Predicting Yelp Ratings with Social Networks

Aojia Zhao - aojia93

December 11 2016

Abstract

In this project, a variety of similarity metrics based on different network graph structures are used to predict user review ratings on the Yelp Dataset. Innovation from this project comes from experimenting with different graph structure setups, including a Business-Business network that seeks to utilize network average properties to bypass user relation based prediction metrics. Results indicate that while business based structures do not outperform user based networks, rating influences can still be observed from this new graph structure.

1 Introduction

Oftentimes, the problem of user review rating prediction is approached by automated systems through an evaluation of prior review preferences of the user and the compatibility of those preferences with the business' logistics. While relatively effective at gauging client sentiments in isolation, it ignores the increasingly important aspect of social network influence on user decisions and resulting reviews. Even though a reviewer could, in the beginning, be opposed to a business entity through a bad first impression, the powerful effect of word of mouth by their social group, close friends and colleagues, could in the end sway their decision and reverse review sentiments. Additionally, external social factors like location, community clustering, and inter-business relationships can also serve as important influences on review outcomes. In this project, a close look at measuring the influence of social networks on Yelp user reviews compared to traditional single user historical tendencies will be undertaken.

2 Related Literature

2.1 Measurement and Analysis of Online Social Networks [1]

In this work, the authors conducted a large-scale, manual web crawl of four, large online social networks: Flickr, LiveJournal, Orkut, and Youtube. In doing so, they were able to evaluate and create rough network graph structures to reason and confirm the general theories hypothesized about real-world social networks. Their important conclusion was that there definitely exists a central core of influential people accounting for about 10% of the population. Additionally, these people exist both smallworld behaviors of interacting in their clustered niches, as well as cascading behaviors of rapidly spreading information to followers. In my project, I utilize this information through creating a metric that predicts review stars using ratings from highly connected individuals in the graph.

2.2 Network-Centric Recommendation: Personalization with and in Social Networks [2]

In this paper, the authors created an online item recommendation system designed to persuad users on Facebook to purchase presented items. To find preferences, they used the users' social networks to find close friends who have purchased items and present similar ones to clients. They resulting success confirmed the hypothesis that users will not only be swayed by their neighbors in social networks, but also that the strength of recommendation heavily depends on how close the user is to those friends. In my project, I implemented a form of this metric in the Jaccard similarity measurement, where the more mutual friends two individuals have, the heavier weight is placed on those recommendations. Similarly, I apply this concept in designing a new, Triad structure based prediction metric between triples of user, business, and business to mimic this shared mutual structure.

2.3 Evaluating Collaborative Filtering Recommender Systems [3]

In this paper, the authors evaluate some properties of network datasets that are necessary empirical conditions for training a successful recommendation system. They found that the type of recommendation being made was an important factor in determining success as well as the density and real world property of the dataset trained on. In my project, I analyzed the Yelp dataset using several techniques to confirm the expected real world properties including Power Law and preferential attachment.

2.4 On the Efficiency of Social Recommender Networks [4]

The authors in this paper investigated the efficiency of relevant information distribution in the Yelp social network. Through a variety of social graph metrics, they concluded that the Yelp dataset is surprisingly efficient given user to user connections. In my project, I first apply several standard metrics to predicting user review ratings based on properties of user network connections. Then, I extrapolate this relationship and construct a similar business to business structure to test if this pattern is preserved across data type.

