Identifying Cascades in Yelp Reviews

Grace Gee gracehg@stanford.edu
Chris Lengerich ctl51@stanford.edu
Emma O'Neill emmaruthoneill@gmail.com

1. Problem Statement

Social media has gained significant influence in the past few years, specifically for businesses looking to leverage the new technology to increase profits, whether by advertising through coupons, or promoting their customer service and good reputation. There has also been an increase in the popularity of review sites, free websites that the public can access to see fellow users' reviews and ratings of businesses. These review sites have great potential to help or hurt a business, based on how visitors perceive other users' reviews and ratings. We investigated whether or not previous reviews and ratings influenced potential patrons and future reviewers of a business (i.e. attempted to identify if cascades exist in business reviews).

Specifically, we used Yelp, a free social review website that aggregates user reviews and ratings of businesses. Yelp receives approximately half a million unique visitors a month, and so could convincingly be vital in helping a business grow. Our project explored the possibility of identifying cascades in the Yelp reviews for restaurants; specifically, identifying if there is a distinguishable trend in the number of positive reviews and ratings in a certain time period, or after a certain review or set of reviews. Our objectives were to (1) provide descriptive statistics of the previously-unstudied Yelp academic dataset and (2) to use this understanding to develop and test a modified cascade model to investigate whether cascades are present in the data. Our studies indicated that a modified herding model would best describe our data, and after applying the herding model to the data set, we found that for approximately 75% of the restaurants under consideration there is no evidence of cascades. This suggests that Yelp reviews in many cases may not be influenced by previous reviews, and in fact represent independent observations of the truth of a restaurant experience. However, this also implies that cascades may exist for as many as 25% of businesses under consideration.

2. Review of Prior Work

In "A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades", Bikhichandani et. al address the topic of cascades, starting from a simple

toy model in which a chain of individuals makes sequential decisions based on a combination of private signals and public information. Bikhichandani et. al demonstrate that cascades can be easy to start, to the extent that once even ten individuals are included in their simple model, the probability of a cascade occurring is greater than 99.9%. Furthermore, they demonstrate that once such a cascade begins, under the conditions of their model, it will continue unless new public information is released, after which point the collective decisions may be quickly reversed. Later on, Bikhichandani et. al proceed to relax some of their initial assumptions, allowing individuals to draw their private signals with heterogeneous precision. This allows the possibility that a high-precision individual later in the cascade can reverse the cascade.

The paper "Patterns of Influence in a Recommendation Network" by Leskovec et al applies this concept of cascades to a large on-line retailer which records recommendations made by purchasers of DVDs, books, music, and video. Leskovec et al demonstrate the existence of cascades, and additionally uncover some of their notable features. They note that cascades tend to be small, though this does not exclude larger occurrences, and that their frequencies vary depending upon the recommended product, and that their sizes reflect a heavy-tailed distribution. In our work, we would like to accomplish similar goals, looking at a different network, one of restaurant recommendations. Our network is not as well defined in a sense, because we do not have specific users targeting other users, but rather a general audience of the entire public who uses Yelp in a particular area. However, in many ways our goals are similar. Like Leskovec et al, we sought to answer questions about what kind of cascades we can discover and how they reflect the properties of their network and what kind of distributions we uncover.

Inspired by the research done by Birkhichandani et. al and Leskovec et al, we addressed the problem of identifying how earlier user reviews on Yelp affect later user reviews (and hence affect the ratings of a business).

3. Data Collection

We used the Yelp Academic Data Set released in September 2011. The data comprises 65,888 users, 6,900 businesses, and 152,327 reviews from the 250 closest businesses to 30 selected universities. The data is stored in JSON format, with each record having the detailed information listed below.

- User records: name, review count, average stars, number of "useful" votes, number of "funny" votes, number of "cool" votes
- Review records: business ID, user ID, stars, review text, date, number of "useful" votes, number of "funny" votes, number of "cool" votes
- Business records: neighborhoods, address, city, state, review count, categories, open, school nearby, URLs

4. Descriptive Statistics and Findings

We used Python and the JSON decoder package to parse the data set and gather statistics on the mean, median, and mode of star ratings. We also looked at the distribution of ratings in order to determine a rating threshold for popular restaurants. Table 1, Table 2, and Figure 1 below show our findings for all businesses.

