Project Title

Finding Influencers for a Business using Social Network Analysis

Problem and Motivation:

Traditionally, companies have employed direct marketing (where the decision to market to a particular individual is based solely on his/her characteristics) or mass marketing (where individuals are targeted based on the population segment to which they belong). These approaches, however, neglect the influence that customers can have on the purchasing decisions of others. For example, consider a person who decides to see a particular movie and persuades a group of friends to see the same film.

We can choose to spend more money marketing to an individual if a person has many social connections. Thus, by considering these interactions between customers, we may obtain higher profits than traditional marketing, which ignores such interactions. This way we can optimize the positive word-of-mouth effect among customers.

The specific problem that we are trying to solve is:

To find how much influence does social circle of a consumer have on his/her choices and ratings for a particular type of business. This influence factor will be later used to determine the most influential customers for that business which can then be used for Viral Marketing.

Data:

The data that we have used for this project is the Yelp data set [1] for academia purposes. It consists of:

4.1M reviews and 947K tips by 1M users for 144K businesses

1.1M business attributes, e.g., hours, parking availability, ambience.

Aggregated check-ins over time for each of the 125K businesses

All this data was in the JSON format in three different files. But, for us to answer the problem we outlined we needed attributes from these different files and co-relate them accordingly. This is where we choose to pick the needed attributes from each JSON file and put them into a CSV file. After this data conversion from one format to another we applied rules such as consecutive rule (that says that there can be no missing values between the lowest and highest values for the attribute, and that all values must also be unique) and null rule (that specifies the use of blanks, question marks, or other strings that may indicate the null condition when a value for a given attribute is not available, and how such values should be handled.) as part of data cleaning.

Key Idea:

We proposed a model that optimizes the amount of marketing money spent on each customer (find the most influential customers). The model considers the following factors that influence a customer's network value. First, the customer should have high connectivity in the network and give the product a good rating. If a highly-connected customer gives a negative review, his/her network value will be lower, in which case, marketing to him/her is not recommended. Second, the customer should have more influence on others than they have on him/her. Third, the recursive nature of this word-of-mouth type of influence should be considered. A customer may influence acquaintances, who in turn, may like the product and influence other people, and so on, until the whole network is reached. The model helps us in incorporating an important consideration: We may pay to lose money on some customers if they are influential enough in a positive way. For example, giving a product for free to a well-selected customer may pay off many times in sales to other customers. This is a big twist from traditional direct marketing, which will only offer a customer a discount if the expected profits from the customer alone exceed the cost of the offer.

Implementation:

To generate the proposed model, we followed the below three approaches:

TEJA KOMMINENI (u1072593)

I have converted the data collected in the CSV format to a graph. I started building a graph for a business. The nodes in the graph represent users who visited this business. Each user is associated with attributes namely, the rating he has given for that business and the date on which he gave the rating. Edges are drawn between two nodes if the following two conditions are satisfied: If the users(A,B) are friends to each other. If the user A has given rating after B has given rating, then we draw an edge directed from A to B and the weight of the edge is the reciprocal of similarity between the ratings given by A and B. This weight on the edges takes care of the first two factors that influence the customer's network value as mentioned in key idea.

Once we have generated this graph I ran the page rank algorithm on this graph to see the relevance of each node in the graph. This shows us how much influence a node has in the graph. Thus, we obtain for each user his influence factor (network value) in the network. The Page Rank algorithm accounts for the third factor that is the recursive nature of the influence factor. The output of page rank when depicted shows the top influencers in the network for a business. Marketing to these customers will highly decrease the costs incurred compared to traditional establishment. Below Fig[1] is the output of my implementation. The graph represents all those customers of Wendy's business. The sizes of the nodes in the graph are proportional to their influence in the network. I have labeled the top 10 influencers for the specific business (Wendy's).

