Network Analysis on Bakeries in PA

Introduction

Yelp is a popular, community driven effort which provides a platform to curate user reviews and ratings for local businesses, e.g. restaurants, salons, auto services, healthcare services etc. Each one of these reviews, tailored to a multitude of businesses serve as an unbiased platform in which the demographic has an equal voice in ranking and comparing local businesses. The quality of a business can be gauged by the average rating of the business, and the number and the connotation of the reviews tend to support the rating.

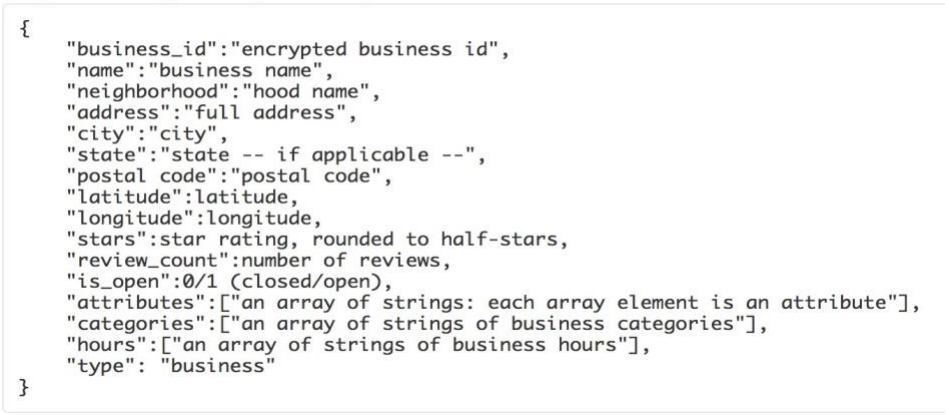
The objective of this project is to explore the bakery network in PA state in order to better recommend bakeries to Yelp users. In this project, the connectivity of bakeries in the network was explored by using the concept of degree and closeness centrality. With information about how certain features of restaurants have impacts on their popularity, the ability to recommend and promote more restaurants to users greatly increases. A close look at what would be the most influential attributes to be considered in recommending bakeries in PA state was also undertaken.

Methods

Data exploration and preparation

The first step for this project is to explore and prepare the data. Datasets from the Yelp dataset challenge were used for this project. There were five datasets with information including business, review, user, check-in, and tip. The files were in json format. Since the objective was to create a bakeries network connected by reviews of same users between bakeries, the analysis only required business and review dataset. Python was used for data exploring. Figure 1 and 2 lists the attribute descriptions for the two datasets used.

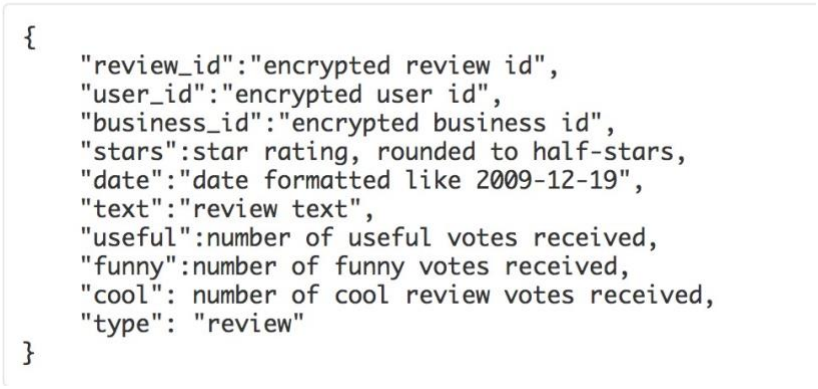
yelp_academic_dataset_business.json



```
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
    "postal code":"postal code",
    "latitude":latitude,
    "longitude":longitude,
    "stars":star rating, rounded to half-stars,
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element is an attribute"],
    "categories":["an array of strings of business categories"],
    "hours":["an array of strings of business hours"],
    "type": "business"
}
```

Figure 1: Attribute description for business dataset

yelp_academic_dataset_review.json



```
{
    "review_id":"encrypted review id",
    "user_id":"encrypted user id",
    "business_id":"encrypted business id",
    "stars":star rating, rounded to half-stars,
    "date":"date formatted like 2009-12-19",
    "text":"review text",
    "useful":number of useful votes received,
    "funny":number of funny votes received,
    "cool": number of cool review votes received,
    "type": "review"
}
```

Figure 2: Attribute description for review dataset

There were 4153150 observations with 10 attributes in the review dataset and 144072 observations with 16 attributes in the business dataset. The following steps were performed during the data understanding and preparation process:

Step 1: Converted files from json format to pandas dataframe. Python's pandas package was used because it is a powerful library in python that can create dataframes for data processing. Figures 3 and 4 shows a part of the business dataset in json and dataframe format.

```
#Preview the json file
yelp_business[:1]
```

```
[{'address': '227 E Baseline Rd, Ste J2',
  'attributes': ['BikeParking: True',
   'BusinessAcceptsBitcoin: False',
   'BusinessAcceptsCreditCards: True',
   "BusinessParking: {'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}",
   'DogsAllowed: False',
   'RestaurantsPriceRange2: 2',
   'WheelchairAccessible: True'],
  'business_id': '0DI8Dt2PJp07XkVvIElIcQ',
  'categories': ['Tobacco Shops', 'Nightlife', 'Vape Shops', 'Shopping'],
  'city': 'Tempe',
  'hours': ['Monday 11:0-21:0',
   'Tuesday 11:0-21:0',
   'Wednesday 11:0-21:0',
   'Thursday 11:0-21:0',
   'Friday 11:0-22:0',
   'Saturday 10:0-22:0',
   'Sunday 11:0-18:0'],
  'is_open': 0,
  'latitude': 33.3782141,
  'longitude': -111.936102,
  'name': 'Innovative Vapors',
  'neighborhood': '',
  'postal_code': '85283',
  'review_count': 17,
  'stars': 4.5,
  'state': 'AZ',
  'type': 'business'}]
```

Figure 3: Business dataset in json format

```
pd_business.head()
```

	address	attributes	business_id	categories	city	hours	is_open	latitude	longitude	na
0	227 E Baseline Rd, Ste J2	[BikeParking: True, BusinessAcceptsBitcoin: Fa...	0DI8Dt2PJp07XkVvIElIcQ	[Tobacco Shops, Nightlife, Vape Shops, Shopping]	Tempe	[Monday 11:0-21:0, Tuesday 11:0-21:0, Wednesda...	0	33.378214	-111.936102	Inn Vap
1	495 S Grand Central Pkwy	[BusinessAcceptsBitcoin: False, BusinessAccept...	LTlCaCGZE14GuaUXUGbamg	[Caterers, Grocery, Food, Event Planning & Ser...	Las Vegas	[Monday 0:0-0:0, Tuesday 0:0-0:0, Wednesday 0:...	1	36.192284	-115.159272	Cu Tas
2	979 Bloor Street W	[Alcohol: none, Ambience: {'romantic': False, ...	EDqCEAGXVGCH4FJXgqtjqg	[Restaurants, Pizza, Chicken Wings, Italian]	Toronto	[Monday 11:0-2:0, Tuesday 11:0-2:0, Wednesday ...	1	43.661054	-79.429089	Piz Piz

Figure 4: Business dataset in pandas dataframe

Step 2: Checked for any missing values. There were 325 records with missing values in the 'categories' column. Since 'categories' was used to target bakeries restaurant, the columns with missing values were dropped. Figure 5 shows the list of null values in the business dataset.

```
address             0
attributes      16910
business_id         0
categories        325
city                0
hours           41608
is_open             0
latitude            0
longitude           0
name                0
neighborhood        0
postal_code         0
review_count        0
stars               0
state               0
type                0
dtype: int64
```

Figure 5: List of null values in business dataset

Step 3: Targeted bakeries by using 'categories' column. There was a total of 2,775 bakeries in

the Yelp business dataset. Figure 6 shows the number of bakeries in each state. It was difficult to find same user reviews on bakeries in different states. Therefore, only PA state was selected for analysis. After filtering records with missing values in attribute column, there were 155 bakeries in PA in the resulting dataset.

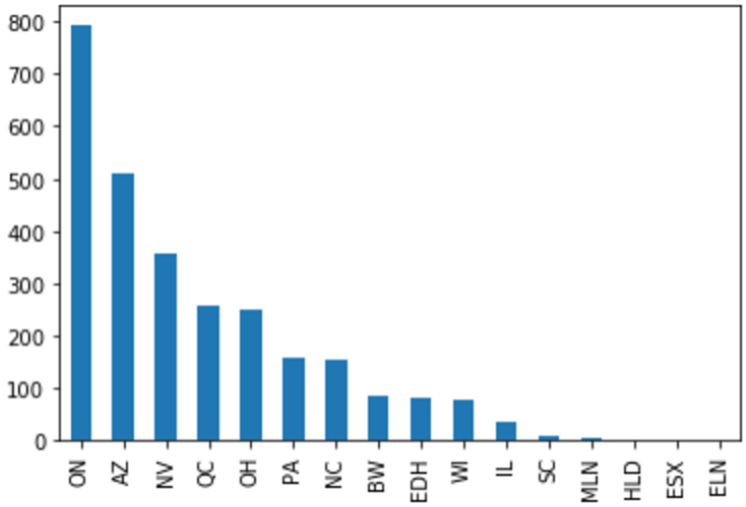


Figure 6: Number of bakeries in each state

Step 4: Extracted information from attributes column. 'Attributes' was a column that provided additional features like bike parking, outdoor seating, and delivery for each bakery. Multiple columns with boolean values were created to add more features for creating the network. Figure 7 shows some of the extracted attributes from the 'Attributes' column.

