Yelp Events: Making Bricks Without Clay?

Jaime Ballesteros, Bogdan Carbunar, Mahmudur Rahman, Naphtali Rishe
Florida International University, Miami, FL

Abstract—**Review based geosocial networks are online social networks centered on the location of venues and users as well as on reviews left by users for visited venues. The popularity and impact of reviews makes them an ideal tool for influencing public opinion. In this paper we study the effects of Yelp Elite events, organized for the benefit of Elite reviewers, on the image of the hosting venues. To this end, we introduce tools for identifying venues receiving abnormally large numbers of reviews in a short time and use them to detect correlations between events and hosting venues. We have implemented a browser plugin that makes users aware of Yelp event manipulations. We use data we collected from Yelp to show that Elite events have a noticeable short-term impact on the rating of hosting venues.**

I. Introduction

Geosocial Networks (GSNs) extend online social networks with the concept of venues, special locations where mobile device users can register their presence. In GSNs such as Yelp [1], Foursquare [2] and Urbanspoon [3], registered venues have accounts where users can leave feedback, in the form of reviews or tips for other users. Reviews have a numerical component, a *star rating*, and venue accounts aggregate their reviews into a single star rating value.

With tens of millions of reviews and monthly unique visitors [4], [5], review based GSNs are playing an increasingly influential part in our lives. As such, they become alluring targets for attacks aiming to bias the public image of venues. Previous work (e.g., [6], [7], [8], [9]) has shown that fake reviews can be commissioned to improve the rating of products and services. The incentive is profit: Anderson and Magruber [10] show that in Yelp, an extra half-star rating causes restaurants to sell out 19 percentage points (from 30% to 49%) more frequently.

In this paper we focus on Yelp [1], a unique geosocial network, and one of its tools, Elite events. Elite events are organized by Yelp for the benefit of "Elite" reviewers: users who not only have many reviews and friends, but enjoy popularity among other users. Yelp creates a new Yelp page with the name of the event and encourages attendees to review the event venue [11]. While the declared goal of the event venues is to prevent unfairness to venues that do not host events, in this paper, we study the impact of events on the venues hosting them. The question we ask is whether Elite events help improve the short and long term venue ratings. If such events have a positive effect, we believe they can be used as an alternative to fake reviews.

Our approach relies on the notion of positive venue *timelines*: the evolution in time of the number of daily positive reviews received by a venue. We use the venue timeline to identify abnormally high numbers of positive reviews received by the venue within a short time interval. This enables us to mark spikes that occur within a short timeframe of an event hosted by the venue. We then compute the impact of the event on the venue, as the difference between the average rating of the venue at a given time following the event and its rating before the event.

We have collected Yelp data from the accounts of more than 10,000 Yelp users, 16,000 venues and 149 events, for a total of more than 1.5 million reviews. We use this data to show that 40% of the venues hosting an event see a short term increase in their star rating (at least 0.5 stars increase).

The contributions of this paper are the following:

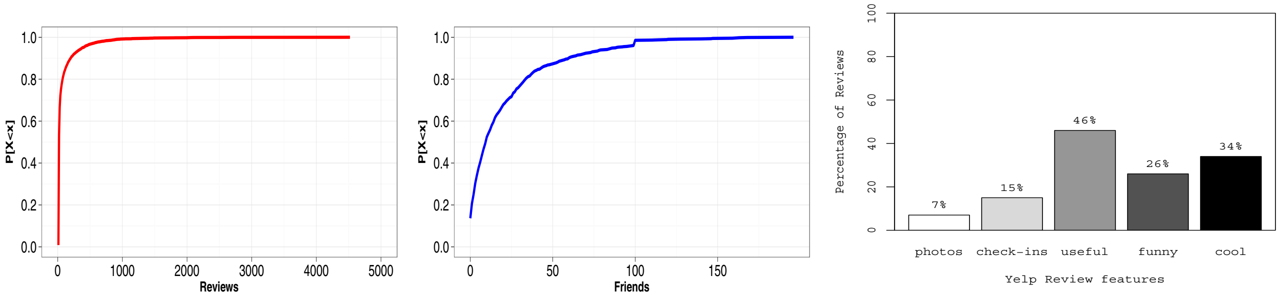
- Introduce SpiKeR , a tool for detecting abnormal numbers of positive reviews received by a venue within a short time interval.
- Analyze the short and long term impact of events on venues, using publicly accessible data collected from Yelp.
- Devise and implement WatchYT (available for download at the project's website [12]), a browser plugin that notifies Yelp users when browsing venues whose average rating has been influenced by events.