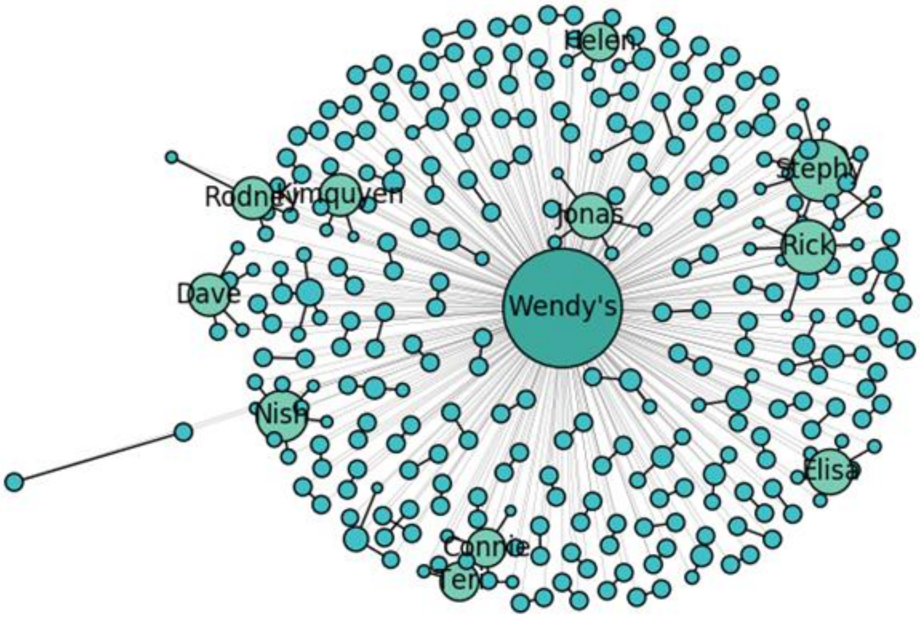

Fig1: Influencers for Wendy's by Page rank

RAM KASHYAP S (u1082810)

I have created an undirected graph by creating weighted edges between two users based on the Jaccard similarity. The intuition behind considering Jaccard similarity is to find the similarity between two users by considering their social connections instead of looking at their user ratings. By considering Jaccard similarity between users, I'm trying to explore friends of friend's approach in a social network.

Given two users x and y, let $\mathrm{F_x}$ be the set of all nodes that have an edge to x and $\mathrm{F_y}$ be the set of all nodes that have an edge to y. Then, Jaccard similarity is given by: $\frac{|\mathrm{F_x} \cap \mathrm{F_y}|}{|\mathrm{F_x} \cup \mathrm{F_y}|}$

While analyzing the graph, I have observed that the connections of the network are very sparse. Approximately 50% of the users don't have friends. I have removed the nodes which don't have more than 2 edges which helped in getting the results faster.

After creating a graph, to find which users are influencing a business more, I ran page rank algorithm and got top influencers in the network for a particular business. By using this approach, for any given business, we can find the top influencers.

I have created a representation of the results by creating a dummy central node for business which has edges to all its customers. The customers with higher influence have node size proportional to their page rank values.

