Reliable Fake Review Detection via Modeling Temporal and Behavioral Patterns

Xian Wu, Yuxiao Dong, Jun Tao, Chao Huang, Nitesh V. Chawla
Department of Computer Science and Engineering, and iCeNSA
University of Notre Dame, Notre Dame, IN 46556, USA
{xwu9, ydong1, jtao1, chuang7, nchawla}@nd.edu*

***Abstract*—Fake reviews have become a pervasive problem in online review systems, wherein fraudulent users manipulate the perception of an object (e.g., a restaurant) by fabricating fake reviews. Extensive work has been devoted to identifying fake reviews via modeling different factors separately, such as user features, object characteristics, and user-object bipartite relations. However, this problem remains challenging due to the fact that more advanced camouflage strategies are utilized by malicious users. In real-world scenarios, spammers may pretend to be normal users by giving fake reviews with the similar score distribution as normal users. To address these issues, we propose to explore the temporal patterns of users' review behavior, because spammers prefer to promote or demote the target businesses in a short period of time. In this work, we present a unified framework *Reliable Fake Review Detection (RFRD)* that explicitly models temporal patterns of users' review behavior into a probabilistic generative model. Moreover, the RFRD framework models users' underlying review credibility and objects' highly-skewed review distributions. We conduct experiments on two Yelp datasets, demonstrating the effectiveness of the proposed RFRD framework.**

***Keywords*-Fake Review Detection; Fraud Detection; Probabilistic Generative Model**

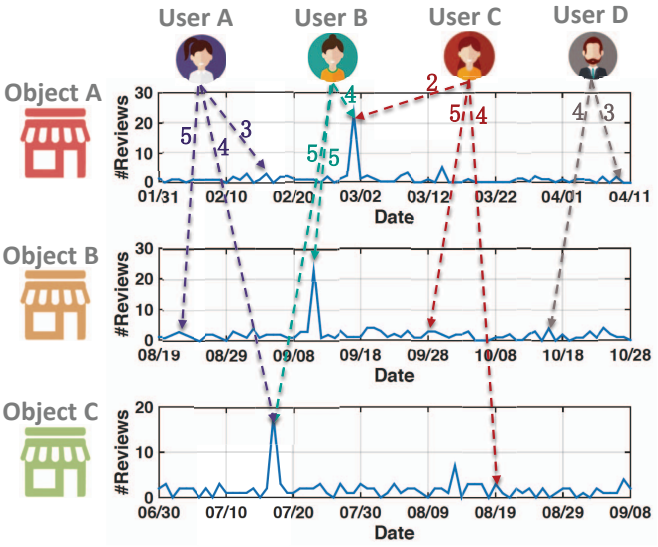


Figure 1. An illustrative example on Yelp

I. Introduction

Online review systems have emerged as an important channel for people to share their opinions on different objects, e.g., products and local businesses [24]. Such online reviews play a critical role in the decision making process of customers' purchase behavior. Unfortunately, some businesses purposely generate opinion spams in order to manipulate the perceptions of their or competitors' products [13]. Therefore, a fundamental problem in online review systems is how to automatically ascertain the unreliable and fake reviews. We refer to this task as the problem of *fake review detection*.

Much work has been proposed to detect fake reviews by leveraging individual factors, including user features and object characteristics, as well as user-object bipartite relations [22], [1], [6], [5]. However, those conventional detection methods fail when simple camouflage strategies are used by spammers. For example, spammers may pretend to be normal users by giving genuine reviews mixed with fake ones [7], making it difficult to distinguish them from each other. First, a fake review may not be detected due to the relatively normal behavior of a camouflaged account. Second, a genuine review may be considered as fake if it is among a burst-sequence of fake reviews. Therefore, individual factor based detection techniques are not capable of offering a reliable detection.

Figure 1 shows an illustrative example from Yelp. Each arrow in this figure represents the review relation between a user and an object in the corresponding timestamp. We can observe that reviews reported by user B are more likely to be fake, since user B always gives reviews during the burst period of each object. Additionally, user A and user C have a high probability to report genuine reviews in burst periods, because user A and user C write only one review in the burst period and the remaining in non-burst period. However, it's difficult to distinguish fake reviews from genuine ones based on the rating scores, due to the score camouflage strategies which might be used by spammers. In this case, the reliable detection of fake reviews requires the joint modeling of user and object features coupled with temporal patterns. In addition to model explicit factors, it would be helpful to characterize both users' and objects latent traits, such as users' review credibility and objects' review reliability.