The remainder of the paper is organized as follows. Section II introduces the Yelp GSN model and basic Yelp statistics. Section III defines venue timelines and introduces the detection tools, SpiKeR and WatchYT. Section IV evaluates SpiKeR and WatchYT and Section V describes implementation details of WatchYT. Section VI presents related work and Section VII concludes.

II. System Model

We model the geosocial network following Yelp's [1] model. It consists of a provider, S, hosting the system along with information on businesses or venues registered, and serving a number of users. Users can subscribe and receive initial service credentials, including a unique user id.

The provider supports a given set of locations, defined in terms of discrete points-of-interests (POIs) or sites: restaurants, concerts, mechanic shops, etc. Users can report their location through *check-ins* at venues where they are present and can share this information with friends. Users are encouraged to leave feedback for the venues they visit, in the form of *reviews*. Reviews consist of a text describing the experience and a numerical component, a *rating* ranging from 1 to 5, with 5 being the highest mark. S associates an *average rating* value for each venue, computed over all the ratings of reviews left by users. Users can leave pre-defined feedback for other reviews i.e., "useful", "funny" and "cool".



(a)

(b)

(c)

Fig. 1. Yelp user stats: Distribution of (a) the number of reviews, (b) the number of friends. (c) Percentage of reviews with feedback.

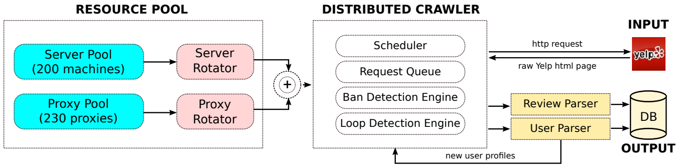


Fig. 2. Crawler architecture.

A. General Yelp Statistics

The crawler. We have developed a crawling engine to automatically collect data from Yelp's user and venue pages. The crawler uses a *resource pool* (see Figure 2) consisting of a set of servers and a set of proxies. For every request, the crawler randomly picks a server from the server pool and pairs it with a proxy from the proxy pool. The request is then made from the server, through the proxy. For each successful request, the crawler fetches the raw HTML page from Yelp and parses the required information. If the request is not successful, a new request is made using a different proxy. A centralized scheduler maintains a request queue to ensure there are no loops in the crawling process, i.e., avoids crawling the same page multiple times if referenced from several sources. When Yelp picks an anomalous proxy, any request made from this IP will return a blank HTML page or a page with error. Our crawler automatically detects this and changes the proxy. Furthermore, to minimize the load on Yelp's servers, and avoid detection, we introduce long inter-request intervals.

Crawling Yelp. In order to collect a representative sample of Yelp data, we used stratified sampling [13]. First, we selected a list of 10 major cities in the U.S. and we collected an initial random list of 100 venues from each of these cities as a seed dataset. It is important to understand that our strata (cities) are mutually exclusive, i.e. venues do not belong to two or more different cities. This way we avoid bias towards high degree nodes, which is a common problem when crawling social networks [14]. We then randomly selected 10,031 Yelp users who reviewed these venues, and collected their data, including their id, location, number of friends and all their reviews, for a total of 646,017 reviews.

Given the list of 10,031 collected Yelp users, we merged the lists of the venues reviewed by those users (to avoid duplicate venues) and we randomly selected 16,199 venues, including venues from cities outside the U.S. (e.g., London, U.K, Vancouver, CA, etc). 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. For each review we extracted the reviewer id, the date the review was written, the number of check-ins performed and the photos uploaded by the reviewer at the venue, as well as feedback received by the review itself (number of users who thought the review was "useful", "funny" or "cool").

Figure 1(a) shows the cumulative distribution function (CDF) of the number of reviews per user. While only 20% of users have more than 100 reviews, the record user has 4,000 reviews. Figure 1(b) shows the CDF of the number of friends per user. Only 15% of users have no friends but 50% of users have more than 10 friends. Furthermore, Figure 1(c) shows the percentage of reviews that have associated photos, check-ins and user feedback. While 15% of reviews have an associated check-in, a respectable 46% of reviews have been labeled as "useful". This shows that Yelp is an active social network, whose users widely embrace its rich features.