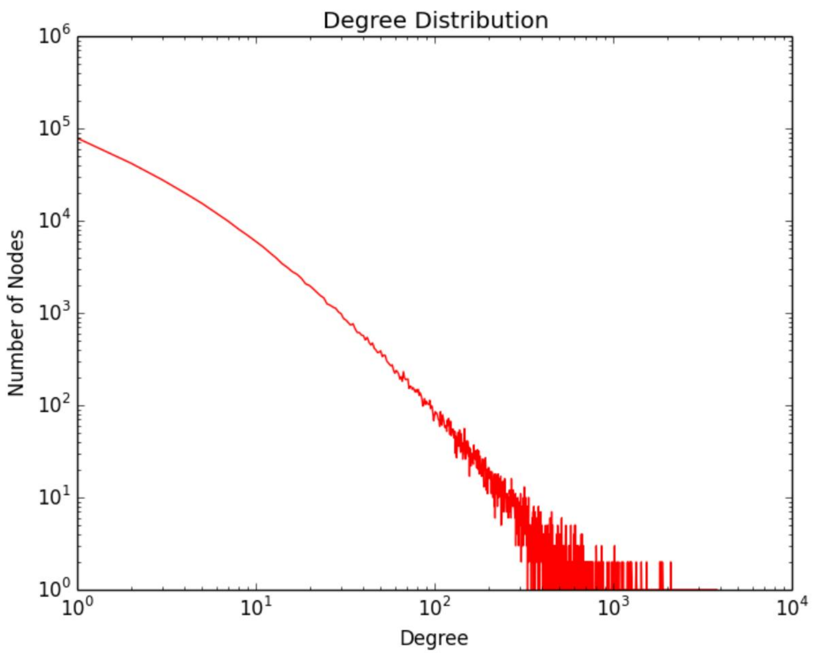
3 Dataset Analysis

3.1 User-Business Metrics - Full Yelp Dataset

Data for this project was acquired through the Yelp! Challenge Dataset for 2016. As the raw data was in Json form, a script was written to convert the relevant fields to csv for easier processing. Below are the information fields parsed out

user.csv	user_id, friends, name
review.csv	business_id, user_id, stars
business.csv	business_id, name, neighborhood, longitude_latitude

This resulted in $686,557$ users, $2,685,061$ reviews, and $85,539$ businesses. To get a better view of the connectivity of the network, a graph of the node degree count with respect to number of nodes was constructed.



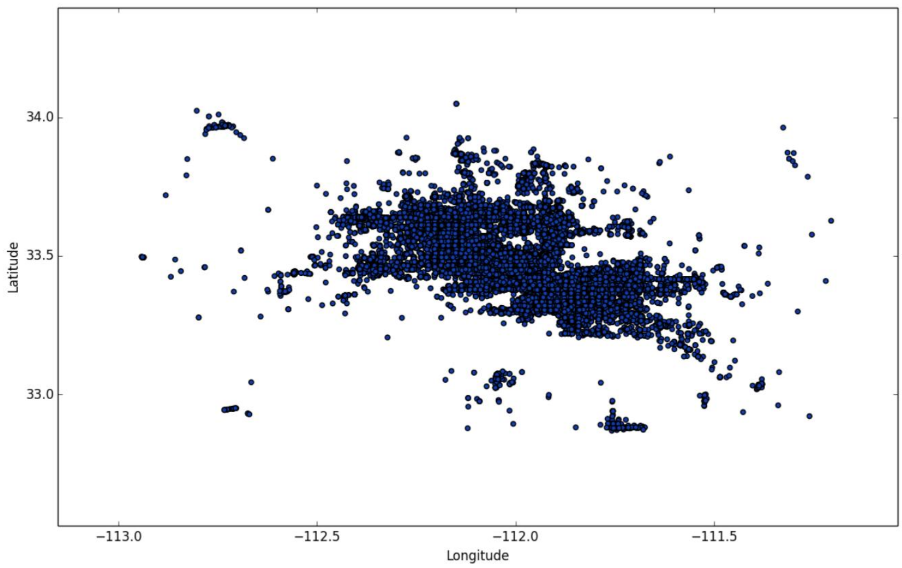
Qualitatively, it is easy to see that the network follows the Power Law as expected of a preferential attachment social network. Using linear least squared estimate, we get $\alpha_{mle} = 1.38615843202$ for the best fit slope. A good intuitive explanation of this property for the Yelp dataset is that the users with high degrees of connection are probably the ones who create the most reviews, giving them the best chance of having at least one of their reviews being read by other users. Those users in turn will friend these highrollers since they see the frequency and quality of reviews.

