# Predicting and Modeling Activity in the Evolving Yelp Network

Calvin Wang, Mirela Spasova, Julius Cheng
{thecw, mirelas, juliusc}@stanford.edu

December 10, 2012

## 1 Introduction

Yelp, the largest and most well-known local search and user review site, has grown so much in size and influence that it now plays a significant role in the success or failure of local businesses. As Yelp becomes the common platform for social evaluation of local businesses, we expect that attributes of the Yelp graph become increasingly better predictors of real-world business performance, including both its quality and its popularity. For this reason, it would be useful to analyze properties of the Yelp graph to discover which attributes serve as good predictors. Additionally, we derive a generative model of Yelp activity, which can be used for analysis of future states.

For this project, we use the Yelp Academic Dataset, which includes Yelp review data as well as business and user info in 30 metro areas near major universities. The dataset is naturally represented as a bipartite graph consisting of users and businesses as nodes, with user reviews as the edges that join them, along with various attributes for each edge and node. We especially make use of the fact that reviews are annotated with date of submission, to induce a series of temporal graphs.

We will present various properties of the temporal graphs, followed by several prediction problems, finished by a model that simulates the evolution of the Yelp network, inspired by results from the previous sections.

## 2 Prior work

One research topic related to our project is link prediction, which is broadly adopted by websites that require some form of recommendation engine. The most common scenario involves products and users that interact with them, which can be modeled as a bipartite graph where there is a link between a user and a product if a user has used/bought/reviewed the product. The next step in the recommendation process is to try to predict links between users and new products. High probability links reveal higher chances of that particular user to use/buy the product on the other side of the edge. The general approach to link prediction (summarized by [1]) involves generating features which describe the relationship between user-entity pairs. [1] distinguish between two different types of features generally used for link prediction in social networks: topological (degree, common neighbors, distance between nodes in the network) and node features based. The next step of developing such models involves training set generation and supervised learning of the predictor.

The goal of our project is not to recommend items to each user, but to predict the expected popularity of a new item in the network. Due to the sparsity of our data, combined with the edge creation restrictions arising from geography, we believe that a more worthwhile approach would be to ignore the occurrence of individual links, and instead focus on aggregate properties such as degree distribution and average clustering coefficient. However, we can still utilize the ideas present in link prediction approaches as we will make use of the network's topological and node specific characteristics. [2] observe the user-item interaction networks from a different perspective. They analyse the properties of the Netflix rating network of videos and users as a bipartite graph. The paper confirms the power-law degree distribution and the exponential growth of nodes and edges, much like our findings in Yelp. It also compares the clustering coefficient of the graph with a randomly generated one with the same degree distributions to study how high degree nodes

influence the particular graph structure. Our overall approach to the analyzing the Yelp data involves both discovering these fundamental graph characteristics and applying regression methods over network features for prediction.

# 3 Problem specification

For any business, we would like to predict its future Yelp performance, which we assume to correlate with real world performance. In particular, we wish to predict the volume of future reviews and project change in star rating, the average of all reviews. In constructing such a predictive model, we may discover which features most influence rating and popularity changes in the Yelp network. Additionally, we would like to produce a model to simulate the evolution of the Yelp network in order to interpolate and analyze future states. But before presenting our prediction and simulation models, we present our findings about the Yelp network that influenced the final model. In particular, we explore the following questions:

- Do the degrees of businesses and users nodes follow a power law distribution?
- Are there businesses that defy prediction by undergoing surges of reviews that are not prompted by some network feature?
- Are future ratings correlated with past ratings, or is the problem better modeled as drawing randomly from a distribution?
- Does the Yelp network have collaborative-filtering like features, where users are likely to rate similarly as their neighbors or to similar users?
- How does geographical location influence user review patterns?

The results of these explorations influence our final models.

# 4 Network specification

The network provided by Yelp is subset of the overall network. This subset contains partial data for activity near 30 universities across the U.S. So the data is sparse, highly clustered, and geographically diverse.