In this work, we present a probabilistic generative model—Reliable Fake Review Detection (RFRD)—to provide reliable review detections in online systems. First, the

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

RFRD framework explicitly explores user and object features, as well as their temporal patterns. Second, it is enabled to learn users' latent review credibility and objects' sparse review distributions. With both explicit and implicit factors, the RFRD model also leverages their inter-dependencies to detect spam reviews. We perform experiments on two Yelp datasets. The experimental results show that our method can accurately detect fake reviews and significantly outperform the state-of-the-art baselines.

The rest of the paper is organized as follows. Section II formally formulates the problem. Section III explains the proposed approach. Section IV presents experimental results that validate the efficacy of our methodology. Section V discusses related work. Finally, Section VI concludes the study.

II. Problem Formulation

We formalize a bipartite graph $G = (U, O, E, R)$ to model the online review data, where $U = \{u\}$ denotes the set of users, $O = \{o\}$ denotes the set of objects (e.g., restaurants or products), and $E \subset U \times O$ represents the set of review relationships between users in U and objects in O. Each relationship $e_{u,o} \in E$ is associated with a set of reviews $R = \{R_{u,o,k}\}$, where $R_{u,o,k}$ is user u's review on object o at time k.

Furthermore, we define $R^T_{u,o,k} = r$ if there are r reviews reported on object o together with u's review $R_{u,o,k}$ at time k. In addition, we define O_u and U_o to represent the set of objects reviewed by user u and the set of users who review object o, respectively. We summarize the notations used in this paper in Table I. Given these definitions, the fake review detection problem is formalized as follows:

Problem: **Fake Review Detection.** Given a bipartite review graph $G = (U, O, E, R)$, the goal of the problem is to estimate the label of each review $\{R_{u,o,k}\} \subset R$, i.e., true or fake.

III. RFRD: Reliable Fake Review Detection

In this section, we present the Reliable Fake Review Detection (RFRD) framework to detect fake reviews in online review platforms. In particular, the RFRD framework presents a probabilistic generative model to explicitly model the distributions of objects' reviews and users' credibility from sparse review data. An Expectation Maximization (EM) based learning algorithm is developed to learn the model parameters.

A. The RFRD Model

To capture the fake reviews in online review systems, we first model each object's type $c \in C$, where C is the credibility category set. $C = \{G, S\}$, where G and S represents genuine and spam category, respectively. Following this rule, users who always report reviews under the spam category S tend to be classified as spammers, while those who tend to review under G category are genuine users. We model objects' two category distributions as uni-variant Gaussian Distributions (i.e., Normal Distributions), since the distribution of review features on an object with a certain type is around a central value (i.e., Central Limit Theorem [3]). In doing so, we define the marginal function $p(R|\mathbf{\Theta})$ of RFRD as follows:

$$p(R|\mathbf{\Theta}) = \prod_{u \in U} \Big(\sum_{c \in C} p(c) p(R_u|\Theta, c) \Big) = \prod_{u \in U} \Big(\sum_{c \in C} p(c) \prod_{o \in O_u} \prod_{k \in K_{u,o}} p(R^T_{u,o,k}|\Theta, c) \Big) \quad (1)$$

where $\mathbf{\Theta}$ is the vector of estimation parameters and $p(R^T_{u,o,k}|c)$ denotes the probability of object o getting the review's temporal feature $R^T_{u,o,k}$ given by user u in the k-th time slot with a certain credibility category c.

One way to identify fake reviews is to maximize the marginal likelihood $p(R|\mathbf{\Theta})$ in Equation (1). However, it is infeasible to use the maximum likelihood estimation (MLE) to infer the review distributions of objects with a limited number of reviews. For example, objects with only one review will have zero variance over distributions, leading to the intractable to calculate $p(R^T_{u,o,k}|c)$, and finally fail to estimate user's credibility. This issue becomes particularly challenging for the Yelp review data that we used, in which more than 10% of objects receive only one review.

