

Interpretable and Effective Opinion Spam Detection via Temporal Patterns Mining across Websites

Yuan Yuan[†] Sihong Xie* Chun-Ta Lu[‡] Jie Tang[†] Philip S. Yu[‡]

[†]Department of Computer Science and Technology, Tsinghua University, Beijing, China

* Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA

[‡]Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

Abstract—Millions of ratings and reviews on online review websites are influential over business revenues and customer experiences. However, spammers are posting fake reviews in order to gain financial benefits, at the cost of harming honest businesses and customers. Such fake reviews can be illegal and it is important to detect spamming attacks to eliminate unjust ratings and reviews. However, most of the current approaches can be incompetent as they can only utilize data from individual websites independently, or fail to detect more subtle attacks even they can fuse data from multiple sources. Further, the revealed evidence fails to explain the more complicated real world spamming attacks, hindering the detection processes that usually have human experts in the loop. We close this gap by introducing a novel framework that can jointly detect and explain the potential attacks. The framework mines both macroscopic level temporal sentimental patterns and microscopic level features from multiple review websites. We construct multiple sentimental time series to detect atomic dynamics, based on which we mine various cross-site sentimental temporal patterns that can explain various attacking scenarios. To further identify individual spams within the attacks with more evidence, we study and identify effective microscopic textual and behavioral features that are indicative of spams. We demonstrate via human annotations, that the simple and effective framework can spot a sizable collection of spams that have bypassed one of the current commercial anti-spam systems.

I. INTRODUCTION

Online reviews and ratings are influential over customer purchasing decisions and business revenues. For example, business revenues on Yelp are positively correlated with the number of stars in ratings [13]; book sales on Amazon.com and Barnesandnoble.com increase as ratings improve [3]. Thus review websites have become the target of spamming attacks that aim at the unjustly manipulating rating of products and businesses. It is imperative to spot such harmful spamming reviews and ratings.

Opinion spams have been intensively studied for over a decade. For example, in [7], [18], [19], textual cues of spamming are derived and analyzed. User behaviors, such as the temporal and spatial distributions of ratings and reviews over products are also utilized for suspicious reviewer detection [8], [16], [17], [12], [4], [15]. Graph-based approaches construct graphs consisting of reviews, reviewers and businesses [28], [22], and use various graph characteristics, such as the degree distribution of neighbors as detection signals. Hybrid methods combine features of texts, behaviors and graphs [20] to provide more detection power. Recent work [14], [24] represents some

endeavors to exploit the rich and complementary information in multiple websites. The intuition is that inconsistency across multiple data sources can indicate abnormal activities. However, the cross-site features adopted by the previous methods are aggregated statistics over time, while statistics local to a time window can exhibit subtle and complicated cross-site patterns that can effectively reveal spamming activities.

We propose a bi-level framework to close the gap. On the macroscopic level, we propose the concepts of cross-site time series anomaly patterns, which define a set of semantically rich signals for the detection and sense-making of many different types of cross-site spamming activities. The motivation is that such patterns should be able to summarize the information in different sources about the same business, and provide more evidence to interpret the scenarios where the attacks happened. For example, one site may present a burst of 5-star reviews in a time window, within which there is no such burst or even a drop in the average rating on the other site. Such co-occurrence of temporal patterns can be regarded as being caused by a potential attack where spammers try to cover up the negative reviews on the first site. The challenge of time series construction is to find comparable metrics that describe similar concepts across sites, such that the co-occurring patterns are semantically meaningful and detectable, see Section III-A for details. To the best of our knowledge, such cross-site patterns have not been fully studied and little is known in the context of spam detection. Thus we study the properties of spams detected by these patterns and quantify their detection power through human annotation. Our annotating results reveal that a lot of spams can be left intact by a current commercial anti-spam system.

On the microscopic level, we study the spams within the time windows detected by the cross-site patterns (Section V) to reveal more evidences for individual spam detection and interpretation. We study the correlations between a comprehensive collection of behavioral and linguistic features and the annotated spams in order to characterize individual spams. The importance of each feature is then quantified and classification models are built to demonstrate that we can detect the spams that have not been previously detected by the commercial anti-spam system. The above macroscopic and microscopic studies provide a full picture of multiple site anomaly detection. Via case studies, we show that by combining the macroscopic and microscopic characteristics, we are able to identify highly

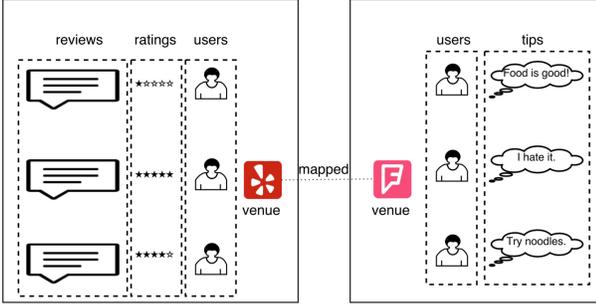


Fig. 1: Graph representation of multiple site review data

suspicious spams with strong evidences for human experts to make accurate and fast final decisions. Our contributions are as follows:

- We propose a novel spam detection framework using time series patterns defined over multiple data sources.
- We perform in-depth studies to reveal a full picture of the defined patterns on two levels. Through a set of human-labeled reviews, we demonstrate the high precision of various macroscopic cross-site patterns (Section III) and microscopic features (Section IV).
- Quantitative and qualitative results demonstrate that the framework can precisely identify and explain attacks that were not previously spotted.