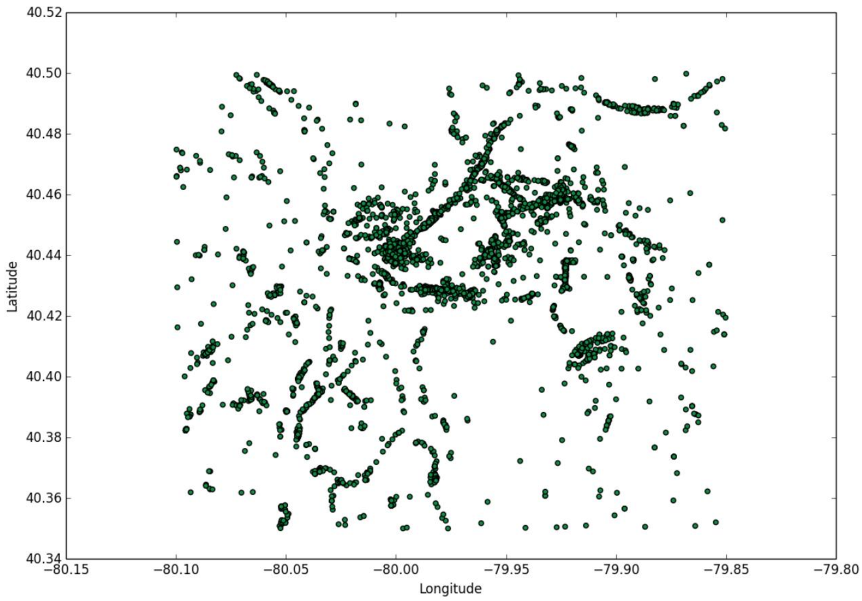
Since a large number of users in the dataset have 0 or only a few connections, they are not good candidates to evaluate social network influences on due to paucity of connections. Thus, for the standard user graph metric evaluations, the group of users was reduced to those with at least 10 connections. This resulted in 77,594 users.

3.2 Business-Business Metrics - Pennsylvania Yelp Dataset

Unlike the previous evaluations on user to user review influences, the primary innovations in this project involve business to business effects, as well as business-business-user triad empirical distribution patterns. To construct such a network, two weighting metrics were applied. First, as Yelp business data provides community cluster information in neighborhoods served, individual business nodes are linked up based on common neighborhood clusters. Second, to evaluate on the physical location aspects, longitude and latitude information was used to gauge relative distances between nodes. Lastly, applying triad structures and properties from other preferential attachment models, a bipartite graph between users and businesses was created to capture (Business, Business, User) triad frequency and patterns. Because all these metrics depend on closely clustering nodes in real world settings, it makes sense to limit the dataset to a smaller setting, while preserving good neighborhood and location diversity. Filtering by US states, Pennsylvania ultimately proved to be a good summarizing of general network patterns, while providing a good diversity of neighborhoods and physical locations. Below are the top 5 states by unique business count, the graph of AZ location distribution by Longitude and Latitude, and that of PA. As can be seen, the physical clustering of PA provides multiple different epicenters, unlike the single large group of AZ.

State	Business Count
AZ	36500
NV	23591
NC	6835
QC	5591
PA	4086





4 Algorithms

4.1 User-Business Models

4.1.1 Baseline Model

For the baseline model, I use the average rating of all other reviews for a particular business as the prediction rating. This is akin to the prior probability distribution of the rating and is a good measure of the dataset's inherent rating for that business. As a vanilla metric, there is no weighting consideration for the neighbors or closely clustered nodes of the predicted node.

$$r_{prediction}(n, b) = \frac{\sum_{n' \in Rev(b)-\{n\}} r_{true}(n', b)}{Count(Rev(b)-\{n\})}$$

4.1.2 Average Neighbor Ratings

For this metric, I implemented average neighbor ratings to mimic level one community influence on the outcome rating of an individual. In real life, as word of mouth recommendation from close friends is normally considered above that of the general public, the hypothesis is that this predictor should generate better results than the baseline. The ratings of neighboring nodes who have previously reviewed the business is averaged together. Weighting is included in this final version, where the weight is the count of number of business that have been reviewed by both the current user and that friend.

$$r_{prediction}(n, b) = \frac{\sum_{n' \in Nbr(b,n)} W_{n'} r_{true}(n', b)}{\sum_{n' \in Nbr(b,n)} W_{n'}}$$

