To Catch a Fake: Curbing Deceptive Yelp Ratings and Venues

Mahmudur Rahman[1], Bogdan Carbunar[1*], Jaime Ballesteros[2], and Duen Horng (Polo) Chau[3†]

[1] *The School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA*

[2] *Here, Nokia Inc. Chicago, IL, USA*

[3] *School of Computational Science & Engineering, Georgia Tech, Atlanta, GA, USA*

Received 13 May 2014; revised 13 November 2014; accepted 3 March 2015
DOI:10.1002/sam.11264
Published online 10 April 2015 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: The popularity and influence of reviews, make sites like Yelp ideal targets for malicious behaviors. We present Marco, a novel system that exploits the unique combination of social, spatial and temporal signals gleaned from Yelp, to detect venues whose ratings are impacted by fraudulent reviews. Marco increases the cost and complexity of attacks, by imposing a tradeoff on fraudsters, between their ability to impact venue ratings and their ability to remain undetected. We contribute a new dataset to the community, which consists of both ground truth and gold standard data. We show that Marco significantly outperforms state-of-the-art approaches, by achieving 94% accuracy in classifying reviews as fraudulent or genuine, and 95.8% accuracy in classifying venues as deceptive or legitimate. Marco successfully flagged 244 deceptive venues from our large dataset with 7,435 venues, 270,121 reviews and 195,417 users. Furthermore, we use Marco to evaluate the impact of Yelp events, organized for elite reviewers, on the hosting venues. We collect data from 149 Yelp elite events throughout the US. We show that two weeks after an event, twice as many hosting venues experience a significant rating boost rather than a negative impact. © 2015 Wiley Periodicals, Inc. Statistical Analysis and Data Mining: The ASA Data Science Journal 8: 147–161, 2015

Keywords: fake review detection; deceptive venue detection; machine learning

1. INTRODUCTION

Online reviews are central to numerous aspects of people's daily online and physical activities. Which Thai restaurant has good food? Which mover is reliable? Which mechanic is trustworthy? People rely on online reviews to make decisions on purchases, services and opinions, among others. People assume these reviews are written by real patrons of venues and services, who are sharing their honest opinions about what they have experienced. But, is that really the case? Unfortunately, no. Reviews are sometimes fake, written by fraudsters who collude to write glowing reviews for what might otherwise be mediocre services or venues [1–4].

In this paper we focus on Yelp [5], a popular social networking and location based service that exploits crowdsourcing to collect a wealth of peer reviews concerning venues and services. Crowdsourcing has however exposed Yelp to significant malicious behaviors: Up to 25% of Yelp reviews may be fraudulent [6].

While malicious behaviors may occasionally be performed by inexperienced fraudsters, they may also be professionally organized. For example, *search engine optimization* (SEO) companies tap into review writer markets [7–9] to offer *review campaigns*. Review campaigns act as "face lift" operations for business owners [10], manipulating venue ratings through multiple, coordinated artificial reviews.

For business owners, profit seems to be the main incentive to drive them to engage in such activities. Studies have shown that an extra half-star rating on Yelp causes a restaurant to sell out 19% more often [11], and a one-star increase leads to a 5–9% increase in revenue [12].

Furthermore, we study the impact of Yelp "elite" events on the ratings of hosting venues. Elite events are organized by Yelp for the benefit of "Elite," influential users, who write popular reviews. Yelp attempts to prevent review "unfairness" by encouraging attendees to review the event instead of the venue. However, the ample warning offered

[*] *Correspondence to:* Bogdan Carbunar (carbunar@cs.fiu.edu)
[†] A preliminary version of this article has appeared in SIAM SDM 2014.

This research was supported in part by DoD grant W911NF-13-1-0142 and NSF grant EAGER-1450619.

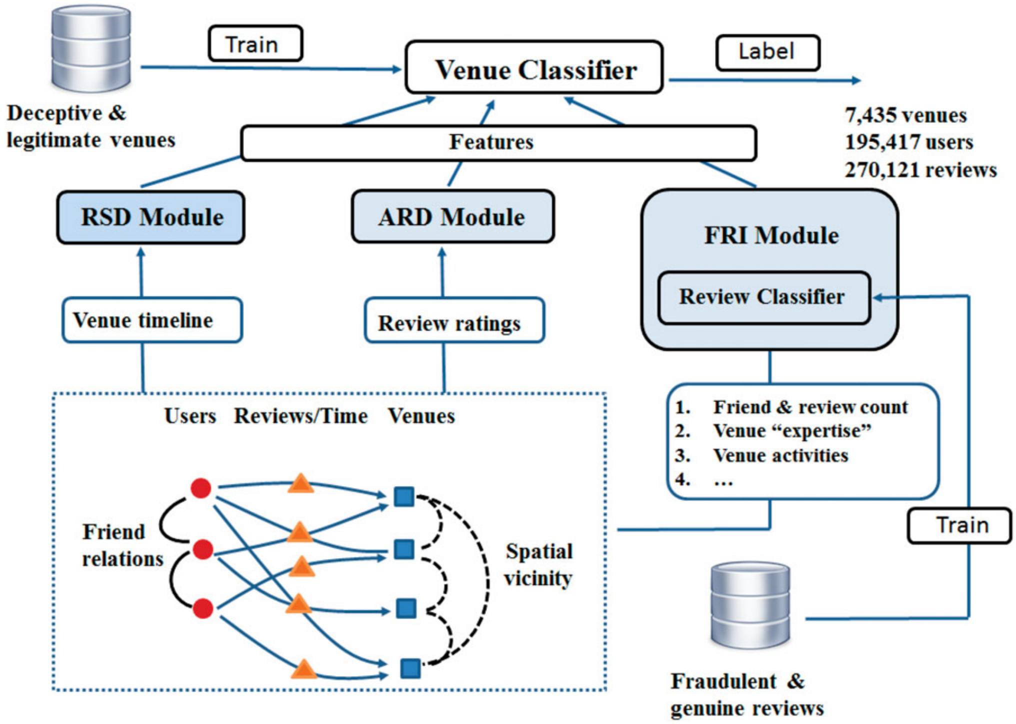

Fig. 1 System overview of *Marco*. Marco relies on social, temporal and spatial signals gleaned from Yelp, to extract novel features. The features are used by the *venue classifier* module to label venues (deceptive vs. legitimate) based on the collected data. Section 5 describes Marco in detail. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to hosts, coupled with the inability of users to accurately follow directions, may be used by adversaries to transform Yelp events into review campaign tools.

We propose *Marco* (MAlicious Review Campaign Observer), a novel system that leverages the wealth of spatial, temporal and social information provided by Yelp, to detect venues that are targets of deceptive behaviors. Marco (see Fig. 1) exploits fundamental fraudster limitations (see Section 5.1) to identify venues with (i) abnormal review spikes, (ii) series of dissenting reviews, and (iii) impactful but suspicious reviews. Marco detects both venues that receive large numbers of fraudulent reviews, and venues that have insufficient genuine reviews to neutralize the effects of even small scale campaigns. Our major contributions include:

- We introduce a *lower bound* on the number of reviews required to launch a review campaign that impacts a target venue's rating, and prove that this bound renders such campaigns detectable. Our theoretical results force fraudsters to compromise between the impact and undetectability of their review campaigns (Section 5).

- We present *Marco*, a system that leverages novel social, spatial and temporal features gleaned from Yelp

- We contribute a novel dataset of reviews and venues, which consists of both ground truth (i.e., objectively correct) and gold standard instances (i.e., selected based on best available strategies); and a large collection of 7,435 venues, 270,121 reviews and 195,417 reviewer profiles (Section 4).

- We demonstrate that Marco is effective and fast; its classification accuracy is up to 94% for reviews, and 95.8% for venues. It flags 244 of the 7,435 venues analyzed as deceptive; manual inspection revealed that they were indeed suspicious (Section 6).

- We collect data from 149 Yelp Elite events throughout the US and use it to study the short and long term impact of Yelp events on the rating of the hosting venues (Section 6.4).

Marco aims to complement legal actions against profitable, fraudulent review activities [10]. Organizations caught red-handed in setting up review campaigns have been shown to pay $1–$10 per fraudulent review. By making the cost of purchasing reviews approach the cost of products and services provided by hiring venues, Marco has the potential to act as an economic counter-incentive for rational venue owners.

2. RELATED WORK

2.1. Research in Detecting Fraudulent Reviews

Jindal and Liu [2] introduce the problem of detecting opinion spam for Amazon reviews. They proposed solutions for detecting spam, duplicate or plagiarized reviews and outlier reviews. Jindal et al. [3] identify unusual, suspicious review patterns. In order to detect "review spam", Lim et al. [4] propose techniques that determine a user's deviation from the behavior of other users reviewing similar products. Mukherjee et al. [13] focus on fake reviewer groups; similar organized fraudulent activities were also found on online auction sites, such as eBay [14]. Mukherjee et al. [15] leverage the different behavioral distributions of review spammers to learn the population distributions of spammer and non-spammer clusters. Li et al. [16] exploit the reviews of reviews concept of Epinions to collect a review spam corpus, then propose a two view, semisupervised method to classify reviews.