II. BACKGROUND

We adopt two popular rating websites, Foursquare and Yelp, as our objects of study. Both websites consist of reviewers, reviews and businesses. On Yelp, the reviews usually have associated integer ratings ranging from 1 to 5, while on Foursquare, a reviewer can post reviews called “tips” without ratings. A business can be rated and reviewed on both Foursquare and Yelp by different groups of users. We represent the above relationships in Figure 1. Formally, let \mathcal{D}^s , $s = 1, 2$ denote the data from the two websites, each of which consists of reviewers, reviews and businesses. In this paper, the superscript s is used to indicate a specific data source. For example, for the first data source, $\mathcal{D}^1 = \{\mathcal{V}^1, \mathcal{U}^1, \mathcal{R}^1\}$, where $\mathcal{V}^1 = \{v_1^1, \dots, v_{n^1}^1\}$ are the n^1 businesses on site 1, $\mathcal{U}^1 = \{u_1^1, \dots, u_{m^1}^1\}$ are the m^1 reviewers, whose reviews are denoted by $\mathcal{R}^1 = \{r_{ij}^1\}$, where r_{ij}^1 is the review given by the i -th reviewer to the j -th business on site 1. Note that $\mathcal{V}^1 = \mathcal{V}^2$ since we aim at anomaly detection for businesses that have data in multiple review websites. We thus ignore the superscripts on the businesses and use $\mathcal{V} = \{v_1, \dots, v_n\}$ instead, where $n = n^1 = n^2$ is the number of common businesses. The review r has its post time, rating, texts and sentiments of the texts, denoted by $\text{time}(r)$, $\text{rating}(r)$, $\text{text}(r)$ and $\text{sentiment}(r)$, respectively. We shift the posting time of all reviews from both sites such that $\text{time}(r) \in [0, \infty)$. Notations are summarized in Table I.

TABLE I: Notations

Symbol	Meaning
\mathcal{U}^s	Reviewers from the s -th site
\mathcal{V}	Businesses that are rated/reviewed by all sites
\mathcal{R}^s	Reviews from the s -th site
n	Number of all businesses
m^s	Number of reviewers from the s -th site
\mathcal{T}	Number of time windows
$\text{time}(r)$	Posting time of the review r
$\text{rating}(r)$	Rating of the review r
$\text{text}(r)$	Texts of the review r
$\text{sentiment}(r)$	Sentiment of the texts of the review r
$ A $	Cardinality of the set A
$\mathbb{1}[\text{Cond}]$	Indicator function of Boolean statement Cond.

TABLE II: Sentimental metrics for time series construction. A star means that the corresponding metric is available for the website on that column.

Metric	Yelp	Foursquare
CR	*	*
AR	*	
FR	*	
LR	*	
AS	*	*
$HPSR$	*	*
NSR	*	*

III. SPAMMING DETECTION VIA CROSS-SITE TEMPORAL PATTERNS

We construct time series of several sentimental metrics, based on which we define cross-site time series patterns that have associated semantics meanings that can be indicative of various types of spamming activities.

A. Design of cross-site time series patterns

Time series anomaly patterns can be indicators of spamming activities, since massive spamming activities are usually reflected by temporal dynamics of various sentimental metrics. For example, in previous works [6], [25], the authors defined a time window with a burst of average rating or number of reviews as an anomaly. However, essentially these works only detected one type of pattern based on a single site, failing to use more comprehensive information across multiple websites to define more anomaly patterns that are actually meaningful. We generalize the previous works to multiple sites and define a wider spectrum of abnormal activities.

1) *Single website time series construction*: We first construct various sentimental time series for individual website. The whole time period when we observe the reviews of a business is divided into smaller equal-sized time windows, denoted by $\mathcal{T} = \{\tau_1, \dots, \tau_T\}$ where T is the total number of time windows, and τ_t is the t -th time window. Each τ_t is of a certain length such as two weeks. By fixing a business and a time window, say τ_t , we can compute the following aggregate sentimental metrics (see Table II). The metrics reflect the reviewer sentiments to businesses which are the target that spammers try to manipulate, and their dynamic will later be used to describe spamming activities. In the following, we

ignore the subscripts of reviewers (i) and businesses (j) and only keep the site indicator for each review (s). The site indicator is necessary as not all the metrics are available on both sites.

- Count of Reviews: number of reviews or tips from the s -th site posted in time window τ_t :

$$CR^s(t) = |\{r^s : \text{time}(r^s) \in \tau_t\}|.$$

This metric captures the volume of reviews for a business. If a business is spammed, regardless of promoting or demoting, the number of reviews tends to go up, as it is hard to manipulate the sentiments of a business via a small number of reviews. This metric ignores reviewers' identities and can thus detect massive posting activities even when a spammer posts very few reviews or ratings [25].

- Average Rating:

$$AR^s(t) = \frac{\sum_{r^s: \text{time}(r^s) \in \tau_t} \text{rating}(r^s) + \alpha \overline{AR}^s}{CR^s(t) + \alpha}.$$

This metric will be indicative when spammers are hired to manipulate business ratings. We use the average rating of the whole time series \overline{AR}^s for Laplace smoothing, where α is a tuning parameter. Smoothing is necessary in the case where a time window contains very few data. The same smoothing is applied to all sentimental metrics in the following.

- Five-star Ratio:

$$FR^s(t) = \frac{\sum_{r^s: \text{time}(r^s) \in \tau_t} \mathbb{1}[\text{rating}(r^s) = 5] + \alpha \overline{FR}^s}{CR^s(t) + \alpha}.$$

To boost the rating of a business, a spammer is more likely to give 5-star reviews, thus the ratio of 5-star reviews within a time window can be indicative of spamming attacks.

- Low-rating Ratio:

$$LR^s(t) = \frac{\sum_{r^s: \text{time}(r^s) \in \tau_t} \mathbb{1}[\text{rating}(r^s) \leq 2] + \alpha \overline{LR}^s}{CR^s(t) + \alpha}.$$

Likewise, to defame the rating of a business, a spammer is more likely to give 1 or 2 star reviews, defined as low-rating reviews, and the ratio of low-rating reviews within a time window can measure such activities.

- Average Sentiment: Sentiment analysis takes the texts of a review as input and assigns to the review a numeric score to indicate the polarity of the review. This metric is necessary although the sentiment of a review is correlated to the review's rating for the following reasons. First, different users will have varied interpretations of the number of stars, (as the personal bias in recommendation systems). In particular, some reviewers can frequently give 5-star reviews while others will reserve the 5-star reviews for the very best businesses. Text sentiments is thus another metric for personal ratings. Second, the text has richer polarity information compared to ratings, which

are single numbers. For example, a 5-star review (highest rating) can contain a mixture of positive and negative sentiments to different aspects of a business, while a 1-star review (lowest rating) can still mention some good aspects. Lastly, not all reviews are associated with a rating (like those on Foursquare), and text sentiment can serve as an alternative way to define review ratings.

