Network Analysis of Yelp Users, Businesses & Reviews

COSC2671: Social Media and Network Analytics

ana

RMIT
UNIVERSITY

CONTENTS

1.Introduction - Yelp

Yelp is a business directory service and crowd-sourced review forum, and a public company of the same name that is headquartered in San Francisco, California. The company develops, hosts and markets the Yelp.com website and the Yelp mobile app, which publish crowd-sourced reviews about businesses. Originally founded to help people find local businesses like dentists, hair stylists and mechanics, it quickly expanded into other business-like restaurants.

Yelp uses automated software to recommend the most helpful and reliable reviews for the Yelp community among the millions we get. The software looks at dozens of different signals, including various measures of quality, reliability, and activity on Yelp. The site has pages devoted to individual locations, such as restaurants or schools, where Yelp users can submit a review of their products or services, using a one to five-star rating system. Businesses can also update contact information, hours and other basic listing information or add special deals.

Yelp had a monthly average of 37 million unique visitors who visited Yelp via the Yelp app and 77 million unique visitors who visited Yelp via mobile web in Q2 2019.

In addition to reviews, Yelp has features to find events, lists and to talk with other Yelpers. Yelp also offers a service to every business owner (or manager) can setup a free account to post photos and message their customers.

In this analysis we try to understand and analyze this networking relationship that Yelpers and generate new business opportunities from this data. The Yelp data is modelled as per below:

There are 6 file types in the Yelp data extracted from https://www.yelp.com/dataset/download

- User (164k) → User details like Name, # Reviews, Friends, # Fans etc.
- Business (192k)→ Business details Name, Categories, Address, Avg Rating etc.
- Reviews (6.7M) → User Id, Business Id, Review Text, Rating
- Tips (1.2M)→ User Id, Business Id and Tips left by user
- Checkin (162k) and Photo (200k) were not used in this analysis

When we do a plot of the spread of business (refer figure 1 below), we see unequal distribution of data and for most of our analysis we pick a business / city that has reasonable size.

Figure 1- State-wise spread of business

When we look at the spread of business reviews by rating, we see the below distribution and most business have more than 3-star rating.

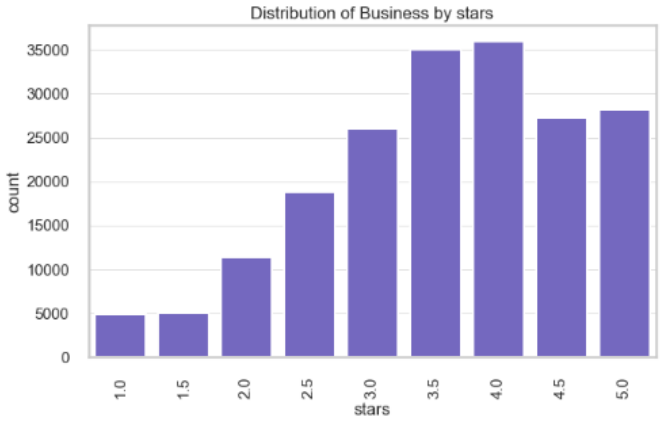


Figure 2- Spread of Business Rating

When we did a distribution of reviews based on stars, we got the following chart:

Figure 3 Distribution of star rating of reviews

3. Location Recommendation for Business

RECOMMEND LOCATIONS TO NEW RESTAURANTS BASED ON THE CUISINE PREFERENCES OF COMMUNITIES OF THE LOCATION

Community structure is a common feature observed in Online Social Networks and the knowledge of the network can be of great business value. It can be used to provide helpful insights into developing more efficient social-aware solutions. In this analysis we aim to detect communities of people in a small geographical location (e.g. a city) based on their interest in different types of cuisines. This will in turn help recommending locations to new restaurants.

We use Yelp data to identify people from a city, create a graph by connecting them based on their friends list and identify communities. After identifying the community, an attempt will be made to identify the restaurant preferences of each community. This would help recommend locations to new restaurants.

3.1. Methodology

In the initial preprocessing stage, the raw data was preprocessed and made into data frames and they were saved to disk as pickle files. The relevant data for this analysis was filtered from these pickle files.

1. The first part of this exercise is fetching users from a city. Since users' location is not available, the users whose majority of reviews are about restaurants of a city are assumed to be residents of that city.
2. After fetching the users, their friends list is collected. These are the first hop friends, implying that they are direct friends of people who reviewed the restaurants.
3. The next task is to identify the second hop friends. These are friends of friends of people who reviewed the restaurants.
4. The next step is to create a graph with nodes as user ID and edges indicating friendship relation. This will be an undirected non-weighted graph. Two nodes will be connected by an edge if the user ids representing the nodes are friends in Yelp.
5. Next step is graph pruning. The nodes with degree less than 2 will be removed from the graph. The pruning assumes that the nodes with degree less than 2 is more likely an inactive user.
6. Run community detection algorithm on the graph.
7. Identify communities and their restaurant preferences.

3.2. Data Pre-Processing

The pre-processing step involves fetching Yelp users from a city and creating a graph with the users and their friends. The Yelp dataset does not have user location. We assume that the users whose majority of reviews of a restaurant are made by residents of that city. The pre-processing stage involves finding a suitable city for analysis and creating a graph of Yelp users of that city and it is explained below.

As explained in methodology, the initial preprocessing was performed on the raw data and the results were saved on disk as pickle files. For the purpose of this analysis, further pre-processing was required. The business data was read from disk and the restaurants were filtered.

There were 59371 restaurants in the filtered dataframe. The figure below shows the restaurants listed in yelp by state.

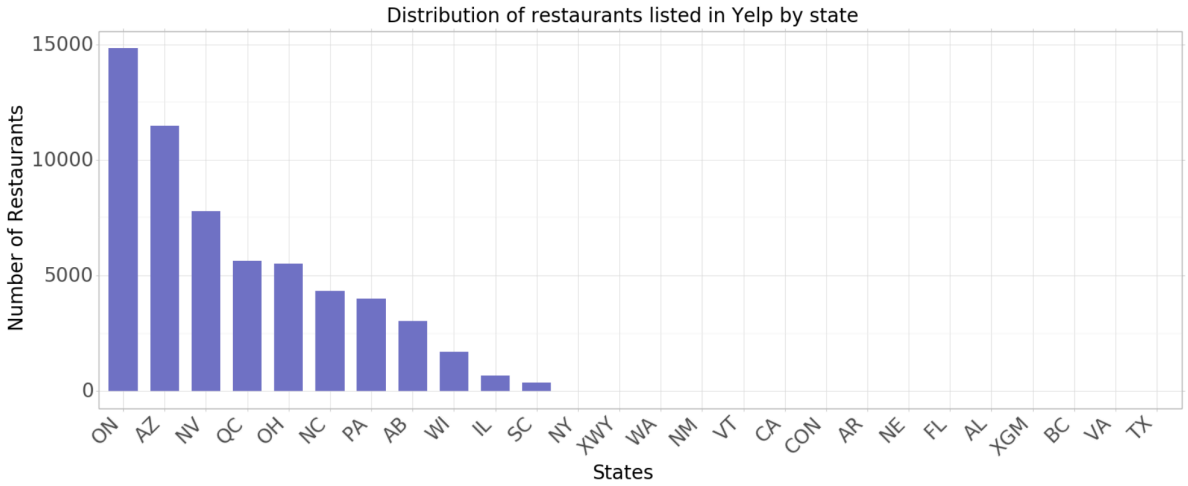


Figure 4: Bar Chart of Restaurants listed in yelp by state.