4.1.3 Jaccard Similarity Ratings

For the Jaccard Similarity Rating, similarity between the current node and a node from the pool of previously reviewing individuals is calculated as the count of common neighbors divided by the count of unique union neighbors. This can be seen in the context of level two community connections, where a third party stranger that is mutual friend with a large number of your friends will be trusted by you as a good recommender. This is because you tend to form high opinions of your neighbor nodes and each mutual connection of them gives you confidance in trusting the third party. In terms of Yelp, this is like a food critic that multiple of your friends upvote, and thus you feel comfortable in trusting his food ratings.

$$r_{prediction}(n, b) = \frac{\sum_{n' \in Rev(b)-\{n\}} \frac{|Nbr(n) \cap Nbr(n')|}{|Nbr(n) \cup Nbr(n')|} r_{true}(n', b)}{\sum_{n' \in Rev(b)-\{n\}} \frac{|Nbr(n) \cap Nbr(n')|}{|Nbr(n) \cup Nbr(n')|}}$$

4.1.4 Betweenness Centrality Rating

To gauge the influence of highrollers in the graph on review ratings, Betweenness Centrality is calculated for all nodes. The top ten nodes ranked by BC who have previously reviewed the business rating being predicted are weighted and taken average of. In real life, this is analogous to trusting some celebrity review simply due to how popular they are, without regards to the substance behind ratings.

$$r_{prediction}(b) = \frac{\sum_{n \in TopBCnodes(b)[0:10]} BC(n) r_{true}(n,b)}{\sum_{n \in TopBCnodes(b)[0:10]} BC(n)}$$

4.1.5 Mean Squared Error

As a numeric metric besides accuracy rate to gauge absolute correctness of predictions, the MSE is calculated between each true and predicted rating. This rating is useful for seeing the standard deviation of each value from the truth.

$$M.S.E = \frac{1}{n} \sum_{i=1}^{n} (r_i - r_{true_i})^2$$

4.2 Business-Business Models

In the following metrics, all true ratings of businesses are the average review ratings over all reviews on that entity.

4.2.1 Common Neighborhood Edge Weighting

One of the attributes provided for each Business is the list of neighborhoods that business serves. In the real world, oftentimes the community that regularly spends at businesses can have preferences that result in skewed rating distributions. For example, if a particular neighborhood consist of mainly minority residents, then it can be expected that businesses geared towards their needs will in general have higher average ratings. In other neighborhoods with different community biases, those stores might not be very popular, and thus garner lower ratings. I created this metric to mimic the effects of friends on user influences, only here the friends are surrounding stores with similar community clusters. For any review of a business, we define

$$r_{prediction}(b) = \frac{\sum_{b' \neq b} Nbrhood(b) \cap Nbrhood(b') (r_{true-by-b'}(b))}{\sum_{b' \neq b} Nbrhood(b) \cap Nbrhood(b')}$$

Weighting for each rating is proportional to the number of common neighborhoods. The hypothesis is that the more common communities two businesses share, the more similar their predicted ratings should be.

4.2.2 Euclidean Longitude-Latitude Distance Weighting

Another feature present in Business dataset is the physical location in terms of (longitude, latitude). This second metric, weighting using inverse Euclidean Distance between businesses, is a more quantitative measurement of how close two location's communities are. One of the problems with the previous setup that could occur is if both are near each other geographically, but the communities being served don't share any common names, due to a variety of physical factors, maybe a river or park is separating the two, etc. We can still expect that on average, the ratings should be fairly well correlated for nearby businesses.

$$r_{prediction}(b) = \frac{\sum_{b' \neq b} \frac{r_{true-by-b'}(b)}{Distance(b,b')}}{\sum_{b' \neq b} Distance(b,b')}$$