Ott et al. [17] integrate work from psychology and computational linguistics to develop and compare several text-based techniques for detecting deceptive TripAdvisor reviews. To address the lack of ground truth, they crowdsourced the job of writing fraudulent reviews for existing venues.

Unlike previous research, we focus on the problem of detecting *impactful* review campaigns. Our approach takes advantage of the unique combination of social, spatial and temporal dimensions of Yelp. Furthermore, we do not break Yelp's terms of service to collect ground truth data. Instead, we take advantage of unique Yelp features (i.e., spelp sites, consumer alerts) to collect a combination of ground truth and gold standard review and venue datasets.

Feng et al. [18] seek to address the lack of ground truth data for detecting deceptive Yelp venues: They introduce three venue features and use them to collect gold standard sets of deceptive and legitimate venues. They show that an SVM classifier is able to classify these venues with an accuracy of up to 75%. In Section 6 we confirm their results on our datasets. We show that with an accuracy of 95.8%, Marco significantly outperforms the best strategy of Feng et al [18].

Li et al. [19] and Ntoulas et al. [20] rely on the review content to detect review spam. Li et al. [19] exploit machine learning methods in their product review mining system. Ntoulas et al. [20] propose several heuristic methods for detecting content based spam and combine the most effective ones to improve results. Our work differs through its emphasis on relationship among reviewers, friends and ratings in the context of Yelp's spatial and temporal dimensions.

Gao et al. [21] target asynchronous wall messages to detect and characterize spam campaigns. They model each wall post as a pair of text description and URL and apply semantic similarity metrics to identify large subgraphs representing potential social spam campaigns and later incorporate threshold based techniques for spam detection. Instead, we focus on temporal and geosocial review context, the where reviewer activity and behavioral pattern are of significant importance.

Wang et al. [22] introduce the concept of heterogeneous review graphs and iterative methods exploring relationship among reviewers, reviews and stores to detect spammers. While we also consider social relations among reviewers we differ on our focus on temporal and spatial dimensions.

2.2. Research in Sybil Detection

Sybil accounts can be used to launch review campaigns, by enabling a single adversary to write multiple reviews for the same venue, each from a different account. Yelp identifies venues that receive multiple reviews from the same IP address (but different user accounts). Tools such as proxies [23] and anonymizers (e.g., Tor [24]) can however be used to avoid detection.

SybilInfer [25], detects Sybil nodes in social networks by using Bayesian inference and knowledge of the social network graph. Sybil tolerant solutions like DSybil exploit the heavy-tail distribution of the typical behavior of honest users and rely on user weights to identify whether the system needs more opinions or not. Similarly, SumUp [26] uses "adaptive vote flow aggregation" to limit the number of fake feedback provided by an adversary to the number of attack edges in the trust network—that is, the number of bi-directional trust edges the attacker is able to establish to other users. Molavi et al. [27] propose to associate weights with ratings and introduce the concept of "relative ratings" to defend against bought ratings and ratings from Sybil accounts. When given access to the perspective of the social network provider, Wang et al. [28] proposed an approach that detects Sybil accounts based on their click stream behaviors (traces of click-through events in a browsing session).

Our work aims to complement Sybil detection techniques. Reviews written from accounts detected to be Sybils may be classified as fraudulent. The number (or percentage) of reviews of a venue written from Sybil accounts can be used as a feature to detect "deceptive" venues. Conversely, user accounts with high numbers of posted fraudulent reviews may be used as candidates for further Sybil detection screenings.

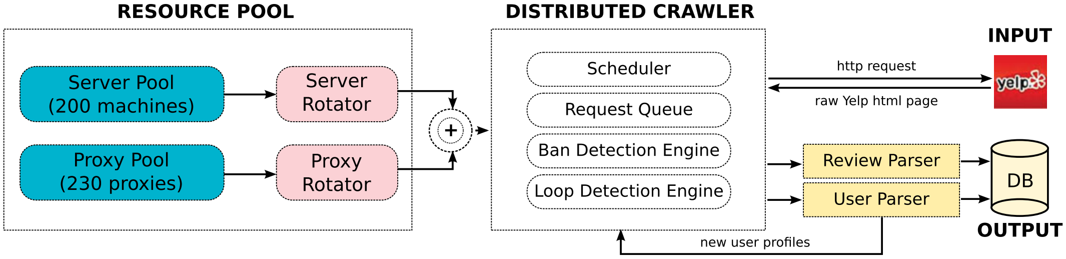

Fig. 2 YCrawl system architecture. YCrawl relies on a pool of servers and proxies to issue requests. The scheduler relies on a request queue to ensure there are no loops in the crawling process. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

3. SYSTEM MODEL

3.1. Yelp's Review System

For this work, we focus on Yelp [5], a review centric geosocial network that hosts information concerning users and venues. Subscribed users ("yelpers") have accounts and can write reviews, befriend other subscribers, report locations and search for venues of interest. We use the term "venue" to represent a business or event with an associated location (e.g., restaurants, shops, offices, concerts).

Reviews have a star *rating*, an integer ranging from 1 to 5, with 5 being the highest mark. In Yelp, an *average rating* value is computed for each venue (rounded to the nearest half star), over the ratings of all the posted reviews. For a review R, let $R.\rho$ denote its rating and $R.\tau$ to denote the time when the review was posted. We say a review is "positive" if its rating is at least 4 stars and "negative" if its rating is 2 stars or fewer. In our analysis we do not consider 3 star reviews. Their impact on the rating of the venue is likely to be small: Yelp denotes a 3 star rating as "A-OK".

3.2. Influential and Elite Yelpers

Users can rate the reviews of others, by clicking on associated buttons (e.g., "useful", "funny" or "cool" buttons). They can upload photos taken at venues reviewed and perform "check-ins", to formally record their real-time presence at the venue. Yelp rewards "influential" reviewers (often peer-recommended) with a special, yearly "Elite" badge.

3.3. Fraudulent Reviews and Deceptive Venues

A review is *fraudulent* if it describes a fictitious experience. Otherwise, the review is *genuine*. We say a venue is *deceptive* if it has received a sufficient number of fraudulent reviews to impact its average rating by at least half a star. Otherwise, the venue is *legitimate*.

Yelp relies on proprietary algorithms to filter reviews it considers fraudulent. See ref. [29] for an attempt to reverse engineer Yelp's filter. Furthermore, Yelp has launched a "Consumer Alert" process, posting "alert badges" on the pages of venues for which (i) people were caught red-handed buying fraudulent reviews, offering rewards or discounts for reviews or (ii) that have a large number of reviews submitted from the same IP address. The consumer alert badge is displayed for 90 days.

3.4. Yelp Events

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.

4. COLLECTED YELP DATA

In this section we describe the Yelp datasets we collected using the *YCrawl* crawler that we developed. Our data consists of: (i) 90 deceptive and 100 legitimate venues; (ii) 426 fraudulent and 410 genuine reviews; and (iii) a large collection of 7,435 venues and their 270,121 reviews from 195,417 reviewers, from San Francisco, New York City, and Miami.

4.1. YCrawl

We have developed YCrawl, a crawling engine for automatically collecting data from Yelp user and venue pages. YCrawl consists of 1820 lines of Python code. It fetches the raw HTML pages of target Yelp user and venue accounts. Fig. 2 illustrates the system design of YCrawl.

Yelp keeps track of requests made from a single IP and suppresses any IP making an exorbitant number of requests within a short time window [1]. To overcome this limitation,

[1] Such IP addresses are suppressed from Yelp's servers and this remains in place for a few weeks (or sometimes forever).

YCrawl uses a pool of servers and IP proxies: For every request, YCrawl randomly picks a server and proxy. 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.

At the time when we performed this data collection, Yelp's filtered reviews could only be accessed by solving a CAPTCHA. In order to collect filtered reviews we used DeathByCaptcha [30] to programatically collect CAPTCHA protected reviews filtered by Yelp.

We used YCrawl to collect a seed dataset of random venue and user accounts, using a breadth-first crawling strategy and stratified sampling [31]. First, we selected a list of 10 major cities (e.g., New York, San Francisco, Los Angeles, Chicago, Seattle, Miami) in the United States and we collected an initial random list of 100 venues from each of these cities as a seed dataset. We note that the strata venues are mutually exclusive, i.e., venues do not belong to two or more different cities. We then randomly selected 10,031 Yelp users who reviewed these venues, and collected their entire Yelp data (the html pages), including all their reviews, for a total of 646,017 reviews. This process enabled us to avoid bias toward high degree nodes (users with many friends, venues with many reviews), which is a common problem when crawling social networks [32]. We have then randomly selected a list of 16,199 venues, reviewed by the previously collected 10,031 Yelp users. We have collected the html pages of the selected the venues, including all their reviews.