The data was further filtered on the state Ontario (ON). There were 14831 rows of data. The figure below shows the restaurant count by city listed in yelp.

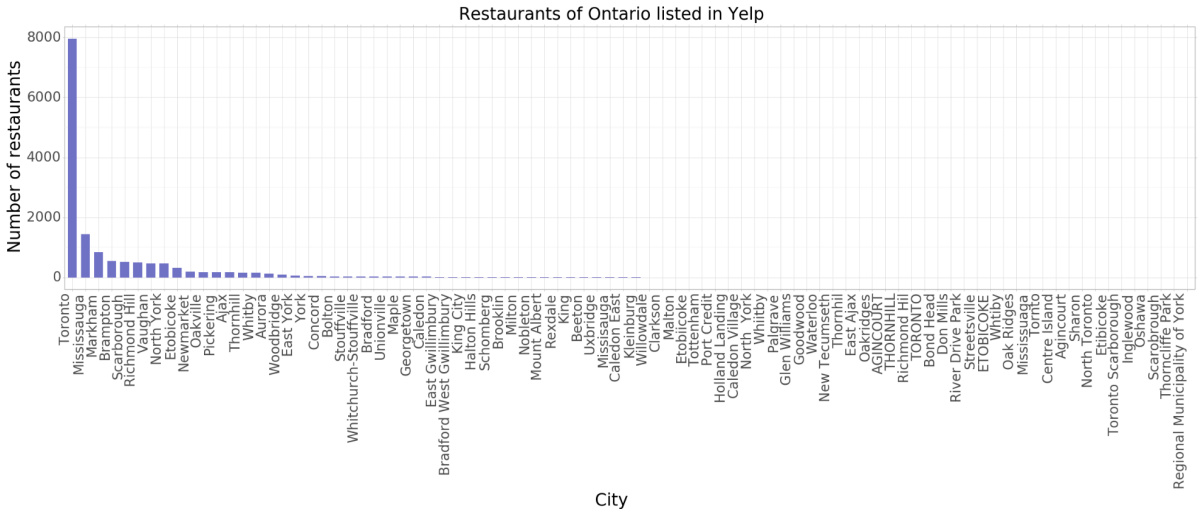


Figure 5: Bar Chart of restaurant count by cities of Ontario listed in yelp.

The data was then filtered on the city Ajax, which had 42 restaurants listed in Yelp. The figure below shows the distribution of review counts of the restaurants located in Ajax. All the restaurants that had more than 20 reviews were retained for this analysis. There were 37 restaurants in this list.

Figure 6: Distribution of review counts for the restaurants located in Ajax.

The figure below shows the restaurant names selected for this analysis. The existence of a variety of restaurants in the city of Ajax is apparent from this word cloud.

Figure 7: Word cloud of Restaurants selected for analysis. The size of the name depends on the number of reviews.

The next step involved identifying the users who reviewed these restaurants. Using reviews data and user data, the users whose majority of reviews are about the restaurants of Ajax city are identified.

A graph is made with the identified users and their friends. The nodes of the graph represent a user and the vertex represents friendship.

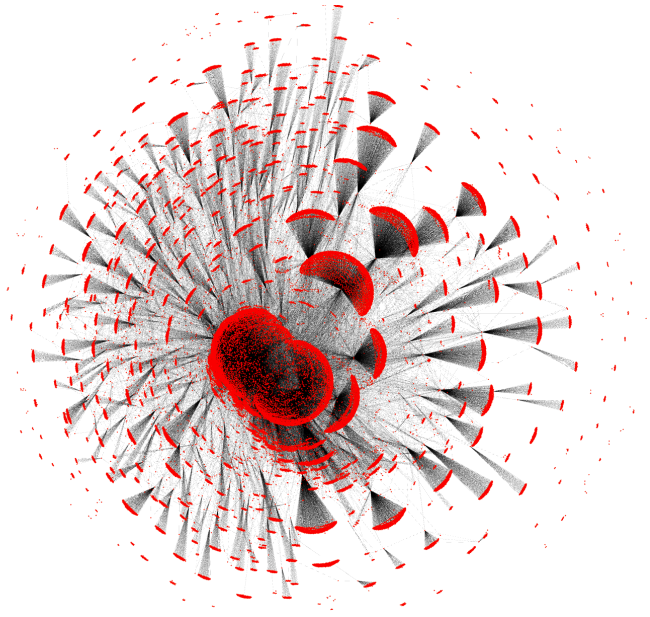


Figure 8: Friend community of Yelp users of Ajax city.

The friends network generated contains 53,961 nodes and 78,188 edges. The Average degree of nodes is ~3.

The degree distribution of the network is shown below. It is evident from the distribution that there is a huge number of nodes with low degree.

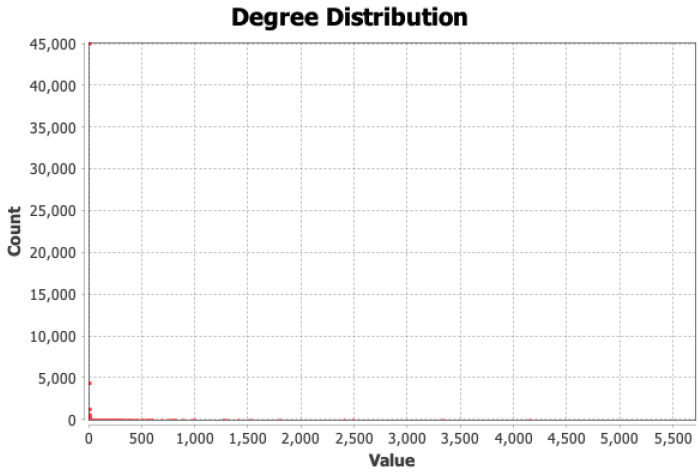


Figure 9: Degree distribution of Friend's network of Yelp users of Ajax city.

