Poster: Towards Detecting Yelp Review Campaigns

Jaime Ballesteros (student)*, Mahmudur Rahman (student)*, Bogdan Carbunar*, Rahul Potharaju (student)†, Radu Sion‡, Nitya Narasimhan§, Naphtali Rishe*
*School of Computing and Information Sciences, Florida International University, Miami, FL
†Department of Computer Science, Purdue University, West Lafayette, IN
‡Department of Computer Science, Stony Brook University,Stony Brook, NY
§Motorola Mobility, Libertyville, IL

I. Introduction

Recently launched geosocial networks (GSNs) such as Yelp [1] and Foursquare [2] extend review-centered sites (e.g., Amazon, TripAdvisor) with social dimensions. Subscribers own accounts where they store public profiles, use them to befriend and maintain contact with other users and provide feedback, in the form of reviews, for visited venues.

With tens of millions of reviews and monthly unique visitors, review based GSNs are playing an increasingly influential part in our lives. Their popularity and impact makes malicious behavior, in the form of fake reviews, a threat not only to their credibility but also to the quality of life of their users. While the review writing process is not rewarded financially, a direct relationship exists between reviews and financial gain: Anderson and Magruber [3] show that in Yelp, an extra half-star rating causes restaurants to sell out 19 percentage points (49%) more frequently.

The impact of the occasional malicious review is likely to be minimal among many reviews. Instead, the goal of this work is to detect *review campaigns*, concerted efforts to bias public opinion: entities that hire groups of people to write fake reviews and dishonestly improve or damage the ratings of target venues.

II. Background

System Model. We model the system after Yelp [1]: The provider hosts information about (i) venues, representing businesses or events with an associated location, e.g., restaurants, shops, concerts, etc, and (ii) user accounts. User accounts store information about friends and reviews written by the user. Besides text, reviews have a numerical component, a *rating* ranging from 1 to 5, with 5 being the highest mark. Yelp associates an *average rating* value for each venue, computed over all the ratings of reviews left by users.

Collected Data. We have randomly collected information from 16,199 venues and 10,031 users from the Yelp website. For each user we have collected the id, location, number of friends and all reviews, for a total of 646,017 reviews. For each venue we have collected its name, location and type, along with all the reviews received, for a total of 1,096,044 reviews.

III. Review Campaigns: A Real Problem?

To verify the feasibility of review campaigns, we created (providing only a name, type and location) three fake Yelp venues, two "located" in Miami (FL) and one in Portland (OR). Following a "verification" step that lasted a few hours, the venues were indexed on Yelp. Subsequently, we posted several HITs (Human Intelligence Tasks) on Amazon Mechanical Turks (MTurk) [4], asking workers to write reviews for one of our venues. The venues are fake, thus so are the reviews. At the completion of the experiment, the venues registered 62 reviews.

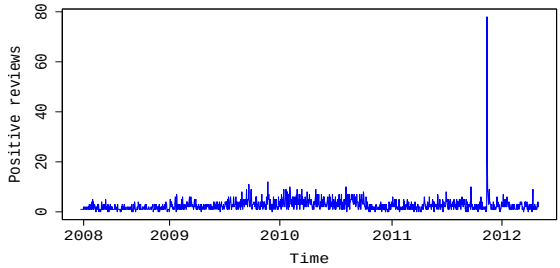

Fig. 1. Venue timeline with positive (4 & 5) reviews.

IV. Detecting Review Campaigns

User ratings. We consider three rating types: positive ($R_i = 1$, for a star rating of 4 or 5), negative ($R_i = -1$, for a star rating of 1 or 2) and neutral ($R_i = 0$, for 3 star rating reviews). $\bar{R}_i$, the average rating of V_i at time T_i, can also take one of three values: -1 if the average rating of venue V_i is below 3, 0 if equal to 3, and +1 if above 3.

We introduce then notion of user ratings, to help identify low quality reviewers. High ratings identify users with expertise in the areas they review, that have active friends. In the following, we call a review *active* if its rating is not neutral (3 stars). Let h denote the number of reviews written by a user U and let $h_a \leq h$ be the number of active reviews of U. Let T_r and T_f be systems parameters defined later, experimentally. We define the *expertise* Exp_V of a user U for a reviewed venue V, as the ratio of the number of reviews written by U in the vicinity (50 miles radius) of V to the number of active reviews of U. Let f_0 denote the number of friends of U that have at least T_r reviews. We define the *rating* of U as:

$$R_U = \begin{cases} 0, & if\ (h < T_r \wedge f_0 < T_f) \\ \frac{\sum_{i=1}^{h_a} sgn(|R_i + \bar{R}_i|) Exp_i}{h_a}, & otherwise \end{cases} \quad (1)$$

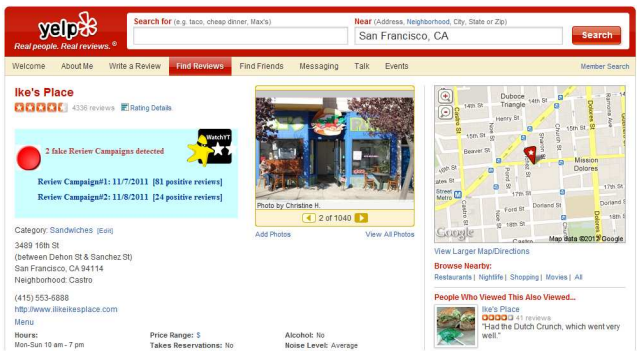

Fig. 2. Snapshot of WatchYT 's plugin functionality for the venue "Ike's Place". The blue rectangle contains the output of SpiDeR .

where sgn is the sign function and Exp_i is U's expertise for the i-th reviewed venue. $sgn(|R_i + \bar{R}_i|)$ can only be 0 or 1. Thus, the rating of a user is defined to be 0 if the user has written less than T_r reviews and has less than T_f friends with at least T_r reviews each. Otherwise, the user's rating is a weighted average (over the length of its active history) of the user's concordance with her reviewed venues' average rating. We observe that $R_U \in [0, 1]$.

