

The Effect of Product Ratings on Viral Marketing

CS224W Project Milestone Report

Stefan P. Hau-Riege, stefanhr@stanford.edu

In network-based marketing, social influence is considered in order to optimize marketing strategies. Social influence in e-commerce has been studied for more than one and a half decades, to a large extent in the context of product endorsements and chains of recommendations. In this work, we study friends' influence in social networks exerted through reviews that include rankings, which means that a review can be anything from a recommendation to disapproval. We use the data from the recent Round 8 Yelp Dataset Challenge and determine the degree to which friends influence the decision to choose and review businesses, broken down by the ranking of the business. We discuss the effectiveness of different figure of merits for this analysis, including the exposure, contagion count, and spreading concentration. We find that encouraging influence by friends is more likely to occur for highly rated businesses than for poorly rated ones. In the final report, we will add a discussion on the extent to which model-generated data corroborate these results for varying degrees of influence.

I. Introduction

An important economic issue for companies is the desire to minimize costs of marketing while maximizing expected profits. Different types of marketing tactics are usually considered, including mass marketing, direct marketing, and network-based marketing. For mass marketing, products are promoted indiscriminately to all potential customers. In contrast, for direct marketing, products are promoted to customers which are likely to be profitable, based on characteristics of individual people. This approach is particularly useful for advertising niche products which, somewhat surprisingly, collectively constitute a significant fraction of the online sales volume. Finally, for network-based marketing, social influence is considered. In this case, customers themselves are leveraged to perform most of promotional effort. Network-based marketing is a more recent and very active field of research. It exhibits similarities to the spread of diseases and is therefore often also referred to as viral marketing. Successful viral marketing is related to the problem of maximizing the spread of influence on a social network, a topic which has been discussed extensively in the course, and which we examine further in this report.

There are multiple key papers in this field, three of which we are discussing in more detail, spanning a time frame of one and a half decades. The oldest paper we consider was authored by Domingos and Richardson in 2001, entitled *Mining the network value of customers* [1]. It is one of the first works to quantify network effects in customer valuation models. Before their pioneering work, it was presumed that consumers make their purchasing decisions individually, independent of other customers and word of mouth. We then consider the paper by Leskovec, Adamic, and Huberman, entitled *The dynamics of viral marketing*, published in 2007 [2]. It builds on the work by Domingos *et al.* and presents an analysis of a person-to-person recommendation network with 4 million people. Finally, we discuss the paper by Vidmer, Medo, and Zhang, entitled *Unbiased metrics of friends'*

influence in multi-level networks, published in 2015 [3]. It critiques the paper by Leskovec *et al.* and points out that large differences in item popularities can cloud the analysis of the influence of friends.

None of the papers that we have mentioned so far considers the nature of the influence explicitly, i.e. whether a customer endorses a product or not. This motivates the goal of this project to extend their work and explore the effect of the degree of enthusiasm in recommendations and reviews on friends' influence. Subsequent to this introduction, we will review previous work, followed by a description of our approach to investigate this issue. We will then proceed to discuss our results by first providing summary statistics of the dataset and then elaborating on the friends' influence in the Yelp dataset. In the final report, we will then discuss friends' influence in our model, but this is not part of this milestone report. Finally, we will conclude this work and provide the references.

II. Review of previous work

The expected value of marketing to a potential customer has two components: The *intrinsic value*, which is the expected profit from sales directly to him/her, and the *network value*, which is the expected profit from sales to other customers that are directly or indirectly influenced by him/her.

A. Domingos' *Mining the network value of customers*

Domingos *et al.* put the goal of finding the best viral marketing strategy on a more solid foundation by being the first to attempt to quantify the network value of a customer. With this information, they answer the fundamental algorithmic question as to which individuals should be targeted to trigger large cascades of product adoption. Domingos *et al.* propose to model social networks as Markov random fields, where each customer's probability of purchasing a product is a function of both the intrinsic desirability of the product for the customer and the influence of other customers [1]. A second major

part of the paper deals with the construction of a framework to mine the required network information. They take a collaborative filtering database as an example for a data source used for mining influence networks.

B. Leskovec’s *The dynamics of viral marketing*

Leskovec *et al.* study the application of disease contagion models to recommender systems. They analyze a person-to-person recommendation network and, for the first time, directly quantify how influential person-to-person recommendations really are by empirically studying an online retailer’s incentivized viral marketing program. This incentive program works by allowing customers that purchase a product to send email advertisements (called recommendations in the paper) to their friends. The first person to purchase the same product through such a referral within a week earns a 10% discount, while the recommender obtains a 10% credit on their purchase.