BusinessAcceptsCreditCards	BikeParking	RestaurantsDelivery	RestaurantsTakeOut	OutdoorSeating	parking_garge	parking_street	p
1	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0

Figure 7: Extracted attributes in business dataset

Step 5: Targeted reviews from bakeries in PA. The business_id from the business dataset was used to target reviews from bakeries in PA. There was a total of 4144 reviews for bakeries in PA.

The star rating in reviews for the bakeries were generally very good. Figure 8 shows the star distribution in reviews.

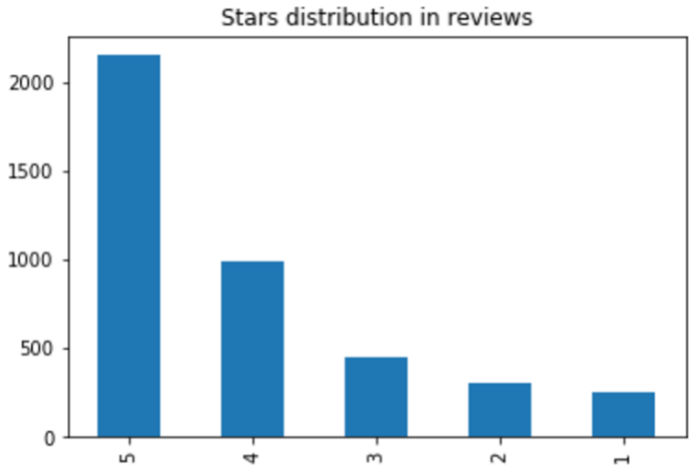


Figure 8: Star distribution in reviews

Graph Construction

After processing the datasets to only include bakeries in PA, the datasets were imported to R to construct the bakeries network. The nodes are the bakeries and the edges are the common reviewers who went to the same bakeries. The edges are undirected because both bakeries are equally related to one another by the common reviewer. The weight of the edges are the number of common reviewers between each bakeries. There were 139 businesses, 238 users, and 1191 reviews in the bakery network.

In order to construct the network, a table that lists all the node connections and the user counts between each node was created. This was done by joining the business and review dataframes on business_id. Then, two of the same dataframes with only business_id, name, and user_id were created. These two dataframes were merged together on user_id. The user counts between the two businesses were aggregated based on the number of common business_id and user_id combination. There were many duplicates in this table, such as business A paired with business A and business A paired with business B is the same as business B paired with business A. All

these duplicates were taken out of the table. A part of the resulting table is shown in Figure 9 and is sorted from the highest to lowest user counts.

name.x	name.y	user_count
La Gourmandine Bakery & Pastry Shop	Gaby Et Jules	30
Prantl's Bakery	La Gourmandine Bakery & Pastry Shop	24
Oakmont Bakery	La Gourmandine Bakery & Pastry Shop	21
La Gourmandine Bakery & Pastry Shop	The Pub Chip Shop	21
La Gourmandine Bakery & Pastry Shop	Jean-Marc Chatellier's French Bakery	18
Prantl's Bakery	Gaby Et Jules	17
The Enrico Biscotti	La Gourmandine Bakery & Pastry Shop	14
Jean-Marc Chatellier's French Bakery	Gaby Et Jules	13
Pink Box Bakery Cafe	Gaby Et Jules	13
Five Points Artisan Bakeshop	La Gourmandine Bakery & Pastry Shop	13
Oakmont Bakery	Prantl's Bakery	13

Figure 9: Sample table describing connected businesses by user count

After this table was created, the graph object was constructed using igraph in R. The node names and edge weights were defined in the graph. The node attributes included in the graph object were city, is open, review count, average stars, business accepts credit card, bike parking, delivery, takeout, outdoor seating, parking garage, parking street, parking validated, parking lot, and parking valet.

Results

Network Overview

The network had a total of 139 nodes and 2252 edges. The network density was determined to be about 23.5%, which means 23.5% of the possible edges between bakeries are present. The average shortest path length was determined to be about 1.8, which means most of the nodes in the network are very close to one another. R was used to create the gexf file and then Gephi was used to explore the bakery network for analysis.