We also assign the each sentence a score between -1 and 1 to indicate the polarity of the sentence. The sign of the score indicates polarity, and the magnitude of the score indicates the strength of the polarity. The sentiment of a review is calculated as the averaged sentiment scores of the constituting sentences, and the sentiment of the time window τ_t is calculated by:

$$AS(t) = \frac{\sum_{r^s: \text{time}(r^s) \in \tau_t} \text{sentiment}(r^s) + \alpha \overline{AS}^s}{CR^s(t) + \alpha}$$

where \overline{AS}^s is the average sentiment scores of the reviews in the s -th site over all time windows.

- Highly Positive Sentiment Ratio: Similar to Five-star Ratio, we define highly positive sentiment ratio as the proportion of reviews with sentiment scores higher than a certain threshold. A review with extremely high positive sentiment score tends to contain or even overuse positive words to describe a business, and previous studies [12] have shown that spamming reviews can contain a purer set of positive or negative sentences or words, while a normal review is likely to have a mixed sentiment. We define the ratio of reviews with extreme positive sentiments in a time window $HPSR^s(t)$ as:

$$\frac{\sum_{r^s: \text{time}(r^s) = t} \mathbb{1}[\text{sentiment}(r^s) > \theta] + \alpha \overline{HPSR}^s}{CR^s(t) + \alpha},$$

where $1 > \theta > 0$ is the threshold, above which the sentiment of a review is considered as highly positive.

- Negative Sentiment Ratio: We define the ratio of reviews with negative sentiment $NSR^s(t)$ as

$$\frac{\sum_{r^s: \text{time}(r^s) = t} \mathbb{1}[\text{sentiment}(r^s) < 0] + \alpha \overline{NSR}^s}{CR^s(t) + \alpha}.$$

Note that we aim at capturing any negative sentiment in the reviews, thus we consider a review as negative if the sentiment score is less than zero. This is useful as the metric is quite sensitive to *any* negative reviews, which the spammers may want to cover in certain cases.

The concatenation of the measurements of each metric over all time windows results in a time series for a business. For example, $CR^s = [CR^s(1), \dots, CR^s(T)] \in \mathbb{R}^T$ is the time series describing the dynamics of the number of reviews over time for the business. Note that fluctuations are quite likely to exist in the above time series, due to insufficient sample size within time windows or stochastic reviewer activities. To eliminate uninteresting fluctuations and focus on the more salient patterns, we adopt the Bayesian change-point (BCP) detection [5] to fit the discrete points on the time series using

smooth curves. An example is shown in the top subfigure of Figure 3 (Section VI). In the sequel, all discussions are based on the fitted curves rather than the discrete points.

We can assign sentimental polarity to the time series. For example, $HPSR^s$ models the changes in positive sentiments, NRR^s expresses the changes in negative sentiments, and CR^s is simply the counts and thus neutral.

2) *Single site time series pattern detection*: To define co-occurring temporal patterns across sites, we first define patterns that capture the sentimental dynamics within a single site. Different from previous works [25], [6] that focus on a single pattern, namely, bursty sections on the time series, we define a more diverse set of patterns including bursts, dives and plateaus, that can be used later to construct semantically richer cross-site patterns.

- **Burst**: a burst over a time series is a small time interval where the defining metric of the time series goes up and then down. The pattern has been adopted to capture sudden massive arrivals of reviews, or sudden increases of sentiments [25], [6]. For example, spammers hired by a business are usually required to post spamming reviews on the review websites within a short time window, leading to a burst on the time series CR^s . Besides, bursts in rating or sentiment related time series (FR^s or $HPSR^s$) are also highly suspicious: if positive sentimental score rises suddenly during a short period, it is likely that someone is attacking the website on purpose to promote the business. We also observe interesting bursts in time series with negative sentiments in the experiments.
- **Dive**: a dive on a time series is a small time interval where the defining metric of the series goes down and then up. For example, using the series CR^s , the downward section of a dive indicates that a store is suffering from small customer traffic. At this time, this store may have a higher motivation to gain some popularity and visibility on the review website by promotions, campaigns or spamming, leading to a mass of reviews and the upward section of the dive.
- **Plateau**: a plateau over a time series is a section where there is no significant change of the metric, and is thus regarded as uninteresting and largely ignored by previous works. However, in the multiple site situation, plateaus on one site can be useful in anomalies on the other site, as we discuss in the following.

The shapes and sentiments together define patterns indicative of spamming activities. To express these combinations succinctly, we use the following coding schema: underlined font indicates the patterns based on any time series with negative sentiments, while normal font means the patterns based on any series with positive or neutral sentiments. For example, $\underline{\mathbf{B}}$ means a burst of negative sentiments (that is the sudden increase in the ratio of negative reviews), and \mathbf{B} means a burst of positive sentiments (the is the sudden increase of the ratio of 5-star reviews). We detect the burst patterns over

a single time series using the following equation:

$$d = \frac{\lambda}{\left(\frac{1}{k_1} - \frac{1}{k_2}\right) \times w + \lambda}, \quad (1)$$

where w is the length of the interval under consideration. An intentional attack is usually finished during a short period, and the smaller the w , the larger the d . $k_1 > 0$ is the slope of the line connecting the starting and the second point of the interval, measuring the upward trend. Similarly $k_2 < 0$ is the slope of the line connecting the second last point and the end point of the interval, measuring the downward trend. We use $\frac{1}{k_1} - \frac{1}{k_2}$ to summarize the magnitude of the burst. The degree of a dive can be defined in a similar way, and the only difference is the signs of k_1 and k_2 . The complete process of discovering significant bursts, dives and plateaus is given in Algorithm 1.