B. Yelp Events

Yelp rewards users that write popular reviews with a special, *Elite* badge status. The Elite badge is awarded to users who not only write many reviews and have many friends, but whose reviews receive significant recognition (e.g., feedback) from other users. The reviews of Elite yelpers are never filtered and are often shown at the beginning of a venue's Yelp page.

Yelp organizes special *Elite events*, at select venues, where only Elite badge holders are invited. For each event, Yelp creates a separate Yelp page, containing the name of the event and the name, address and information for the hosting venue. Attendees are encouraged to review the event account, which then lists the reviews, just like a regular venue.

C. Yelp Event Collection

We have collected Yelp events from 60 major cities covering 44 states of USA. The remaining states had no significant

Yelp events or activities (WY, VT, SD, NE, WV, ND). After identifying an Elite event, we identified the hosting venue through either its name or address. We used the crawler previously described to collect a majority of the available Yelp events and hosting venues, for a total of 149 pairs.

For each Yelp event and corresponding venue, we have collected their name, number of reviews, star rating and all their reviews. For each review, we have collected the date when it was written, the rating given and the available information about the reviewer, including the Elite status, number of friends and number of reviews written. In total, we have collected 24,054 event/hosting venue reviews.

D. Review Campaigns

The popularity and impact of Yelp [10] act as incentives for malicious behavior, in the form of fake reviews. A few fake reviews, possibly dispersed in time, as well as "neutral" (e.g., 3 star rating) reviews are likely to have a small impact on the average rating of a venue. Of particular concern then are the more impactful "review campaigns": entities that hire people to perform a ballot-stuffing (undeserved 4 and 5 star reviews) or bad-mouthing attack (1 and 2 star reviews) to alter the average rating of a target venue.

III. Yelp Campaigns

We conjecture that Yelp events can be used as review campaigns. Our hypothesis is based on several observations. First, the process of choosing the venues hosting Yelp events is not public. Second, a venue hosting an event is given ample warning to organize the event. Third, only Elite yelpers attend this event. While the attendees are encouraged to review the event's Yelp account, we have identified Yelp events that impacted the ratings of the corresponding host venues. We call such events, *Yelp campaigns*.

In the following, we introduce several tools we use to detect Yelp campaigns. In Section IV we use the tools to evaluate our conjecture.

A. Venue Timelines

We organize the 5 rating types (1-5 stars) available in Yelp into 3 categories: positive (for a star rating of 4 or 5), negative (for a star rating of 1 or 2) and neutral (3 star rating reviews). We associate a *timeline* with each venue. We define the timeline of a venue V, $H_V = \{(p_i, n_i, T_i) | i = 1..v\}$, to be the succession of daily (T_i) number of positive (p_i) and negative (n_i) reviews received by V. v denotes the number of days the venue V has been active, starting with the day when the venue has received its first review and ending with the current day (or the day when the venue was closed).

B. Identifying Yelp Campaigns

In order to verify our conjecture, we introduce SpiKeR , an algorithm for detecting abnormal reviewing behaviors, then WatchYT, a tool that detects correlations between Yelp events and review spikes in the hosting venues.

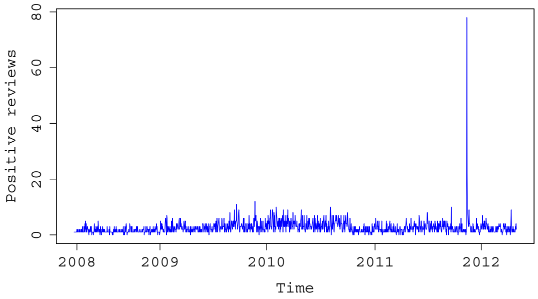


Fig. 3. Positive review venue histogram.

SpiKeR : Detecting abnormal review behavior. In a first step, we propose to detect abnormal reviewing activity by analyzing the histogram of each venue. We introduce SpiKeR , an algorithm that retrieves ranges of abnormal activity – spikes in a venue's histogram. Spikes are outliers in the distribution of the data. For instance, Figure 3 shows the evolution in time of the number of daily positive reviews for a venue called "Ike's Place" in San Francisco, CA [1], whose first review was registered in 2008. The number of daily positive reviews seldom approaches 15, however, on Nov. 7, 2011, the venue records a spike of 78 positive reviews. With a total of 3169 positive reviews in 1220 active days, Ike's Place has an average of 2.59 daily reviews.