To overcome the above challenge, we instead formulate the fake review detection problem as the maximum a posteriori (MAP) estimation task, which allows us to incorporate the *priori belief* (conjugate priors) into the parameter inference process to avoid zero variance. We set the prior distribution of Multinomial distribution $p(c|\boldsymbol{\pi})$ as Dirichlet distribution. Notice that the mean and variance in the Normal distribution $p(R^T_{u,o,k}|\mu_{o,c}, \sigma^2_{o,c})$ are unknown beforehand; thus we model the conjugate prior of Normal distributions as Normal-inverse-chi-squared distribution (*NIX*) [14]. Formally, we summarize our generative model

Table I
Symbols and Definitions

Symbol	Interpretation
U, O	the user set, the object set
u, o	a user, an object
$R^T_{u,o,k}$	the number of reviews reported on object o together with review $R_{u,o,k}$
O_u	the set of objects rated by user u
U_o	the set of users giving rating scores on object o
$K_{u,o}$	the set of times that user gave rating scores on object o
c	index of credibility category (G, S)
$\boldsymbol{\pi}$	parameters of review credibility category distribution
$\boldsymbol{\mu}, \boldsymbol{\sigma^2}$	parameters of all objects' review distribution
$\boldsymbol{\alpha}$	hyperparameters of Dirichlet distribution
$\mu_0, \kappa_0, \nu_0, \sigma^2_0$	hyperparameters of Normal-inverse-chi-squared distribution

as follows:

$$\begin{aligned} c &\sim \textit{Multinomial}(c; \boldsymbol{\pi}) \\ \boldsymbol{\pi} &\sim \textit{Dirichlet}(\boldsymbol{\pi}; \boldsymbol{\alpha}) \\ R^T_{u,o,k} &\sim \textit{Normal}(R^T_{u,o,k}; \mu_{o,c}, \sigma^2_{o,c}) \\ \mu_{o,c}, \sigma^2_{o,c} &\sim \textit{NIX}(\mu_{o,c}, \sigma^2_{o,c}; \mu_0, \kappa_0, \nu_0, \sigma^2_0) \end{aligned} \tag{2}$$

where the variable on the left side of the semi-colon is assigned a probability under the parameter on the right side. $\boldsymbol{\alpha}$ are hyperparameters of Dirichlet Distribution, and μ_o, κ_o, ν_o, and σ^2_o are hyperparameters of Normal-inverse-chi-squared Distribution (*NIX*). Specifically, $\mu_{o,c}$ and $\sigma^2_{o,c}$ are the mean and variance of object o's review distribution under category c. The types of user u obeys a Multinomial distribution with parameters $\boldsymbol{\pi}$. The posterior distribution of the RFRD model is formally defined as follows:

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}|R) \propto p(R|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\sigma^2}) \tag{3}$$

By incorporating Equation (1) into the posterior distribution, we have the following log likelihood objective function:

$$\begin{aligned} \mathcal{O}(\boldsymbol{\Theta}) = & \log \Big(p(R|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\sigma^2}) \Big) \\ = & \sum_{u \in U} \log \Big(\sum_{c \in C} p(c|\boldsymbol{\pi}) \prod_{o \in O_u} \\ & \times \prod_{k \in K_{u,o}} p(R_{u,o,k}|\mu_{o,c}, \sigma^2_{o,c}) \Big) \\ & + \log p(\boldsymbol{\pi}) + \sum_{o \in O} \sum_{c \in C} \log p(\mu_{o,c}, \sigma^2_{o,c}) \end{aligned} \tag{4}$$

B. Learning Algorithm

In Equation (4), the summation over all credibility category C is within the *log* operation, making it difficult to be differentiated. To address this issue, we apply the Expectation Maximization (EM) algorithm [2] to move the *log* operation into the summation over C. In doing so, we convert the maximization of Equation (4) to the problem of maximizing the auxiliary function $\mathrm{Q}(\boldsymbol{q}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$—the lower bound of the joint distribution in Equation (4) is guaranteed by the Jensen's inequality. The auxiliary function can be given as follows:

$$\begin{aligned} \mathrm{Q}(\boldsymbol{q}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = & \sum_{u \in U} \sum_{c \in C} q_{u,c} \log \frac{p(R_u|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, c)}{q_{u,c}} \\ & + \sum_{o \in O} \sum_{c \in C} \log p(\mu_{o,c}, \sigma^2_{o,c}) + \log p(\boldsymbol{\pi}) \end{aligned} \tag{5}$$

where $q_{u,c} \in \boldsymbol{q}$ with $\sum_c q_{u,c} = 1$.