4.2.3 Business-Business-User Triad Pattern

The last metric is inspired by the Triad pattern social structures present in preferential attachment models. In this case, we create hypothetical triads between two business that have both been reviewed by a common user. The edges between user and those businesses have weights calculated as $r_{user} - r_{business}$. The weight on the edge between two businesses is calculated as $r_{business-being-predicted} - r_{other-business}$. In all cases, the edge weights can be positive or negative, thus forming 8 unique triad patterns. To predict rating for a business, we define

$$r_{prediction}(b) = r_{true}(b) + \frac{\sum_{b' \in Business_{reviewed-by-user}} Ratio(b,b') Ratio(user,b') Weight(user,b)}{\sum_{b' \in Business_{reviewed-by-user}} Ratio(b,b') Ratio(user,b')}$$

$$Ratio(b, b') = \frac{Weight(b,b')}{Weight_{Triad-Average}}$$

Here, the triad average is the average weight for that corresponding edge of that particular triad pattern. For example, if the triad edge weights are $(user - b : -0.56, user - b' : 1.21, b - b' : -0.56)$, then the triad pattern is (-,+,-) and we use the average edge weights of all triads of this pattern to calculate ratio.

5 Results

5.1 Full Dataset

Metric	Accuracy	MSE
Baseline	0.328052697479	1.957559745089
Average Neighbor	0.544326278402	1.51238461257
Jaccard Similarity	0.613751278679	1.48349364843
Betweenness Centrality	0.462511975316	1.78462946344

Although absolute accuracy for correctly predicting ratings 1 through 5 for each business is not very promising, the expected trend for relative success is present.

5.1.1 Baseline

For the baseline, an accuracy of 32% is definitely better than the 20% expected of a random predictor given output space of 5. This confirms that the inherent rating of the general public does have sway on the final rating given by individual users. Though not a huge influence, this can be partly explained by the Mob Mentality. Besides being in a good mood if the general public gave a restaurant 4-5 stars average, there is inherent social pressure in conforming to the ratings. In the event an individual fails to do so, their peers will ask them questions and expect stronger arguments backing up their lower ratings than usual.

5.1.2 Average Neighbor

In the case of the average neighbor rating, it is reasonable to observe a stronger influence by your friend's ratings for a business as compared to some stranger's rating. Compared to the average of all ratings for a business, the averaging of just your friends' ratings gives a marked boost of around 12%. In absolute terms, the Mean Squared Error decreased by about 0.2, meaning that on average, the deviation from the final rating is less than the general population.

5.1.3 Jaccard Similarity

For the Jaccard Similarity Metric, the accuracy jumped drastically to around 61%, an increase of over 15% on the neighbor average. This can be explained by the fact that although each individual mutual acquaintance only increases your confidence in the recommender by a slight amount, the multiple number of mutual friend small boosts in trust together sum to be a bigger factor than the observed ratings of a direct friend can influence you. This is especially true for the highly connected nodes representing celebrities or food critics on Yelp. When hundreds of your friends follow and support the opinion of those individuals, you also tend to trust their opinions more than that of a single connected friend. In terms of raw error, the MSE also dropped drastically to 1.48, a decrease of nearly 0.3 deviation.