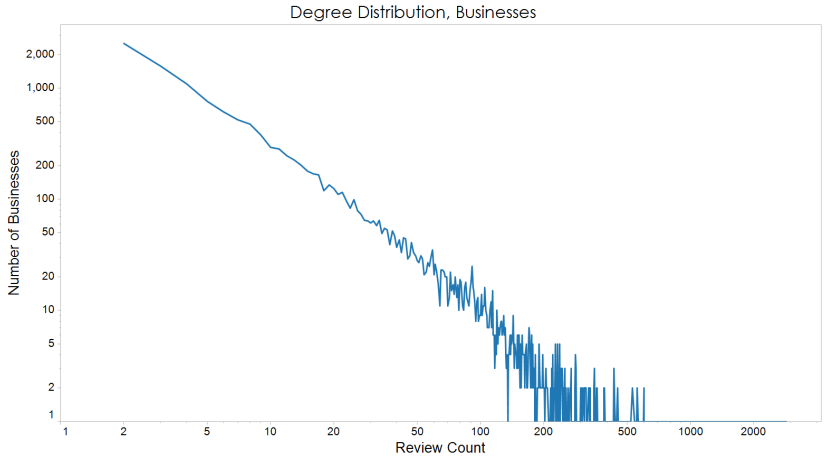
The network is most naturally represented as a bipartite graph, with businesses and users as nodes, and reviews as edges. Each of these entities is specified by a number of non-graph attributes, such as the city for businesses and the "funny"/"useful"/"cool" ratings of each review. Crucially, each review is given a review date, the only time-based attribute we are given access to.

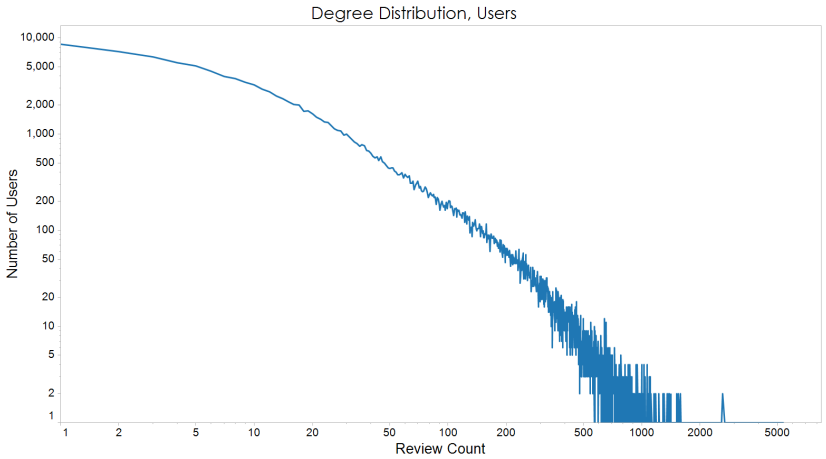
Since we are attempting to predict changes to the network not only based on snapshots of the network at certain times, but the changes over periods of time, we define a temporal model for the data. At some time $t$, the entities in the network are limited to reviews that have occurred on or before $t$, and the businesses and users joined by these reviews. Therefore, the graph "grows" new businesses and users over time, and once new entities appear, they never disappear.

We will use the following notation. A graph $G$ of the network is defined with $\{B, U, R, t\}$, where $B$ is the set of business nodes, $U$ is the set of user nodes, $R$ is the set of review edges, and $t$ is the time of the graph. $b_i, u_i, r_i$ represent specific elements in each of the sets, and edge $r_i$ is defined as $r_i = \{b_j, u_k, d_l\}$, with $d_l$ being a date. When ambiguous, we use a representation like $B^{t_1}$ to represent the set of businesses on the graph at time $t_1$.

# 5 Fundamental analysis

We first examine whether Yelp graph follows a power law distribution. The following graphs illustrate the edge distribution of each node type in a log-log scale:




These figures show that businesses are distributed according to the power law, while users are not really; even at the log scale the distribution follows an exponential pattern. However, we will find later that in our simulation of Yelp network evolution dynamics, assuming the power law as "close enough" for users yields useful results.