The network was pruned with the assumption that the users having low degree is not important. The resultant graph is given below.

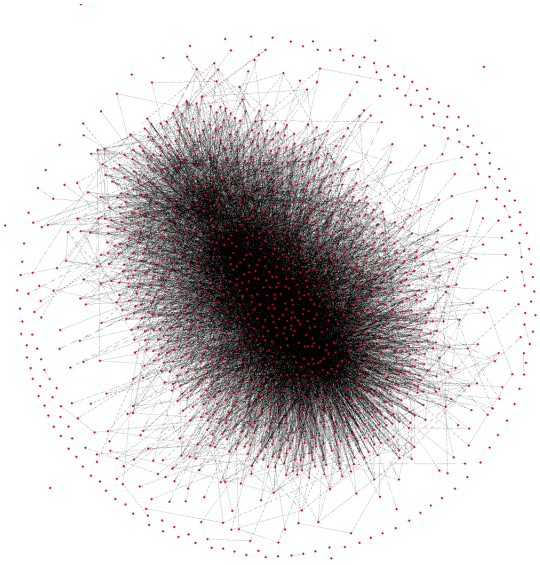


Figure 10: Pruned Friend's network of Yelp users of Ajax city.

The pruned graph has 1,579 nodes and 13,792 edges. The average degree of the new graph is 17.46 which is a well connect network.

The degree distribution of pruned Friend's network of Yelp users of Ajax city is given below.

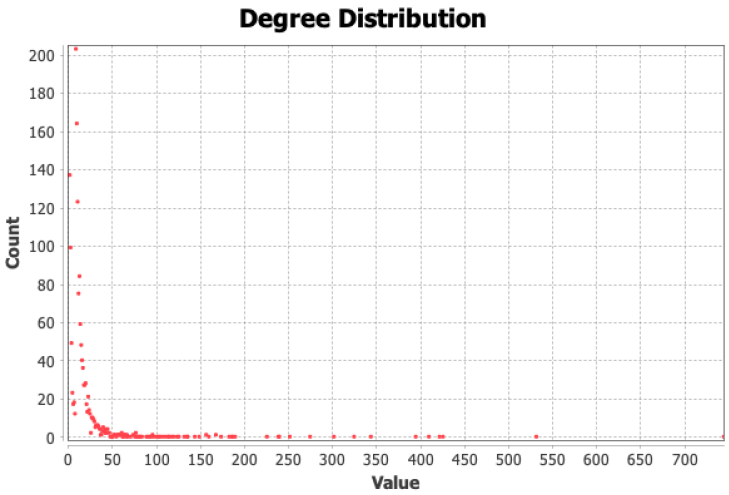


Figure 11: Degree distribution of the pruned Friend's network of Yelp users of Ajax city.

The pruned network was used for community detection. The analysis part will be done in next stage.

3.3. Analysis

Modularity based community detection algorithm was applied on the pruned network and the algorithm identified 160 modularity classes. Given below are the size distribution of the modularity classes identified. The size distribution hints the presence of 4 major communities.

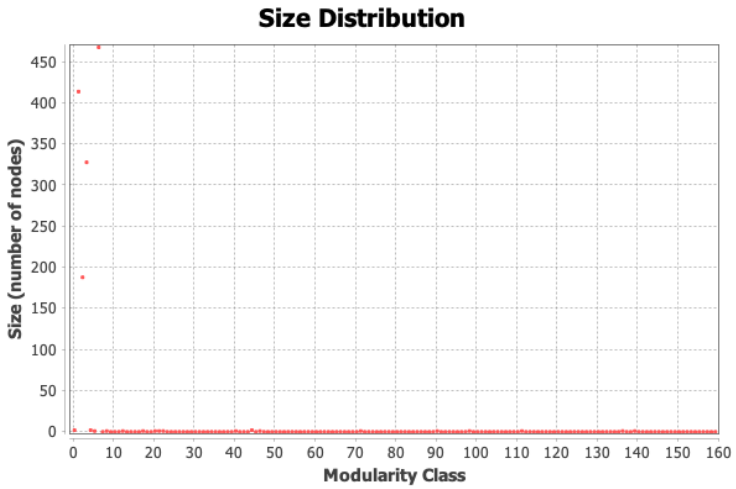


Figure 12: Size distribution of the modularity class identified from pruned Friend's network of Yelp users of Ajax city. Resolution: 1.0.

The identified communities are color coded in figure 13.

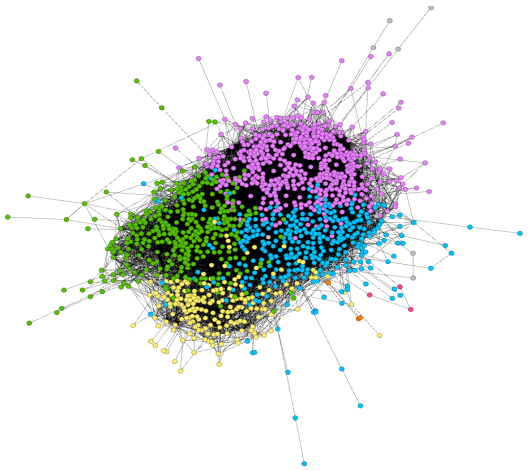


Figure 13: Communities detected from pruned Friend's network of Yelp users of Ajax city.

The four major communities are displayed separately in figure 14

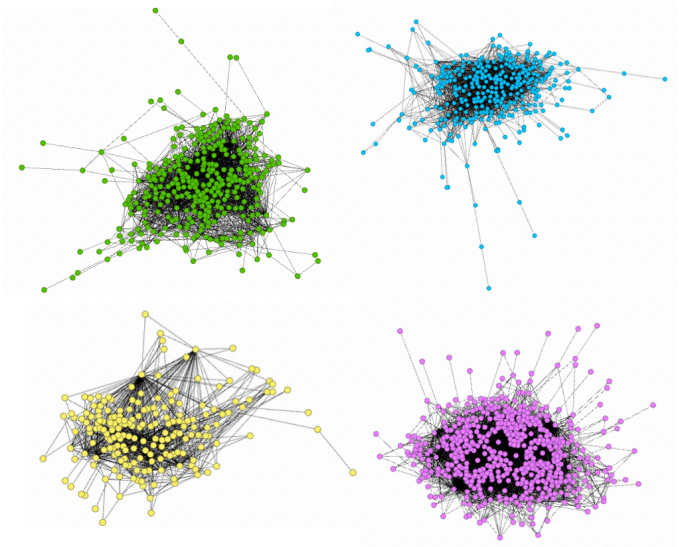


Figure 14: Communities detected from pruned Friend's network of Yelp users of Ajax city.