A. SpiDeR : Spike Detection Ranges

We define timeline of a venue V to be the set of tuples $H_V = \{(U_i, R_i, T_i)|i = 1..v\}$, the chronological succession of reviews R_i written for V by users U_i at time T_i. We exploit the observation that in order for a review campaign to have an impact on the aggregate rating of a subject, it needs to contain sufficient numbers of reviews. Figure 1 shows the evolution in time of the number of positive reviews (4 and 5 star) for a venue called "Ike's Place" in San Francisco, CA [5], exhibiting an abnormal number (78) of positive reviews on Nov. 7, 2011.

We use Box-and-Whisker plots [6], relying on quartiles and interquartile ranges (IQRs), to detect such outliers: Given a venue V, we first compute the quartiles and the IQR of the positive reviews from V's timeline H_V (negative reviews are handled similarly). We then compute the upper outer fence (UOF) value using the Box-Whiskers plot [6]. For each day d during V's active period, let P_d denote the set of positive reviews from H_V written during day d. If $|P_d| > UOF$, we output P_d, i.e., a spike has been detected. For instance, "Ike's Place" has a UOF of 9 for positive reviews: any day with more than 9 positive reviews is considered to be a spike.

We introduce SpiDeR (Spike Detection Ranges), an algorithm that identifies review campaigns by combining detected review spikes with user ratings. For each spike, SpiDeR counts the number of reviews written by users with low ratings. It flags as suspicious spikes made up by more than a threshold percentage T_p of reviews written by reviewers whose rating is below a threshold w_r.

V. Evaluation

WatchYT Implementation: We have prototyped SpiDeR as part of our system WatchYT (Watch Yelp Timelines), that we made publicly available [7]. Figure 2 shows the output of WatchYT for the venue "Ike's Place".

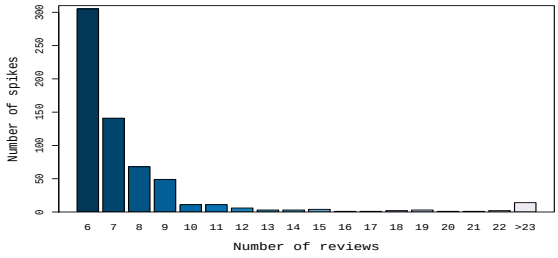

Fig. 3. Distribution of review spikes when using Box-and-Whisker plots.

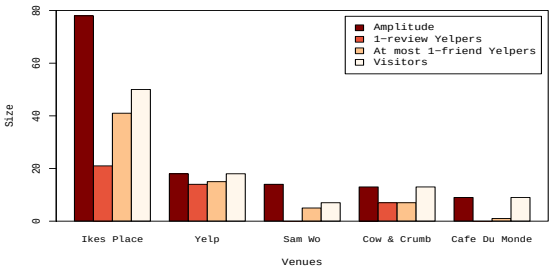

Fig. 4. SpiDeR output. Zoom-in of Figure 3.

SpiDeR Evaluation.: Figure 3 shows the output of the Box-and-Whisker plot detection technique (see Section IV) when applied to the positive reviews of the 16,199 venues collected across the U.S.: the distribution of the amplitude (the number of reviews) of the spikes detected. It shows that the amplitude has a long-tail distribution. Figure 4 zooms in into several spikes from Figure 3, showing, for each venue, the spike's amplitude, the number of reviewers that have only one review, the number of reviewers that have at most one friend and the number of out-of-town reviewers. Thus, SpiDeR detects the spikes of these venues even with low parameter values: $T_r = 2$, $T_f = 2$, $w_r = 0$ and $T_p = 0.25$.

VI. Related Work

Of notable importance is the work of Ott et al. [8] which focuses on the text of fake reviews in TripAdvisor. Our research relies on social and geographic dimensions to address the same issue in Yelp. However, unlike TripAdvisor, Yelp provides us with access to locations and friends of reviewers.

References

[1] Yelp. http://www.yelp.com.
[2] Foursquare. https://foursquare.com/.
[3] Michael Anderson and Jeremy Magruder. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *Economic Journal*, 122(563):957–989, 2012.
[4] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome.
[5] IKE's Place. http://www.yelp.com/biz/ikes-place-san-francisco.
[6] A. C. Tamhane and D. D Dunlop. *Statistics and data analysis: From elementary to intermediate*. Upper Saddle River, NJ: Prentice Hall, 2000.
[7] WatchYT: Watch Yelp Timelines. http://users.cis.fiu.edu/~mrahm004/watchyt.
[8] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Human Language Technologies*, HLT '11, 2011.