The authors are able to draw several important conclusions: First, the number of additional purchases due to recommendations is generally small, suggesting that viral marketing is not as epidemic as vendors have hoped. Further, the likelihood that a product is purchased by a customer increases with the number of recommendations. However, this is true for only the first few recommendations and then quickly saturates at a relatively low probability. This is inconsistent with traditional epidemic and innovation diffusion models. Also, when the number of recommendations between two customers increases, the recommendations are more likely to be ignored (diminishing returns). This means that high-degree nodes tend not to be effective recommenders. Finally, they observe homophily, which is the tendency of similar customers to purchase similar products.

In addition to this analysis, Leskovec *et al.* propose a model to identify products for which viral marketing is likely to be effective. The model predicts the general tendency that recommendation chains are relatively short, and that only a few larger cascades occur.

C. Vidmer’s *Unbiased metrics of friends’ influence in multi-level networks*

Vidmer *et al.* show that currently-used measures for the influence of friends are biased when very popular products are present. They propose three metrics to better measure the influence of friends in networks:

1. A *normalized exposure* to better distinguish the influence of friends from preferential attachment,
2. A *normalized contagion count* to correct contagion count for highly-active users, and
3. A *spreading coefficient* to measure topological features of item spreading.

Vidmer *et al.* proceed to apply these three metrics to datasets from Yelp and Digg.com. To assess the usefulness of these metrics in measuring the influence of friends, they compare their values in actual graphs with values obtained in randomized graphs for which the influence of friends, supposedly, has been removed. They find that their proposed metrics perform better at distinguishing original and randomized networks than conventional metrics.

To further drive home this point, Vidmer *et al.* develop a simple artificial network model based on preferential attachment which allows setting the degree of friends’ influence explicitly. Again, they are able to demonstrate the superiority of their new metrics compared to conventional metrics to assess the influence of friends.

D. Critique

Neither Domingos *et al.* nor Leskovec *et al.* have access to a measure for consumer satisfaction; customers do not rank their purchases, and so it is not clear if they really recommend a product to their friends. Domingos *et al.* derived a network from collaborative filtering without ranking products, and Leskovec *et al.* base their endorsement on an incentive program which strongly biases the recommender to promote a product to his/her friends, even if the product has not even been received yet. The Yelp dataset used by Vidmer *et al.* contains explicit ranking information but the authors chose to ignore it. Clearly, it is expected that a negative review is not likely to ensue a cascade of purchases. In this project, we critically build on the work by Vidmer *et al.* and include the ratings of businesses from the Yelp dataset. We explore the effectiveness of various summary statistics in analyzing rating-specific friends’ influence on customer decision.

III. Approach

A. Dataset

We use the data from the recent Round 8 Yelp Dataset Challenge [4] as it is particularly suitable for this work. It features a user-user social network with friendship edges. In addition, it provides a list of businesses as well as reviews written by the users, from which a user-business bipartite graph can be constructed. In this graph, users and businesses form the nodes, and the reviews are the edges. The reviews include one-to-five-stars ratings of the businesses.

The original dataset contains 2.7M reviews, 687K users, and 86K businesses, along with 566K business attributes. The original friends network contains 687K users and 4.2M social edges. We retain only users with at least one friend and only businesses with at least one review. Also, for uniformity, we limit our analysis to restaurant businesses, which constitute the vast majority. This results in 1.1M reviews, 219K users, and 26K businesses. The modified friends network contains 219K users and 1.6M social edges.

B. Analysis methods

We use various metrics to analysis friends' influence, including exposure, contagion counts, and spreading concentration. We use data randomization techniques to verify the degree to which the influence of friends matters. To this end, we use the configuration model to randomize the user-business network while keeping the user-user network the same. This means that the set of businesses that a user reviews is randomized while the degrees of both the items and the users stay unchanged. This way, the influence of friends has been removed from the network.