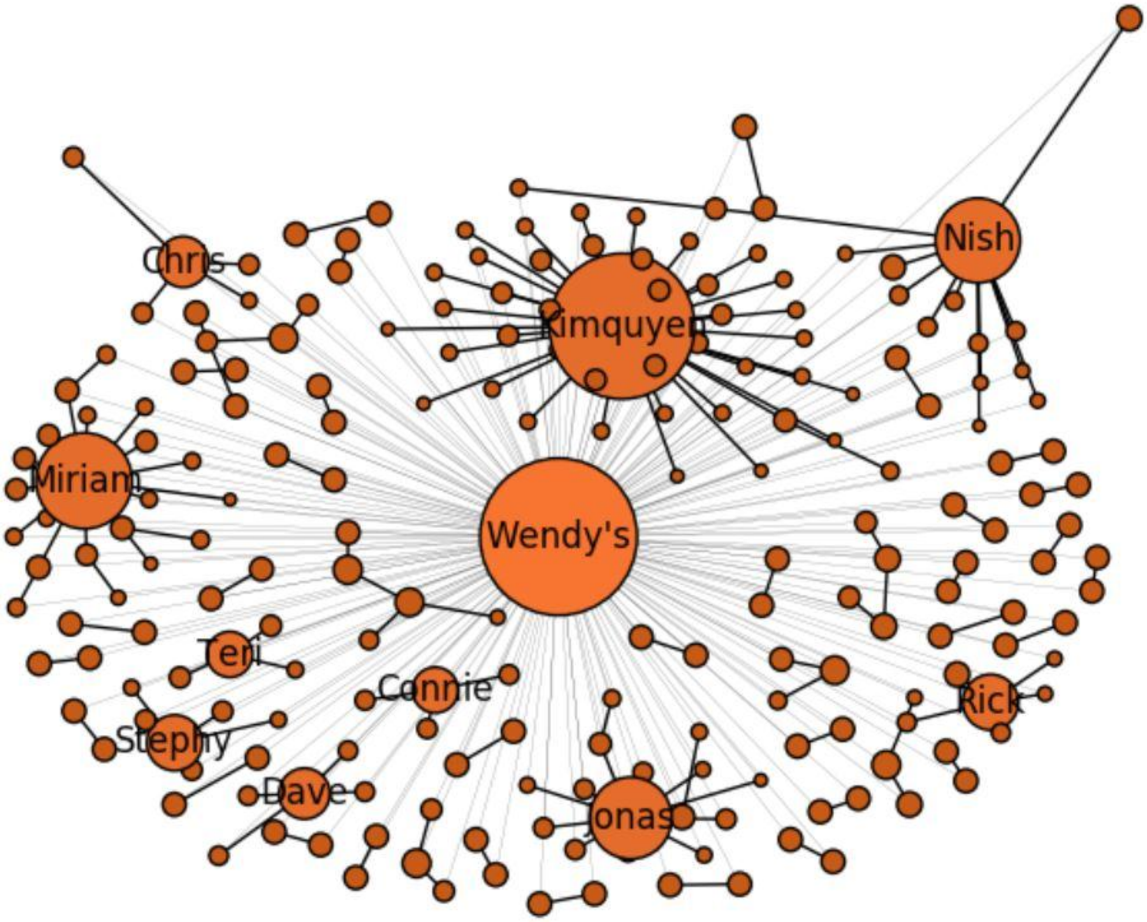

Fig2: Influencers in Wendy's customer network created using Jaccard Similarity

PRASHANTH PADMANABAN (u1076652)

The graph structure is similar to what Teja has mentioned above. The nodes are the users and edges have weights, which is nothing but the inverse of difference between the ratings that the users have given. Say, there are two users, user1, user2 and both are friends. There exists an edge only when user2 has visited after user1 and the weights will be the one described above.

Now, to find which user is influencing more, I experimented with Closeness Centrality. Once the graph is constructed with constraints as mentioned above, for every node we see shortest path distance to every other node. So this gives us the information that, more central the nodes are, the more quickly we will reach other nodes. In our context, the more central a node is, the more it is influencing. The centrality is calculated as inverse of the average shortest path length. Hence the higher the centrality value the higher central the node is. One thing that we discussed with the professor is about the connectedness of a graph. There are situations were a user just reviewed the business and that user is not a friend of other users. Hence the graph is not completely connected. Since such users had no friends they cannot be the influencers and hence I removed them from our graph. Removal of nodes also made our code to run faster. But there are situations where there are many strongly connected components in the graph. So the centrality closeness algorithm has to be run for each connected part separately.