4.2. The Data

We use the term "ground truth" set to denote data objectively known to be correct. We use the term "gold standard" to denote data selected according to the best available strategies. We collect such data following several stringent requirements, often validated by multiple third-parties.

Ground truth deceptive venues. We relied on Yelp's "Consumer Alert" feature to identify deceptive venues. We have used Yelp and Google to identify a snapshot of all the 90 venues that received consumer alerts during July and August, 2013.

Gold standard legitimate venues. We have used the collected list of 16,199 venues previously described to first selected a preliminary list of venues with well known consistent quality, e.g., the "Ritz-Carlton" hotel. We have then manually verified each review of each venue, including their filtered reviews. We have selected only venues with at most one tenth of their reviews filtered by Yelp and whose filtered reviews include a balanced amount of positive and negative ratings. While Yelp tends to filter reviews received from users with few friends and reviews, Feng et al. [18] showed that this strategy is not accurate. In total, we selected 100 legitimate venues.

In addition to collecting the html pages of all the reviews of the selected deceptive and legitimate venues, we have also collected the html pages of all the users who wrote reviews for them, and the html pages of all the reviews written by these reviewers. This data enables us to extract the features that we introduce in the following sections.

For the 90 deceptive venues we have collected their 10,063 reviews written by 7,258 reviewers. We have collected all the reviews (311,994 in total) written by the 7,258 reviewers of the 90 deceptive venues. In addition, we have collected the 9,765 reviews, written by 7,161 reviewers, of the 100 legitimate venues. We have then collected all the reviews written by these 7,161 reviewers, for a total of 530,408 reviews. Thus, for these 190 venues, we have collected more than 840,000 reviews. Note how the 90 deceptive venues have received more reviews than the 100 legitimate venues. However, the total number of reviews written by reviewers of legitimate venues significantly exceeds those written by the reviewers of deceptive venues.

Gold standard fraudulent reviews. We have used spelp (Spam + Yelp) sites (e.g., refs. [33,34]), forums where members, often "Elite" yelpers with ground truth knowledge, reveal and initiate the discussion on fraudulent Yelp reviews. While in theory such sites are ideal targets for fraudulent behavior, the high investment imposed on fraudsters, coupled with the low visibility of such sites, make them unappealing options. Nevertheless, we have identified spelp reviews that (i) were discussed by and agreed upon by *multiple* other Yelp users, (ii) were written from accounts with no user photo or with web plagiarized photos (identified through Google's image search), and (iii) that were short (less than 50 words). From this preliminary set, we have *manually* selected 410 generic reviews, that provide no venue specific information [35].

Specifically, each "spelp" review we collected was posted by a Yelp users, and discussed and agreed upon by *multiple* other Yelp users.

Gold standard genuine reviews. Given the seed user and venue datasets previously described, we have extracted a list of 410 genuine reviews satisfying a stringent test that consists of multiple checkpoints. In a first check we used Google (text and image search) to eliminate reviews with plagiarized text and reviewer account photos. In a second check we discarded short (less than 50 words), generic reviews, lacking references to the venue. Third, we gave preference to reviews written by users who

- Reached the "Elite" member status at least once.
- Participated in forums e.g., Yelp Talk.

- Garnered positive feedback on their reviews.
- Provided well thought out personal information on their profile.

We have collected the 54,213 reviews written by the writers of the 410 genuine reviews. We have also collected the 1,998 reviews written by the writers of the 426 fraudulent reviews.

Large Yelp Data Set. We have used YCrawl to collect the data of 7,435 car repair shops, beauty & spa centers and moving companies from San Francisco, New York City and Miami. The collection process took 3 weeks. Of the 7,345 venues, 1928 had no reviews posted. We have collected all their 270,121 reviews and the data of their 195,417 reviewers (one user can review more than 1 of these venues). Table 8 shows the number of venues collected for each venue type and city. Yelp limits the results for a search to the first 1000 matching venues. Entries with values less than 1000 correspond to cities with fewer than 1000 venues of the corresponding type.

Yelp event collection. We have collected Yelp events from 60 major cities covering 44 states of United States. 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 YCrawl 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.

While we are unable to make public these datasets, due to possible legal action from Yelp, we recommend researchers to contact us with questions concerning this data.

5. MARCO: PROPOSED METHODS

We present Marco, a system for automatic detection of fraudulent reviews, deceptive venues and impactful review campaigns. We begin with a description of the adversary and his capabilities.

5.1. Adversarial Model

We model the attacker following the corrupt SEO (Search Engine Optimization) model mentioned in the introduction. The attacker $\mathcal{A}$ receives a contract concerning a target venue V. $\mathcal{A}$ receives a finite budget, and needs to "adjust" the rating of V, i.e., either increase or decrease it by at least half a star.

Table 1. Table of notations.

Notation	Definition
$\mathcal{A}$	Adversary
V	Target venue
H_V, ΔT	V's timeline and active interval
$\rho_V(T)$	Rating of V at time T
δr	Desired rating increase by $\mathcal{A}$
δt	Review campaign duration
q	Number of fraudulent reviews by $\mathcal{A}$
R, $R.\rho$, $R.\tau$	Review, its rating and its posting time
n	Number of genuine reviews of V
σ	Sum of ratings of all genuine reviews
p	Number of genuine positive reviews

We assume $\mathcal{A}$ controls a set of unique (IP address, Yelp Sybil account) pairs and has access to a market of review writers. Sybil accounts [36] are different Yelp identities controlled by $\mathcal{A}$. $\mathcal{A}$ uses these resources to launch a "review campaign" to bias the rating of V: post one review from each controlled (IP address, Yelp Sybil account) pair and/or hire (remote) review writers, with valid Yelp accounts, to do it.

The number of reviews $\mathcal{A}$ can post is limited by the number of unique (IP address, Yelp Sybil account) pairs it controls as well as by the budget received in the contract (minus $\mathcal{A}$'s fee) divided by the average cost of hiring a review writer.

5.2. Overview of Marco

Marco, whose functionality is illustrated in Fig. 1, consists of 3 primary modules. The Review Spike Detection (RSD) module relies on temporal, inter-review relations to identify venues receiving suspiciously high numbers of positive (or negative) reviews. The Aggregate Review Disparity (ARD) module uses relations between review ratings and the aggregate rating of their venue, at the time of their posting, to identify venues that exhibit a "bipolar" review behavior. The Fraudulent Review Impact (FRI) module first classifies reviews as fraudulent or genuine based on their social, spatial and temporal features. It then identifies venues whose aggregate rating is significantly impacted by reviews classified as fraudulent. Each module produces several features that feed into a venue classifier, trained on the datasets of Section 4.2. Table 1 shows the notations used by Marco.

The approach used in Marco leverages manually labeled data, including fraudulent and genuine reviews, as well as deceptive and legitimate venues, to classify reviews and venues. Marco does not require knowledge of all the data and can classify new data in an online manner. A drawback of this approach stems from the difficulty of acquiring ground truth and gold standard data. While it is also difficult to identify relevant features that are hard to bypass by adversaries, we note that Marco introduces a trade-off for attackers, between impact and detectability.

An alternative approach is to use unsupervised outlier detection solutions [37–41]. While such solutions do not require labeled data, they require knowledge of the entire dataset. This approach is thus suitable for the providers (i.e., Yelp). We note however that an adversary with sufficient knowledge of the data can attempt to bypass this approach, by determining and introducing fraudulent data that would not be classified as outlier.

5.3. Review Spike Detection (RSD) Module

A review campaign needs to adjust (e.g., increase) the rating of its target venue, by posting (fraudulent) reviews that compensate the negative ratings of other reviews. The RSD module detects this behavior by identifying venues that receive higher numbers of positive (or negative) reviews than normal.

In the following, our first goal is to prove that review campaigns that impact the ratings of their target venues are detectable. For this, let q denote the total number of fraudulent reviews that $\mathcal{A}$ posts for the target venue V. We focus on the typical scenario where an attacker attempts to increase the rating of V (ballot stuffing). Attempts to reduce the rating of V (bad mouthing) are similar and omitted here for brevity.

$\mathcal{A}$ can follow any strategy, including (i) *greedy*, by posting all q reviews in a short time interval and (ii) *uniform*, by spreading reviews over a longer time interval. While a greedy strategy is likely to quickly impact the venue, a uniform strategy seems more likely to pass unnoticed. However, we show in the following that, if the review campaign is successful, it becomes detectable.