SpiKeR relies on measures of dispersion, i.e., quartiles and interquartile ranges (IQR) [13], to detect outliers. SpiKeR takes as input argument a time range ΔT. Given a venue V, SpiKeR first computes the quartiles and the IQR of the positive reviews from V's timeline H_V. It then computes the upper outer fence (UOF) value using the Box-Whiskers plot [13]. For each interval d of length ΔT during V's active period, let P_d denote the set of positive reviews from H_V written during day d. If $|P_d| > UOF$, SpiKeR outputs P_d, a spike has been detected. For instance, the aforementioned 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.

WatchYT: event/spike correlations. We introduce WatchYT (Watch Yelp Timelines), an algorithm that relies on SpiKeR to detect correlations between Yelp events and increased review activity concerning the venues hosting the events. Algorithm 1 shows the pseudocode of the approach. Specifically, given a set of Yelp events ($events$) and a time interval ΔT (system parameter), WatchYT determines the set of venues that benefit from an event within an interval ΔT of the event's date. WatchYT processes each Yelp event separately (lines 4-12). It first retrieves the date of the event, as representing the date when the first review was written for the event (line 6). It then retrieves the venue hosting the event (line 7), collects its reviews and reconstructs its timeline (line 8). WatchYT runs SpiKeR to detect abnormal review behavior over the timeline (line 9). If a spike occurs within an interval ΔT from the date

[1] http://www.yelp.com/biz/ikes-place-san-francisco.

Algorithm 1 WatchYT: Yelp campaign detection tool.

```
1.WatchYT(events[] : YelpEvent, ΔT : Time)
2.  campaings[];      #campaigns detected
3.  campaigns := newVenue[];
4.  for i := 0 to events.size() do
5.      YelpEvent e := events[i];
6.      Date eDate := e.getDate();
7.      Venue V := e.getVenue();
8.      Timeline $H_V$ := V.getTimeline();
9.      TimeRange[] spikes := SpiKeR($H_V$);
10.     if (spikes.correlated(eDate, ΔT)) then
11.         campaigns.add(V); fi
12. od
13 return campaings;
```

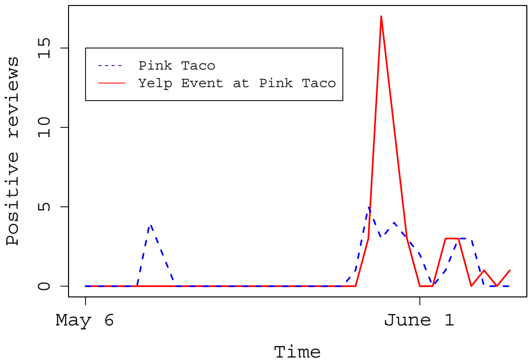


Fig. 4. The timeline of “Pink Taco 2” (Los Angeles) and of the Yelp event for this venue. Note the correlation between the two.

of the event (line 10), it adds the venue to the list of detected campaigns (line 11).

Figure 4 shows an example of venue and event timelines, correlated in time, for the venue “Pink Taco 2” (Los Angeles). Note how the venue’s latest two spikes coincide with the spikes of the event.

IV. Evaluation

We have evaluated WatchYT over the event and venue data described in Section II-C. We first evaluate SpiKeR ’s ability to detect spikes, as a function of ΔT, the interval around an event’s occurrence date. Figure 5 plots this dependence, when ΔT ranges from 1 to 5 weeks. For instance, when ΔT is 14 days, SpiKeR detected 36 spikes on the 149 venues. Some venues had more than one spike within the 14 days. The total number of venues with at least one spike is 24, accounting for around 17% of the venues. While for ΔT = 35 SpiKeR detected 47 spikes, we prefer a shorter interval: the correlation between the event and spikes may fade over longer intervals. In the following we use ΔT=14.

We now focus on determining the influence of Yelp events on the overall rating of a venue. First, we compute the *2-week impact* of the Yelp event on the venue. We define the 2-week impact as the difference between the rating of the venue two weeks after the event and the rating of the venue before the event. We compute the rating of a venue at any given time T as the average over the ratings of all the reviews received by the venue before time T. Figure 6 shows the distribution of the 2-week impact of the Yelp event on the venue. While 55 (of the 149) venues show no impact, 60 venues show at least a 0.5 star improvement, with 3 at or above 2 star improvements. 32 venues are negatively impacted. Thus, almost twice as many venues benefit from Yelp events, when compared to those showing a rating decay.

