadhain, J., Hutchins, J., & Smyth, P. (2005). Prediction and ranking algorithms for event-based network data. ACM SIGKDD explorations newsletter, 7(2), 23-30.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (pp. 285-295).

Sawant, S. (2013, December). Collaborative filtering using weighted bipartite graph projection: a recommendation system for yelp. In Proceedings of the CS224W: Social and information network analysis conference (Vol. 33).

Shang, M. S., Fu, Y., & Chen, D. B. (2008, December). Personal recommendation using weighted bipartite graph projection. In 2008 International Conference on Apperceiving Computing and Intelligence Analysis (pp. 198-202). IEEE.