Let T_s and T_e denote the start and end times of the campaign, the times when the first and last fraudulent reviews initiated by $\mathcal{A}$ are posted. $\delta t = T_e - T_s$ is the campaign duration interval. Let n denote the number of genuine reviews V has at the completion of the campaign (time T_e). We assume V receives fraudulent reviews only from $\mathcal{A}$. We prove the following lower bound on the number of reviews that $\mathcal{A}$ needs to write in order to impact the rating of V.

Claim 1 *The minimum number of reviews $\mathcal{A}$ needs to post in order to (fraudulently) increase the rating of V by half a star is $q = n/7$.*

Proof: Let $R_1, R_2, ..., R_n$ denote the n genuine reviews of V. Let $\sigma = \sum_{i=1}^{n} R_i.\rho$. According to Yelp semantics, $R_i.\rho \in [1, 5]$, thus $\sigma \in [n, 5n]$. The "genuine" rating of V is $\rho_V^g = \frac{\sigma}{n}$. In order to minimize q, $\mathcal{A}$ has to write only 5 star reviews. Let δr be the increase in the rating of V generated by $\mathcal{A}$'s review campaign. Note that $\delta r \in [0.5, 4)$. Furthermore, $\frac{\sigma}{n} + \delta r \leq 5$, as the final rating of V cannot exceed 5. Hence,

$$\frac{\sigma + 5q}{n + q} = \frac{\sigma}{n} + \delta r,$$

Thus, $q = \frac{n^2\delta r}{5n-\sigma-n\delta r}$. Given that $\sigma \geq n$, we have $q \geq \frac{n\delta r}{4-\delta r}$. When $\delta r = 1/2$, this results in $q \geq n/7$. For $\delta r = 1$, $q \geq n/3$, when $\delta r = 2$, $q \geq n$, etc. ■

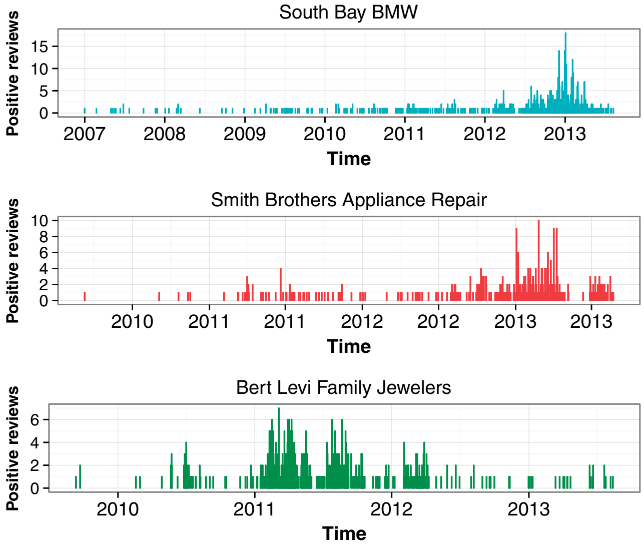

Fig. 3 Timelines of positive reviews of three deceptive venues (see Section 4.2). Each venue has several significant spikes in its number of daily positive reviews. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We say a review campaign is *successful* if it increases the rating of the target venue by at least half a star ($\delta r \geq 1/2$). We introduce now the notion of venue timeline:

DEFINITION 1: The timeline of a venue V is the set of tuples $H_V = \{(U_i, R_i) | i = 1...n\}$, the list of reviews R_i received by V from users U_i, chronologically sorted by the review post time, $R_i.\tau$. Let $\Delta T = T_c - T_1$ denote the **active interval** of the venue, where T_c denotes the current time and $T_1 = R_1.\tau$.

Fig. 3 illustrates this concept, by showing the evolution of the positive review (four and five star) timelines of three venues selected from the ground truth deceptive venue dataset (see Section 4.2). Let p denote the number of positive reviews received by V during its active interval, ΔT. We now show that:

Claim 2 *Assuming a uniform arrival process for genuine positive reviews, the maximum number of genuine positive reviews in a δt interval is approximately $\frac{p\ \delta t}{\Delta T}(1 + \frac{1}{\sqrt{c}})$, where $c = \frac{p\ \delta t}{\Delta T\ log \frac{\Delta T}{\delta t}}$.*

Proof: The distribution of reviews into δt intervals follows a balls and bins process, where p is the number of balls and $\Delta T/\delta t$ is the number of bins. It is known (e.g., [42,43]) that given b balls and B bins, the maximum number of balls in any bin is approximately $\frac{b}{B}(1 + \frac{1}{\sqrt{c}})$, where $c = \frac{b}{B\ log B}$. Thus, the result follows. ■

We introduce now the following result.

THEOREM 1: If $n > 49$, a successful review campaign will exceed, during the attack interval, the maximum number of reviews of a uniform review distribution.

Proof: Let p denote the number of positive, genuine reviews received by the target venue at the end of the review campaign. $p < n$, where n is the total number of genuine reviews at the end of the campaign. According to Claim 1, a successful review campaign needs to contain at least $n/7$ positive (5 star) reviews. Then, since the expected number of positive genuine reviews to be received in a δt interval will be $\frac{p\delta t}{\Delta T}$, following the review campaign, the expected number of (genuine plus fraudulent) positive reviews in the attack interval will be $\frac{n}{7} + \frac{p\delta t}{\Delta T}$.
The maximum number of positive genuine reviews posted during an interval δt, assuming a uniform distribution, is, according to Claim 2, approximately $\frac{p\ \delta t}{\Delta T} + \sqrt{\frac{p\delta t\ \log\frac{\Delta T}{\delta t}}{\Delta T}}$. Thus, the number of positive reviews generated by a review campaign exceeds the maximum positive reviews of a uniform distribution if

$$\frac{n}{7} + \frac{p\delta t}{\Delta T} > \frac{p\delta t}{\Delta T} + \sqrt{\frac{p\delta t\ \log\frac{\Delta T}{\delta t}}{\Delta T}}.$$

Since $n > p$, this converts to $\frac{n}{49} > \frac{\log\frac{\Delta T}{\delta t}}{\frac{\Delta T}{\delta t}}$. Since $\Delta T > \delta t$, we have that $\frac{\log\frac{\Delta T}{\delta t}}{\frac{\Delta T}{\delta t}} < 1$. Thus, the above inequality trivially holds for $n > 49$. ■

Theorem 1 introduces a tradeoff for attackers. Specifically, an attacker can choose to either (i) post enough reviews to impact the rating of a venue (launch a successful campaign) but then become detectable (exceed the maximum number of reviews of a uniform distribution) or (ii) remain undetected, but then do not impact the rating of the venue.

Detect abnormal review activity. We exploit the above results and use statistical tools to retrieve ranges of abnormal review activities. In particular, our goal is to identify spikes, or outliers in a venue's timeline. For instance, each venue in Fig. 3 has several significant review spikes. The RSD module of Marco uses the measures of dispersion of Box-and-Whisker plots [31] to detect outliers. Specifically, given a venue V, it first computes the quartiles and the inter-quartile range IQR of the positive reviews from V's timeline H_V. It then computes the upper outer fence (UOF) value using the Box-and-Whiskers plot [31]. For each sub-interval d of set length (in our experiments $|d|$ = 1 day) in V's active period, let P_d denote the set of positive reviews from H_V posted during d. If $|P_d| > UOF$, the RSD module marks P_d, i.e., a spike has been detected. For instance, the "South Bay BMW" venue (see Fig. 3) has a UOF of 9 for positive reviews: any day with more than 9 positive reviews is considered to be a spike.

We note that a different empirical approach, proposed by Fei et al. [44] is to use Kernel Density Estimation (KDE) to estimate the probability distribution function of the reviews of a venue.

The RSD module outputs two features (see Table 3): $SC(V)$, the number of spikes detected for a venue V, and $SAmp(V)$, the amplitude of the highest spike of V, normalized to the average number of reviews posted for V during an interval d.

5.4. Aggregate Rating Disparity

A venue that is the target of a review campaign is likely to receive reviews that do not agree with its genuine reviews. Furthermore, following a successful review campaign, the venue is likely to receive reviews from genuine users that do not agree with the venue's newly engineered rating.