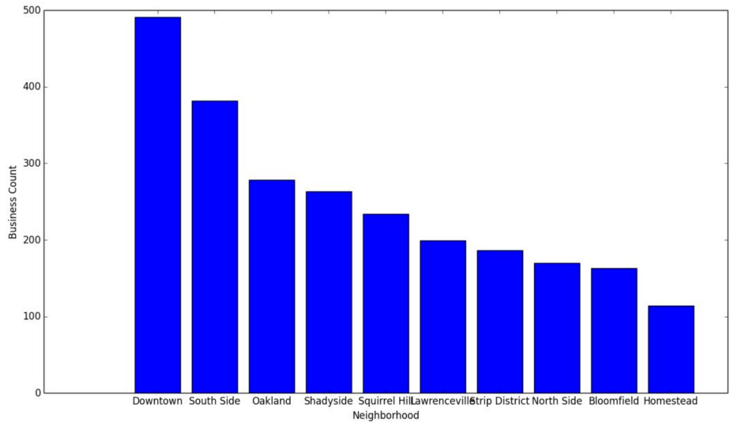
5.1.4 Betweenness Centrality

Lastly, with the Betweenness Centrality similarity measure, we capture the most well-known nodes, those individuals with high degrees of shortest path intermediary. Surprisingly, the results show lower accuracy for this metric than that of Average Neighbor and Jaccard Similarity. Accuracy of classification dropped to 46% while MSE increased to 1.78. Initially, the hypothesis for an expected good performance comes from the fact that highly central individuals would be good information and influence middleman, thus whatever ratings they harbored should have strong weight in effecting other nodes in the network. A few problems with assumption that may explain the discrepancy in results are: dataset contains distantly connected clusters meaning some high centrality nodes don't pass influence to user in question, strongest influence on review ratings come from primary or secondary connections instead of central representative nodes, and only top ten nodes provide a small enough sample per review that large variations in provided ratings can cause noisy predictions. With these factors in mind, we proceed to evaluations on a more centralized dataset next.

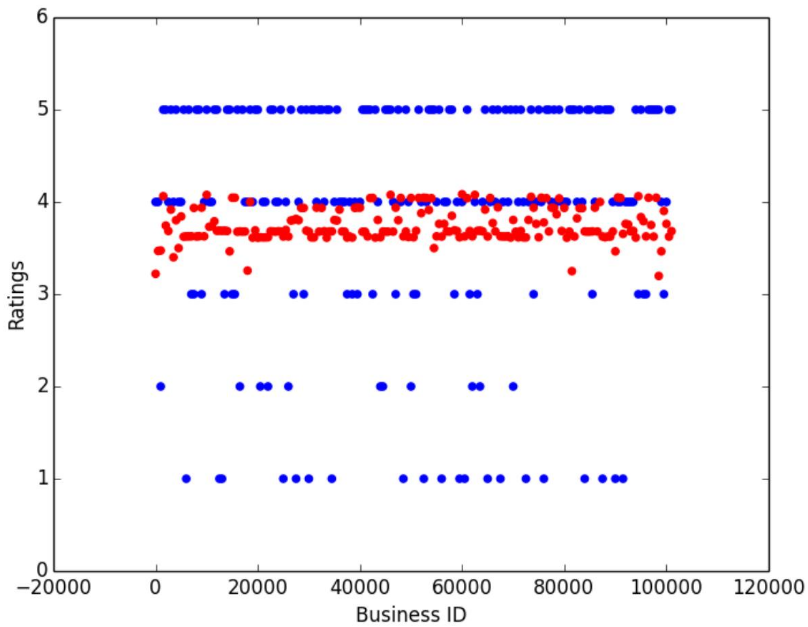
5.2 Pennsylvania Dataset

5.2.1 Common Neighborhood

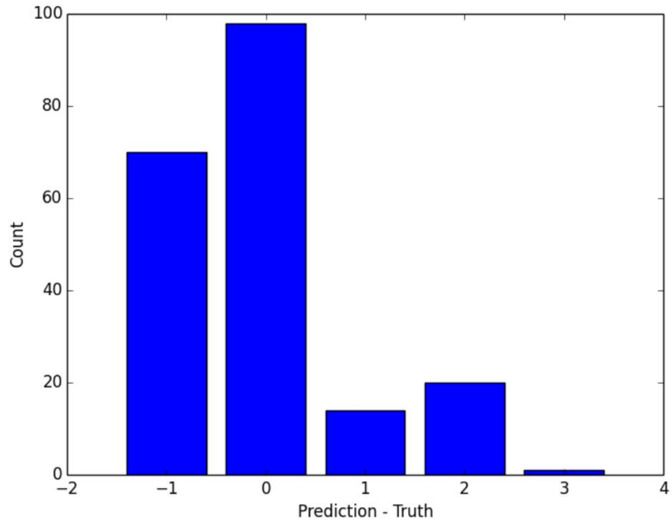
As a context setter, below is the top 10 neighborhoods in the PA dataset by number of businesses serving. On average, the spread is pretty diverse and shows good promise for having large amounts of community overlap.