Algorithm 1 Suspicious Pattern Detection

Input: time series \mathbf{c} , threshold θ , tunable parameter λ, l
Output: suspicious time periods set \mathcal{X}_{burst} , \mathcal{X}_{dive} and $\mathcal{X}_{plateau}$

- 1: $m \leftarrow \max(\mathbf{c})$
- 2: $\mathbf{c} \leftarrow \mathbf{c}/m$
- 3: **for** i **in** all time points **do**
- 4: **for** j **from** $i + 1$ **to** $i + l$ **do**
- 5: $k_1 \leftarrow \mathbf{c}_{i+1} - \mathbf{c}_i$
- 6: $k_2 \leftarrow \mathbf{c}_{j+1} - \mathbf{c}_j$
- 7: $w \leftarrow j - i + 1$
- 8: Calculate d using Eq. (1).
- 9: **if** $k_1 > 0$ and $k_2 < 0$ and $d > \theta$ **then**
- 10: $\mathcal{X}_{burst} \leftarrow \mathcal{X}_{burst} \cup [i, j]$
- 11: **else if** $k_1 < 0$ and $k_2 > 0$ and $d < -\theta$ **then**
- 12: $\mathcal{X}_{dive} \leftarrow \mathcal{X}_{dive} \cup [i, j]$
- 13: **else**
- 14: $\mathcal{X}_{plateau} \leftarrow \mathcal{X}_{plateau} \cup [i, j]$
- 15: **end if**
- 16: **end for**
- 17: **end for**

3) *Cross-site time series pattern design and construction*: Having defined single site time series patterns, we can put them together to construct more complex cross-site patterns. These patterns will have the shape and sentiment properties, and thus shall be interpretable to human inspectors in the sense that they reflect various real world spamming scenarios. We focus on the following representative patterns (Table III) that can be easily explained by real world spamming activities. These patterns are not studied in previous works but are shown to be effective in finding spams in our experiments. The pattern names in the table are abbreviations. For example, “BB” means and Burst (on Yelp) Burst (on Foursquare), and “DP” means Dive (on Yelp) and Plateau (on Foursquare).

- **BB**: there is a co-occurrence of positive bursts in both sites. There can be a few explanations for this pattern. It can be caused by the grand-opening of a new business and

there was a promotion or campaign. It can also be caused by an organized spam attack on both sites to maximally boost the fame of the business.

- **BB**: there is a burst of negative sentiments in one site and a burst of positive sentiments in the other. For example, customers may complain about foods on one site while there is a burst of positive reviews on the other, then it is likely that the positive reviews are spams that try to cover the negative ones. Note that spammers may want to cover the negative reviews in the same site where the negative reviews are posted. Since we have a time series that is sensitive to any negative sentiment (LR^s and NSR^s), the positive reviews on the same site will not be able to totally cover the negative ones and the negative burst can still be discovered by our algorithm.
- **BP**: this conflicting cross-site pattern is intuitively suspicious. For example, on the one site, there is a large volume of arrivals of reviews or a sudden rise of positive sentiments, while on the other site there is no such positive burst, then the burst is likely caused by spam attacks rather than a grand-opening campaign or promotion.
- **BD**: similar to **BP**, but the cross-site conflict is more severe. For example, the sentiments of the reviews on one site become worse than its usual level, while the ratings on the other site suddenly go up. One explanation is that customers were complaining on one site, while the business hired spammers to rescue its reputation on both sites (that is the upward section of a dive and the sudden rise in the burst, while the downward section of the dive will not be covered due to the sensitivity of our pattern detection algorithm).
- **DP**: there is no sudden change of the time series on one site, which can be regarded as neutral, while there is a downward trend on the other site, which can be caused by customer complaints or malicious spam attacks trying to defame a business.
- **DD**: the consistent dives on both sites render this pattern less suspicious, as it can be explained by the sudden drop in quality such as room service of a hotel or the leaving of a famous chef in a restaurant. We do not exclude the possibility of an organized spam attacking trying to defame a business on both sites.
- **PP**: this is the most common pattern that occurs, since most often there is no special event for a business. Note that if the plateaus on both sites are in their high positions, it is possible that there is a lasting spamming activity going on.

Here we introduce a simple method to detect the above cross-site patterns. For a time series on each website, we label each time window as B, D or P (if this period is detected to be included in both a burst and a dive, we label it as B). By combining the desired pairs of single site labels for the same time window over different time series (within or across websites), we obtain cross-site patterns.

We discuss the time and space complexity to find single

site patterns. For convenience, we define C_v^s as the number of reviews of venue v from the s -th site, \mathcal{T} as the number of time windows and \mathcal{S} as the number of websites. First, we need to calculate each time series defined in Section III-A. After enumerating, all reviews of a business, average values for each time series are obtained. Then for each business, measurements of each metric can be calculated by the equations in Section III-A. The time complexity for this step is $O(C_v^s)$ and space complexity is $O(1)$. Second, lines 3 and 4 in Algorithm 1 indicate that the time complexity of the algorithm is $O(lT)$, where l is a small constant and can be ignored. Space complexity in this step is $O(\mathcal{T})$ (for storing \mathcal{X}_{burst} , \mathcal{X}_{dive} and $\mathcal{X}_{plateau}$). The last step is to combine multiple websites. Both time and space complexity are $O(\mathcal{PST})$, where \mathcal{P} is the number of cross-site patterns. In sum, our method for detecting cross-site patterns is efficient, with time complexity of $O(\mathcal{PST} + C_v^s)$ and space complexity of $O(\mathcal{ST})$.

IV. EMPIRICAL STUDIES OF CROSS-SITE PATTERNS

A. Experiment setup

For tips in Foursquare, we crawled 301,717 venues. Information of each venue contains its full name, location (longitude and latitude), tips and users who post the tips. As for Yelp, we use businesses in Las Vegas and Pittsburg from the Yelp challenge dataset¹. It also provides the full names and locations of the businesses. We also crawled the reviews filtered by Yelp’s anti-spam system.