Degree Centrality Analysis

The degree centrality of the bakery network was explored. The degree is the amount of edges a node has in the network. A node with a higher degree is more highly connected in the network. Bakeries with higher degrees have an advantage because bakeries with more edges are more central in the network. High degree bakeries are the most important or most influential bakeries in the network. Figure 10 shows the degree distribution of the network. Generally, as the degree value increases, the number of bakeries decreases.

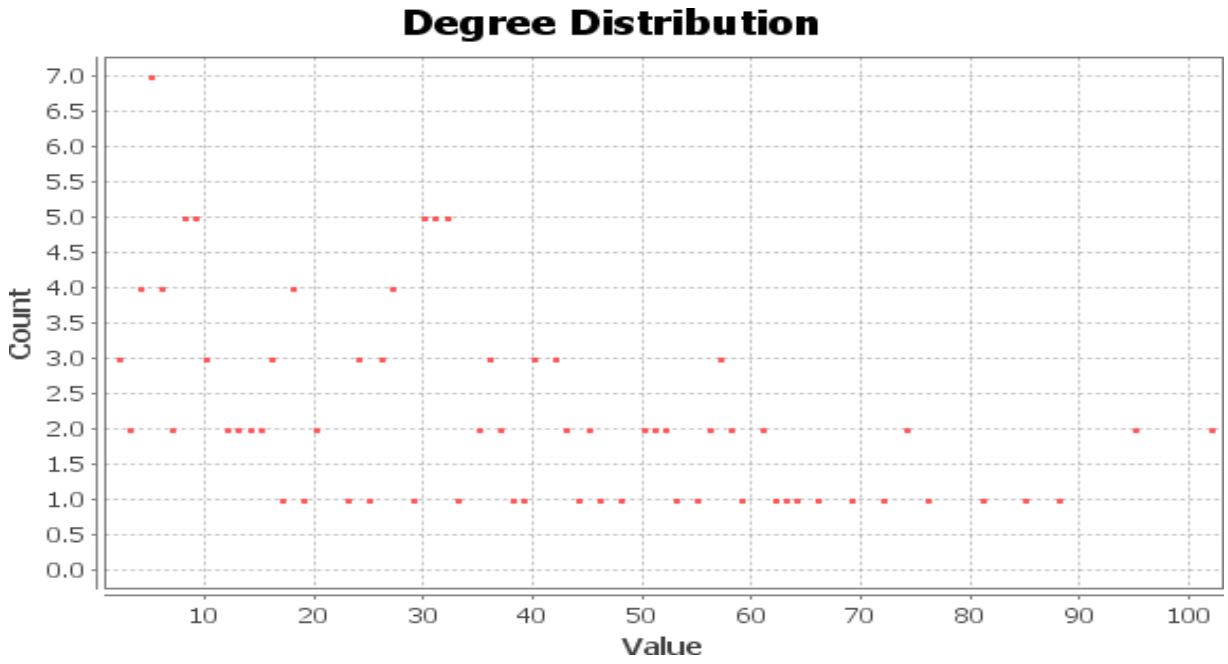


Figure 10: Degree distribution of bakeries

It would be interesting to find which bakeries are the most central in the network. Figure 11 shows the bakeries with the top 8 degree centrality. These are the most influential bakeries in PA. This means that it would be good for bakery businesses to be connected to them in order to help gain more customers.

Figure 11: Bakeries with top 8 degree centrality

Location and Star Rating on Degree Centrality

The effect of the bakery's location in PA were explored to see how location can affect the connectivity of the bakery in the network. There were about 70% in Pittsburgh and 30% in other parts of PA. In the network graph in Figure 12, the green represents bakeries located in Pittsburgh and the pink represents bakeries located elsewhere in PA. The size of the nodes and the text is proportional to the degree of the node, so the higher the degree, the larger the node and text. The layout used was Fruchterman Reingold so nodes of higher degree tend to be in the center. The network graph shows that most of the highly connected bakeries were in Pittsburgh. This is probably because Pittsburgh is more populated and commercialized than the other cities.