The EM algorithm estimates the values of unknown parameters that are not directly observable from the data iteratively. Its iteration usually covers two key steps: the expectation step (E-step) and maximization step (M-step). In E-step, it computes the expectation of the auxiliary function using the current estimates of the model parameters. In M-step, it derives the new solutions of parameters that maximize the expected auxiliary found in the E-step. Specifically, the two steps are introduced below.

E-step: We aim to maximize the $\mathrm{Q}(\boldsymbol{q}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$ with respect to $\boldsymbol{q}$ by setting $q^t_{u,c} = p(c|R_u, \boldsymbol{\pi}^{t-1}, \boldsymbol{\mu}^{t-1}, \boldsymbol{\sigma^2}^{t-1})$. We define t as the iteration index. In this way, $q^t_{u,c}$ in t-th iteration can be updated as follows:

$$\begin{aligned} & q^t_{u,c} = p(c|R_u, \boldsymbol{\pi}^{t-1}, \boldsymbol{\mu}^{t-1}, \boldsymbol{\sigma^2}^{t-1}) = \\ & \frac{\pi_c \prod_{o \in O_u} \prod_{k \in K_{u,o}} p(R^T_{u,o,k}|\mu^{t-1}_{o,c}, \sigma^{2\ t-1}_{o,c})}{\sum_{c' \in C} \pi_{c'} \prod_{o \in O_u} \prod_{k \in K_{u,o}} p(R^T_{u,o,k}|\mu^{t-1}_{o,c'}, \sigma^{2\ t-1}_{o,c'})} \end{aligned} \tag{6}$$

M-step: In the M-step, we derive the solutions of $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma^2}$, which maximize $\mathrm{Q}(\boldsymbol{q}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$, with the new estimates of $\boldsymbol{q}^t$ in E-step. We can get solutions of model parameters as follows:

$$\begin{aligned} \pi^t_c &= \frac{\sum_u q^t_{u,c} + \alpha_c - 1}{|U| + \sum_{c'} \alpha_{c'} - |C|} \\ \mu^t_{o,c} &= \frac{\sum_{u \in U_o} \sum_{k \in K_{u,o}} q^t_{u,c} R^T_{u,o,k} + \kappa_0 \mu_0}{\sum_{u \in U_o} \sum_{k \in K_{u,o}} q^t_{u,c} + \kappa_0} \\ \sigma^{2\ t}_{o,c} &= \frac{\sum_{u \in U_o} \sum_{k \in K_{u,o}} q^t_{u,c} (R^T_{u,o,k} - \mu^t_{o,c})^2}{\sum_{u \in U_o} \sum_{k \in K_{u,o}} q^t_{u,c} + 3 + \nu_0} \\ & \quad + \frac{\nu_0 \sigma^2_0 + \kappa_0 (\mu_0 - \mu^t_{o,c})^2}{\sum_{u \in U_o} \sum_{k \in K_{u,o}} q^t_{u,c} + 3 + \nu_0} \end{aligned} \tag{7}$$

where $|U|$ is the number of users and $|C|$ is the number of rating criteria. Since $C = \{G, S\}$ in this work, $|C| = 2$.

Output: the probability of each review as fake or genuine is shown as follows:

$$p(c|R^T_{u,o,k}) = \frac{p(c|R_u, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})(R^T_{u,o,k}|\mu_{o,c}, \sigma_{o,c})}{\sum_{c'} p(c'|R_u, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) p(R^T_{u,o,k}|\mu_{o,c'}, \sigma_{o,c'})} \tag{8}$$

The proposed framework solves a MAP estimation problem where the users' credibility is modeled as latent variables. In particular, the framework jointly assigns credibility to users and distribution parameters to objects through EM. Specifically, in E-step, it estimates each user's credibility based on the current estimates of the model parameters (e.g., the mean and variance). In M-step, it computes the new estimates of the model parameters that maximize the expected log posterior function in E-step. The time complexity of RFRD framework is $O(|R|)$ which is linear to the number of reviews.