We propose the following algorithm to link businesses to both websites. We first build a KD-tree for longitude and latitude of businesses in Foursquare. Then for each business on Yelp, the nearest 20 venues in Foursquare are retrieved from the KD-tree. Lastly, we search in these 20 venues. If the full name of a venue on one website is a substring of the name the other venue, the venues from the two websites are matched. We employed 95 matched businesses to conduct our experiments, with a total of 68,517 reviews posted by 31,092 reviewers on Yelp and 15,004 tips posted by 12,147 reviewers on Foursquare. We divide the duration within which the reviews are posted (from May 28, 2010 to Dec 30, 2014) into 120 periods to construct various time series.

B. Macroscopic characterizations of the cross-site patterns

We study the macroscopic properties of the proposed cross-site patterns to demonstrate their significance.

1) *Basic statistics of cross-site patterns*: Table IV summarizes the basic statistics of certain cross-site patterns, including the number of reviews and reviewers detected on Yelp and Foursquare. From the table, we can see that there is a remarkable size of businesses, reviews and reviewers that fall into the time window detected by the proposed cross-site patterns that represent potential spam attacks (like **BB** and **PB**). In the column “# related reviews”, we show the number of reviews that are posted by the same reviewers who have posted any reviews within the detected time window. The idea

¹http://www.yelp.com/dataset_challenge

TABLE III: Cross-site time series patterns, with solid lines depicting the trends on the first website and dashed lines depicting the trends on the second.

CS pattern names	BB, <u>BB</u>	BP	BD	DP	DD	PP
Shapes						

TABLE IV: Cross-Site pattern statistics

Pattern (Y-F)	Yelp					Foursquare		
	#business	#review	#reviewer	#related reviews	filtered ratio	#business	#review	#reviewer
BB	7	181	179	19133	27.07%	9	89	83
BP	27	821	772	127427	26.31%	27	200	186
BD	8	295	290	41713	18.98%	9	122	114
PB	51	3795	3187	636679	13.68%	52	1154	1089
PP	95	59830	23509	9364943	11.99%	95	12152	9491
PD	33	3024	2589	548993	15.41%	34	1036	943
DB	4	76	76	10321	21.05%	6	79	74
DP	10	303	300	23822	48.18%	9	73	71
DD	4	192	190	21059	28.13%	6	99	96

is that, if a reviewer is found to be suspicious by the cross-site patterns, then any of his/her other reviews not detected by these patterns are also not trustworthy, and we gain more detection power. We also show the ratio of reviews detected by Yelp’s spam detector within those detected by the proposed cross-site patterns. We can see that most of the cross-site patterns find subsets of reviews that do not overlap that much with reviews detected by Yelp’s filter. How likely a review in the detected time windows is suspicious? Can we trust the Yelp’s filter? We next hire human annotators to label sampled reviews falling in the detected windows to answer the above questions, and demonstrate the effectiveness of various cross-site patterns.

2) *Human evaluation*: Note that it is impossible to exhaust all spams due to the large number of reviews posted. Therefore recall is not available and we use precision as the evaluation metric. We label a subset of the reviews, with the following constraints over how we sample the reviews. First, as boosting the rating or sentiments is more profitable [13], [3], we only sample the 5-star reviews. Second, in order to demonstrate the inadequacy of Yelp’s filter, the set of sampled reviews does not intersect with the reviews filtered by Yelp. Lastly, as a prototype, we detect the BB, BP, PB, PP and PD patterns on the pair of time series consisting of averaged rating on Yelp and the averaged sentiment on Foursquare (the other patterns detect too few reviews).

Three human annotators independently label the sampled reviews using 3 levels of suspiciousness, namely, 1: *not suspicious*, 2: *likely suspicious* and 3: *very suspicious*. To find any suspicious traces, the annotators are asked to not only look into the sampled reviews, but also the historic reviews

posted by the same reviewers who posted the sampled reviews. The suspicious score of each review is averaged over all three annotators to obtain the final scores as ground truths. We only have the reviews from Yelp labeled since there are more reviews and reviewer footprints to provide more evidence for human investigations, leading to more confident labeling results.

In Table V, we summarize the annotation results. The column “Avg Scores” are the averaged final suspicious scores. “ $Prec(> t)$, $t = 1, 2$ ” are the precision of detection made by the patterns, such that reviews with averaged suspicious scores greater than the threshold t will be considered as spams. These cross-site patterns are compared with a baseline pattern, the single-site burst pattern B^* that detects bursts of averaged rating on Yelp [25]. The baseline simply collects the reviews falling in the union of the time windows detected by cross-site patterns BB, BP.

From the table, we can observe the following. First, the cross-site patterns have higher precision than the single-site pattern. When $t = 1$ and a review is considered as a spam with slight suspiciousness, the BP pattern reaches the highest precision of over 98%, which is quite effective. When $t = 2$, a review has to be highly suspicious to qualify as a spam, the cross-site pattern BB achieves the best precision of 44%, which is much higher than the runners-up. Besides, the cross-site patterns can be used to interpret the types of attacks. For example, a large portion of highly suspicious reviews ($t = 2$) detected by the BB pattern may correspond to an organized spam attack over multiple review sites, where the spammers try to holistically boost up the positive sentiment of a business; the spams detected by the BP patterns can be explained as a spam attack on the first site and there is no or a lasting spamming attack (depending on the elevation of the plateau) on the second site. However, there is no such explanation can be derived from the single site detection method.

Observe that although the cross-site pattern PB does not detect bursts of positive sentiment on Yelp, it has the highest averaged suspicious score, and its detection precision is also reasonably high. The plateaus on Yelp may be further analyzed according to the elevations of the plateaus: 5-star reviews in high-rising plateaus can also be suspicious too, as they may correspond to long-lasting spamming attacks, while 5-star reviews in low plateaus shall be considered as less harmful ones, as it is seldom considered as an effective spamming strategy by spammers.

We return to the sufficiency issue of the Yelp spam filter. We see that the ratio of true spams in a uniform sample of the 5-

TABLE V: Human annotation results

Patterns	# reviews	Avg Scores	$Prec(> 1)$	$Prec(> 2)$
B*	93	1.9785	0.9677	0.3871
BB	18	1.9074	0.8889	0.4444
BP	75	1.9956	0.9867	0.3733
PB	68	2.0098	0.8971	0.3824
PP	55	1.8606	0.9091	0.2909
PD	14	1.7857	0.7857	0.2857

star reviews is pretty high in time windows detected by various cross-site patterns. Further notice that none of these detected spams is included in the Yelp’s filtered reviews. Therefore, we are almost sure, at least on the set of 5-star reviews, that Yelp’s filter system is not doing a very good job, and the proposed method can complement the detection power of Yelp’s system.