But after running centrality algorithm I observed that most of the nodes shared the same centrality value. For example while taking top 15 users based on centrality value, almost 8 had the same centrality value and interestingly those 8 are influenced by another user with slightly higher centrality value in the same component. Therefore, I could not conclude if top 15 are the influencers of the entire business that was under consideration. [Fig 3]

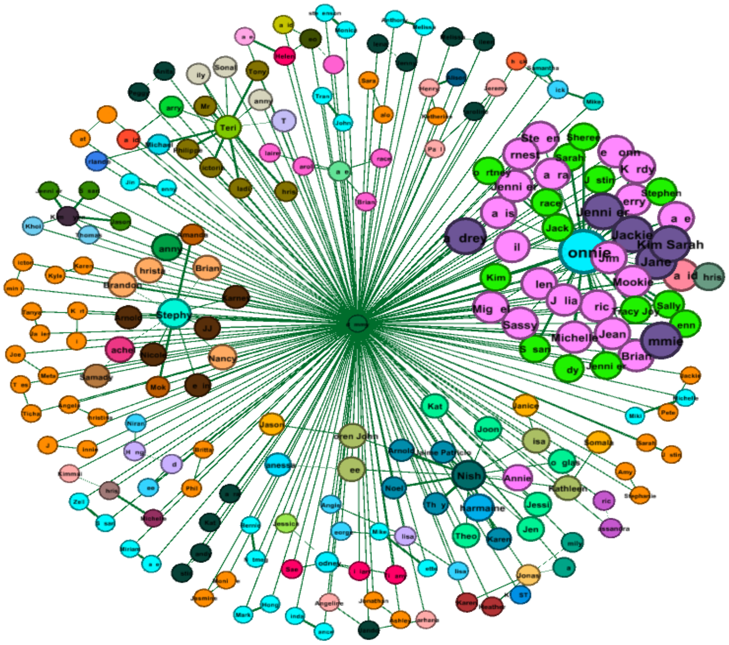

Fig3: Influencers for Wendy's by Closeness Centrality

From the above figure as you can see on the right side I cannot really figure out who’s influencing more.

Hence I experimented with Betweeness Centality based on nodes/vertices with same constraints and design as mentioned above. Betweenness centrality of a node `v` is the sum of the fraction of all-pairs shortest paths that pass through `v`. The results of this algorithm is convincing as it showed the influencers in every components. And when visualizing this it helped me spot the influencers in each components distinctively. [Fig 4]