Total businesses:	6900
Average number of reviews:	23
Mean star rating:	3.6
Median star rating:	4.3
Mode star rating:	3.5

Table 1. Yelp Academic Data Set Ratings Statistics

1.0	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**4.0**	**4.5**	**5.0**
2%	2%	4%	9%	15%	22%	21%	15%	10%

Table 2. Yelp Academic Data Set Ratings Distribution

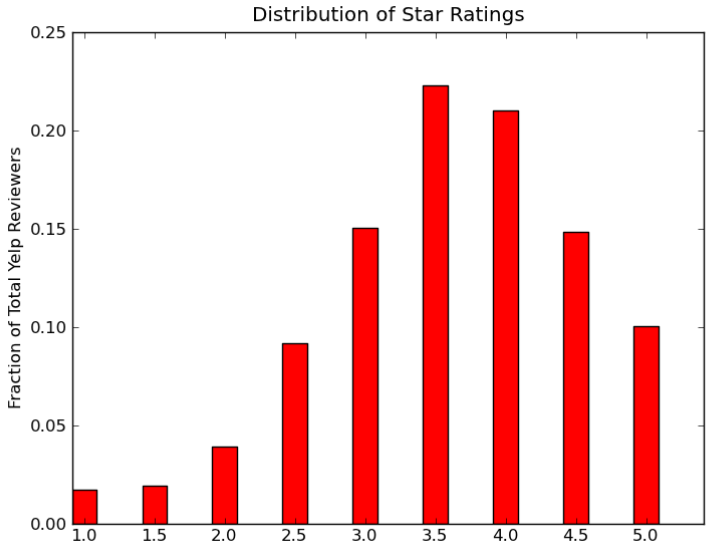

Figure 1. Yelp Academic Data Set Ratings Distribution

These statistics and the right skew of Figure 1 indicate that generally, users give more positive reviews. Hence, we decided to choose a threshold of 3.5 stars and above to indicate that a restaurant is actually "good."

Many of the findings that we discussed above for all of the businesses in our Yelp data set also apply to a subset of all these businesses: restaurants, upon which we are focusing our research. Examining this data set, we note a couple of preliminary statistics which are fairly illuminating. We are interested in restaurants which have received enough reviews over time that we can recognize trends in their ratings. Furthermore, we are particularly interested in restaurants with high ratings because we expect to see cascades in the reviews of these restaurants. There are 6 restaurants total, out of 2564 restaurant businesses, with more than 50 reviews and a rating less than or equal to 2. We find 38 restaurants total with more than 200 reviews and a rating greater than or equal to 4. There are 7 restaurants with more than 200 reviews and a rating greater than or equal to 4.5. We conclude that people are more likely to write reviews when they want to give a restaurant a good rating. This initial study indicates hope for our goal of identifying cascades.

Another notable feature of our data set is that very few restaurants have a 5 star rating, and none of the ones with a 5-star rating have very many reviews. The maximum number of reviews for a 5-star restaurant is 15. We see, then, that while people are in general hesitant to review restaurants of which they have a poor opinion, they are also unlikely to announce the perfection of a restaurant.

We conclude that it may be the case that very small deviations in restaurant reviews may be very telling. The distinction in caliber between a 3.5 star-rated restaurant and a 4.5 star-rated restaurant may be fairly wide due to the overwhelming positivity of reviews.

In order to begin looking for cascades, we looked at a very narrow subset of the total data set; constraining our initial restaurant set to those that had over 200 reviews and an average rating greater than or equal to 4.5. (There were 7 such restaurants.) First, we looked at the individual star ratings over time and the average star rating over time.