The Statistics of the major communities are displayed in the table below. The communities are named green, blue yellow and pink.

	GREEN COMMUNITY	BLUE COMMUNITY	YELLOW COMMUNITY	PINK COMMUNITY
NUMBER OF NODES	350	323	194	532
NUMBER OF EDGES	2101	1842	791	3503
AVERAGE DEGREE	12	11	8	13

Table 1: Statistics of major communities detected from Friend's network of Yelp users of Ajax city.

The nodes of these communities represent Yelp user. For each of these users we extract their restaurant preferences from the reviews data. We look at the restaurant preference for communities (refer word cloud in figure 12). It is clear from the word cloud that Pink community prefers Asian cuisine and Blue community prefers fast food. Bars and nightlife are common among all communities and yellow community prefers western cuisine more.

Figure 15: Communities detected from Friend's network of Yelp users of Ajax city.

3.4. Conclusion

A friend's network of Yelp users of city of Ajax, Ontario, Canada was created for the purpose of detecting communities. Modularity based community detection algorithm was applied of the friend's network and 4 major communities were identified. The restaurant preference of the detected communities was identified from the reviews data.

IDENTIFYING MUTUALLY BENIFITING BUSINESSES

As we saw earlier in our pre-processing, Yelp data has details of 192k businesses across 1204 cities and 36 states across US and Canada. But we can see that not all states have equal distribution and not all businesses from all cities are part of this data extract. We know that the Yelp business data has a feature called categories that defines what type of business there are – as an example for the business - Saks Fifth Avenue, there are attributes like - Men's Clothing, Women's Clothing, Shopping, Fashion. Whilst we can group business based on these attributes it does not necessarily mean that a user will visit all the business that have Men's Clothing, Women's Clothing, Shopping, Fashion.

Also, if a user has a taste he will most likely stick to his favorite choice. So, an entrepreneur does not know if a new clothing franchise is warranted. Maybe all users who shop at Saks will like Fast Food or maybe they like Italian. There are significant opportunities to identify such trends using Yelp data.

4.1. Methodology

In this module we are trying to identify business that have more things in common and build a graph that shows relationship between the different business. This will help us identify what business to open in new city based on comparison of other cities from Yelp data.

1. We want to understand the trends by Identify a city for analysis and filtering all business that meets the pre-processing criteria
2. Once we know a definite list of cities, we filter all the reviews left by users of that business. Because we want to build a recommendation type engine, we want users and their review rating of 5 only.
3. From the filtered set of users, we Identify unique users that left reviews for more than 1 business and use this to again filter the unique business that they reviewed.
4. Generate graph using the below methodology:
   a. Each business from the city we had identified becomes a node
   b. The edges between businesses are drawn if multiple users have reviewed both the business with a 5-star rating. For e.g. if a customer gave a review for Fast-food chain Subway and McDonalds we draw an edge between Subway and McDonalds.
   c. The weight of the edges is the number of customers who have reviewed both businesses.
5. From the generated graph we do Community detection which will give us the unique business that are highly rated together.
6. For the largest communities identify the common words by looking at the business categories.

4.2. Data Pre-Processing

For the detailed analysis we have picked a small city – Brampton in Ontario Province. When we look at the distribution of business by their average rating, we can see that the average review is 3.5. Refer figure.

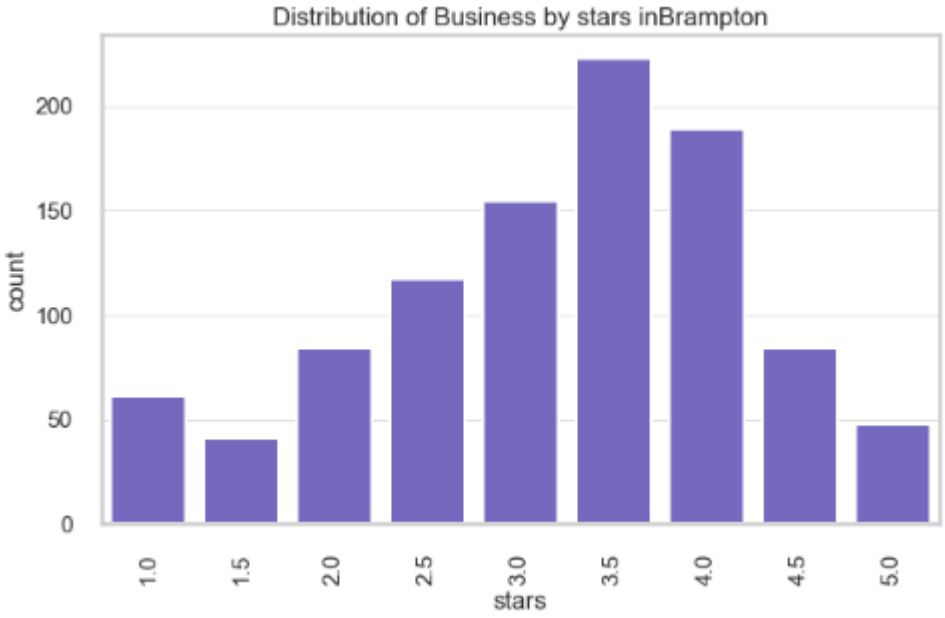


Figure 16: Distribution of Business by stars in Brampton

Next we pick all business that have more than 3.5 stars to avoid negative recommendations. This leaves us with 544 business in Brampton that we want to carry forward in our analysis. A quick word cloud of these business and their names shows some prominent words:

Figure 17- Word Cloud of 'Categories' of business and word cloud of business 'names'

Whilst we see Restaurants prominently in the categories, we cannot infer much since this dataset from Yelp is filtered and may be causing bias. The word cloud of names shows that it's a smattering of restaurants with most common word Brampton since we believe these are business that are local business with the name 'Brampton' in them.

For this exercise we will pick the reviews that are left by users against these businesses. We have 8.1k reviews of which ~3k are 5-star reviews of the 544 Brampton businesses. Refer below figure.

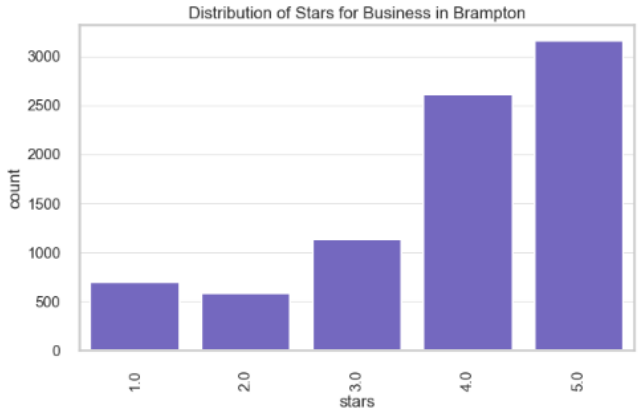


Figure 18: Distribution of Starts for business

We can further keep only reviews that have received certain number of 'cool', 'useful' or 'funny' rating but these are sparsely populated and hence we don't use them.