Let $\rho_V(T)$ denote the average rating of a venue V at time $T \in [T_1, T_c]$. We define the rating disparity of a review R written at time $R.\tau$ for V to be the divergence of R's rating from the average rating of V at the time of its posting, $|R.\rho - \rho_V(R.\tau)|$. Let $R_1, ..., R_N$, $N = n + q$, be all the reviews received by V (both genuine and fraudulent) during its active interval ΔT. We define the aggregate rating disparity (ARD) score of V to be the average rating disparity of all the reviews of V:

$$ARD(V) = \frac{\sum_{i=1}^{N} |R_i.\rho - \rho_V(R_i.\tau)|}{N}$$

By influencing the average rating of a venue, a review campaign will increase the rating disparity of both fraudulent and of genuine reviews. This is illustrated in Fig. 4, that plots the evolution in time of the average rating against the ratings of individual reviews received by the "Azure Nail & Waxing Studio" (Chicago, IL). The positive reviews (1 day has a spike of 19, five-star reviews, shown in red in

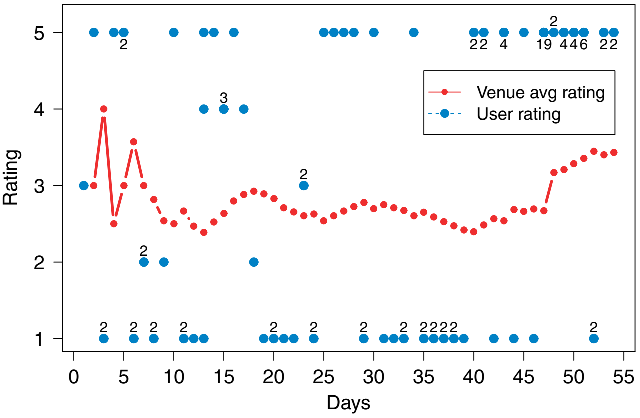

Fig. 4 Evolution in time of the average rating of the venue "Azure Nail & Waxing Studio" of Chicago, IL, compared against the ratings assigned by its reviews. The values in parentheses denote the number of reviews that were assigned a corresponding rating (shown on the y-axis) during one day. The lack of consensus between the many low and high rated reviews raises a red flag. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the upper right corner) disagree with the low rated reviews, generating a high ARD value. The ARD module contributes one feature, the ARD score, see Table 3.

We note that Jindal and Liu [2], Lim et al. [4], Mukherjee et al. [13] and Mukherjee et al. [15] proposed a feature similar to ARD. However, the ARD feature we introduce differs, in that the disparity is between the rating of a review and the rating of the venue at the time when the review was written. Previous work considers a formula where the disparity is computed at the present time.

5.5. FRI Module

Venues that receive few genuine reviews are particularly vulnerable to review campaigns (see also Theorem 1). Furthermore, long term review campaigns that post high numbers of fraudulent reviews can re-define the "normal" review posting behavior, flatten spikes and escape detection by the RSD module. They are also likely to drown the impact of genuine reviews on the aggregate rating of the venue. Thus, the ARD of the campaign's target venue will be small, controlled by the fraudulent reviews.

We propose to detect such behaviors through fraudulent reviews that significantly impact the aggregate rating of venues. For this, in a first step, the FRI module uses machine learning tools to classify the reviews posted for V as either fraudulent or genuine. It uses features extracted from each review, its writer and the relation between the review writer and the target venue (see Table 2). Specifically, let R denote a review posted for a venue V, and let U denote the user who wrote it. In addition to the friend and review count of U, we introduce the concept of *expertise* of U around V. $Exp_U(V)$ is the number of reviews U wrote for venues in the vicinity (50 mile radius) of V. Furthermore, FRI uses the number of activities of U recorded at V, the feedback of R, counting the users who reacted positively to the review, and the age of U's account when R was posted, $Age_U(R)$. Section 6.1 shows that the Random Forest tool achieves 94% accuracy when classifying fraudulent and genuine reviews.

Table 2. Features used to classify review R written by user U for venue V.

Notation	Definition
$f(U)$	The number of friends of U
$r(U)$	The number of reviews written by U
$Exp_U(V)$	The expertise of U around V
$c_U(V)$	The number of check-ins of U at V
$p_U(V)$	The number of photos of U at V
$feedback(R)$	The feedback count of R
$Age_U(R)$	Age of U's account when R was posted

In a second step, the FRI module introduces the notion of *FRI*, to model the impact of fraudulent reviews on the final rating of the venue. Let $\rho_V^g = \frac{\sigma}{n}$ denote the genuine rating of V, computed as an average over its n genuine reviews. Then, $FRI(V) = \rho_V(T_c) - \rho_V^g$, where $\rho_V(T_c)$ is the average rating of V at current time T_c. Note that $FRI(V)$ can be negative, for a bad-mouthing campaign. The FRI module contributes two features, $FRI(V)$, and the percentage of reviews classified as fraudulent for V, $CF(V)$ (see Table 3).

5.6. Venue Classification

In addition to the features provided by the RSD, ARD and FRI modules, we also use the rating of V, ρ_V, its number of reviews N, its number of reviews with associated user check-ins, $cir(V)$, and with uploaded photos, $pr(V)$, and the current age of V, $Age(V)$, measured in months since V's first review. Table 3 lists all the features we selected. Section 6.2 shows that the features enable the Random Forest classifier to achieves 95.8% accuracy when classifying the venue sets of Section 4.2.

6. EMPIRICAL EVALUATION

In this section we show that Marco is scalable as well as efficient in detecting fraudulent reviews and deceptive venues. We have implemented Marco using (i) Python, to extract data from parsed pages and compute the proposed features and (ii) the statistical tool R, to classify reviews and venues. We used MySQL to store collected data and features.

6.1. Review Classification

We investigated the ability of the FRI module to classify reviews, when using five machine learning tools: Bagging, k-Nearest Neighbor (kNN), Random Forest (RF), Support Vector Machines (SVM) and C4.5 Decision Trees (DT). We used tenfold cross-validation over the fraudulent and 410 genuine reviews of Section 4.2. Fig. 5(a) shows the receiver operating characteristic (ROC) curve for the top three performers: RF, Bagging and DT.

The overall accuracy ($\frac{TPR+TNR}{TPR+TNR+FPR+FNR}$) of RF, Bagging and DT is 93.8%, 93.6% and 93.2% respectively. TPR is the true positive rate, TNR is the true negative rate, FPR the false positive rate and FNR the false negative rate. The (FPR, FNR) pairs for RF, Bagging and DT are $(7.0\%, 5.3\%)$,$(6.3\%, 6.6\%)$ and $(5.1\%, 8.6\%)$ respectively (shown in Table 4). In the remaining experiments, the FRI module of Marco uses the RF classifier.

The top 2 most impactful features for RF are $r(U)$ and $Exp_U(V)$. Fig. 5(b) compares the distribution of the $r(U)$ feature for the 426 fraudulent and the 410 genuine reviews. We emphasize their symmetry: few fraudulent review writers posted a significant number of reviews, while few genuine review writers posted only a few reviews. Fig. 5(c) compares the distribution of the $Exp_U(V)$ measure. The distributions are also almost symmetric: most writers of

Table 3. Features used to classify a venue V as either deceptive or legitimate.

Notation	Definition
$SC(V)$	The number of review spikes for V
$SAmp(V)$	The amplitude of the highest spike
$ARD(V)$	Aggregate rating disparity
$FRI(V)$	The FRI of V
$CF(V)$	Count of reviews classified fraudulent
ρ_V	The rating of V
N	The number of reviews of V
$cir(V)$	The number of reviews with check-ins
$pr(V)$	The number of reviews with photos
$Age(V)$	The age of V