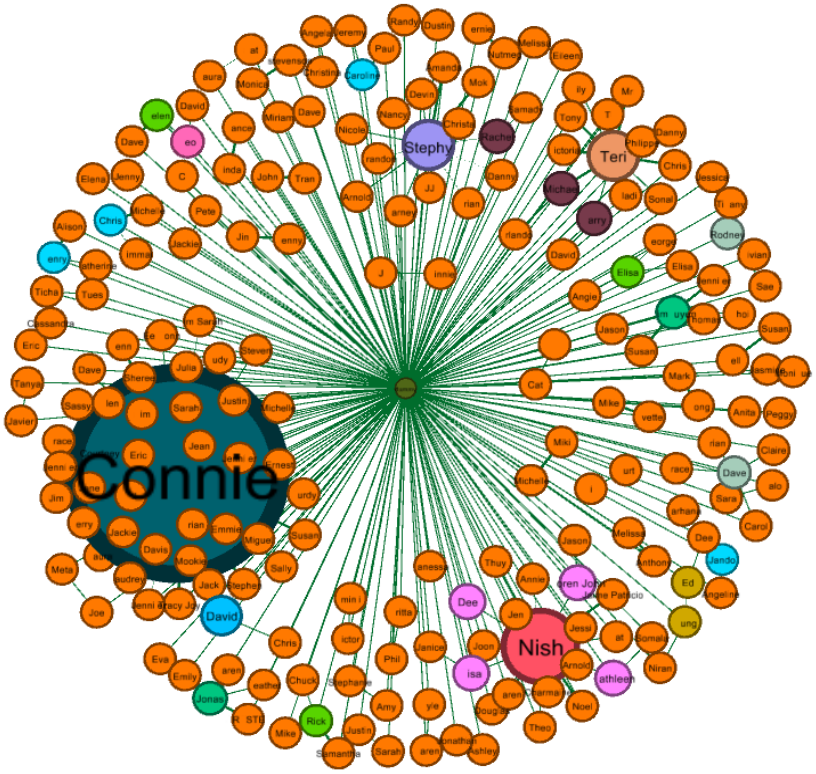

Fig4: Influencers for Wendy's by Betweeness Centrality

Initially I took ‘a’ business to analyze which users are influencing more with centrality closeness algorithm. This was slow as there were many single nodes and hence there were many components and algorithm was run on all such components. This also made representation of the graph difficult as it was clumsy. After the removal of such nodes the code was faster and I was able to represent the graph to fit in a window. To make the graph look connected I created a dummy node and edges were drawn from this node to every other node in the graph. This is done for visual representation and also it made easier to infer the influencing users in all connected components for the business that was considered.

Comparison:

After generating these graphs, we compared the top 10 influencers. Below is our comparison.

Rank	Closeness Centrality	Betweeness Centrality	PageRank	PageRank with Jaccard Similarity
1	Connie	Connie	Connie	Connie
2	Jackie	Nish	Nish	Nish
3	Jane	Stephy	Stephy	Teri
4	Emmie	Teri	Teri	Stephy
5	Audrey	David	Jonas	Kimquyen
6	Kim Sarah	Kathleen	Kimquyen	Dave
7	Jennifer	Dee	Dave	Chris
8	David	Loren John	Rodney	Jonas
9	Davis	Lisa	Elisa	Rick
10	Jerry	Rachel	Rick	Miriam

We observed the correlation between our approaches and found that betweeness centrality and PageRank are highly correlated. Below is the scatteplot matrix for different approaches.