With the large number of businesses present in reviews, a direct graph of the true and predicted ratings would cover up much of the trends due to density of sample points. In order to alleviate this problem, reviews are collected from every 500 to be plotted. In the graph below, blue represents true rating by the user in the review, while red represents predicted rating for the business. In general, the prediction tends to be clustered around 3.5 to 4, corresponding to the general network propensity for numerous 4 and 5 star ratings as can be seen from the heavy blue lines. However, there are a few outlier dots reaching close to 3 for a few businesses. These points, when checked in data, came from business that shared fewer common neighborhoods with other businesses. Thus, more extreme outliers in these cases can cause predictions that differ stronger from the network average.

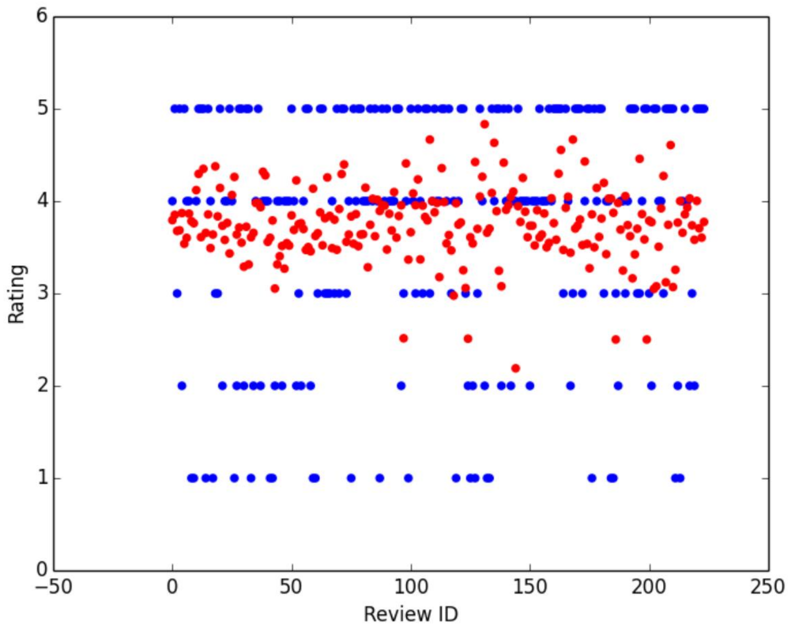


As a more quantitative way to view the trend of prediction, the below graph shows count of differences between prediction and truth, where the difference is rounded to nearest integer. As expected, with the large number of 4 and 5 star ratings, the difference count is heavily skewed right, with over 85% of predictions being below the actual rating by a margin of 0 to -1.

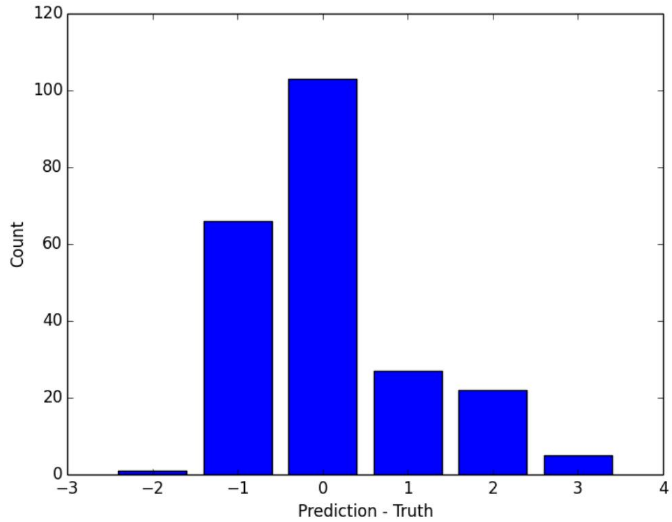


5.2.2 Euclidean Longitude-Latitude Distance

Graphing the results of inverse Euclidean weighting, we notice a stronger scattering effect as compared to common neighborhoods. This can be explained by the fact that while the previous weighting is discretized by presence of certain community names, thus allowing for the presence of empty sets, Euclidean distance, no matter how small, always contribute to part of the rating prediction. This allows influence from businesses further away to spread from local rating averages.