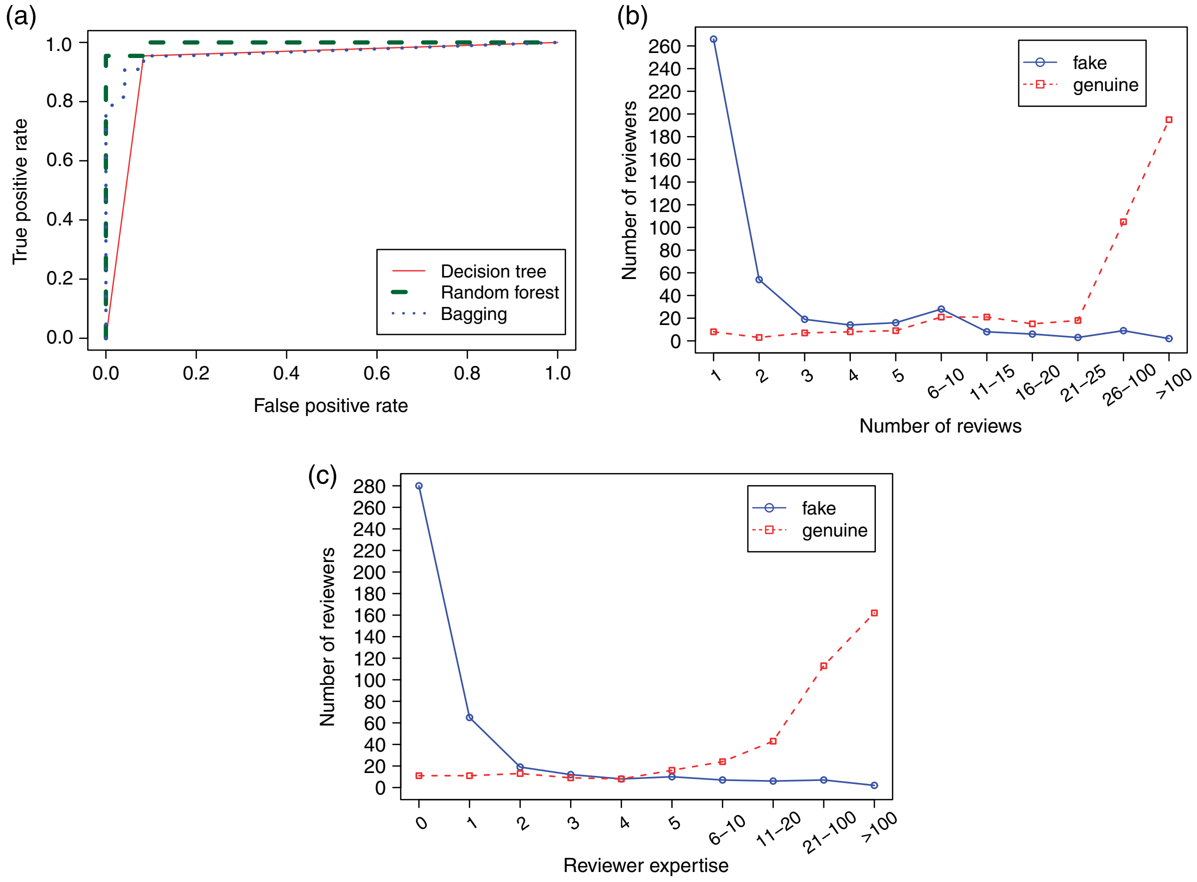

Fig. 5 (a) ROC plot of Random Forest (RF), Bagging and C4.5 Decision Tree (DT) for review classification (426 fraudulent, 410 genuine). RF performs best, at 93.83% accuracy. (b) Distribution of reviewers' review count: fraudulent versus genuine review sets. (c) Distribution of reviewers' expertise levels: fraudulent versus genuine sets. Note their symmetry: unlike genuine reviewers, fraudulent reviewers tend to have written only few reviews and have low expertise for the venues that they reviewed. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 4. Review classification: comparison of machine learning algorithms. RF performs best, at 93.83% accuracy.

Classifier	TPR (%)	FPR (%)	FNR (%)	Acc (%)
Random Forest	94.71	7.0	5.29	93.83
Bagging	93.45	6.28	6.55	93.59
Decision tree	91.44	5.07	8.56	93.22
SVM	89.92	9.66	6.04	92.11

Table 5. Significance test: pairwise comparison of machine learning algorithms using McNemar's test. With the exception of the (Bagging, RF) pair, for all other pairs McNemar's test produces a χ^2 value with one degree of freedom, highly significant with a confidence level of more than 95.0%.

Compared Classifiers	χ^2 value	p-value
Bagging-DT	11.6452	0.0006437
RF-DT	13.5	0.0002386
Bagging-RF	0.0476	0.8273
RF-SVM	4.8983	0.0268
Bagging-SVM	5.2258	0.0222
DT-SVM	5.1142	0.0237

Table 6. Marco versus the three deceptive venue detection strategies of Feng et al. [18]. Marco shows over 23% accuracy improvement over $dist\Phi$.

Strategy	FPR	FNR	Accuracy
Marco/RF	$5/90 = 0.055$	$3/100 = 0.3$	95.8%
$avg\Delta$	$33/90 = 0.36$	$31/100 = 0.31$	66.3%
$dist\Phi$	$28/90 = 0.31$	$25/100 = 0.25$	72.1%
$peak\uparrow$	$41/90 = 0.45$	$37/100 = 0.37$	58.9%

genuine reviews have written at least 4 reviews for other venues in the vicinity of the venue of their selected review.

Furthermore, we tested the null hypothesis that the classifiers used in review classification are equivalent i.e., the difference in performance metrics of different classifiers is not significant. As the classifiers are trained and tested on the same dataset, we used McNemar's test which tabulates the outcomes of every two classifiers used for review classification. The results are shown in Table 5. With the exception of the test that compares Bagging and RF, all other tests produce a χ^2 value with one degree of freedom, highly significant with a confidence level of more than 95.0% (the p-value is <0.05). Thus, we reject the null hypothesis, which means that the differences in performance metrics of DT, RF, Bagging and SVM models are statistically significant.

6.2. Venue Classification

We have used tenfold cross-validation to evaluate the ability of Marco to classify the 90 deceptive and 100 legitimate venues of Section 4.2. Fig. 6(a) shows the ROC curve for Marco when using the RF, Bagging and C4.5 DT classifiers on the features listed in Table 3. The overall accuracy for RF, Bagging and DT is 95.8%, 93.7% and 95.8% respectively, with the corresponding (FPR,FNR) pairs being $(5.55\%, 3\%), (8.88\%, 4\%)$ and $(5.55\%, 3\%)$ respectively.

Fig. 6(a) shows the distribution of $SC(V)$ for the 190 venues. Only one legitimate venue has a review spike, while several deceptive venues have more than 10 spikes. Furthermore, 26 deceptive venues have an FRI score larger than 1; only one legitimate venue has an FRI larger than 1. *Comparison with state-of-the-art.* We compared Marco with the three deceptive venue detection strategies of Feng et al. [18], $avg\Delta$, $dist\Phi$ and $peak\uparrow$. Table 6 shows the FPR, FNR and overall accuracy of Marco, $avg\Delta$, $dist\Phi$ and $peak\uparrow$. Marco achieves a significant accuracy

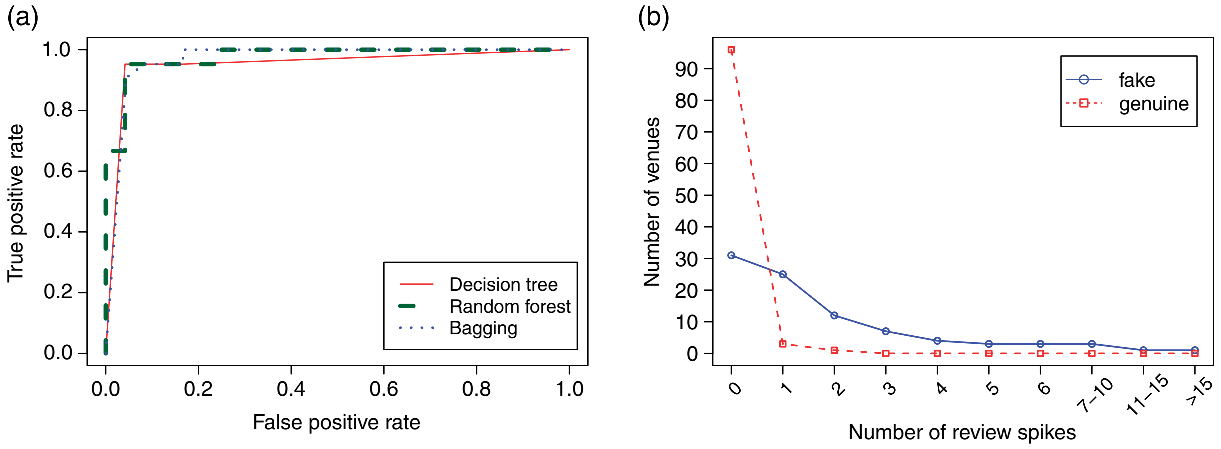

Fig. 6 (a) ROC plot of RF, Bagging and C4.5 DT for the 90 deceptive/100 legitimate venue datasets. RF and DT are tied for best accuracy, of 95.8%. (b) Distribution of SC(V), for the 90 deceptive and 100 legitimate venues. 60 deceptive venues have at least one review spike. One legitimate venue has one spike. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 7. Marco performance on new, unpopular venues: comparison of machine learning algorithms. RF and DT perform the best.

Classifier	TPR (%)	FPR (%)	FNR (%)	Acc (%)
Random Forest	96.07	20.0	3.92	94.64
Bagging	94.12	20.0	5.88	92.15
Decision tree	94.12	0.0	5.88	94.64

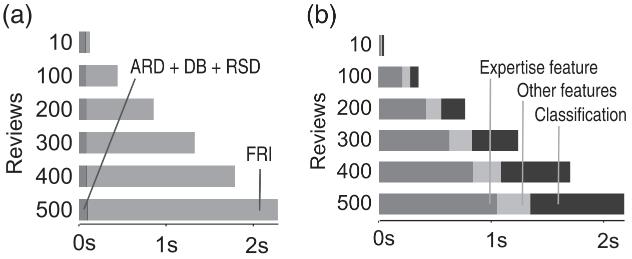

Fig. 7 (a) Marco's per-module overhead: FRI is the most expensive, but under 2.3 s even for venues with 500 reviews. (b) Zoom-in of FRI module overhead. Computing the $Exp_U(V)$ feature takes the most time.

improvement (95.8%) over $dist\Phi$, the best strategy of Feng et al. [18] (72.1%).