As a next step we want to identify users who gave these reviewers and a unique count shows that there are 426 unique users who have had at-least 2 reviews and if we filter reviews for these users, we end up with 1.5k reviews.

As a next step we want to create an edge between all these businesses. To create an edge list, we follow this algorithm:

1. Identify each user from our unique user list.
    a. Identify all reviews from this user and get all business that the user reviewed.
    b. For each business in the list.
        - Create an edge to every other business.
        - Avoid drawing self-loops.
    c. Perform this operation for all the business.
2. Perform this for each user.

As an example, say a user U1 reviews 3 business B1, B2, B3, we draw 3 edges between B1-B2, B1-B3 and B2-B3. We avoid drawing circular loops of B1-B1, B2-B2 and B3-B3. We end up with 3869 non-unique edges for this graph. We define weights as the number of times the business are connected. Each time 2 business are reviewed by one user we increase the edge weight by 1.

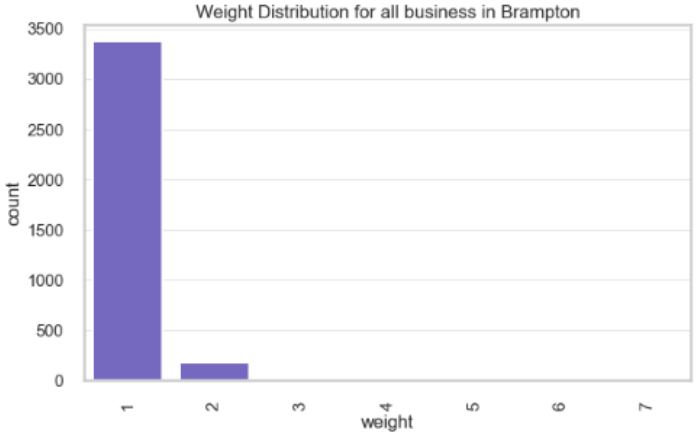


Figure 19: Weight distribution of edges in network generated

4.3. Analysis

We use the Louvain algorithm to identify community in large networks and for visualization we use Gephi. For our analysis we set the Modularity to 0.404 and color code the communities. The resultant network looks like the below:

Figure 20: Distribution of Starts for business

From the characteristics of the graph we can see that

- Average Node degree is 17
- Network Diameter is 6
- Modularity of 0.404 and Resolution of 1.75

When we did the community distribution, we see the below:

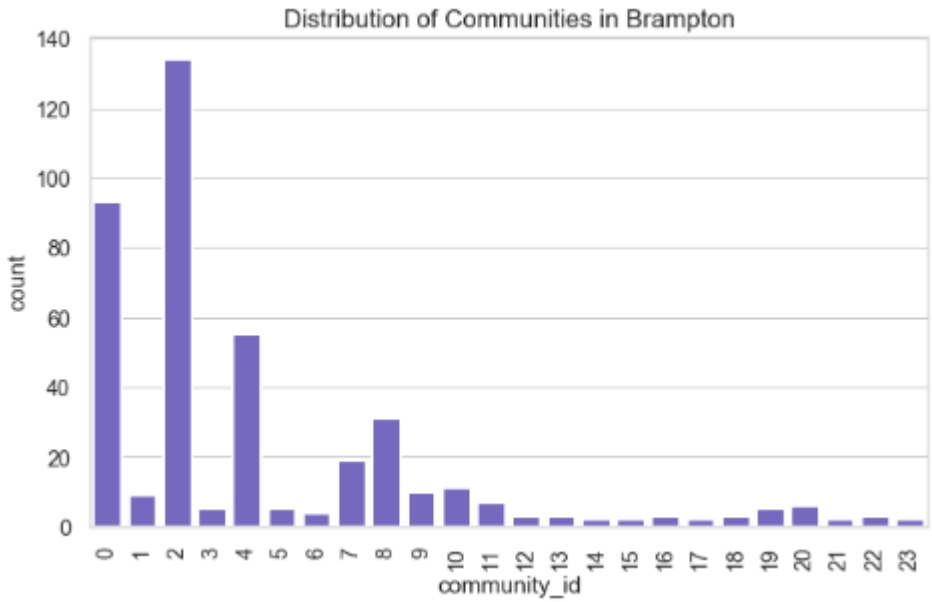


Figure 21: Distribution of Business Communities in Brampton

We see from the above diagram that 80% of the business fall under communities – 0,2,3. These communities are represented as Blue, Orange and Purple respectively in the below analysis.

Figure22 - Communities (0-Blue, 2-Orange,3-Purple)

	BLUE COMMUNITY	ORANGE COMMUNITY	PURPLE COMMUNITY
NUMBER OF NODES	117	122	55
NUMBER OF EDGES	561	834	157
AVERAGE DEGREE	10	14	6

Table 2 Network Analysis of Business in Brampton

When we filter all these businesses and do word cloud of these business, we see that there are distinct set of words between each of them. The below diagram shows the 3 communities and their prominent categories

Figure 23: Distribution of word clouds for business communities

From all this analysis we infer

- If a city is similar to Brampton, opening an Indian Restaurant will more likely get customers from 2 of the largest communities.
- we can see that if a breakfast and Chinese business is famous in a city it is more like to be famous for Caribbean restaurants as well.
- The Purple Community seems to be made of more other business, like gifts and florists etc.

4.4. Conclusion

An example city of Brampton was picked for this analysis and we used Community detection algorithms to identify the most prominent communities. We can expand this analysis and If we had all the cities and all the business, we can compare cities and also identify which cities are similar to Brampton and then recommend business for each city.

RECOMMEND USERS TO BUSINESS

Today in the era of social media, a review online can make huge difference in the business being viewed in positive or negative light. Generally online connections are from reviewers/celebrities/influencers to their followers. Businesses can use the reach/connectivity/centrality of a reviewer/celebrity or an influencer, to promote their business and also be watch closely for any negative reviews by critical reviewer/celebrity or an influencer to control the damage to the business.

Snapchat lost more than $1.3 billion in share values, when Kyle Jenner gave a bad review about its chat platform. Kylie Jenner Tweet: Snapchat 1 billion market loss

Yelp data contains reviews by users and their friends list. The idea is to use the reviewers of a business and create a community and identify users who are central in the community and reach out to them to promote or rectify in case of negative review. Further Topic modelling can highlight the themes of discussion about the business.