C. Model discussion

Vidmer *et al.* proposed a model which allows us to vary the friends' influence continuously. It keeps the original Yelp user-user network but generates an artificial user-business network by creating global and local (friend) review links stochastically. The global influence is modeled with a preferential attachment model with decaying relevance, and the local (friends') influence is assumed to be proportional to the number of friends $f_i^\alpha(t)$ of user i that have reviewed business α at time t . The algorithm is shown in 1. Here, n_u is the number of Yelp users, n_b is the number of businesses, k_i^{Yelp} is the number of friends of user i in the Yelp network, $k[\alpha]$ is the current number of reviews of item α , m is the ratio of the number of Yelp reviews and n_b , and $R_\alpha(t) = R_0[\alpha] \exp(-\beta(t-t_\alpha))$, where $R_0[\alpha]$ is drawn from the exponential distribution, and t_α is the time business α was added. The original algorithm by Vidmer *et al.* constrained $\nu = 1$, but we allow $\nu \neq 1$ to be able to model both stronger ($\nu > 1$) and weaker ($\nu < 1$) influences of high-degree nodes. We fit this model to the Yelp user-business networks, separated by review ratings, to match user and business degree distributions. Comparisons of the model fitting parameter values provides further insight into the user-business network. This artificial network model can be used to determine the performance of the new metrics for friends' influence.

Algorithm 1 Create user-business bipartite graph

```

1: procedure MODELUSERITEMGRAPH
2:   input  $p, \nu, \lambda, \alpha, k_i^{\text{Yelp}}, n_u, n_b$ 
3:   for  $t \leftarrow 0, \dots, n_b - 1$  do
4:     pick  $R_0[t]$  from PDF  $p(x) = \lambda e^{-\lambda x}$ ,  $x \in [1, \infty)$ 
5:      $k[t] = 1$ 
6:     for  $l \leftarrow 0, \dots, m - 1$  do
7:       pick a user  $u$  with PDF  $p(u) \propto k_i^{\text{Yelp}}$ 
8:       with probability  $p$ 
9:       pick business  $\alpha$  with PDF  $p_l(\alpha) \propto f_i^\alpha(t)$ 
10:      else  $\triangleright$  with probability  $(1-p)$ 
11:        pick business  $\alpha$  with PDF  $p_g(\alpha) \propto (k_\alpha[t])^\nu R_\alpha(t)$ 
12:      if review  $(u, \alpha)$  does not exist yet then
13:        add review  $(u, \alpha)$  with timestamp  $t$ 
14:         $k[\alpha] \leftarrow k[\alpha] + 1$ 

```

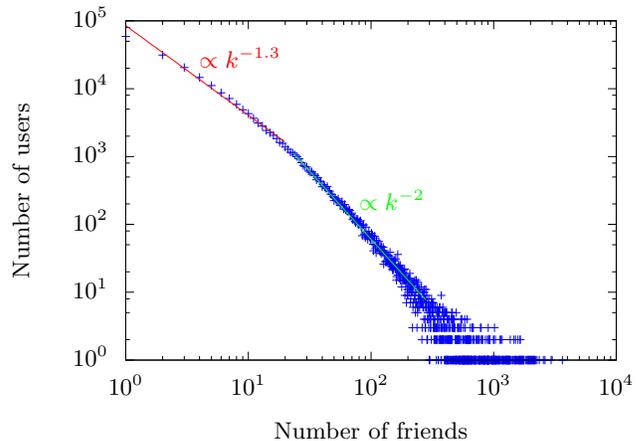


FIG. 1. Degree distribution in the user-user network.

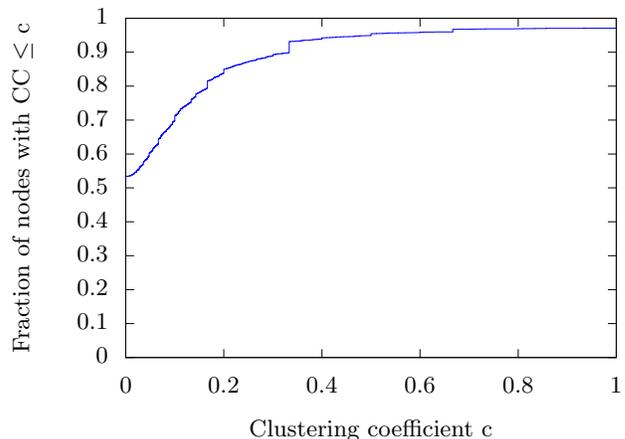


FIG. 2. Distribution of clustering coefficients in the user-user network.