Next, we observe that when stratifying the networks into 6-month time periods, the rate of change in edges is very similar. From July 2009 to July 2011, the network grows close to 25% of its existing edges every 6 months. Furthermore, the graph consistently increase its total users by about 25% and total businesses by 6% in the same time period. We do not achieve this kind of consistency when using 3-month time periods. At the 1 year period, the change is of course even smoother. Since we achieve maximum granularity while still retaining predictive power at 6 months, we adopt it as the standard time period for the remainder of the project.

# 6 Geographic considerations

We then examined the influence of location factor in determining users reviews. At the basic level, businesses receive new reviews in future periods in one of the three scenarios:

1. A user has reviewed some business in the same region in the past period,
2. A user who has reviewed businesses, but none in the region, such as a traveller.
3. A user who is new to Yelp, which we use the appearance of first review to indicate.

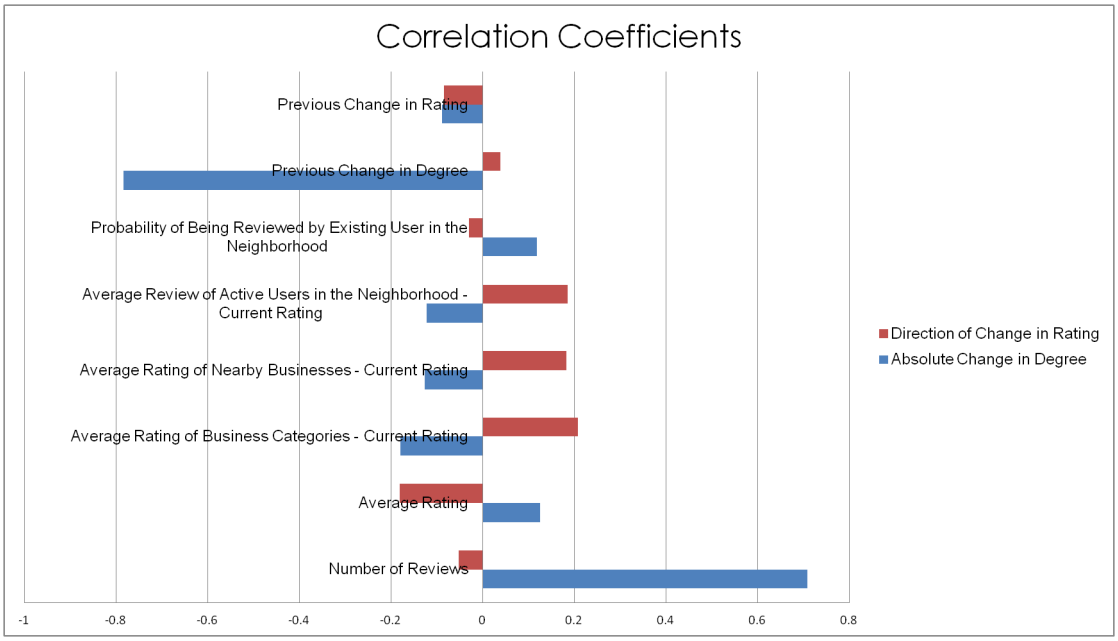
For each of the 30 locations, we estimated the probabilities of those three events using the entire review data from 2008 to 2012. Our results indicate that the probability that a review comes from a new user to

range from 30-42% depending on the region. Businesses are reviewed by locals with a 56-78% chance and "travelling" users link to nodes in a region with 2-3% probability. This distribution is similar to the "small world phenomenon", where the majority of the edges between nodes are close range but there exist random, long distance edges with a probability $q$.

Consequently, in predicting user behavior, we are able to say that a user will review locally with high probability. In link prediction, this information can be used to drastically narrow the choices. In simulation, we can produce edges according to physical location.