IV. Evaluation

In this section, we perform extensive experiments to evaluate the performance of *Reliable Fake Review Detection (RFRD)* framework on real-world datasets collected from a online review system (i.e., Yelp). We first introduce experimental settings and then present the evaluation results.

Table II
DATASET STATISTICS

Dataset	#Users	#Objects	#Reviews	Spams%	Density%
CLT	8,870	3,130	29,640	41.7	0.107
PHX	24,506	6,167	62,803	36.5	0.042

A. Experimental Setup

Dataset Description. In order to evaluate the proposed method in real-world settings, we adopt the online review platform Yelp as testbeds. Yelp has emerged as a new experiment platform where massive reviews are uploaded voluntarily from crowds about the quality assessments of the businesses [11]. The reported reviews on Yelp may be fake or genuine due to the open data collection environment and unvetted data sources [13]. However, this noisy nature of Yelp actually provides us a good opportunity to investigate the performance of our scheme and the state-of-the-art techniques in real-world scenario.

In this study, we use two real-world datasets collected from Yelp. Specifically, we employ the review data spanning between Jan. 2013 and Dec. 2013 from two cities: *Charlotte (CLT)* and *Phoenix (PHX)* released by the Yelp Dataset Challenge[1]. We target these cities due to their urban diversity, reflective of different degrees of review activities. Each review in the Yelp datasets is formatted as: (user ID, business ID, review, timestamp). The statistics of the collected datasets are summarized in Table II. Figure 2 reports the distributions of #reviews per user and #reviews per object, respectively. We can observe that more than 10% of users (objects) report (receive) only one review in both two cities, which suggests the sparsity of online review datasets.

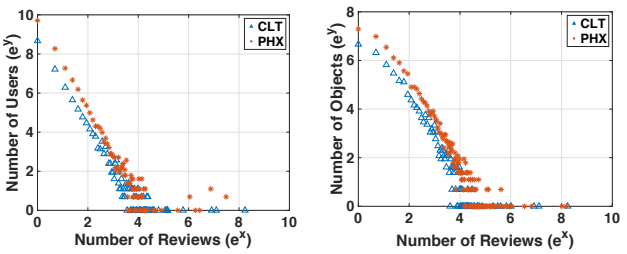
Evaluation. Yelp has its own proprietary algorithm to detect fake/suspicious reviews [16]. These reviews are available and can be accessed in Yelp website [2]. While the detection results of Yelp filtering system may not be perfectly accurate, it has been shown to be effective to identify fake reviews [19]. Thus, in our evaluation, we regard those filtered reviews as *near* ground truth to evaluate the accuracy on fake review detection results of all compared algorithms. In particular, the review filtered by Yelp system will be considered as *fake review*. Otherwise, we consider the review as *genuine review*. Similar evaluation rubrics have been used in recent research work on fake review detection [16].

1) Evaluation Metrics: In the experiments, we use two categories of evaluation metrics (i.e., *Classification* and *Ranking-based* Metrics) to evaluate the performance of all compared schemes.

Classification Metrics. The first set of metrics are used to evaluate the accuracy of different techniques in terms of detecting fake reviews. These metrics include *Precision (Pre)*, *Recall (Rec)*, *F1-Score (F1)* and *Accuracy (Accu)*. In our experiment, the True Positives and True Negatives are the reviews that are correctly classified by a particular scheme as fake or genuine, respectively. The False Positives and False Negatives are the genuine and fake reviews that are misclassified to each other, respectively.

[1]https://www.yelp.com/dataset_challenge

[2]https://www.yelpblog.com/2010/03/yelp-review-filter-explained



(a) Users' Review (b) Objects' Review

Figure 2. Review Distributions on Yelp

Ranking-based Metrics. The second set of metrics are used to evaluate the ranking performance of all compared algorithms. We use *Average Precision (AP)* and *Area Under Curve (AUC)* to quantify the performance of all compared schemes by computing the area under the curves. In specific, the higher AP and AUC values means more accurate fake review detection results.

2) Baselines: In this evaluation, we compare our RFRD framework with the state-of-the-art fake review detection and fact-finding techniques.