IV. Results and discussion

A. Summary statistics

1. User-user network

Fig. 1 shows that the user-user social network exhibits a power-law degree distribution of node degrees, i.e. $P(k) \propto k^{-\gamma}$, so it is a scale-free network. For our dataset, γ ranges from 1.3 to 2.0. Fig. 2 shows the distribution of clustering coefficients in this user-user network on a cumulative plot. We find that more than half of the users have a clustering coefficient of zero, in agreement with the relatively weak friend influence which we will discuss later on. Fig. 3 shows the average clustering coefficient as a function of the node degree, i.e. the number of friends. As it is typical for a scale-free network, the clustering coefficient decreases with increasing degree. We further found that 97% of the users belong to a single weakly-connected component, which is nearly the whole network.

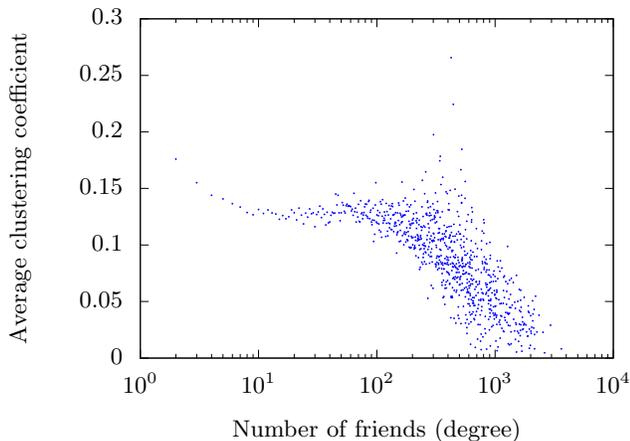


FIG. 3. Distribution of clustering coefficients in the user-user network as a function of node degree (number of friends).

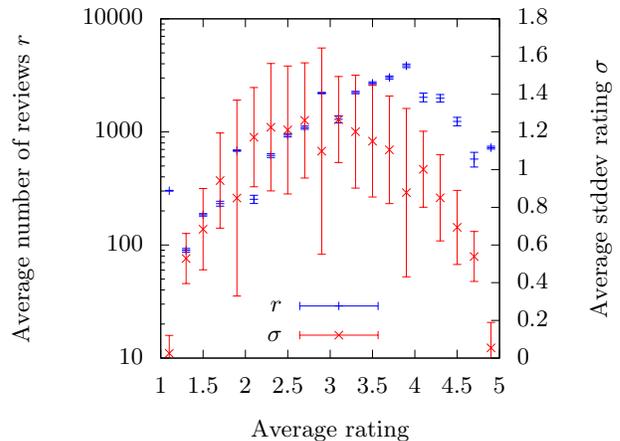


FIG. 5. Average number of reviews and average standard deviation of the ratings as a function of the average rating.

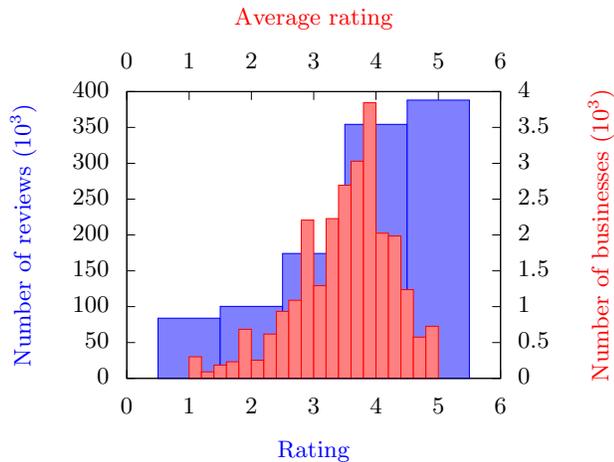


FIG. 4. Histogram of individual ratings (blue) and average business ratings (red).

2. Business reviews

A user can submit a review for a business with a rating of one to five stars. Fig. 4 shows a histogram of the individual ratings in blue. The vast majority of votes have 4 and 5 stars. From that information we can compute the average rating of each business. The distribution of these average ratings is also shown in Fig. 4 in red. The most likely average rating is around 4. Fig. 5 shows the distribution of the average number of reviews as a function of the average rating. Businesses with an average rating around 4 obtain the most reviews. Also shown in Fig. 5 is the average standard deviation of the business rating as a function of the average business rating. It measures the controversy over the rating of a business. As expected, extreme ratings have the smallest standard deviation because, by design, they require good agreement. The standard deviation for businesses with average ratings around 3 is larger.