It is worth noting that while the scattering effect caused a larger deviation from true ratings, percentage wise correct predictions is approximately the same as previously. Evaluating over similar numbers of reviews, this distance metric also achieved around 100 rating predictions with discrepancy close to 0. This is around 50% of the predictions being reasonably close. For the general shape, though still right skewed, shows flattening out towards normal distribution Gaussian shape.



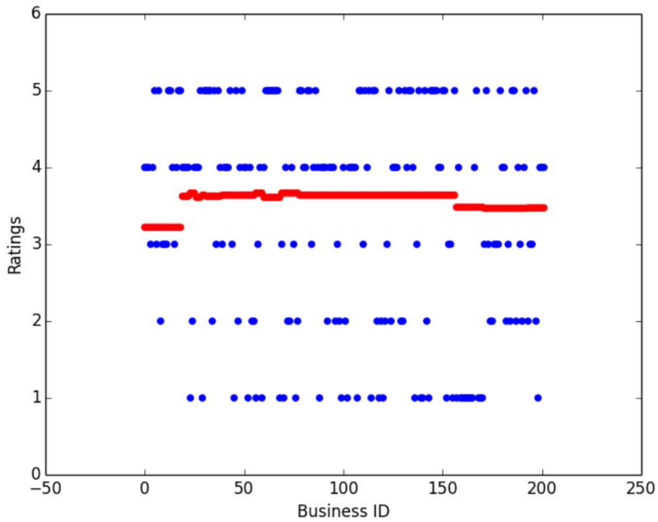
5.2.3 Business-Business-User Triad

For this metric, the relative proportion of various types of triads is a good indicator to general patterns in the network.

u-b	u-b'	b-b'	Percentage
+	+	+	24.3%
+	-	+	5.2%
+	+	-	11.5%
+	-	-	7.9%
-	+	+	8.6%
-	-	+	12.9%
-	+	-	2.1%
-	-	-	27.5%

Many of the higher percentage share patterns do make sense intuitively. For example, if the user's rating is higher than b rating average, and user rating is also higher than b' rating average, then the occurrence, whether b-b' is positive or negative, is very likely due to consistency in this user overrating in their review. On the other hand, if user rates b below average and rates b' above average, the chance of b-b' being negative should be low, since this pattern requires the user to dislike the first business and like the second business contrary to general network trends, usually an oddity.

Included below is the rating prediction in red and true ratings in blue. Unfortunately, from the pattern it seems like the implementation was not completely accurate, as the prediction follows a near constant pattern. The expected pattern, based on distribution of triads defined above, should have two distinct bands of ratings, one above the average, and one below the average. This is because the triad patterns identify the majority of users as either overall harsh in rating, or overall pleasant in giving high scores. A properly implemented resulting graph should probably show clusterings of prediction around 4 and 3.3.



6 Conclusion

In this project, I applied a variety of standard similarity measurement metrics in aiding user social network based review rating predictions. Moving beyond classic methods, I explored effects of a business-business graph structure: creating metrics that mimic community clustering effects in the common Neighborhood Average, quantitatively evaluate geographical location effects on rating prediction with Euclidean Distance weighting, and apply triad structure pattern analysis to intuitive expected rating differential combinations. Overall, though not highly accurate in correctly predicting review ratings, results do indicate social network connections have noticeable influences on individual review decisions. In particular, large common friend groups exhibited in Jaccard Similarity showed the highest prediction accuracy at 61.4%. For future work, a good direction to explore is incorporating star ratings provided by the tip dataset, shorter reviews with summarizing blurbs on why businesses are rated certain ways. In conclusion, social network structures can provide valuable insight on how communities influence individual review ratings as exhibited in the Yelp Dataset.

References

[1] Mislove, Alan, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. "Measurement and Analysis of Online Social Networks."Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement - IMC '07 (2007): n. pag. Web.

[2] Sharma, Amit, and Dan Cosley. "Network-Centric Recommendation: Personalization with and in Social Networks." 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (2011): n. pag. Web.

[3] Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS) 22(1), 5-53.

[4] Wong, Felix Ming Fai, Zhenming Liu, and Mung Chiang. "On the Efficiency of Social Recommender Networks." 2015 IEEE Conference on Computer Communications (INFOCOM) (2015): n. pag. Web.