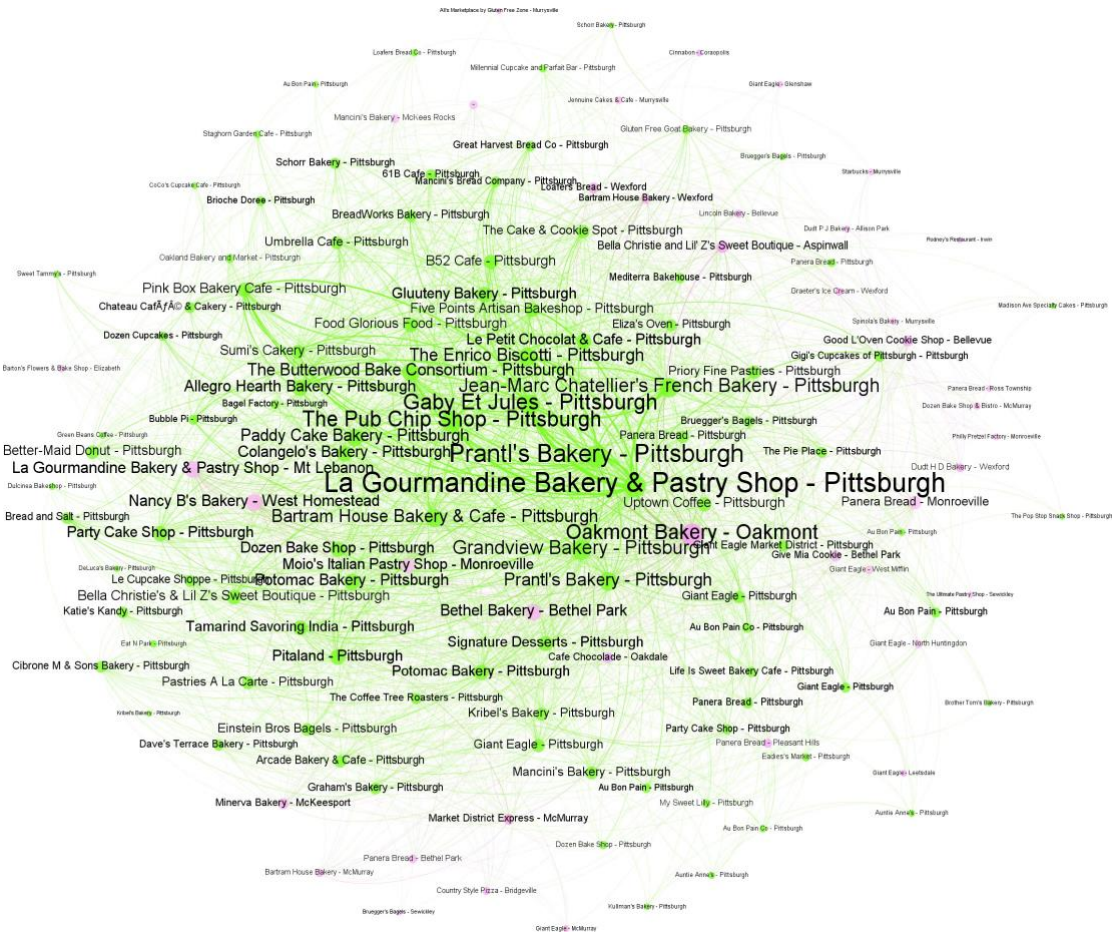


Figure 12: Bakery network based on city in PA

Another question explored was if people tend to go to the highly rated bakeries. Figure 13 shows a network graph where green represents 5 stars, blue represents 4 or 4.5 stars, pink represents 3 or 3.5 stars, and grey represents under 3 stars. Again, the size represented the degree of each node and the layout used was Fruchterman Reingold. Most of the bakeries in the center of the graph had a 4 or 4.5 star rating and most of the bakeries that had a 3 or 3.5 rating were towards the border of the network. This indicates that the star rating may affect the degree centrality of a bakery, though the highest star rating does not mean the bakery will have the highest degree. Therefore, it is important for bakeries to have a good rating to have high degree centrality, but a good rating does not guarantee high degree centrality.