V. CLASSIFICATION

After finding the spams within the suspicious time windows obtained in the macroscopic level, in this section, we further investigate individual spam activity in the microscopic level.

Based on user behaviors (B) and review texts (T), we can extract 12 different features, which are summarized in Table VI. These features are termed as “microscopic” features as they characterize individual reviews and reviewers, in contrast to the macroscopic cross-site temporal patterns that describe collective dynamics. Among these 12 features, three new features (DC, DS, and MP) are proposed in this paper to reveal the reviewers who posted many reviews on one day, and the rest have been studied in [14]. We consider these three features because if one reviewer posted reviews in the same days for businesses in different cities or states, he/she is quite suspicious. To get a quick glimpse of the effectiveness of these features, we first use the human annotated dataset to find the correlations between each feature and the labels provided by annotators. As shown in the second column of Table VI, the three new features (DC, DS and MP) are more correlated to spamming activities in the detected time windows, while the traditional features, such as the length of a review, seem to be less correlated.

The first six features in Table VI are based on user behaviors and the rest are based on review texts.

To demonstrate the usefulness of these microscopic features in spam detection, we implement two classification models. The first model is based on class prior [20]. It serves as an unsupervised classifier. ROC curves of the model using 3 different sets of features (behavioral+textual, behavioral only and textual only) are shown in Figure 2(a), with the actual AUC shown in the legend. We can see that, without text features, AUC is even higher. We can draw a similar conclusion from the precision-recall curve in Figure 2(b).

To take into account different contributions of the microscopic features, we implement a linear regression model, with the coefficients for each feature being the feature’s correlation to spamming activity. Training and testing are based on 5-fold cross-validation. The resulting curves are shown in Figure 2(c) and 2(d). Compared Figure 2(c) with Figure 2(a), we can see

TABLE VI: Microscopic features of reviewers and reviews, and their correlations with the ground truths

Feature	Corr.	Description
DC(B)	+0.252	Proportion of days when a reviewer posts reviews on businesses in different cities.
DS(B)	+0.230	Proportion of days when a reviewer posts reviews on businesses in different states.
MP(B)	+0.183	Proportion of days when a reviewer posts 3 or more reviews.
LRR(B)	-0.148	Proportion of reviews with 1 or 2 stars posted by a reviewer.
FRR(B)	+0.121	Proportion of reviews with 5 stars posted by a reviewer.
RC(B)	+0.086	Sum of reviews posted by a reviewer.
WC(T)	-0.006	Sum of words in a review.
LC(T)	-0.010	Sum of letters in a review.
CWR(T)	+0.106	Proportion of ALL-CAPITAL words. (“I” excluded)
CLR(T)	+0.065	Proportion of capital letters.
1PP(T)	-0.034	Proportion of first person pronouns.
2PP(T)	+0.094	Proportion of second person pronouns.
EX(T)	+0.032	Proportion of exclamation.

that the linear regression model outperforms the first method. These two methods shows that, after detecting suspicious patterns via macroscopic signals, individual spamming reviews can be detected using the microscopic features, with the most significantly correlated ones best describing the spams. By taking this two-step approach, we can better understand the spamming activities in both the macroscopic and microscopic levels.

VI. CASE STUDIES

We use case studies of a business and two reviews to show the effectiveness of the proposed framework.

A. Case study of a business

We focus on a restaurant in Las Vegas. There are 552 reviews with rating of 3.0/5.0 on Yelp², and 150 tips with rating of 7.4/10 based on online voting on Foursquare³, indicating that this business is not that satisfactory. As shown in Figure 3, at around the 75-th time window (from March 15 to May 23, 2013), with 22 reviews and 13 tips posted on Yelp and Foursquare respectively, several cross-site patterns present in the period, as highlighted by the boxes:

- on the pair of curves of the count of reviews (CR) on Yelp (red curve) and Foursquare (orange curve), a plateau-burst pattern is detected,
- on the pair of curves of five-star ratio (FR) on Yelp (purple curve) and highly positive sentiment ratio $HPSR$

²<http://www.yelp.com/biz/gold-and-silver-pawn-shop-las-vegas>

³<https://foursquare.com/v/gold-silver-pawn-shop/4b2a869ef964a520cdaa24e3>

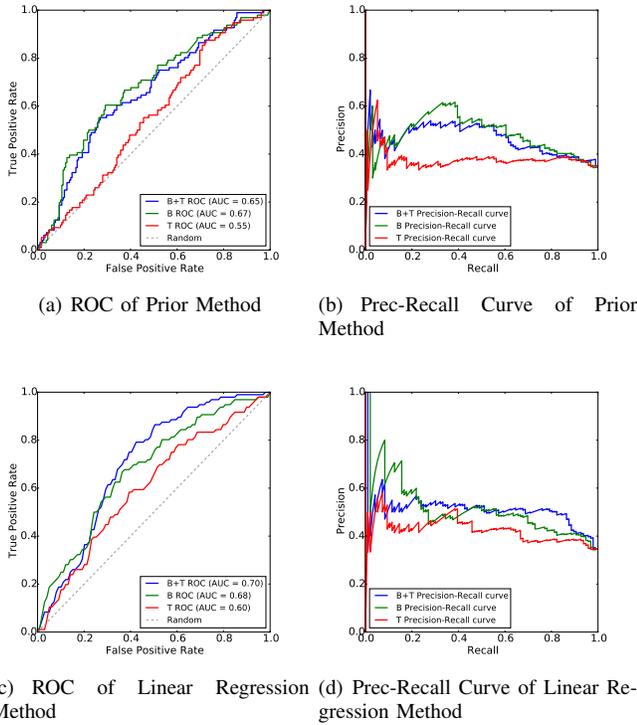


Fig. 2: ROC and Prec-Recall Curves for Two Methods

on Foursquare (green curve), a burst-plateau pattern is detected,

- on the pair of low rating ratio (*LR*) on Yelp (light grey curve) and negative sentiment ratio *NSR* on Foursquare (black curve), there is a dive-burst pattern.