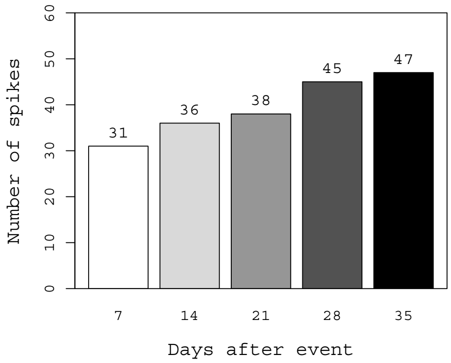


Fig. 5. Yelp events: Spike count as a function of ΔT.

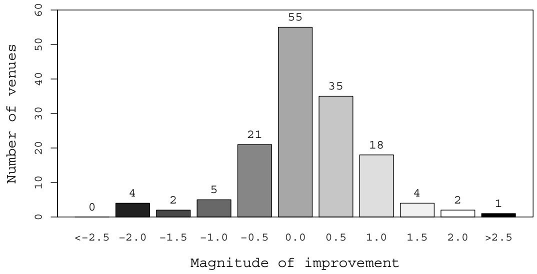


Fig. 6. Distribution of the short term impact (2 weeks) of Yelp events on venue ratings.

We then study the possibility of a relation between the number of reviews of a venue and the short term impact an event has on the venue. We observe that the impact of an event is quantified with fractions of rating, which means that we are dealing with a categorical variable. Therefore, we cannot use methods for linear or non-linear association, e.g. correlation coefficient. Instead, we tested the hypothesis of independence, using a χ^2 test [13], between the rating impact and the number of reviews, a discrete variable. The test gave us a $\chi^2 = 58.6837$ with 36 degrees of freedom, which is highly significant with a p-value of 0.009854. Thus, we reject the hypothesis of independence.

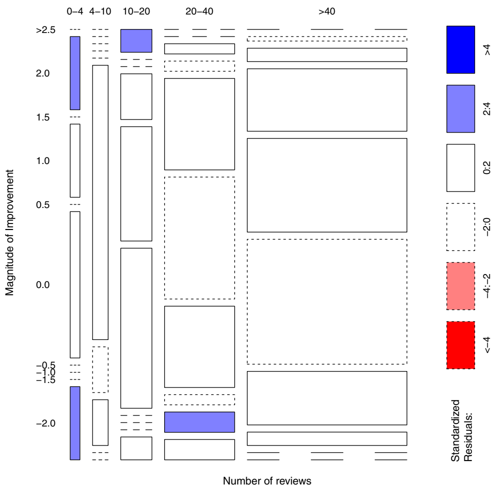


Fig. 7. Mosaic plot: dependency between the short term rating change of venues due to events and their number of reviews. The standardized residuals indicate the importance of the rectangle in the χ^2 test.

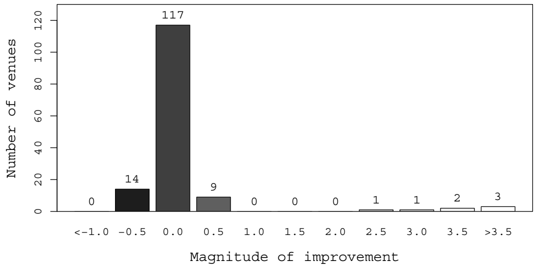


Fig. 8. Yelp events: Distribution of the improvement due to events

Figure 7 shows the mosaic plot depicting this relation. Each rectangle corresponds to a set of venues, that have a certain review count range (the x axis) and having been impacted by a certain measure within two weeks of an event (the y axis). The shape and size of each rectangle depict the contribution of the corresponding variables, so a large rectangle means a large count in the contingency table. Blue rectangles indicate that they are more than two standard deviations above the expected counts. Then, the figure shows that more than half of the (149) venues have more than 40 reviews. Moreover, we notice that the venues having more than 40 reviews set the trend of Figure 6: while roughly one third of the venues show no impact, twice as many venues show a positive impact vs. a negative one.