5.1. Methodology

The following describes the steps taken to perform the analysis and is formed using **positive reviews > 4 stars** and **negative reviews < 3 stars** separately.

1. Select a business randomly and its reviews.
2. Create a list of reviewers(users).
3. Create a graph using reviewers and friends and friends of friends.
4. Prune the graph and remove users with degree less than 2.
5. Run the centrality algorithm and identify important members of the community.
6. Perform topic modelling to identify recurring themes.

5.2. Data Pre-Processing

We randomly selected a business, **Ichiban Fish House**, which has 92 reviews. The business is a restaurant, which specializes in *sea food and sushi bars*. Each reviewer in the review allocates a star rating for the business between 1-5, with 5 being most positive and 1 being the most negative. Given below is the distribution of the star ratings for the Ichiban Fish House.

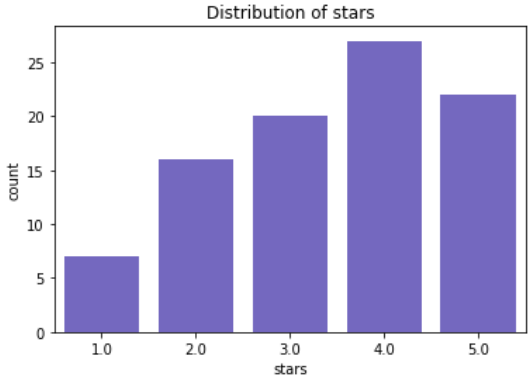


Figure 24- Rating Distribution for Ichiban Fish House

When we do a word cloud of the business categories, we see the following

Figure 25 Word Cloud of Positive and Negative Reviews

For the 2 different types of reviews (positive and Negative) we form a basic network of users. For this purpose, we wanted to look at non-Elite users only and below is the graph representation of each of the users and their friends and friends of friends. The graph based on the positive reviews contains 8777 nodes and 9033 edges. The graph based on the negative reviews contains 3516 nodes and 4440 edges.

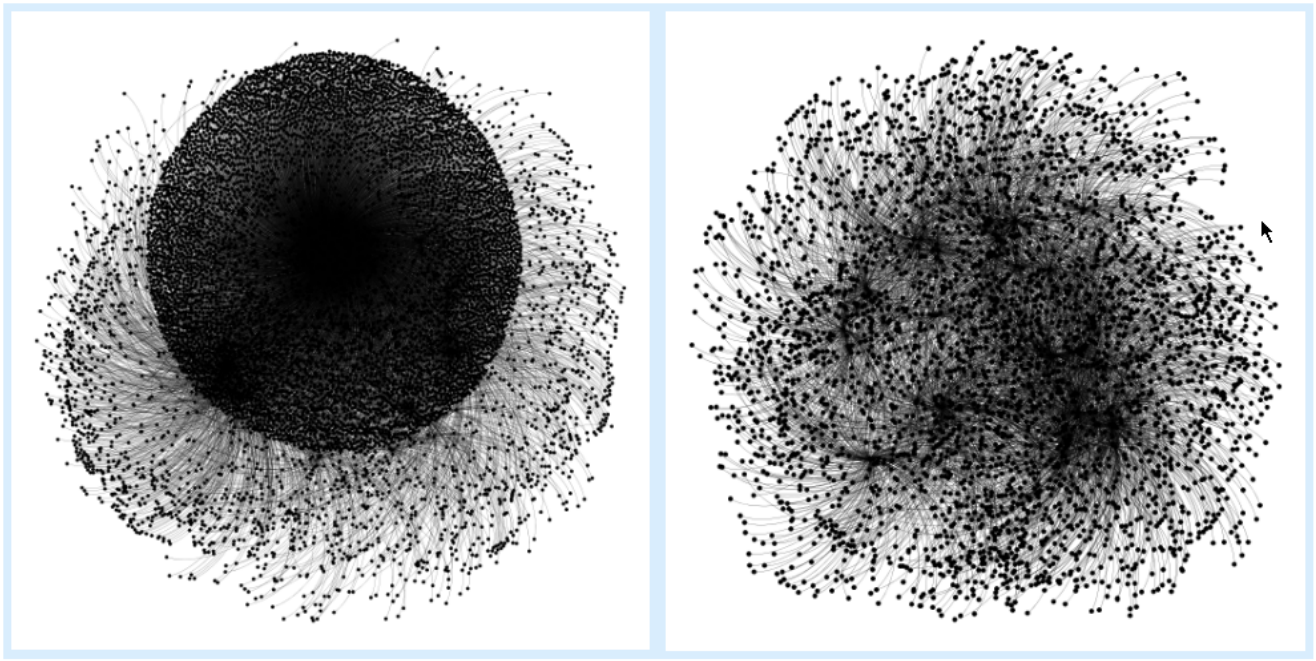


Figure 26 Psotive Reviews (Left) and Negative Reviewers (Right) and their friend network

5.3. Analysis

Centrality identifies the most important vertices in the graph. Centrality comes in different flavors and each flavor or a metric defines importance of a node from a different perspective and further provides relevant analytical information about the graph and its nodes.

For graph/network based on positive reviews, the central users can be used as influencers or promotors. New campaigns launched by the business can be marketed using these central users at fraction of the cost when compared with traditional means. The endorsement by influencers have positive impact to the image of the business.

Similarly, in a graph based on negative reviews, identifying central figures will help reduce the damage to the brand as the business can take corrective measures on the negative feedback immediately. It will be a great win for the business if a central figure from this network can be converted over to positive side. It can improve brand value and business will get new customers who would have otherwise would not have used the services based on negative reviews.

In our analysis we use **Eigen Vector Centrality** and **Betweenness Centrality**.

Eigen Vector Centrality measures the importance of a node in a graph as a function of the importance of its neighbors. If a node is connected to highly important node, it will have a higher Eigen Vector Centrality score as compared to a node which is connected to lesser important nodes. This helps us identify the users whose opinion will be important as to improve brand image.