Consider the restaurant “East Side Pockets” near Brown University in Providence, RI. This restaurant has received 209 reviews, and has an average star rating of 4.5. In the following figure (2), we see that for this example, the restaurant receives many more high reviews than low reviews, but it does not exclusively receive high ratings; even later than December 2010, it receives a rating of 3. Not unexpectedly, this is evidence of some noise in the dataset. (The proportion of good ratings is high.) We note that the average star rating becomes very stable as time passes. The distribution of reviews is somewhat random initially, but the change in the average star rating becomes very small as time passes. Partly, this is due to agreement of restaurant-goers regarding their restaurant experiences. It is also a result of the high number of reviews, however; once there are a large number of reviews, each subsequent review has less impact on the average.

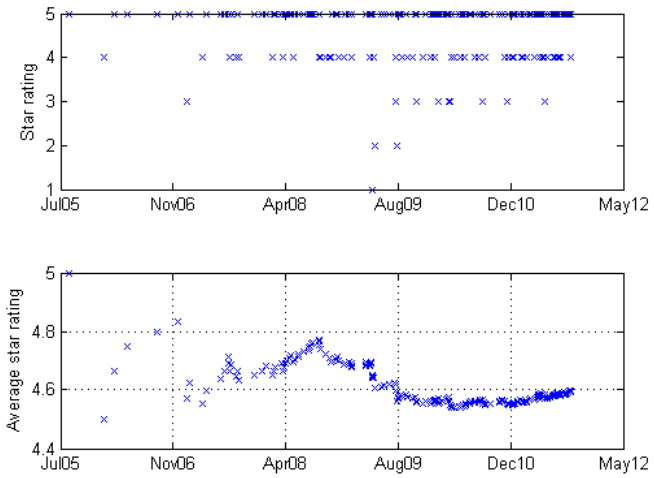

Figure 2. "East Side Pockets", RI, 209 reviews, average star rating of 4.5
Star rating and average star rating vs. time

The next plot shows results for the same data, this time showing a moving average of the rating of the restaurant and the moving standard deviation of the restaurant rating. If we are to identify a cascade, we expect that later points will correlate better with new reviews than earlier reviews because reviewers are beginning to ignore their personal restaurant experiences and to assign ratings based on the previous ratings. The moving average and standard deviation allow us to cluster reviews that are more closely spaced in time. In this way, we can also account for a lower volume of reviews in the earlier days of Yelp and for noise. Here, we see what we suspect is a cascade beginning in August 2009. After this time, the standard deviation between clustered reviews becomes smaller and smaller, indicating agreement among reviewers of the high quality of this restaurant that may or may not match the reality of their restaurant experiences.

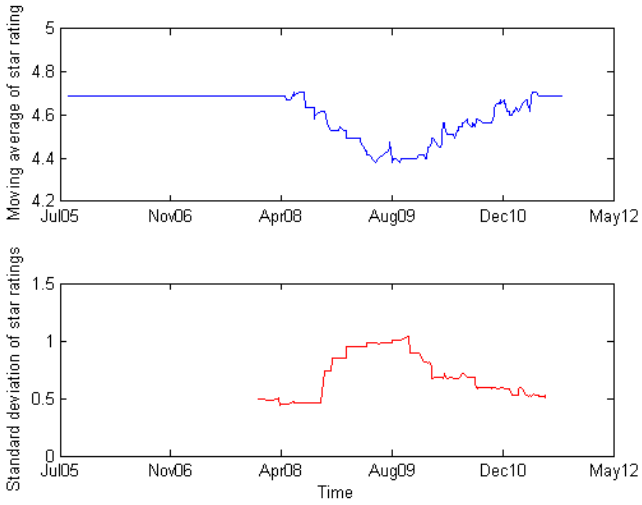

Figure 3. "East Side Pockets", RI, 209 reviews, average star rating of 4.5
Moving average and standard deviation of star rating vs. time

We do not see this behavior for all restaurants with a high rating and many reviews. In fact, many of them, despite a very constant average rating, have a moving standard deviation that is inconsistent and does not seem to represent any particular trend. Despite the consistency of the star rating, reviews do not demonstrably conform to the previous reviews. Furthermore, we note that cascades in the data may be present but interrupted. For example, this is the case for the restaurant "Veggie Heaven" in Austin, TX which has received 223 reviews and has an average star rating of 3.5. We do not see an overall trend in the moving average of the star rating nor in the moving standard deviation of the star rating, but we do see a convergence in popular opinion for a period of time; from August 2009 till September 2010. This provides evidence in the data for brief cascades.