Marco performance for new venues. We have also evaluated the performance of Marco to classify relatively new venues with few genuine reviews. Specifically, from our set of 90 deceptive and 100 genuine reviews, we selected 51 deceptive and 5 genuine venues that had less than 10 genuine reviews when we collected them. The overall accuracy of RF, Bagging and DT on these 56 venues is 94.64%, 92.15% and 94.64% respectively. The (FPR, FNR) pairs for RF, Bagging and DT are (20.0%, 3.92%), (20.0%, 5.88%) and (0.0%, 5.88%), respectively (Table 7).

6.3. Marco in the Wild

Marco takes only a few seconds to classify a venue, on a i5@2.4GHz, 4 GB of RAM Dell laptop. Fig. 7(a) shows the per-module overhead of Marco (averages over 10 experiment runs), as a function of the review count of the venue classified. While the FRI module is the most time consuming, even for venues with 500 reviews the FRI overhead is below 2.3 s. The RSD and ARD modules impose only a few ms (6 ms for 500 reviews), while DB access and data retrieval take around 90 ms. Fig. 7(b) zooms-in into the FRI overhead. For 500 reviews, the most time consuming components are computing the user expertise, $Exp_U(V)$ ($\approx$ 1.1 s), computing *all* the other features ($\approx$ 0.4 s) and classifying the reviews ($\approx$ 0.8 s).

In order to understand the ability of Marco to perform well when trained on small sets, we have trained it on 50 deceptive and 50 legitimate venues and we have tested it on the remaining 40 deceptive and 50 legitimate venues. On average over 10 random experiments, Marco achieved an FPR of 6.25% and an FNR of 3%.

Table 8. Collected venues organized by city and venue type. Values between parentheses show the number of venues detected by Marco to be deceptive. San Francisco has the highest percentage of deceptive venues.

City	Car Shop	Mover	Spa
Miami, FL	1000 (6)	348 (8)	1000 (21)
San Fran., CA	612 (59)	475 (45)	1000 (42)
NYC, NY	1000 (8)	1000 (27)	1000 (28)

We have used Marco to classify the 7,435 venues we collected from Miami, San Francisco and New York City. We have divided the set of 7,435 venues into subsets of 200 venues. We trained Marco on the 190 ground truth/gold standard venues and tested it separately on all subsets of 200 venues. Table 8 shows the total number of venues collected and the number of venues detected to be deceptive, between parentheses. San Francisco has the highest concentration of deceptive venues: Marco flags almost 10% of its car repair and moving companies as suspicious, and upon our manual inspection, they indeed seemed to engage in suspicious review behaviors. While the FRI of San Francisco's collected genuine venues is at most 1, 60% of its deceptive venues have an FRI between 1 and 4.

6.4. Detecting 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*. Fig. 8(a) 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.

To detect correlations between Yelp events and increased review activity concerning the venues hosting the events, we use Marco's RSD module as follows. Specifically, given a Yelp event and a time interval ΔT (system parameter), we determine of the hosting venue experiences a positive review spike within an interval ΔT of the event's date.

For the events and hosting venues collected (see Section 4.2), Fig. 8(b) plots the number of positive review spikes detected within ΔT days, when ΔT ranges from 1 to 5 weeks. For instance, when ΔT is 14 days, Marco

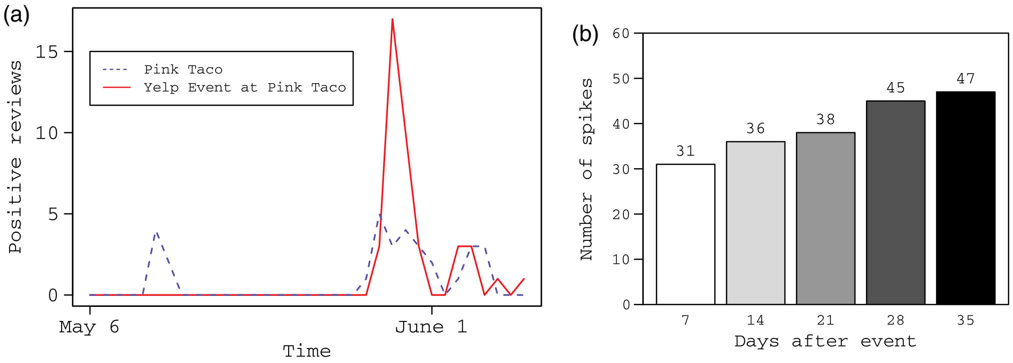

Fig. 8 (a) The timeline of “Pink Taco 2” (Los Angeles) and of the Yelp event for this venue. Note the correlation between the two. (b) Yelp events: Positive review spike count as a function of ΔT. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

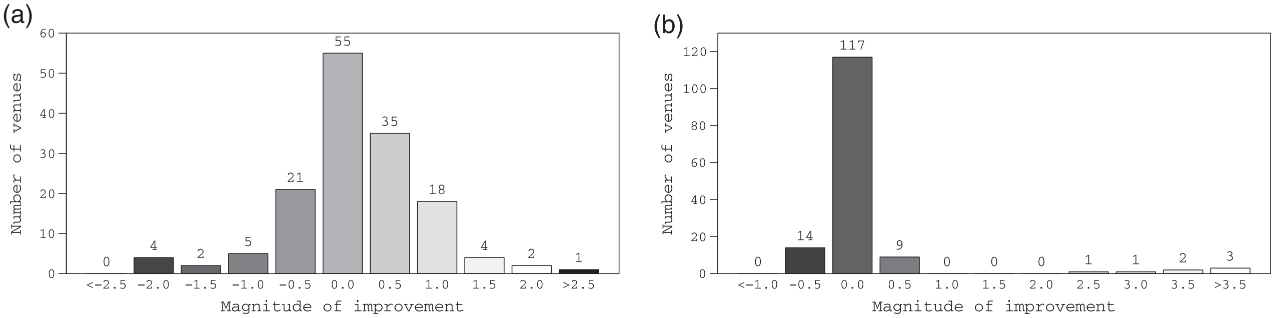

Fig. 9 (a) Distribution of the short term impact (2 weeks) of Yelp events on venue ratings. (b) Yelp events: Distribution of the improvement due to Elite events.

detected 36 spikes on the 149 venues. Some venues have 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 $\Delta T = 35$ Marco 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.

Furthermore, we focused on determining the influence of Yelp events on the overall rating of a venue. First, we computed 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. Fig. 9(a) 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.

This result raises the question of whether there exists a relation between the number of reviews of a venue and the short term impact an event has on the venue. The impact of an event is a categorical variable, as it is quantified with fractions of a star (integer). The number of reviews however is a discrete variable. Therefore, we cannot use methods for linear or nonlinear association, e.g., correlation coefficient. Instead, we tested the hypothesis of independence between the rating impact and the number of reviews, using a χ^2 test [31]. The test produced a $\chi^2 = 58.6837$ with 36 degrees of freedom, which is highly significant (the p-value is 0.009854). Thus, we reject the independence hypothesis.

Fig. 10(a) 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 Fig. 9(a): while roughly one third of the venues show no impact, twice as many venues show a positive impact versus a negative one.