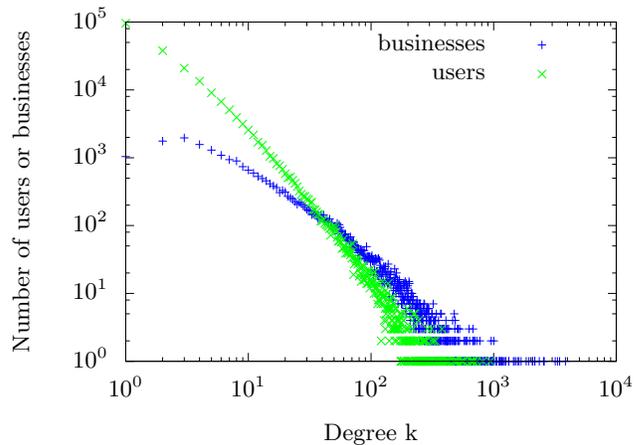


FIG. 6. Degree distribution of businesses and users in the user-business network.

3. User-business network

Fig. 6 shows the degree distribution of businesses and users in the bipartite user-business network. Like the user-user social network, the node degrees follow a power-law distribution in this network too. As we discussed in Section III C on our modeling approach, such a network can be generated by preferential attachment. The number of businesses is larger than the number of users, so the business degrees are larger than the user degrees.

Fig. 7 shows the distributions of average business ratings. We group the businesses by their average ratings as shown in Table I throughout the paper. For example, Fig. 8 shows the degree distribution of businesses in user-business network for various rating groups. Businesses with a poor rating in the range [1, 3] tend to have fewer reviews than businesses with better ratings in the range (3, 5]. A likely explanation is that customers are deterred from frequenting businesses with poor ratings.

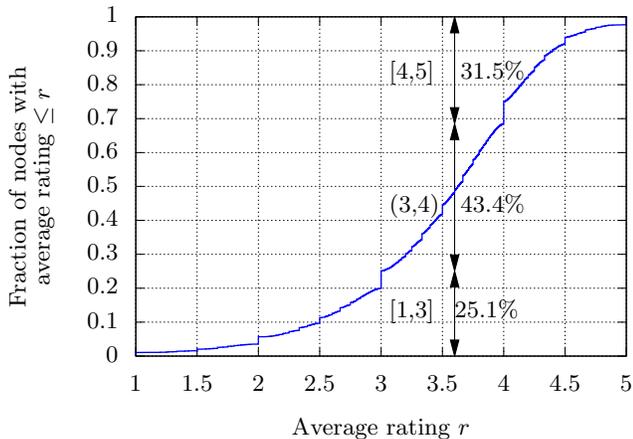


FIG. 7. Distribution of business ratings.

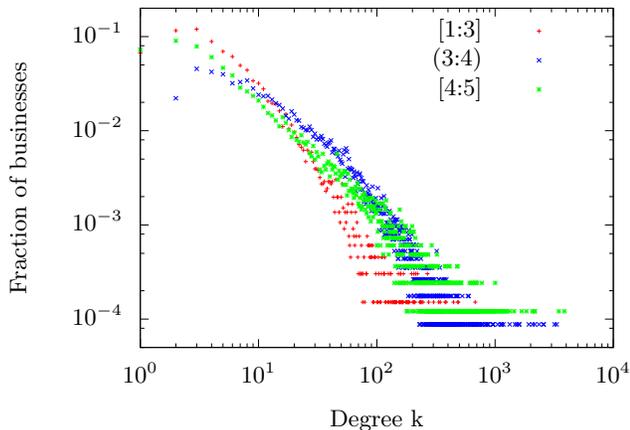


FIG. 8. Degree distribution of businesses in the user-business network for various rating groups.

B. Friends' influence in the Yelp dataset

We are using multiple metrics to measure the effect of ratings on friends' influence in the Yelp dataset.