# 7 Predicting rating and degree change

We gathered multiple features of the graph, and correlated them two variables of interest: whether a business' average rating will increase or decrease over time, and how much the degree will increase by, or the number of new reviews. It turns out that the average star rating has slightly decreased over time, which is worth noting when discussing these correlation coefficients. The types of features and their correlation with both the degree and rating change target variables are illustrated in the following figure:



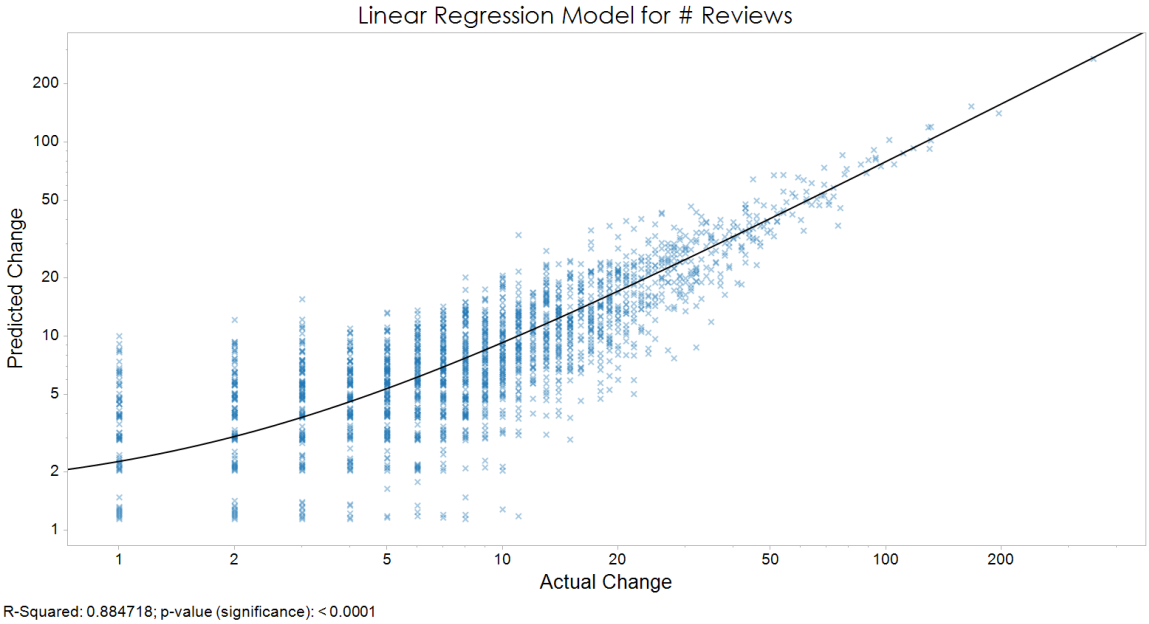

There is roughly a 0.18 positive correlation between the average rating of users with links to a location of a business minus its current rating and the change in its rating in the next period. This implies that if a business has greater than average review for the region (the difference in rating becomes negative), it is likely to get lower score for the next period and if the business is not doing well, its rating tends to increase in certain cases. This may be caused by two independent scenarios: the business' average rating was extreme due to biased draws, and was simply converging to an average, or the tendency for users to overrate to "correct" for others ratings which they perceive as biased.

We also generated features for the average rating of the business's category, and average rating of neighboring business. Features like current degree and previous changes in rating did not correlate to the output.

We attempted logistic regression to classify whether a rating would increase or decrease over a period. Ensembling the most highly correlated features, we reached 0.64 average test accuracy. The results indicate that the described features are somewhat predictive, but not enough to be useful for any serious use.

In another trial, we applied linear regression on a set of features to predict the change of degree over a period. We ensembled a mixture of basic features such as the current total degree, change in degree over the previous period, business category, and star rating squared (squared to compound the effects of a high rating), with network features such as neighborhood size, neighborhood degree change, and clustering coefficient. After performing a complete search on the set of features (an $\mathcal{O}(2^n)$ task, but computed quickly

enough), the algorithm determined that only current degree, change in degree, and star rating squared were highly predictive. We show the results of the linear regression here:




Since the data is in 4 dimensions, we instead show a graph marking each of the predictions versus the actual degree change. The line shows the mean prediction for each degree change bucket. Our mean curve has a 0.885 coefficient of determination with the actual values, which shows promise. But ultimately, this reinforces the obvious: the rate at which a business gets reviews currently is the best predictor for many many reviews it will receive later.

# 8 Formulating a generative model

In the above, we learned several important features of the graph. We now turn to an alternative approach to understanding the Yelp network: constructing mechanisms that emulate its features to better understand the underlying processes.