As revealed in the experiments, those patterns are important macroscopic signals that suggest suspicion. Representative reviews on Yelp and Foursquare are listed in Table VII. In sum, an increasing number of 5-star reviews were posted on Yelp, many of them casting doubt on previous negative reviews. At the same time on Foursquare, negative tips were still dominating.

Let's focus on the 73rd time window. During that time, only two reviews were posted and both of them are 5-star, giving rise to a significant burst in FR and dive in NR. These 5-star reviews are suspicious. Take one of the reviewers, Sharttle who posted reviews in the time window, as an instance. Firstly, he has little social connection on Yelp with only one friend, and he should be considered as an inactive user on Yelp. However, he has posted a large number of reviews (47 in total). Secondly, he has posted 10 reviews on March 27, 2013, seven of which are 5-star. It is unlikely for a normal reviewer to recall all the details of ten businesses on the same day after the visit. However, on Foursquare, there is a slight burst in count of tips and significant burst in negative sentiment ratio, suggesting that this business may be favorably reviewed at that time. Most of the tips mentioned that this business was overpriced and going there was waste of time. The other reviews in the detected period on the two websites shows conflicting opinions, making the 5-star reviews on Yelp more

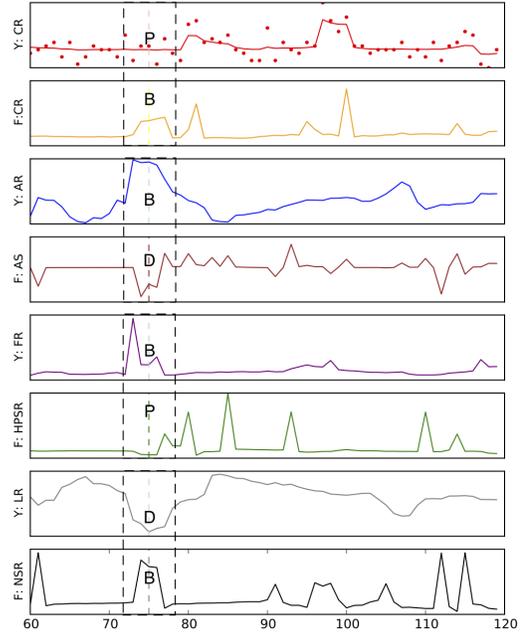


Fig. 3: A detected suspicious time window of a business

suspicious. This case study shows the effectiveness of spam detection via comparisons of various metrics on multiple sites and the microscopic features.

B. Case study of reviews

The first review⁴ in Table VII gives 5 stars while the average score of such business is 3.5/5.0. From the text, we can see that it is short and contains lots of mistakes. It speaks highly of the business without evidence and detail. Behavior of the reviewer⁵ is also suspicious. At first, she did not upload her photograph and she has only seven friends, which means this tends to be an account exclusively for spamming. Besides, 70 out of the 93 reviews she has posted are 5-star. This is an indicator that she is likely to be a spammer posting opinion reviews to promote businesses. Moreover, based on our statistical analysis, 25% of days when she posted at least one review are related to different cities, and in 15% of days she posted at least three reviews when she has a post. On May 23, she even posted reviews on businesses in Los Angeles, Burbank, North Hollywood, Van Nuys and Glendale. In fact, it is impossible for a normal user to visit and review so many places in one day. There is another suspicious review⁶, which is regarded by the annotators as spam because of its unconvincing text. It mentioned prices several times, and used the all-capital word “FREE” and the dollar sign “\$”. It also

⁴<http://www.yelp.com/biz/orleans-hotel-and-casino-las-vegas-2?hrid=ixvLjZaFQ1zB4TWR34g3Bg>

⁵http://www.yelp.com/user_details?userid=LC-IqRqt5e5sfzIDfp3jJg

⁶<http://www.yelp.com/biz/orleans-hotel-and-casino-las-vegas-2?hrid=JZLz5kDISx8m46mm-RnPia>

TABLE VII: Case study: representative reviews (the codes under the site names indicate detected patterns)

Representative reviews	
Yelp	(5 stars)... really was awesome to be there. I don't know why people are complaining, ...
CR: P	
AR: B	(5 stars) Ignore the negative reviews... that part was fun in itself!
FR: B	
LR: D	(5 stars) ... I don't know why people are complaining, they don't even have to have it opened, but they do. Enjoy it!
	(5 stars) ... parking is FREE... they have items on display from \$100,000 and more to magnets of the cast for \$8.00...
Foursquare	Waste of a trip!
CR: B	They are way over priced on every-thing, including there franchised items from the show.
AS: D	
HPSR: P	Extremely overpriced, they got famous on TV and now screw everyone with high prices!
NSR: B	An exhilarating experience. I find going to dumps and almost getting murdered exhilarating.
	Waste of time!!!

uses more second personal pronouns like “you” (account for 8%) than first personal pronouns like “I” (only account for 1.6%), which makes itself more like a business promoter.

VII. RELATED WORK

Spam detection on review websites has been an important research area. Based on the features used for the detection, there are approaches using textual, behavioral and network features. [16], [21] used text similarity as a sign of spamming. Reviews sharing similar words or even topics are believed to be written by professional spammers. [19] collected spam reviews written by a crowdsourcing service, and studied the collected reviews to reveal several textual features that are predictive of spams. In [7], the authors studied spams from an NLP perspective. Irrelevance of review texts is also considered as an indicative feature of spams [23], [26]. Behavioral features are proved to be quite important in spam detection, since spammers have learned to write reviews that sound more realistic, rendering the traditional text-based detection algorithms less helpful. [15] studied rating behaviors features that are useful for spam detection, such as extreme rating, rating deviation, early time frame and rating abuse. [2] studied a new type of spams that are caused by organized spammers targeting at the same group of products or brands. Network based spam detection construct features from the network that connecting reviewers, reviews and products. [27] used heterogeneous pairwise features to find conflicts between behaviors and linguistic patterns to detect spam. Such features include neighbor diversity and self-similarities of the nodes

in the network [28]. In [22], [20], the authors propose to use supervised and unsupervised label propagation to define a score of spamming of all nodes in the constructed network. A semi-supervised learning method is presented in [11], to avoid manual labeling for possible suspicious reviews. All these studies can be considered to complementary to the proposed pipeline here, as one can use the previous work as further evidences for the detection and investigation of the spams found by our method.