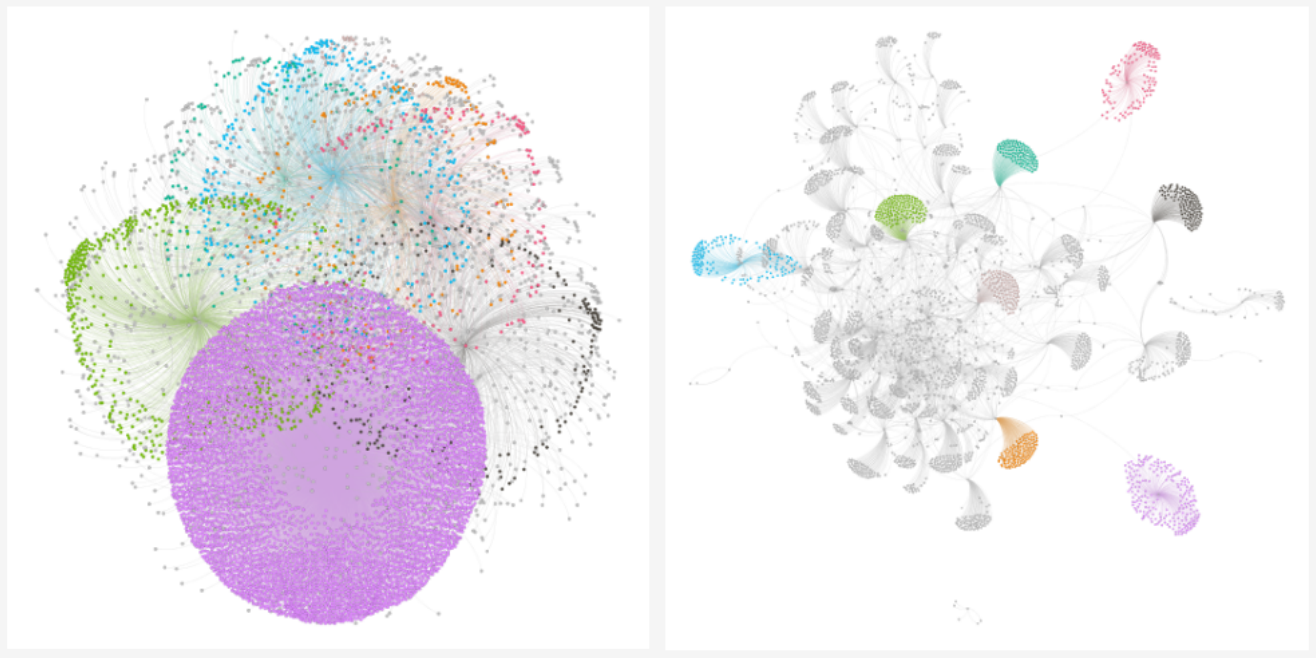


Figure 27 Eigenvector Centrality Graphs created based on positive(left) and negative(right) reviews.

Betweenness Centrality defines and measures the importance of a node in a network based upon how many times it occurs in the shortest path between all pairs of nodes in a graph. This helps identify the reviewers/users who can be used as influencers or should be looked at in case of negative reviews. These users are be able to communicate information quickly across the network. In the below diagram all the central users have colored vertices and all non-significant users have been greyed out.

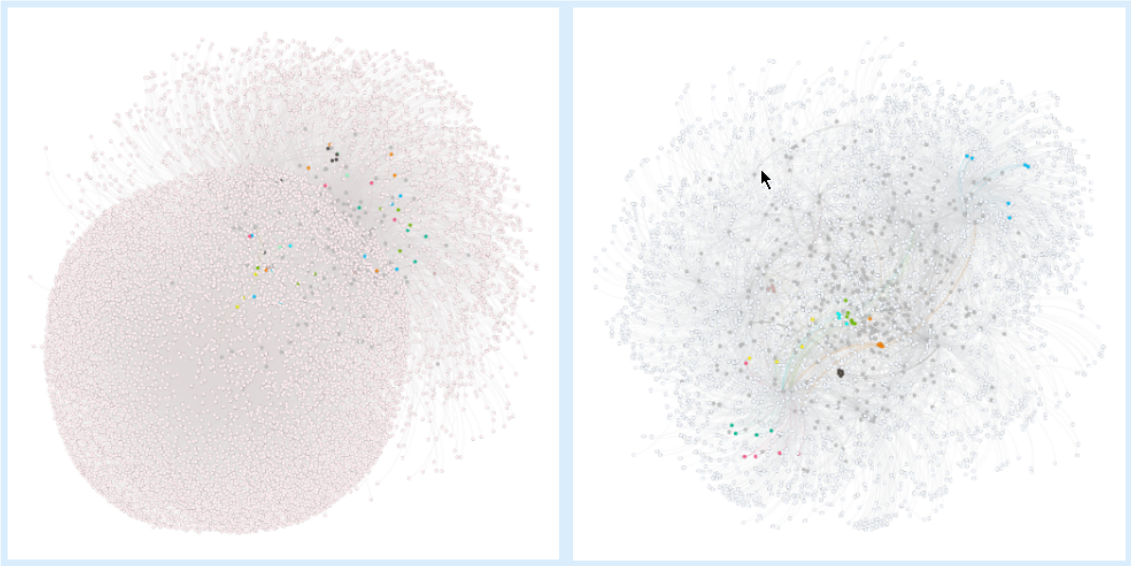


Figure 28 Betweenness Centrality Graphs created based on positive(left) and negative(right) reviews.

As the next step we want to do Topic Modelling on the reviews to identify the recurring themes for the business using topic modelling. Below given are the top ten topics being discussed in positive reviews.

1. place food fresh price japanese super delicious really reasonable lobster
2. favourite love thing quaint expensive orders platters selections leave reasonable
3. sashimi pretty japanese place quality order people orders little come
4. good fish place sushi fresh salmon small actually quick pretty
5. sushi place fresh roll sashimi time ichiban great service food
6. fish sashimi time like great order fresh sushi good point
7. people orders pretty photo sashimi really japanese delicious raw order
8. good food quality japanese rolls sushi price fish ichiban place
9. sushi restaurants eat selections baked scallop average raw price rolls
10. like place dishes boxes try best say sushi menu markham

Below given are the top ten topics being discussed in positive reviews.

1. literally chinese fishy owner dive photo better green mouth flavour
2. literally going okay chinese grade come closed dessert menu friends
3. dive going bento bad lot experience box excuse getting photos
4. dessert close definitely okay high menus large area better going
5. deep okay come menus menu plate coming like closed decent
6. come album overall pretty literally grade going japanese dive experience
7. come ordered closed green chinese owner fishy food high maki
8. going come close dessert coming ended chinese price boat japanese
9. okay korean asked lunch closed bad chinese green going high
10. bad average coming know okay close bento literally chinese owner

5.4. Conclusion

This analysis clearly shows the power of social media and how the social media platforms can be used for marketing and brand management at fraction of the cost when compared to traditional media. It also shows how the communities and businesses can interact and follow each other on the social network. The analysis gives the tools and if used wisely, it can benefit both businesses and users.