Earlier, we showed that some properties of the Yelp network remain strikingly constant over time, especially the rate at which new businesses form, the rate at which users review for the first time, and the total increase in new reviews. These even hold true for individual cities. We also showed that users have a high tendency to review businesses within a particular locality, as one might expect. With these, it seems possible to create a generative model that respects these properties while plausibly simulating Yelp network evolution. A few challenges arise: for one, we know that the degrees of businesses respect the power law, so it seems natural to involve some form of preferential attachment. Another difficulty is that our notion of time is different than in common models of evolutionary graph generation; we have many edges and nodes being generated in one time step, whereas a model like Barabási-Albert creates one new node and one new edge at each time step.

Nonetheless, we have formulated a generative model that approximates Yelp network behavior in a plausible way. To reduce the size and complexity of the simulation, we limit the network to only include businesses located in a particular city and users who have reviewed any of those businesses. We wish to simulate the state of the network at time $t_n$, assuming we only have information on the network at states $t_1, ..., t_{n-1}$, and so on until the beginning of the network. First, we fix the number of edges to be generated to be the average rate of increase over each of the time periods. We can express this is as $\Delta|R| = \sum_{i=2}^{n-1} \frac{|R^{t_i}| - |R^{t_{i-1}}|}{|R_t|}/(n-2)$. As shown before, this rate is very steady per city, so for simplicity, we can take the rate of change from the most recent time period only.

Next, since the rates at which new businesses and new users arrive are also fairly steady, we specify a

probability $p_{nb}$ that a new edge is connected to a new business, and a probability $p_{nu}$ that the new edge is connected to a new user. We set each probability such that the expected value of the rate of change in each category is equal to the rate of change in the previous time period. This is expressed as:

$$\begin{aligned}
\frac{p_{nu}\Delta|R||R^{t_{n-1}}|}{|U^{t_{n-1}}|} &= \frac{|U^{t_{n-1}}| - |U^{t_{n-2}}|}{|U^{t_{n-2}}|}, \\
p_{nu} &= \frac{|U^{t_{n-1}}|(|U^{t_{n-1}}| - |U^{t_{n-2}}|)}{|U^{t_{n-2}}|\Delta|R||R^{t_{n-1}}|} \\
p_{nb} &= \frac{|B^{t_{n-1}}|(|B^{t_{n-1}}| - |B^{t_{n-2}}|)}{|B^{t_{n-2}}|\Delta|R||R^{t_{n-1}}|}
\end{aligned}$$

With these, we have that the number of edges is fixed to accord to the known rate of change, and the creation of new businesses and users is centered around the known rates of change, with very small binomial distribution variances $\Delta|R|p(1-p)$.

Finally, we must specify actions when a new edge does not induce a new node. Noticing that the previous rate of change is by and large the best predictor of the current rate of change, we should adopt preferential attachment such that the chance of a new edge forming with a node is proportional to how many new edges were formed with that node in the previous time period. So, the probability $P(b_1)$ that a new edge forms with any particular business $b_i$ is:

$$P(b_i) = (1 - p_{nb})\frac{\sum_{r \in R^{t_{n-1}} - R^{t_{n-2}}, r=\{b_i, u_j\}} 1}{|R^{t_{n-1}} - R^{t_{n-2}}|}$$