- *Singular Value Decomposition (SVD)* [20]: it aims at discovering deceptive opinion spam by using Singular Value Decomposition approach through projecting high dimensional data into a lower dimension space.
- *TruthFinder (TF)* [23]: it aims to infer reviewers' reliability using a heuristic based pesudo-probabilistic model. We regard the reviews from users with low reliability inferred by TruthFinder as fake reviews.
- *Average_Log (AL)* [15]: it detects a reviewer's credibility by considering the structure of the generated bipartite graph between reviewers and objects. Reviews reported from untrustworthy reviewers are regarded as fake.
- *HITS* [8]: it estimates a hub and an authority score to find popular webpages. The reviews from users with low credibility are regarded as fake.
- *FRAUDEAGLE* [1]: it is a propagation-based fake review detection algorithm which explores the network effect between users and objects in online review platforms.

B. Performance Validation

Table III presents the evaluation results on Charlotte datasets. In this table, we can observe that the RFRD scheme

outperforms the compared algorithms in most of the evaluation metrics. We repeated the above experiments on Phoenix datasets. The evaluation results are presented in Table IV. In those tables, we observe that RFRD continuously outperform all compared baselines in most of evaluation metrics. In the occasional cases that RFRD misses the best performance, it still achieves very competitive results.

Furthermore, the results across two cities with different degrees of data sparsity (see Table II) demonstrate that RFRD is robust to the data sparsity issue, which is largely due to our incorporation of the priori belief into the parameter inference process. The performance improvements of RFRD are achieved by: i) RFRD models the complex temporal patterns of users' review behaviors and complex user-object relations based on a probabilistic generative model. ii) RFRD carefully handles the sparse online review data as we discussed in Section III.

Table III
PERFORMANCE OF ALL COMPARED METHODS ON CHARLOTTE

Algorithm	Pre	Rec	F1	Accu	AP	AUC
SVD	0.392	0.796	0.525	0.399	0.765	0.662
TF	0.416	0.752	0.536	0.459	0.740	0.626
AL	0.358	0.637	0.458	0.366	0.428	0.358
HITS	0.367	0.690	0.479	0.402	0.812	0.749
FRAUDEAGLE	0.569	**0.835**	0.677	0.537	0.816	**0.797**
RFRD	**0.743**	0.659	**0.699**	**0.763**	**0.820**	0.786

Table IV
PERFORMANCE OF ALL COMPARED METHODS ON PHOENIX

Algorithm	Pre	Rec	F1	Accu	AP	AUC
SVD	0.390	**0.791**	0.522	0.471	0.521	0.339
TF	0.313	0.654	0.424	0.361	0.681	0.417
AL	0.302	0.588	0.399	0.317	0.463	0.442
HITS	0.267	0.581	0.366	0.306	0.709	0.665
FRAUDEAGLE	0.292	0.625	0.398	0.309	0.676	0.672
RFRD	**0.670**	0.659	**0.665**	**0.756**	**0.807**	**0.821**

V. RELATED WORK

Online Review Mining. Online review systems have emerged as a new sensing paradigm of collecting feedback about objects from human sources at scale. Recent research work addresses various problems in online review systems, such as opinion and sentiment mining [26], review rating prediction [25] and review summarization [12]. Different from the above works, we focus on fake review detection from online review data by exploring user and object features, as well as their temporal patterns.

Fake Review and Spammer Detection. Our work is closely related to the works that address the fake review or anomaly detection problem by exploring different dimensional information from data [1], [17], [10], [21]. For example, Akoglu et al. proposed a network-based approach to tackle the fake review detection problem in online review data by considering the graph structure among users and objects [1]. Shah et al. developed an adversarial approach to discover fraudsters by grouping similar users in the graph [17]. Lim et al. developed an aggregated behavior scoring method to identify spamming behaviors [10]. Different from the above works, our framework jointly considers credibility of users, the temporal feature of reviews received by objects, and their inter-dependencies. We further handle the sparsity of data by converting the MLE problem to a MAP problem.

Fact-Finding. There exists a good amount of work on the topics of *fact-finding*. Fact-finding techniques aim to infer the source reliability and object's correctness [23], [15], [4], [18]. For example, Yin et al. introduced TruthFinder as a transitive voting based fact-finder for trust analysis [23]. Other fact-finding methods [15], [9] enhanced the basic framework in which the truth detection step and source weight estimation step are iteratively conducted until convergence. Furthermore, [4] studied the problem of fact-finding by estimating the correct label of objects from noisy annotations. However, the above solutions aim at estimating the credibility of objects. Instead, we develop a principled approach to detect fake reviews which are reported from users on different objects.