We now focus on the long term impact of Yelp events. For this, we compare the current ratings of the 149 venues with their ratings before the events. Figure 8 shows the distribution (over the 149 venues) of the difference between the current rating of the venues and their rating before the events. 78% of venues show no improvement. Furthermore, we see a balance between the number of venues showing an improvement versus a negative impact (16 positive vs. 14 negative). However, we emphasize that the negative impact is only half a star, while the positive impact reaches up to 3.5 stars.

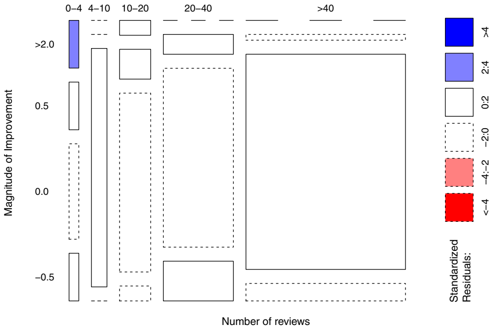


Fig. 9. Mosaic plot: the dependency between the long term rating change of venues due to events and their number of reviews. The standardized residuals indicate the importance of the rectangle in the χ^2 test.

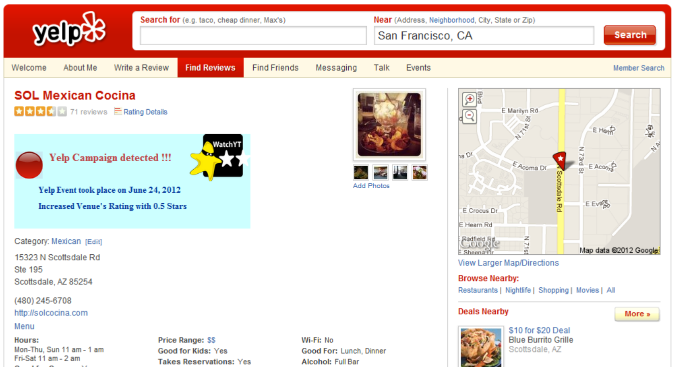
We conduct a χ^2 test to verify the dependence of the long term impact of events on venues on the number of ratings of the venues. The test was highly significant with $\chi^2 = 29.2038$, 12 degrees of freedom and a p-value of 0.003674. Figure 9 shows the mosaic plot: a vast majority of the venues having more than 40 reviews have no impact on the long term. This shows that review spikes have a smaller impact on constantly popular venues.

Conclusions. On the long term, events do not seem to impact the ratings of hosting venues. We believe this is because high numbers of regular reviews tend to overwhelm the impact of event spikes. However, Yelp events show a noticeable short term positive impact. Even a short term increase in popularity may act as a motivation to host such events [10].

V. WatchYT Implementation

We have implemented WatchYT as a web server and a browser plugin running in the user's browser. We have used Apache Tomcat 6.0.35 to route requests (exposed to the client through a REST API interface) to our server-side component. The server-side component relies on the latest servlet v3.0 which offers additional features including asynchronous support, making the server-side processing much more efficient. We implemented the browser plugin for the Chrome browser using HTML, CSS and Javascript. The plugin interacts with Yelp pages and the web server, using content scripts (Chrome specific components that let us access the browser's native API) and cross-origin XMLHttpRequests. The plugin is available for download at the project's website [12].

The browser plugin becomes active when the user navigates to a Yelp page. For user and venue pages, the plugin parses their HTML file and retrieves their reviews. We employ a stateful approach, where the server's DB stores all reviews of pages previously accessed by users. This enables significant



San Francisco, CA

SOL Mexican Cocina

71 reviews

Yelp Campaign detected !!!

Yelp Event took place on June 24, 2012

Increased Venue's Rating with 0.5 Stars

Category: Mexican [Edit]

15323 N Scottsdale Rd
Ste 195
Scottsdale, AZ 85254

(480) 245-6708
http://solcocina.com
Menu

Hours: Mon-Thu, Sun 11 am - 1 am
Fri-Sat 11 am - 2 am

Price Range: $$
Good for Kids: Yes
Takes Reservations: Yes

Wi-Fi: No
Good For: Lunch, Dinner
Alcohol: Full Bar

$10 for $20 Deal
Blue Burrito Grille
Scottsdale, AZ

Fig. 10. Snapshot of WatchYT in action.

time savings, as the plugin needs to send to the web server only reviews written after the date of the last user's access to the page. Given the venue's set of reviews, the server employs SPIKER to determine spikes in the venue's timeline. It then retrieves information about any Yelp events hosted by the venue. If an event is discovered, the server determines the impact of the event and sends this back, along with the date of the event, to the plugin. The plugin displays this information in the browser. Figure 10 shows WatchYT's extension to the Yelp page of the venue "Sol Mexican Cocina" (Scottsdale, AZ) in the central-left blue rectangle.