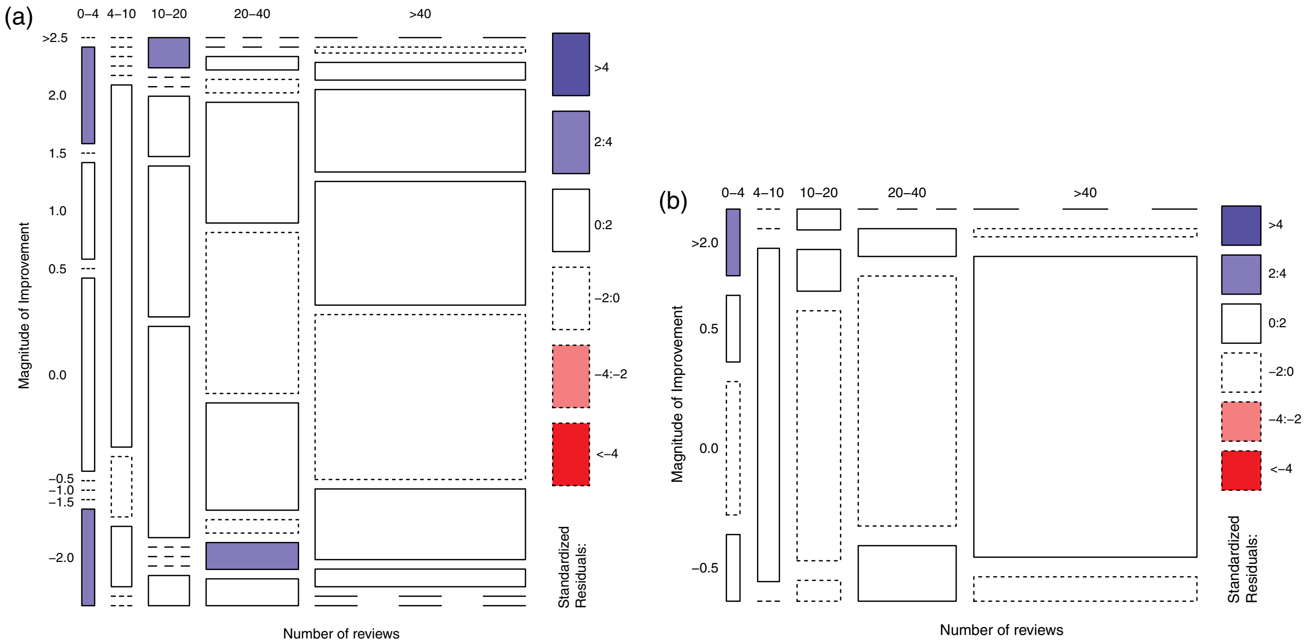

Fig. 10 Mosaic plots: The standardized residuals indicate the importance of the rectangle in the χ^2 test. (a) The dependency between the short term rating change of venues due to events and their number of reviews. (b) The dependency between the long term rating change of venues due to events and their number of reviews. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Second, we study the long term impact of Yelp events. For this, we compare the current ratings of the 149 venues with their ratings before the events. Fig. 9(b) 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.

We conducted 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. Fig. 10(b) 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.

7. CONCLUSIONS

We presented Marco, a system for detecting deceptive Yelp venues and reviews, leveraging a suite of social, temporal and spatial signals gleaned from Yelp reviews and venues. We also contribute a large dataset of over 7K venues, 270K reviews from 195K users, containing also a few hundred *ground-truth* and *gold-standard* reviews (fraudulent/genuine) and venues (deceptive/legitimate). Marco is effective in classifying both reviews and venues, with accuracies exceeding 94%, significantly outperforming state-of-the-art strategies. Using Marco, we show that two weeks after an event, twice as many venues that host Yelp events experience a significant rating boost, when compared to the venues that experience a negative impact. Marco is also fast; it classifies a venue with 500 reviews in under 2.3 s.

REFERENCES

[1] D. Segal, A Rave, a Pan, or Just a Fake? The New York Times www.nytimes.com/2011/05/22/your-money/22haggler.html, 2011.

[2] N. Jindal and B. Liu, Opinion spam and analysis, In Proceedings of the International Conference on Web Search and Web Data Mining, WSDM '08, New York, ACM, 219–230, 2008.

[3] N. Jindal, B. Liu, and E-P. Lim, Finding unusual review patterns using unexpected rules, In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, 1549-1552, 2010.

[4] E-P. Lim, V-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, Detecting product review spammers using rating behaviors, In Proceedings of the International Conference on Information and Knowledge Management (CIKM), 939–948, 2010.

[5] Yelp, http://www.yelp.com.

[6] Yelp admits a quarter of submitted reviews could be fake, BBC, www.bbc.co.uk/news/technology-24299742.

[7] Sponsored Reviews, www.sponsoredreviews.com/, 2013.

[8] Posting Positive Reviews, www.postingpositivereviews.blogspot.com/, 2013.

[9] Pay Per Post, https://payperpost.com/, 2013.

[10] A. G. Schneiderman announces agreement with 19 companies to stop writing fake online reviews and pay more than $350,000 in fines, www.ag.ny.gov/press-release/ag-schneiderman-announces-agreement-19-companies-stop-writing-fake-online-reviews-and.

[11] M. Anderson and J. Magruder, Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. Econ J, 122 (2012), 957–989.

[12] M. Luca, Reviews, reputation, and revenue: the case of Yelp.com, www.hbswk.hbs.edu/item/6833.html.

[13] A. Mukherjee, B. Liu, and N. Glance, Spotting fake reviewer groups in consumer reviews, In Proceedings of the International Conference on World Wide Web, 2012.

[14] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos, Netprobe: a fast and scalable system for fraud detection in online auction networks, In Proceedings of the International Conference on World Wide Web, ACM, New York, 201–210, 2007.

[15] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, Spotting opinion spammers using behavioral footprints, In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’13, 632–640, 2013.

[16] F. Li, M. Huang, Y. Yang, X. Zhu, and X. Zhu, Learning to identify review spam, In IJCAI, pages 2488–2493, 2011.

[17] 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), 2011.

[18] S. Feng, L. Xing, A. Gogar, and Y. Choi, Distributional footprints of deceptive product reviews, In Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM), 2012.

[19] F. Li, M. Huang, Y. Yang, and X. Zhu, Learning to identify review spam, In Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI ’11, 2488–2493, 2011.

[20] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, Detecting spam web pages through content analysis, In Proceedings of the 15th international conference on World Wide Web, WWW ’06, 83–92, 2006.

[21] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y.. Zhao, Detecting and characterizing social spam campaigns, In Proceedings of the 10th Annual Conference on Internet Measurement, IMC ’10, 35–47, 2010.

[22] G. Wang, S. Xie, B. Liu, and P. S. Yu, Review graph based online store review spammer detection, In ICDM ’11, 1242–1247, 2011.

[23] Hide My Ass! Free Proxy and Privacy Tools, http://www.hidemyass.com/.

[24] R. Dingledine, N. Mathewson, and P. F. Syverson, Tor: The second–generation onion router, In USENIX Security Symposium, 303–320, 2004.

[25] G. Danezis and P. Mittal, Sybilinfer: Detecting sybil nodes using social networks, In Proceedings of the Network and Distributed System Security Symposium (NDSS), 2009.

[26] N. Tran, B. Min, J. Li, and L. Subramanian, Sybil-resilient online content voting, In NSDI’09: Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation, Berkeley, CA, USENIX Association, 2009, 15–28.

[27] A. Molavi Kakhki, C. Kliman-Silver, and A. Mislove, Iolaus: Securing online content rating systems, In Proceedings of the Twenty-Second International World Wide Web Conference (WWW’13), Rio de Janeiro, Brazil, May 2013.

[28] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, You are how you click: Clickstream analysis for sybil detection, In Proceedings of USENIX Security, 2013.

[29] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, What yelp fake review filter might be doing, In Proceedings of the International Conference on Weblogs and Social Media, 2013.

[30] Death By Captcha. www.deathbycaptcha.com/.

[31] A. C. Tamhane and D. D Dunlop. Statistics and Data Analysis: From Elementary to Intermediate. Upper Saddle River, NJ, Prentice Hall, 2000.

[32] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, Walking in Facebook: a case study of unbiased sampling of OSNs, In Proceedings of IEEE INFOCOM ’10, San Diego, CA, 2010.

[33] Spelp. www.yelp.com/topic/boston-spelp-9.

[34] Flelp. www.yelp.com/topic/miami-flelp-we-rock.

[35] 3 tips for spotting fake product reviews - from someone who wrote them. MoneyTaksNews, www.moneytalksnews.com/2011/07/25/3-tips-for-spotting-fake-product-reviews–from-someone-who-wrote-them.

[36] J. R. Douceur, The Sybil attack, In Revised Papers from the First International Workshop on Peer-to-Peer Systems, 2002.

[37] V. J. Hodge and J. Austin, A survey of outlier detection methodologies, Artif Intell Rev 22 (2004), 85–126.

[38] A. Zimek, E. Schubert, and H.-P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Stat Anal Data Min 5 (2012), 363–387.

[39] K. Yamanishi, J-I. Takeuchi, G. Williams, and P. Milne, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’00, ACM, New York, 2000, 320–324.

[40] K. Zhang, S. Shi, H. Gao, and J. Li. Unsupervised outlier detection in sensor networks using aggregation tree, In Advanced Data Mining and Applications, New York, Springer, 2007, 158–169.

[41] Z. Ferdousi and A. Maeda, Unsupervised outlier detection in time series data, In Proceedings of 22nd International Conference on Data Engineering Workshops, x121–x121, IEEE, 2006.

[42] P. Berenbrink, A. Brinkmann, T. Friedetzky, and L. Nagel, Balls into bins with related random choices, In Proceedings of the Symposium on Parallelism in Algorithms and Architectures (SPAA), 2010.

[43] M. Raab and A. Steger. Balls into bins: a simple and tight analysis, In Proceedings of Randomization and Approximation Techniques in Computer Science (RANDOM), 159–170, 1998.

[44] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, Exploiting burstiness in reviews for review spammer detection, In ICWSM, 2013.