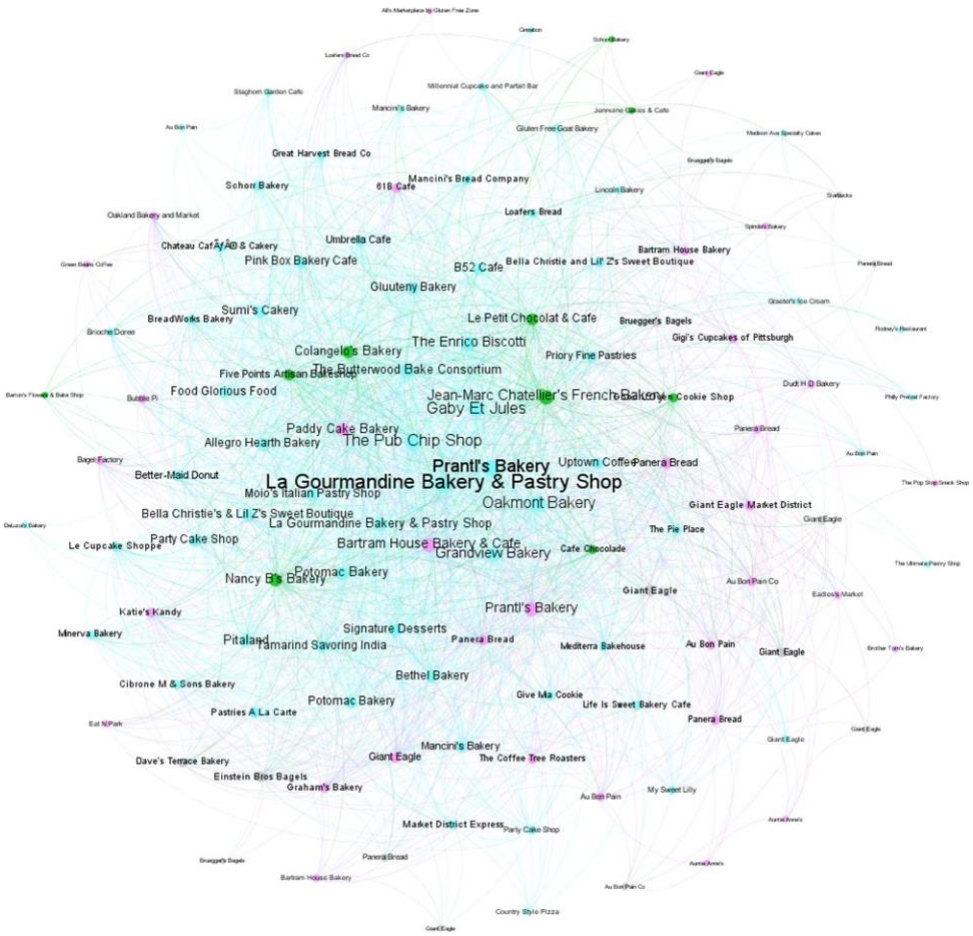


Figure 13: Bakery network based on star ratings

The effect of location and star rating on bakery degree centrality was further explored by filtering bakeries with a degree centrality of 60 and above. Figure 14 shows this network, where color represented star rating and each bakery is labeled with their name and city. The La Gourmandine Bakery and Pastry Shop has locations in Mt Lebanon and Pittsburgh, but the one in Pittsburgh had a higher degree centrality, even though the star rating was in the same range. Most bakeries were located in Pittsburgh and had a 4 or higher rating, which is very good. The three bakeries with a 3 or 3.5 rating were located in Pittsburgh. This suggests that the city and star rating of the bakery important for bakeries to have a high degree centrality.