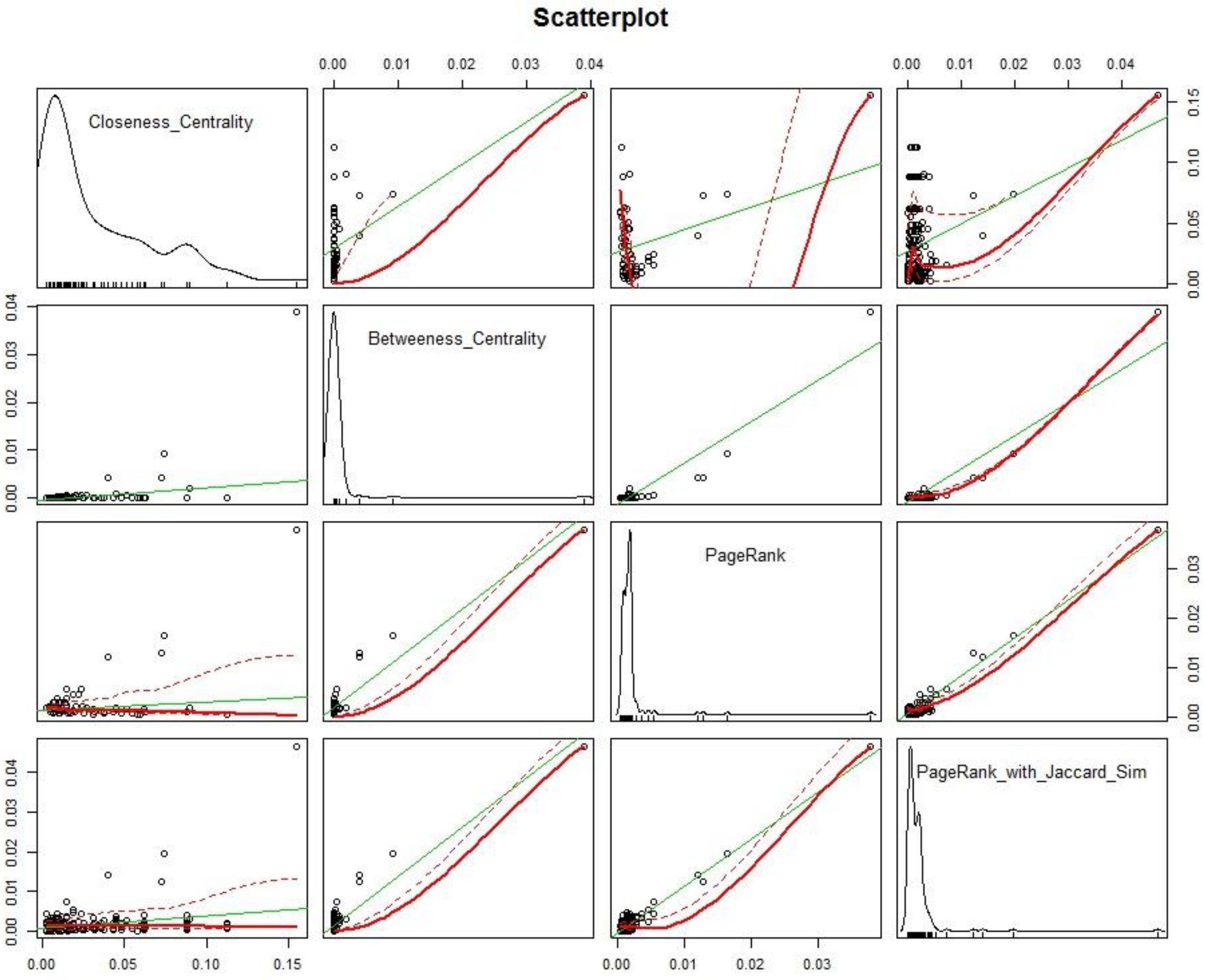

Fig5: ScatterPlot Matrix

From third and fourth plots in row 2 we observe the positive correlation between PageRank and Betweeness Centrality. From below table we can infer the high correlation between PageRank and Centrality.

	Closeness Centrality	**Betweeness Centrality**	PageRank	PageRank with Jaccard Sim
Closeness Centrality	1	0.300146	0.17158	0.267972
Betweeness Centrality	0.300146	1	0.93388	0.928903
PageRank	0.171583	**0.933887**	1	0.953648
PageRank with Jaccard Sim	0.267972	**0.928903**	0.95364	1

Learnings:

Technologies: Python NetworkX (For graphs), PyPlot, Gephi (For Visual Representation), R

Concepts: Social Graph Mining, Page Rank, Centrality, Jaccard Similarity

The data preparation phase took more time than we anticipated.

Having a diverse team helped in our project. This is because we were able to think different problems in different ways and apply different approaches to solving them.

Reference:

[1] https://www.yelp.com/datasetchallenge/dataset

[2] PlotLy - https://plot.ly/

[3] Viral Marketing -http://homes.cs.washington.edu/~pedrod/papers/iis04.pdf

[4] Network graphs https://www.grafxnetwork.com/

[5] Network graphx https://networkx.github.io/

[6] Page rank -https://en.wikipedia.org/wiki/PageRank

[7] Fruchterman Reingold layout https://github.com/gephi/gephi/wiki/Fruchterman-Reingold

[8] Gephi https://gephi.org/users/

[9] Do (kelvindo@), Rotimi Opeke (ropeke@), James Webb (jmwebb@). *Predicting Yelp Ratings From Social Network Data Predicting Yelp Ratings From Social Network Data.*

[10] Wenqing Yang (wenqing), Yuan Yuan (yuan125), Nan Zhang (nanz) *Predicting Yelp Ratings Using User Friendship Network Information*

[11] ScaterPlot http://www2.unb.ca/~ddu/6634/Lecture_notes/Lecture_centralitymeasure.pdf

[12] Visualization of Networks http://www.datasciencecentral.com/profiles/blogs/visualize-your-social-media-analytics