VI. CONCLUSION

This paper presents an effective and efficient framework (i.e., RFRD) to solve the fake review detection problem in online review platforms. The new framework takes the feature of temporal report activities of all users into account in the fake review detection solution. The proposed approach jointly estimates the reliability of users as well as the truthfulness of reviews using expectation maximization scheme. We evaluate the RFRD framework using the real-world datasets collected from Yelp. The results demonstrate that our solution achieved significant performance gains compared to the state-of-the-art baselines.

ACKNOWLEDGMENTS

This work is supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 and the National Science Foundation (NSF) grant IIS-1447795. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. *ICWSM*, 13:2–11, 2013.

[2] M. Collins. The naive bayes model, maximum-likelihood estimation, and the EM algorithm. *Lecture Notes*, 2012.

[3] A. M. Committee et al. Uncertainty of measurement: implications of its use in analytical science. *Analyst*, 120(9):2303–2308, 1995.

[4] L. Duan, S. Oyama, M. Kurihara, and H. Sato. Crowdsourced semantic matching of multi-label annotations. In *IJCAI*, pages 3483–3489, 2015.

[5] S. Günnemann, N. Günnemann, and C. Faloutsos. Detecting anomalies in dynamic rating data: A robust probabilistic model for rating evolution. In *KDD*, pages 841–850. ACM, 2014.

[6] B. Hooi, N. Shah, A. Beutel, S. Günnemann, L. Akoglu, M. Kumar, D. Makhija, and C. Faloutsos. Birdnest: Bayesian inference for ratings-fraud detection. In *SDM*, pages 495–503. SIAM, 2016.

[7] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. Fraudar: Bounding graph fraud in the face of camouflage. In *KDD '16*, pages 895–904. ACM, 2016.

[8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632, 1999.

[9] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD*, pages 1187–1198. ACM, 2014.

[10] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *CIKM*, pages 939–948. ACM, 2010.

[11] M. Luca. Reviews, reputation, and revenue: The case of yelp. com. 2016.

[12] R. Mason, B. Gaska, B. Van Durme, P. Choudhury, T. Hart, B. Dolan, K. Toutanova, and M. Mitchell. Microsummarization of online reviews: An experimental study. In *AAAI*, pages 3015–3021, 2016.

[13] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance. What yelp fake review filter might be doing? In *ICWSM*, 2013.

[14] K. P. Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16, 2007.

[15] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING*. ACM, 2010.

[16] S. Rayana and L. Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *KDD*, pages 985–994. ACM, 2015.

[17] N. Shah, A. Beutel, B. Gallagher, and C. Faloutsos. Spotting suspicious link behavior with fbox: An adversarial perspective. In *ICDM*, pages 959–964. IEEE, 2014.

[18] B. Shi and T. Weninger. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104:123–133, 2016.

[19] K. Weise. A lie detector test for online reviewers. *Bloomberg Business Week*, 2011.

[20] G. Wu, D. Greene, and P. Cunningham. Merging multiple criteria to identify suspicious reviews. In *Recsys*, pages 241–244. ACM, 2010.

[21] J. Ye and L. Akoglu. Discovering opinion spammer groups by network footprints. In *ECML/PKDD*, pages 267–282. Springer, 2015.

[22] J. Ye, S. Kumar, and L. Akoglu. Temporal opinion spam detection by multivariate indicative signals. *ICWSM*, 2016.

[23] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *TKDE*, pages 796–808, 2008.

[24] R. Zhang, C. Sha, M. Zhou, and A. Zhou. Exploiting shopping and reviewing behavior to re-score online evaluations. In *WWW*, pages 649–650. ACM, 2012.

[25] W. Zhang, Q. Yuan, J. Han, and J. Wang. Collaborative multi-level embedding learning from reviews for rating prediction. In *IJCAI*, 2016.

[26] W. X. Zhao, J. Wang, Y. He, J.-R. Wen, E. Y. Chang, and X. Li. Mining product adopter information from online reviews for improving product recommendation. *TKDD*, 10(3):29, 2016.