In our initial survey of the Yelp data, we noted that the statistics, particularly the tendency of reviewers to give positive star ratings, provided us with evidence for cascades in public opinion on Yelp. Upon examining in greater detail the review trends for certain restaurants, we discovered examples of cascades, not always long-lasting cascades, but cascades nonetheless. In further project work, we formalized our interpretation of a cascade in the context of Yelp, based on our herding network model and did a more in-depth search for the phenomenon in our restaurant data set.

5. Model Selection

Overview

We decided to fit a herding-like model to the Yelp data, as it seemed to be the best-fit decision model. The reviews are sequential, and each review is a public indication of a person's decision (here, the decision is whether or not to write a positive review of a restaurant). Reviewers make their decision based on previous reviews they read and their own personal experience. We would like to see if we can model reviewers' decisions about a restaurant as a cascade (here, we are looking at cascades of good reviews), and use a random Bernoulli model as the baseline to judge whether or not these reviews truly do represent a cascade, or if they are more random in nature, which would indicate that people do not really consider others' decisions about a restaurant when they are making their own decision.

However, when applying the herding model to Yelp data, we must take into account that it is very rare for a restaurant to have exclusively good reviews even after a number of positive results. The herding model assumes that once one type of decision has been made at least two more times than the other decision, a cascade has started and every decision thereafter will simply choose the majority decision. However, reviews on Yelp almost never work in this fashion -- intuitively, and as evidenced by our data (see previous section) people take into account their personal experience at a restaurant rather than only the reviews from other people (strangers). Hence, it is not uncommon to see reviews in such a sequential fashion:

good bad bad good good good good bad good good good

If we used the unaltered herding model to predict future reviews, there would only ever be good reviews expected after the number of positive reviews outweighed the negative by at least two.

In order to take this fragility of the herding model into account, we tried to address it in three different ways:

1. Using the average of the past k consecutive reviews, k >2, at each step in order to make sure the "true" review average dominate any outlier reviews
2. Using the moving average of the past reviews at each step, also to override any outlier reviews
3. Using a noise term α, where the probability of a reviewer's decision (i.e. type of review -- good or bad) is flipped with probability $1 - \alpha$

After implementing the herding model in all three ways, we adopted the third method, finding it to be the most robust and mathematically sound. The other two ways, which tried to dilute the influence of any noisy reviews, ended up skewing the review data in such a way that sometimes the model would predict future reviews with 100% accuracy (most likely if the restaurant had very high reviews overall).

We describe the herding model we implemented as well as the relevant baseline in more detail in the subsection below.

A. Random Model

For the baseline random model, we treated each review as one drawn from a Bernoulli random variable with mean θ. Then the probability of a series of good and bad reviews is

$$l(x_n, x_{n-1}, \dots x_1; \theta) = \theta^k (1-\theta)^{n-k}$$

where $k = \sum_{i=1}^{n} 1\{x_i = 1\}$. Here, we assumed that the order of the reviews matters for its likelihood as we wanted to compare the likelihoods of this model to the likelihoods of the herding model, for which order matters. The maximum likelihood estimate of theta is simply the mean of the distribution, that is:

$$\theta^* = \frac{k}{n}$$

Therefore, we can estimate the maximum likelihood of a sequence of reviews under the assumption of randomness as:

$$max_\theta l(x_n, \dots x_1; \theta) = \theta^k (1-\theta)^{n-k}$$

$$\text{with } \theta = \frac{k}{n}$$

B. Herding Model

We developed a modified version of the herding model. For this model, individuals leave sequential reviews after receiving private and public signals about the restaurant (their experience at the restaurant and the preceding public reviews, respectively).