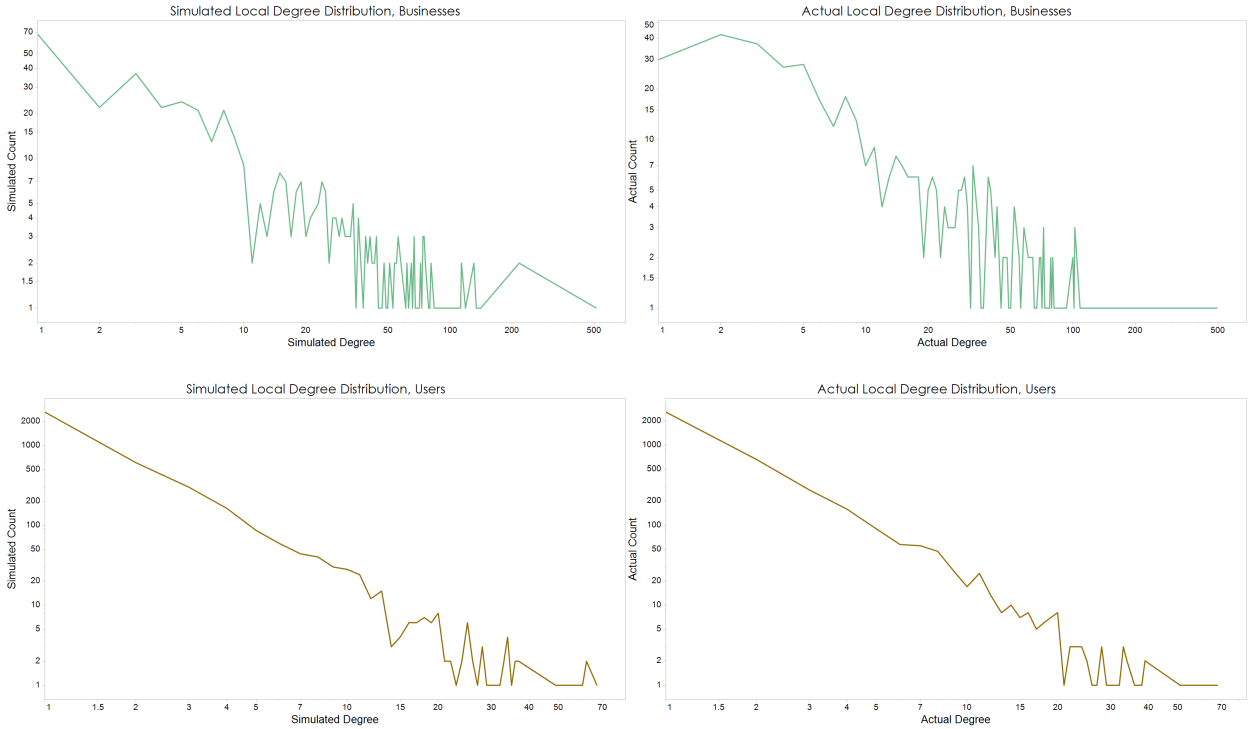
where the coefficient $(1 - p_{nb})$ expresses the case that the new edge does not form with a new businesses instead. The probability that a new edge connects to any given user is similar to the above, with some terms exchanged.

Note that if a new edge is chosen that already exists, we discard the result and try again, although we can also count it towards new edges to model a possible decay for when the graph is too dense. But no Yelp network currently exhibits this property, so we ignore it.

We then use this algorithm to predict a future state of the graph. We induce the Seattle subgraph over time periods $t_1, t_2, t_3, t_4$ where $t_1$ is July-January 2009, $t_2$ is January-July, and so on for half-year periods. We set $t_1, t_2, t_3$ to be the "training periods", and $t_4$ to be "test", where we first predict a graph at $t_4$ with only prior information, and then compare it against the real one. For time $t_3$, we retrieved the parameters $\Delta|R| \approx 0.200, p_{nb} \approx 0.018, p_{nu} \approx 0.422$. Immediately, it's clear that new businesses appear at a very low rate, especially relative to the rate of new users. After executing the algorithm and generating a simulated graph at time $t'_4$, we show some side-by-side statistics:

| | $t_3$ | $t_4$ | $t'_4$ |
|---|---|---|---|
| total businesses | 405 | 432 | 428 |
| total users | 3333 | 4104 | 4095 |
| total reviews | 8184 | 9906 | 9985 |

We see that the algorithm is very accurate in the number of added entities. But we must examine if it holds up to scrutiny in more complex network features. First, we examine the degree distributions for businesses and users in the simulated and actual networks, with simulated networks on the left, and businesses nodes on the top:

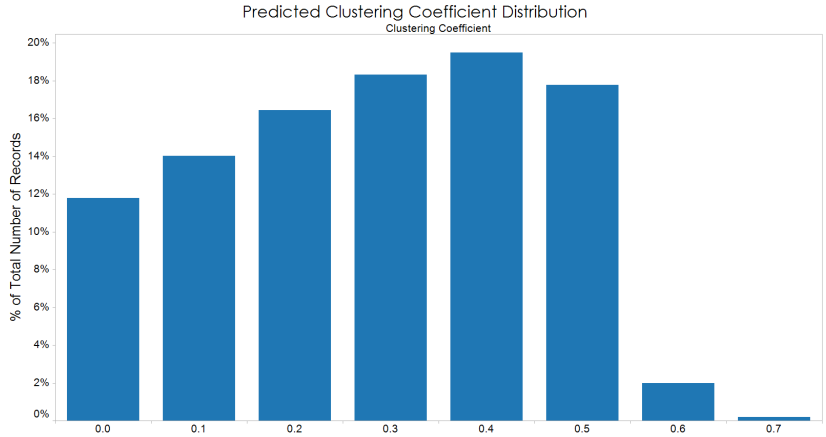


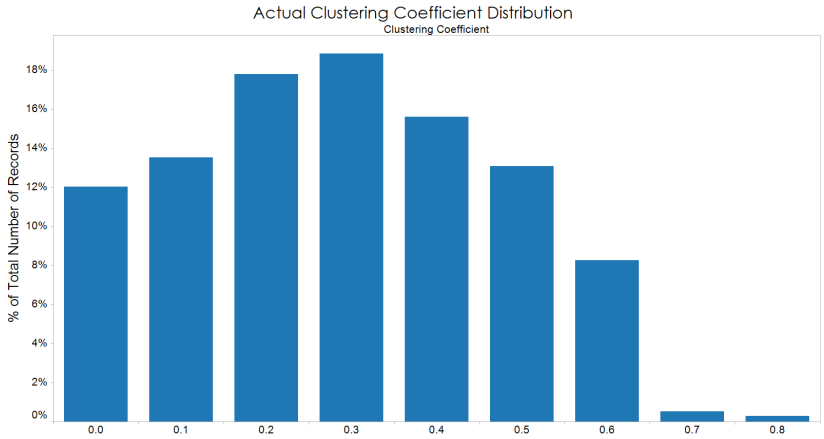

Indeed, the degree distributions are quite similar. Our mechanism generates distributions for both businesses and users according to the power law, since each edge gained by a node in the previous period boosts the likelihood of gaining new edges. Earlier, we showed that real Yelp businesses are distributed by power law, and users approximately so. So our model accounts for this phenomena.

But a network feature such as clustering coefficient is not captured by our model. We preferentially attach, however, we do not explcitly ensure that the nodes are attached in a way that preserves clusters. For bipartite graphs, the standard definition of clustering does not work, since bipartite graphs have no triangles. Instead, we use a definition given by [3]:

$$C(u) \quad = \quad \frac{\sum_{v \in N(N(u))} \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}}{|N(N(u))|}$$

where $N(u)$ retrieves the neighbors of $u$. Roughly speaking, this coefficient is a measure of the average shared connections between a node and its degree-2 neighbors. We compare the clustering coefficients of our simulated graph versus the real one:






Unfortunately, we see that our simulated clustering coefficients skew higher than in the Yelp graph. Since the number of nodes and edges between the graphs are the same, and the degree distribution is very similar, this leaves that the nodes are simply configured differently. One possibility is that the preferential attachment mechanism of our model tends to collect clusters and leave isolated nodes alone. So while we have seen that this simulation mechanism can capture many important aspects of the Yelp network, it is does not simulate this particular feature.

# 9 Conclusion

In this project, we have proposed a methodology for analzying the Yelp network as an evolving temporal graph. We then used it to identify some of the fundamental features of the Yelp network, including predictors for how a business will be rated in the future and how many reviews it will receive. We then used this knowledge to construct a mechanism for evolving the network with only prior information. This novel simulation method was able to create plausible future states, which can be used for further analysis.

# References

[1] Nesserine Benchettara, Rushed Kanawati, and Celine Rouveirol. Supervised machine learning applied to link prediction in bipartite social networks. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '10, pages 326–330, Washington, DC, USA, 2010. IEEE Computer Society.

[2] Mariano Beguerisse Di?, Mason A. Porter, and Jukka-Pekka Onnela. Competition for popularity in bipartite networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4):043101, October 2010.

[3] Matthieu Latapy, Clemence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31 – 48, 2008.