1. Exposure

The first metric we consider is related to the exposure $N_{i\alpha}$, defined as the number of friends of user i that have reviewed business α . Fig. 9 shows an estimate of the probability $P_{\text{review}}(N)$ that a user i will review business α , given that $N_{i\alpha} \geq N$. We estimated this probability for all $N_{i\alpha} > 0$ by counting the number of people m that were exposed to at least N reviews and then reviewed business α , and by counting the number of people n that were exposed to at least N reviews and did not write a review of business α over the whole observation period. We then estimate $\hat{P}_{\text{review}}(N) = m/(m+n)$ [2]. As expected, the review probability increases with increasing N and eventually saturates, as it has been observed before by others [2]. The new observation which we have

Range of average ratings	Number of businesses	Fraction of businesses
[1, 5]	26,239	100%
[1, 3]	6,597	31.5%
(3, 4)	11,387	43.4%
[4, 5]	8,255	25.1%

TABLE I. Number of businesses in different average rating groups.

made is that the review probability increases with increasing rating values, suggesting that positive reviews are an incentive in particular for *friends* to visit and review businesses with high ratings.

Fig. 9 also shows that when we randomize the user-business network, the review probability is significantly smaller than for the unperturbed network, at least for $N \leq 15$. For larger N , the probability keeps increasing. Vidmer *et al.* attribute this behavior to the heterogeneity of item popularity: The probability for a user i to review business α increases with increasing $N_{i\alpha}$ because for a randomized user-business network, $N_{i\alpha} \propto f_i k_\alpha / U$, where f_i is the number of friends of user i , and U is the number of users. So for larger $N_{i\alpha}$, k_α is larger, and, assuming preferential attachment, the review probability increases. To circumvent this issue, they propose to use the *normalized exposure* instead, defined as

$$n_{i\alpha} = \max_{t < t_{i\alpha}} \left(\frac{f_{i\alpha}(t)}{f_i} \frac{k_{\min}}{k_\alpha(t)} \right), \quad (1)$$

where $t_{i\alpha}$ is the time user i reviewed business α . To limit noise effects for small f_i and k_α , we consider only businesses with currently at least $k_{\min} = 10$ reviews and users with at least $f_{\min} = 5$ friends. Fig. 10 shows the review probability as a function of the normalized exposure. In agreement with our expectations, the probability increases with n for ratings between 3 and 4. Also, better than for the metric based on the exposure N , the review probability decreases with increasing n for poor ratings, and the randomized user-business networks generally show lower probabilities. It is somewhat surprising that the review probability decreases with increasing n for very good reviews between 4 and 5.

2. Spreading concentration

If friends' influence is of significance in increasing the number of reviews of businesses, then we expect that the business is more likely to be spread within clusters of friends than between users without a friends relationship. We characterize this behavior by using the spreading concentration $r_\alpha \in [0, 1]$, defined as the ratio of the number of friends who have reviewed business α over the total number of friends,

$$r_\alpha = \frac{1}{k_\alpha} \sum_{i \in \mathcal{N}(\alpha)} \frac{\sum_{j \in \mathcal{F}(i)} a_{j\alpha}}{f_i}. \quad (2)$$

Here, $\mathcal{N}(\alpha)$ is the set of users that reviewed business α , $\mathcal{F}(i)$ is the set of friends of user i , and $a_{j\alpha}$ is 1 if user j has

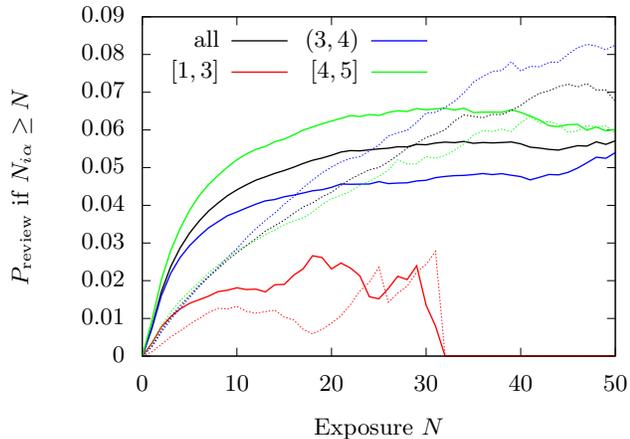


FIG. 9. Probability that a user writes a review if his/her exposure (the number of friends who have written a review $N_{i\alpha}$) is at least N . The dashed lines show the corresponding results for the randomized data.