In all of the following exposition, we take $\emptyset$ as a parameter representing whether a restaurant is actually good ($\emptyset = 1$) or actually bad ($\emptyset = 0$).

Private Signal

First, let individual n's private signal s_n be drawn from a Bernoulli distribution which represents the true nature of the restaurant with probability β. So we have:

$$P(s_n; \emptyset) = \beta^{\emptyset^{sn}(1-\emptyset)^{1-sn}}(1-\beta)^{(1-\emptyset)^{sn}\emptyset^{(1-sn)}}$$

Public Signal

Next, let X_n be a random variable representing a public signal. We assume that X_n is generated by distorting a latent variable G_n which represents an individual n's guess as to the nature of the restaurant at the time they wrote their review. This distortion takes the form of flipping G_n from its original value with probability $(1-\alpha)$. This distortion represents noise in our review data and allows us to model sequences of observations.

Latent Variable

Finally, define G_n to be a random variable representing n's guess as to whether a restaurant is good or bad, which is derived from his/her private signal, s_n, and public signals $\{x_{n-1}, x_{n-2}, \dots\ x_1\}$. Then this guess can be defined as:

$$G_n = argmax_{\emptyset} P(s_n, x_{n-1}, x_{n-2}, \dots x_1;\ \emptyset)$$

Because G_n can only take on values of 0 or 1, we can represent this as:

$$G_n = 1\{\ P(s_n, x_{n-1}, x_{n-2}, \dots x_1;\ \emptyset = 1) >\ P(s_n, x_{n-1}, x_{n-2}, \dots x_1;\ \emptyset = 0)\}$$

Derivation of Log-Likelihood of Data

We want to derive an expression for the likelihood of our data, $[x_n, x_{n-1}, \dots x_1]$, in terms of $\alpha, \emptyset$ and β to make this expression tractable in code. Dropping the dependency on α and β in the notation for convenience, we have:

$l(\emptyset) = P(x_n, x_{n-1}, x_{n-2}, \dots x_1;\ \emptyset)$

$$= P(x_n|x_{n-1}, x_{n-2}, \dots x_1;\ \emptyset) * P(x_{n-1}|x_{n-2}, \dots x_1;\ \emptyset) * \dots * P(x_1;\emptyset)$$

We will show that we can represent $P(x_n, x_{n-1}, x_{n-2}, \dots x_1;\ \emptyset)$ in terms of the parameters and the data via induction.

First, we find an expression for the base case. For $n = 1$, we have:

$$P(X_1 =\ x_1;\emptyset) = P(G_1 =\ x_1;\emptyset) * P(no\ flip) + P(G_1 =\ x_1;\emptyset) * P(flip)$$

by the definition of $X_1$and G_1. Therefore, we have an expression for probability in terms of the parameters and the data, which satisfies our assertion. Now, we consider the inductive case. By the inductive hypothesis, we have that:

$\forall i, i < n$:

$l(x_{i-1}, \dots, x_1)$ can be expressed in terms of the parameters and the data

$\rightarrow P(x_{i-1}|x_{i-1}, \dots, x_1)$ can be expressed in terms of the parameters and the data

Then we simply want to show that $P(x_{i-1}|x_{i-1}, \dots, x_1)$ can be expressed in terms of the preceding probabilities to prove our assertion. By the definition of G_n, we have:

$$P(x_n|x_{n-1}, x_{n-2}, \dots, x_1;\ \emptyset)$$
$$= P(G_n =\ x_n|x_{n-1}, \dots, x_1;\ \emptyset) * (\alpha) + P(G_n =\ \neg x_n|x_{n-1}, \dots, x_1;\ \emptyset) * (1 -\ \alpha)$$