There are some other less traditional features used for spam detection. Time series anomalies are proposed as a macroscopic spam detector [25], [6]. The basic idea is that a sudden rise in positive sentiment can be the indicator of spamming activities. However, these studies only focus on time series patterns defined on single data source. Our work here is based on the similar idea, but we extend previous work to consider data from multiple review websites. Ranking consistency is also proposed in [1] to define ranking and rating anomalies. In that paper, rankings of products are collected from a wide range of review websites to provide a good product ranking for detection.

In terms of utilizing multiple websites for anomaly detection, there are much research [14], [24], [10], [9]. In [14], [10], [9], the authors construct cross-site features based on the pairwise interactions of single-site features, and there is no time series involved for the macroscopic detection. In [24], the authors detect businesses that have different rating behaviors across websites, but did not provide a way to find out actual spams on the microscopic level. More importantly, inconsistency is not the only suspicious sign of cross-site spamming, as shown in our experiments.

VIII. CONCLUSION

In this paper, an opinion spam detection framework is proposed. We design several cross-site sentimental time series patterns to capture suspicious activities in multiple review websites. An efficient detection algorithm is proposed to find such patterns. We characterize the patterns from the macroscopic and microscopic aspects to fully understand the patterns using human annotated dataset. We found that the extracted patterns can detect review spams that were not filtered out by a commercial anti-spam system. Novel features are designed to enable automatic classifications of the spams. Case studies show that we can use the cross-site patterns to find convincing evidences of spamming activities.

ACKNOWLEDGEMENT

This work is supported in part by NSF through grants IIS-1526499, and CNS-1626432. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

REFERENCES

- [1] An effective and economic bi-level approach to ranking and rating spam detection. DSAA, 2015.
- [2] Mukherjee A, Liu B, and Glance N. Spotting fake reviewer groups in consumer reviews. WWW '12.

- [3] Judith A Chevalier and Dina Mayzlin. The Effect of Word of Mouth on Sales: Online Book Reviews. 2006.
- [4] Lim E-P, Nguyen V-A, Jindal N, Liu B, and Lauw H W. Detecting product review spammers using rating behaviors. CIKM '10.
- [5] Chandra Erdman, John W Emerson, et al. bcp: an r package for performing a bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13, 2007.
- [6] Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Exploiting burstiness in reviews for review spammer detection. ICWSM, 2013.
- [7] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. ACL, 2012.
- [8] Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. Distributional footprints of deceptive product reviews. ICWSM, 2012.
- [9] Liang Ge, Jing Gao, Xiaoyi Li, and Aidong Zhang. Multi-source deep learning for information trustworthiness estimation. KDD, 2013.
- [10] Liang Ge, Jing Gao, Xiao Yu, Wei Fan, and Aidong Zhang. Estimating local information trustworthiness via multi-source joint matrix factorization. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 2012.
- [11] Bing Liu HuayiLi, Arjun Mukherjee, and Jidong Shao. Spotting fake reviews using positive-unlabeled learning. *Computación y Sistemas*, 18(3):467–475, 2014.
- [12] Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. Learning to identify review spam. IJCAI, 2011.
- [13] Luca M. Reviews, reputation, and revenue: The case of yelp.com. In *Harvard business school working papers, Harvard Business School*, 2011.
- [14] Amanda J. Minnich, Nikan Chavoshi, Abdullah Mueen, Shuang Luan, and Michalis Faloutsos. Trueview: Harnessing the power of multiple review sites. WWW, 2015.
- [15] Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Spotting opinion spammers using behavioral footprints. KDD, 2013.
- [16] Jindal N and Liu B. Opinion spam and analysis. WSDM '08.
- [17] Jindal N, Liu B, and Lim E-P. Finding unusual review patterns using unexpected rules. CIKM '10.
- [18] Myle Ott, Claire Cardie, and Jeff Hancock. Estimating the prevalence of deception in online review communities. WWW, 2012.
- [19] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. HLT, 2011.
- [20] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. KDD, 2015.
- [21] Vlad Sandulescu. *Opinion spam detection through semantic similarity*. PhD thesis, MSc thesis (Unpublished), Technical University of Denmark, 2014.
- [22] Guan Wang, Sihong Xie, Bing Liu, and Philip S Yu. Identify Online Store Review Spammers via Social Review Graph. *ACM Trans. Intell. Syst. Technol.*, 2012.
- [23] Jing Wang, Clement T. Yu, Philip S. Yu, Bing Liu, and Weiyi Meng. Diversionary comments under political blog posts. CIKM, 2012.
- [24] Houping Xiao, Yaliang Li, Jing Gao, Fei Wang, Liang Ge, Wei Fan, Long H. Vu, and Deepak S. Turaga. *Believe It Today or Tomorrow? Detecting Untrustworthy Information from Dynamic Multi-Source Data*. 2015.
- [25] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S Yu. Review Spam Detection via Temporal Pattern Discovery. KDD, 2012.
- [26] Sihong Xie, Jing Wang, Mohammad S.Amin, Baoshi Yan, Anmol Bhasin, Clement Yu, and Philip S.Yu. A context-aware approach to detection of short irrelevant texts. DSAA, 2015.
- [27] Chang Xu and Jie Zhang. Combating product review spam campaigns via multiple heterogeneous pairwise features. *SDM. SIAM*, 2015.
- [28] Junting Ye and Leman Akoglu. Discovering Opinion Spammer Groups by Network Footprints. ECML/PKDD, 2015.