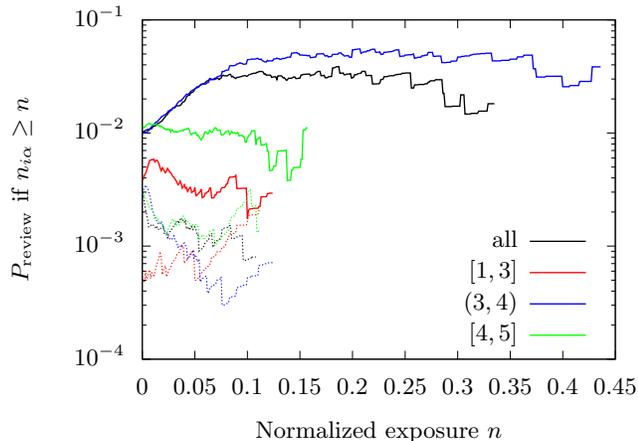


FIG. 10. Probability that a user writes a review if his/her normalized exposure $n_{i\alpha}$ is at least n . The dashed lines show the corresponding results for the randomized data.

reviewed business α and 0 otherwise. Fig. 11 shows the spreading concentration for various rating groups. For about 40% of the businesses the spreading coefficient is zero, explaining that the cumulative curves appear to start at a fraction of 0.6 instead of 1.0. For $r < 0.2$, the spreading coefficient for businesses with poor ratings in the range of 1 to 3 is significantly lower than for businesses with higher ratings. For larger r values, businesses with higher ratings in the 4 to 5 range have the largest spreading coefficients. This is direct evidence that well-rated businesses are more likely to be reviewed by the immediate friends, suggesting stronger friends' influence than for poorly rated items. Randomizing the user-business bipartite network leads to very small values for the spreading coefficients, further suggesting that the concept of spreading coefficient is useful.

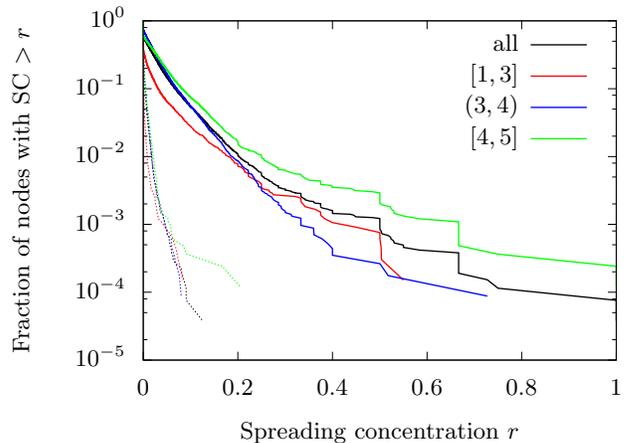


FIG. 11. Cumulative plot of spreading concentration for various rating groups. The dashed lines show the corresponding results for the randomized data.

3. Contagion count

We define the contagion count n_{ij} as the number of businesses first reviewed by user i and later reviewed by user j , where i and j are friends. A large contagion count suggests that user i has influence over user j . Fig. 12 shows a cumulative plot of contagion counts for various rating groups. In this plot, we normalized the contagion count by the number of businesses n_b in each category. The ordering of the normalized contagion count curves coincides with the ordering of average ratings, which is understandable because if a friend is giving a poor review for a business, it is unlikely that a close friend will try or review that restaurant (and vice versa). When we randomize the user-business network, the contagion counts drop significantly, as can also be seen in Fig. 12, demonstrating that this is not primarily a global but a local (friends' influence) effect.

Vidmer et al. suggested another normalization of the contagion count, based on the idea that if two friends are very prolific in writing reviews, then it is more likely that they have two businesses in common than for less active users. Based on this idea, they proposed to normalize n_{ij} by the expected number of businesses reviewed by both users i and j in the fully random case,

$$c_{ij} = \frac{k_{ij}}{k_i k_j} k_{\min}, \quad (3)$$

where only users are considered that have reviewed at least k_{\min} businesses. A cumulative plot of c_{ij} is shown in Fig. 13, where we again normalize these values by the number of businesses in the different rating categories. The counter-intuitive ordering of these curves suggest that the normalized contagion count suggested by Vidmer *et al.* is not very useful.

C. Friends' influence in model datasets

This section will be added in the final report.