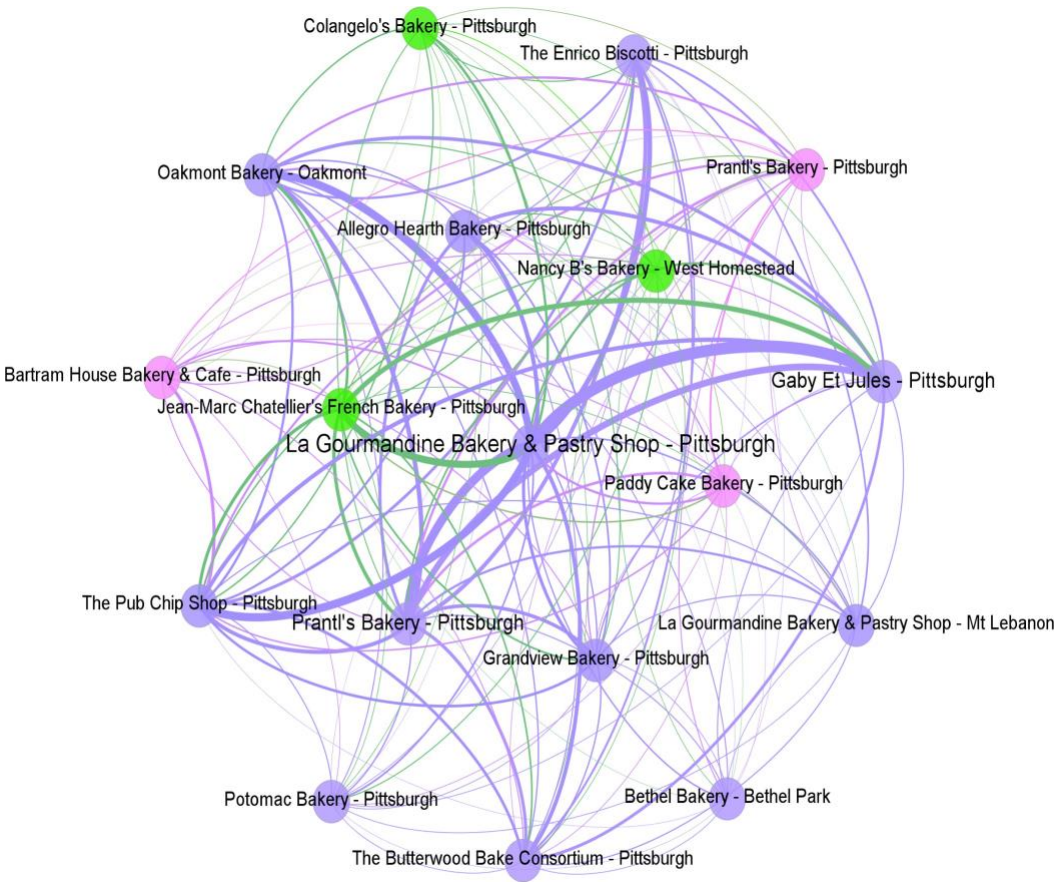


Figure 14: Network based on star ratings of bakeries with high degree centrality

Community Detection

Community is formed by individuals such that those within a group interact with each other more frequently than with those outside the group. Community is also known as group, cluster, cohesive subgroup, module in different contexts. What community detection does is discover groups in a network where the individual's group memberships are not explicitly given. Community detection is key to understanding the structure of complex networks, and ultimately extracting useful information from them. Network interaction provides rich information about the relationship between users and also provides basic information for other tasks such as recommendation, which is the objective in this project.

Modularity is one measure of the structure of networks or graphs. The modularity is, up to a

multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random. It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. The modularity can be either positive or negative, with positive values indicating the possible presence of community structure. Thus, one can search for community structure precisely by looking for the divisions of a network that have positive, and preferably large, values of the modularity. Modularity is often used in optimization methods for detecting community structure in networks. This is the methods used in the project to detect the community.

Gephi was used to divide bakeries into three communities by modularity. When using modularity, randomization and edge weights were used and the resolution was 1. The modularity of the network was determined to be 0.17. The three group community graph is shown in Figure 15.