6. Project Conclusion

The project involved analyzing data from the Yelp business website released as part of the Round 13 dataset challenge that ran from 15 Jan 2019 to 31 Dec 2019. We did 3 different pieces of analysis–

1. Recommend locations to new restaurants based on the communities
2. Identifying mutually benefiting businesses
3. Recommend users to business

These results were then analyzed using Gephi and various summary charts were drawn giving us actionable insights. We also did analysis of elite users, influential reviewers, page rank algorithms for identifying users etc. However, they have not been presented here as they need a little more work.

7. Limitations

- We are constrained by limitations of Yelp dataset – it does not contain all cities and all reviews as well. It also does not have all the user details so building a network of users is limited by the data given.

- We do not have a mechanism to validate these findings as they would need A-B testing of various scenarios.

- We were limited by computing power and limitations of Gephi software for visualization and hence we had to pick smaller cities for more visual reasons. As a next step we would want to use Graph databases (Neo4j) and other powerful software for visualization.

8. References

- Needham, M. and E. Hodler, A. (2019). Graph Algorithms: Practical Examples in Apache Spark and Neo4j. 1st ed.

- Cochrane, E., Ali, N., Yu, Y. and Zeng, Z. (2019). Yelp-Data-Analysis. [online] GitHub. Available at: https://github.com/NazifALI/Yelp-Data-Analysis/blob/master/YelpDatasetAnalysis.ipynb [Accessed 11 Oct. 2019].

- Yelp.com. (2019). Yelp Dataset. [online] Available at: https://www.yelp.com/dataset/challenge [Accessed 11 Oct. 2019].

- En.wikipedia.org. (2019). Yelp. [online] Available at: https://en.wikipedia.org/wiki/Yelp [Accessed 11 Oct. 2019].

- Zafarani, R., Ali Abbasi, M. and Liu, H. (2019). Social Media Mining: An Introduction. 1st ed.

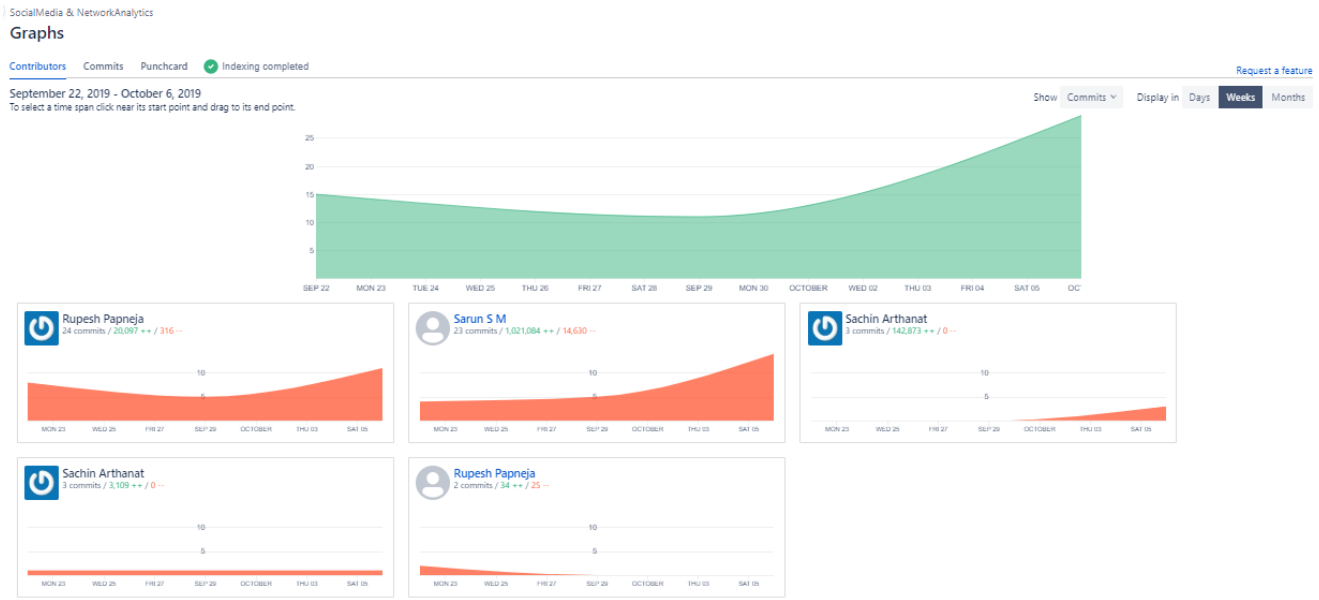
- Aksakalli, C. (2019). Network Centrality Measures and Their Visualization. [online] Aksakalli.github.io. Available at: https://aksakalli.github.io/2017/07/17/network-centrality-measures-and-their-visualization.html [Accessed 11 Oct. 2019].

9. Team Contribution

Our team consisting of Sarun, Sachin and Rupesh started working on the assignment around 16th September 2019. We started searching for open source data that may be useful for the assignment and forming ideas around the datasets. This involved looking the data and preparing basic network graphs and analyzing the features present in the dataset. We prepared the list of top ideas and voted on them as a team. Each of us gave our preference order and based on the result we decided to go with yelp datasets.

For the yelp dataset we prepared the possible list of analysis which may preformed. We noted down all the ideas and based on the input from the Professor and Tutor we further kept refining the ideas. We setup a git repository in bitbucket on 22nd September. We created individual branches for each team member and maintained golden copy of the code in master branch.

Each team member owned an idea and did analysis and collaborated to complete the analysis. We provided feedback to each other and kept refining the output. We used Google hangouts for regular communication and discussions. Given below is the graph generated from bitbucket showing git commits and contributions from each team member.



10. Code Execution Guidelines

1. The code is based on yelp dataset. Please download the data from below link. https://www.yelp.com/dataset/download
2. Download/clone the code from git master branch and go the code folder.
3. The data is in json format. We constructed data frames and created pickles files. Please execute the below notebooks in the given order to create pickle file.

    a. 02_1_tips_users_pickle.ipynb
    b. 02_2_reviews_pickle.ipynb
    c. 02_3_business_pickle.ipynb

4. Place the pickle files under data/pickles folder.
5. Run the individual analysis using the below python notebooks.

    a. 03_00_location_recommendation_business.ipynb
    b. 04_00_business_communities.ipynb
    c. 05_00_important_network_users.ipynb

6. Below is the list of graph files used for analysis in gephi.

    a. 03_1_bluecommunity.net
    b. 03_2_greencommunity.net
    c. 03_3_pinkcommunity.net
    d. 03_4_yellowcommunity.net
    e. 03_5_firsthopfriends_Ajax.gefx
    f. 04_1_business_community_graph.gexf
    g. 05_1_positive_reviews_users_graph.gexf
    h. 05_2_negative_reviews_users_graph.gexf