Examining the conditional probabilities on the right hand side of the above equation, we have that:

$$P(G_n =\ x_n|x_{n-1}, \dots, x_1;\ \emptyset)$$
$$= P(s_n = 1;\emptyset) * 1\{P(s_n = 1, x_{n-1}, \dots, x_1;\ \emptyset =\ x_n) > P(s_n = 1, x_{n-1}, \dots, x_1;\ \emptyset =\ \neg x_n)\}$$
$$+ P(s_n = 0;\emptyset) * 1\{P(s_n = 0, x_{n-1}, \dots, x_1;\ \emptyset =\ x_n) > P(s_n = 0, x_{n-1},\ \dots, x_1;\ \emptyset =\ \neg x_n)\}$$

However, by the conditional independence of our private and public signal, we have that the terms inside the indicator function can be split up. Taking the left hand side of the first condition above, we have:

$$P(s_n = 1, x_{n-1}, x_{n-2}, \dots, x_1;\ \emptyset =\ x_n) =\ P(s_n = 1;\emptyset =\ x_n) * P(x_{n-1}, x_{n-2}, \dots, x_1;\ \emptyset =\ x_n)$$

However, $P(s_n = 1;\ \emptyset =\ x_n)$ can be expressed in terms of our parameters and data by our initial assumptions and

$$P(\ x_{n-1}, x_{n-2}, \dots, x_1;\ \emptyset =\ x_n) = l(\ x_{n-1}, x_{n-2}, \dots, x_1;\ \emptyset =\ x_n)$$

which can be expressed in terms of our parameters and data by our inductive hypothesis. By similar methods, we find that $P(G_n = \neg x_n | x_{n-1}, \dots x_1; \emptyset)$ and $P(s_n = 0, x_{n-1}, x_{n-2}, \dots, x_1; \emptyset = x_n)$ can be expressed in terms of the data and parameters. Therefore, we can calculate $l(x_n, x_{n-1}, \dots, x_1; \emptyset)$ in terms of the data and parameters and so we can create a tractable expression for the likelihood of our data by using recursion.

Windowing

To the above model, we added one additional assumption, namely that our conditional probabilities follow a modified Markov property. In particular, we assumed that for a window of size k, we have the following property:

$$P(x_n | x_{n-1}, \dots, x_1; \phi) = P(x_n | x_{n-1}, \dots, x_{n-k}; \phi)$$

This represents the assumption that an individual may only read the k most recent reviews on a restaurant before writing a review. Notice that this does not change any of the steps in the derivation of the log-likelihood of the data. However, it does limit the depth of the recursive probabilities that must be calculated for any conditional probabilities.

Selection of Parameters

To select the parameters for $\boldsymbol{\alpha, \beta}$ we iterated through the space in intervals of 0.025 and selected the parameters that maximized the likelihood. The window was tested with several different values and will be explicitly reported where it appears.

7. Model Evaluation and Discussion

In all of the following data sets, we classified star ratings as $1\{rating \geq 3.5\}$.

Verification of the Cascade Model

To verify the cascade model performed as expected, we tested the model on two data sets:

1. Review chains from restaurants with average rating $\geq \mathbf{3.5}$ and over 200 reviews
2. From these same restaurants, a 20-review moving average of star ratings. The average was then classified according to the preceding indicator function

For a window size of 5, in data set #1, the cascade model performed better than the random model for 12 out of the 73 businesses. However, for data set #2, this figure increased to 47 out of 73 businesses. This therefore serves to support that the model behaves as expected as data set #2 has less entropy and would therefore be expected

to have a higher likelihood under the cascade model. Similar results were achieved for other window sizes.

Likelihoods of Cascade and Random Models

As a starting point, we selected chains of reviews from restaurants based on several criteria:

1.) The restaurant had an average rating of 3 and above, inclusive
2.) The restaurant had over 200 reviews

which resulted in a collection of 85 restaurants. These reviews were selected so that we would have a long chain of reviews and so that in most cases, the number of variables we were conditioning on would be determined by the window size, not by the length of the review chain.