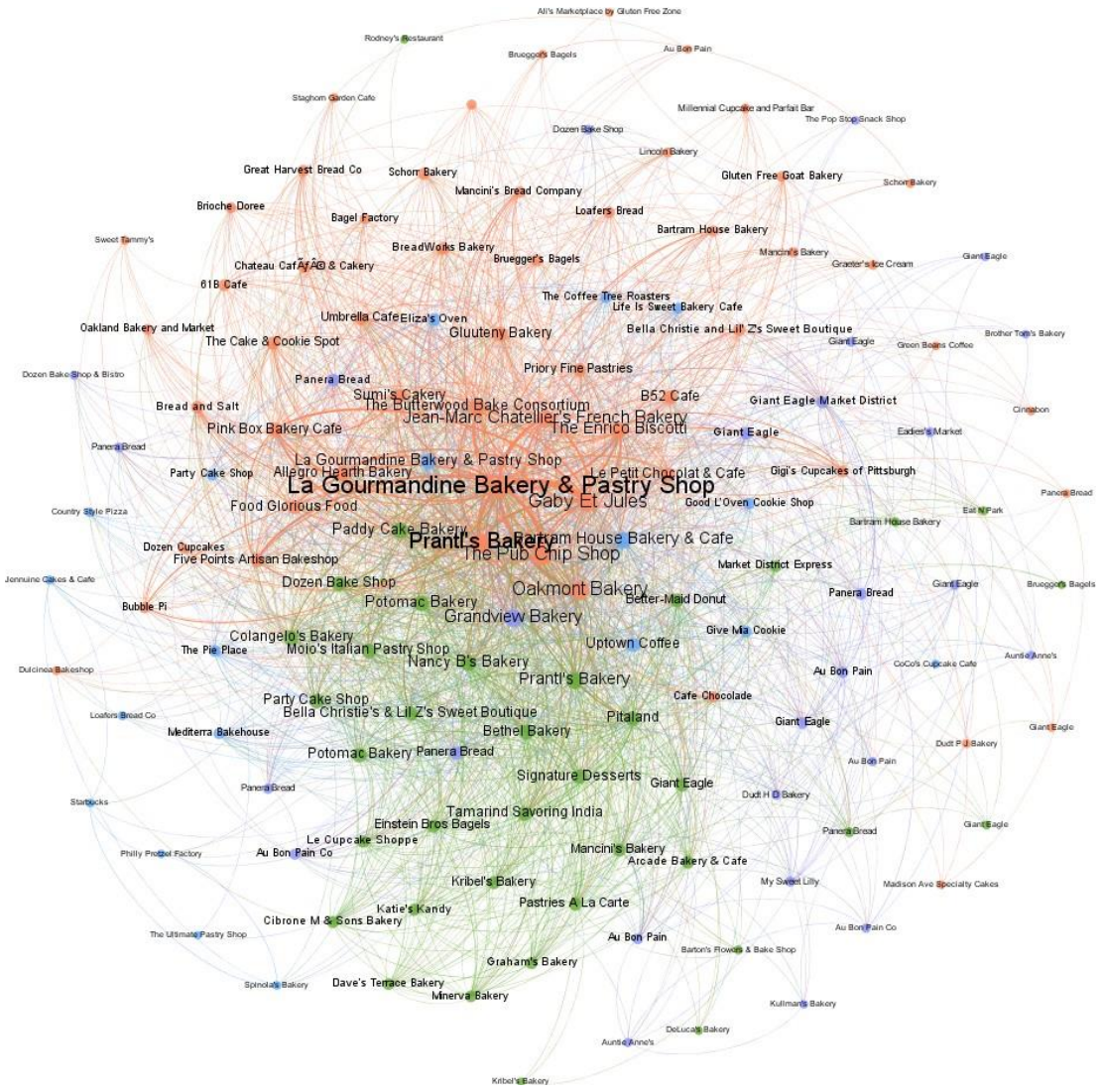


Figure 15: Full community detection using modularity

This community detection result (Figure 15) looks quite comprehensive, however, it is also a bit messy for companies to target. Thus, what was done next was filtering the edges with weights lower than 10. Since the number of common users was the edge weight, the edge weight lower than 10 was unrepresentative for analysis. The communities with high edge weights is shown in Figure 16.

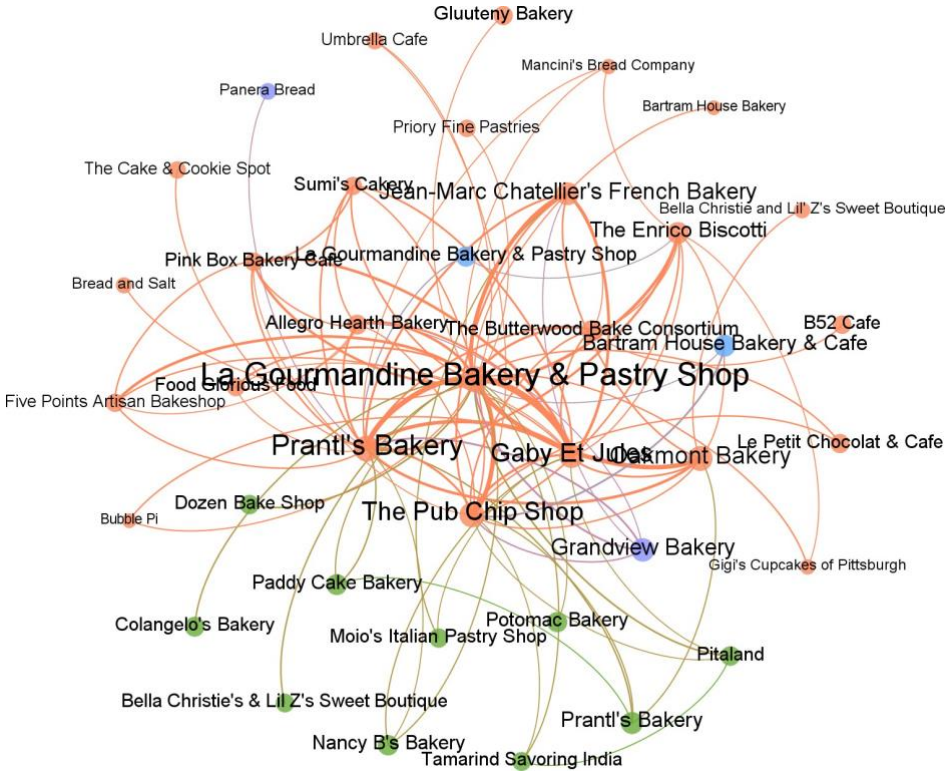


Figure 16: Community detection using modularity with high edge weight

Conclusion

In this project, the effect of location and star rating on bakeries and the highly connected bakeries were in Pittsburgh are examined by modest extrinsic factors surrounding a bakery and several intrinsic factors that lie within a bakery. For future work, a good direction to explore is incorporating star ratings provided by the tip dataset, shorter reviews with summarizing blurbs on why bakeries are rated certain ways. The social network structures can provide valuable insight on how communities influence individual review ratings as exhibited in the Yelp Dataset.