VI. Related Work

Ott et al. [6] created a database of fake hotel reviews in TripAdvisor, then integrated work from psychology and computational linguistics to develop and compare three text-centric approaches to detecting deceptive opinion spam. While the conclusions of this work can be used by attackers to refine the text of their reviews and escape detection, the follow up work of Feng et al. [15] relies on the J-shaped distributions of review ratings received by most venues to identify venues that receive too many 5 star reviews from single-time users.

Jindal and Liu [7] introduced the problem of detecting opinion spam in the context of product reviews. The techniques proposed in the context of Amazon reviews, include detecting spam, duplicate or plagiarized reviews and outlier reviews. Jindal et al. [8] extend this work to identify unusual review patterns that can represent suspicious reviewer behavior. They formulate the problem as finding unexpected domain independent rules; they test their solution on Amazon reviews. In the context of review spam, Lim et al. [9] propose techniques that determine a user's deviation from the behavior of other users reviewing similar products.

Instead, in this work we focus on a different geosocial network, Yelp, and its unique event mechanisms. We do not study fake reviews but reviews likely to be real, written by people who attended Yelp events. We propose the conjecture that events created by Yelp at select venues, impact positively the rating of the venue. Our results can be used in conjunction with previous work, to provide a comprehensive defense against manipulation of venue ratings.

VII. Conclusions

In this paper we have studied the problem of review campaigns organized by Yelp and involving Elite yelpers. We have proposed SPIKER , an approach that identifies positive review spikes in the timelines of venues. We have introduced WatchYT, a browser plugin that finds venues that have spikes correlated with events they organized. We have used venue and event data collected from Yelp to investigate the impact of Yelp events. We have shown that while a short term positive effect can be seen, in the long run, the effects of events are normalized by the reviews of regular users. In future work, we plan to investigate the impact of regular events, that venues organize and post on Yelp.

VIII. Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Nos. CNS-0821345, CNS-1126619, HRD-0833093, IIP-0829576, CNS-1057661, IIS-1052625, CNS-0959985, OISE-1157372, IIP-1237818, IIP-1215201, IIP-1230661, IIP-1026265, IIP-1058606, IIS-1213026.

References

[1] Yelp. http://www.yelp.com.

[2] Foursquare. https://foursquare.com/.

[3] Urbanspoon. http://www.urbanspoon.com.

[4] Matthew Lynley. Yelp CEO: IPO window is still open, Yelp on track. Venture Beat http://venturebeat.com/2011/09/13/yelp-ipo-on-track-disrupt/, September 2011.

[5] Kira Cochrane. Why TripAdvisor is getting a bad review. The Guardian, http://www.guardian.co.uk/travel/2011/jan/25/tripadvisor-duncan-bannatyne, January 2011.

[6] 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 Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[7] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 219–230, New York, NY, USA, 2008. ACM.

[8] Nitin Jindal, Bing Liu, and Ee-Peng Lim. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1549–1552, New York, NY, USA, 2010. ACM.

[9] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 939–948, New York, NY, USA, 2010. ACM.

[10] 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.

[11] Yelp. Yelp Elite Events: What's the deal? http://officialblog.yelp.com/2009/03/yelp-elite-events-whats-the-deal.html, 2009.

[12] WatchYT: Watch Yelp Timelines. http://users.cis.fiu.edu/~mrahm004/watchyt.

[13] A. C. Tamhane and D. D Dunlop. *Statistics and data analysis: From elementary to intermediate*. Upper Saddle River, NJ: Prentice Hall, 2000.

[14] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proceedings of IEEE INFOCOM '10*, San Diego, CA, March 2010.

[15] Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. Distributional footprints of deceptive product reviews. In *Proceedings of the Sixth International Conference on Weblogs and Social Media (ICWSM)*, 2012.