We simulated the random baseline on the data using the MLE for $\boldsymbol{\theta}$. We also tested the cascade model on the data, using an iterative search for the MLEs of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

For a window of 8, we have the following differences between the log-likelihood of the cascade model and the log-likelihood of the random model

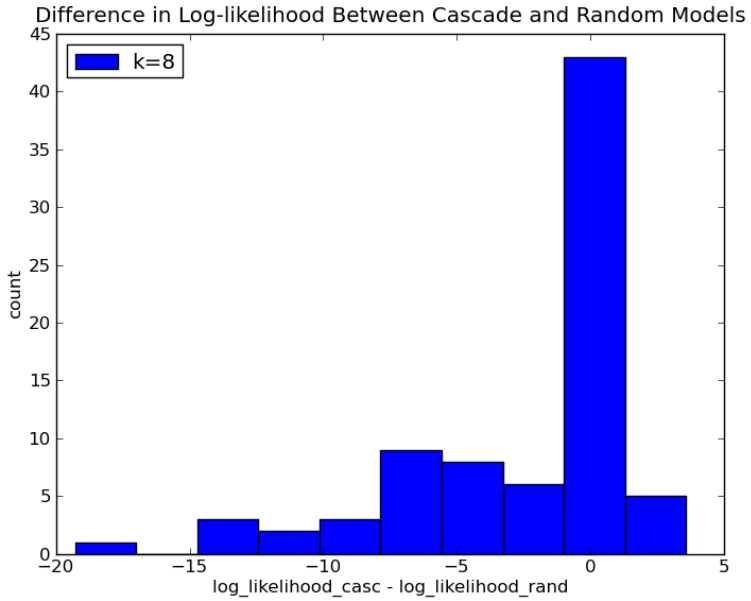

Figure 4.

The distribution of these differences suggests that in a data set with very long review chains, in most cases, the random model performs better (about 75% of the time), however in some cases the cascade model performs better (about 25% of the time).

This suggests that while most review chains do not exhibit cascades, there may be some review chains for which there are cascades.

Comparison of Different Window Sizes

Wishing to cross-validate against our window size, we ran our data with windows of varying sizes on the data set described above. We found that the best window size for our model is approximately k = 8 :

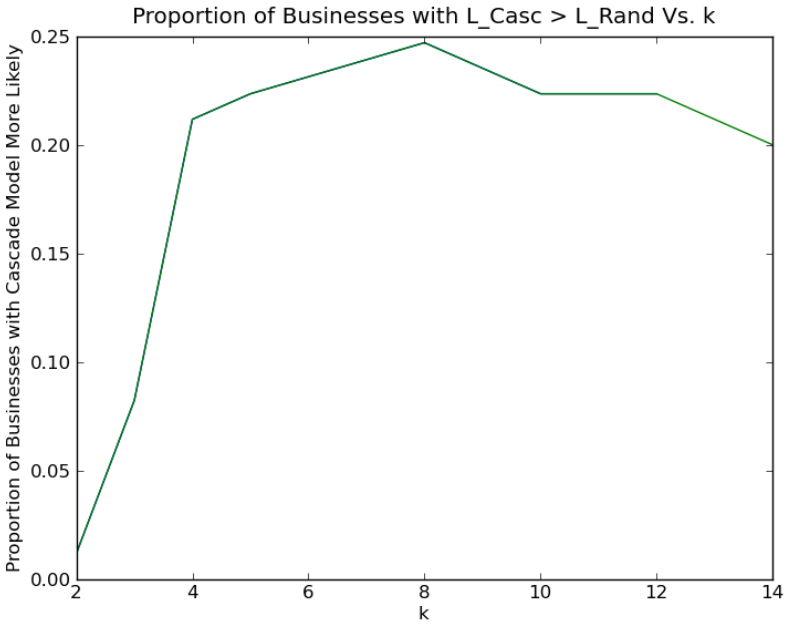

Figure 5.

This suggests that increasing the window size can drastically improve the performance of our model.

Expansion of the Dataset

We selected k = 8 which maximized the likelihood of our cascade model in the 85 reviews and then expanded our dataset to include businesses with greater than 40 reviews (without filtering based on star rating). This represents 647 restaurants, which is a statistically significant part of the 2564 restaurants included in total data set.