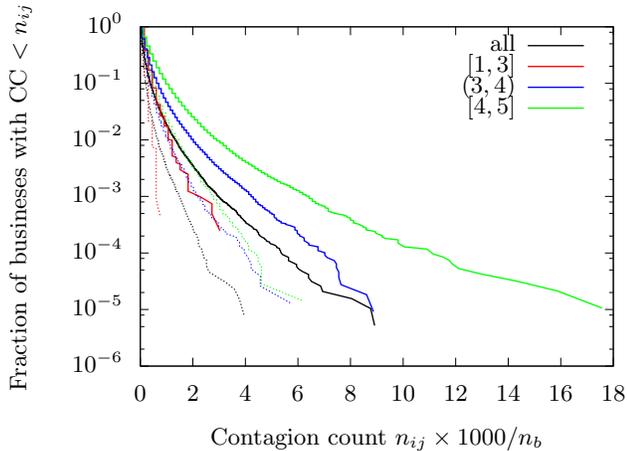


FIG. 12. Cumulative plot of contagion counts n_{ij} normalized by the number of businesses n_b in the various rating groups. The dashed lines show the corresponding results for the randomized data.

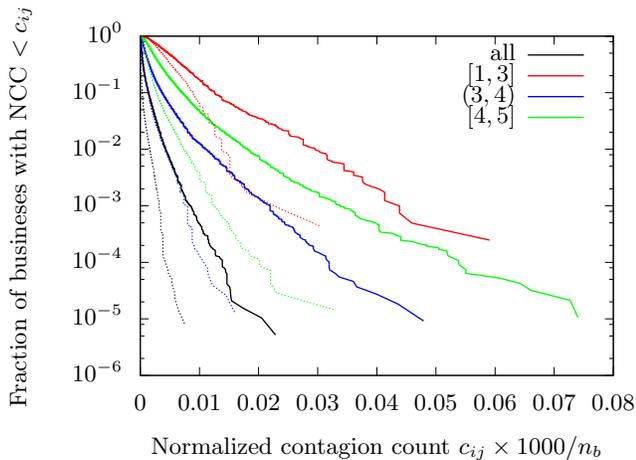


FIG. 13. Cumulative plot of normalized contagion count for various rating groups. The dashed lines show the corresponding results for the randomized data.

V. Summary and conclusions

Using the data from the recent Round 8 Yelp Dataset Challenge, we determined the degree to which friends influence the decision to choose and review businesses,

broken down by the average rankings of the businesses.

Our summary statistics showed that the Yelp user-user social network is a scale-free network with mostly small clustering coefficients, in agreement with a relatively weak friend influence. The business reviews are generally skewed to larger ratings, and businesses with poor ratings in the range $[1, 3]$ tend to have fewer reviews than businesses with better ratings in the range $(3, 5]$, presumably because customers are deterred from businesses with poor ratings.

To assess friends' influence for businesses with different ratings, we explored the effectiveness of various metrics, including the exposure, contagion count, and spreading concentrations. Coupling this to data randomization techniques allowed us to determine the degree to which the influence of friends matters. We found that the review probability increases with increasing exposure, but, counter-intuitively, the probability keeps increasing for the randomized network. This shortcoming is remedied by introducing a normalized exposure. The spreading concentration gives direct evidence that well-rated businesses are more likely to be reviewed by immediate friends, suggesting stronger friends' influence than for poorly rated items. Randomizing the user-business bipartite network leads to very small values for the spreading coefficients, further suggesting that this is a well-defined quantity. The contagion count normalized by the number of businesses shows that for businesses with good ratings, it is likely that a close friend will try and review that restaurant, as well (and the other way around). The normalized contagion count proposed by Vidmer *et al.*, on the other hand, gives unreasonable results.

In summary, we found that it is advantageous to categorize reviews by ratings to understand the influence on social marketing better. We found that friends' influence is significantly stronger for businesses with high ratings than for poorly rated ones. However, in absolute numbers, viral marketing is still not as epidemic as businesses might have hoped.

VI. Future work

In the final report, we will discuss if model-generated data corroborate these results for varying degrees of influence. We are going to use the model described in Section III C.

[1] P. Domingos and M. Richardson, *Mining the Network Value of Customers*, In Proc. KDD, 2001.
 [2] J. Leskovec, L. Adamic, and B. Huberman, *The Dynamics of Viral Marketing*, TWEB, 2007.

[3] A. Vidmer, M. Medo, and Y.-C. Zhang, *Unbiased metrics of friends' influence in multi-level networks*, EPJ Data Science 4 (20), 2015.
 [4] https://www.yelp.com/dataset_challenge