In this expanded dataset, we found that the cascade model does better in 167 out of the 647 = 25.8% of the review chains for the businesses. Interestingly, this is a very similar probability to that suggested by the smaller dataset and the difference in the log-likelihood between the two models also appears similar. This suggests that the

difference in the length of the review chains does not affect the relative order of the model rankings.

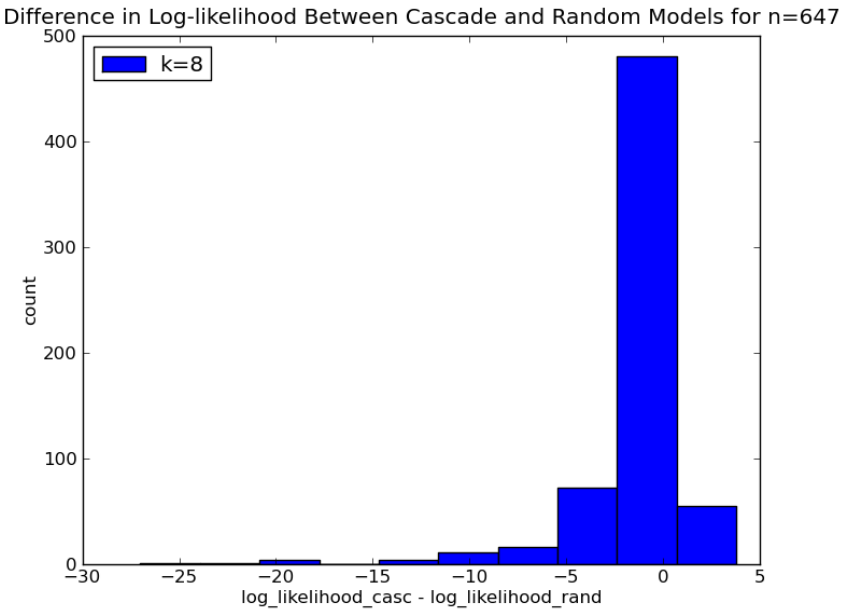

Figure 6.

Taken together, our findings support the hypothesis that while there may be some cascades in the Yelp data, for the majority of the data, the Yelp data is better or equally well modeled by a random distribution. This suggests that Yelp reviews, on a whole, are not influenced by previous reviews, and in fact represent independent observations of the truth of a business.

Nevertheless, the fact that 25% of the reviews are better fit by the cascade model suggests that there may be some chains of reviews in which cascades may be occurring. From a simplified modeling perspective, the characteristics of these reviews are fairly simple - they have less entropy and have long continuous chains of 0s or 1s. From a real-world perspective, however, it is less clear exactly what characteristics of the restaurants cause that particular chain of reviews to appear. While it is beyond the scope of this project to attempt a descriptive analysis of the restaurant reviews which generate these cascades, this topic could be a fruitful area for further research.

8. Conclusion

Inspired by the research by Bikhchandani et al and Leskovec et al, as well as the recent release of the academic Yelp dataset, we set out to investigate the possibility of

cascades in restaurant reviews. Our statistical and graphical findings suggested that reviews may follow trends and that reviewers were more likely to give positive ratings than negative ratings. This indicated the likelihood of finding cascades, and we modified a herding model to test our hypothesis. We compared our results for the herding cascade model with a random Bernoulli model, and found that there are some restaurants for which the cascade model is a better model than the random model; however, for 75% of the restaurants, the random model is indeed a better model than the cascade model. This suggests that Yelp reviews in many cases may not be influenced by previous reviews, and in fact represent independent observations of the truth of a restaurant. Only for a minority of restaurants does the cascade model represent the trends of review ratings well.

Works Cited:

S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *J. of Political Economy*, (5), 1992.

P. Desikan and J. Srivastava. Mining temporally evolving graphs. In WebKDD, 2004.

J. Leskovec, A. Singh, and J. Kleinburg. Patterns of Influence in a Recommendation Network. *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006.

Z. Wang. Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews. *The B.E. Journal of Economic Analysis and Policy*, 10(1) (Contributions), Article 44, 2010.