An Inferable Representation Learning for Fraud Review Detection with Cold-start Problem

Qian Li, Qiang Wu, Chengzhang Zhu, Jian Zhang
Faculty of Engineering and Information Technology
University of Technology Sydney, Australia
{Qian.Li-7, Chengzhang.Zhu}@student.uts.edu.au
{Qiang.Wu, Jian.Zhang}@uts.edu.au

Wentao Zhao
College of Computer
National University of Defense Technology, China
wtzhao@nudt.edu.cn

***Abstract*—Fraud review significantly damages the business reputation and also customers' trust to certain products. It has become a serious problem existing on the current social media. Various efforts have been put in to tackle such problems. However, in the case of cold-start where a review is posted by a new user who just pops up on the social media, common fraud detection methods may fail because most of them are heavily depended on the information about the user's historical behavior and its social relation to other users, yet such information is lacking in the cold-start case. This paper presents a novel Joint-bEhavior-and-Social-relaTion-infERable (JESTER) embedding method to leverage the user reviewing behavior and social relations for cold-start fraud review detection. JESTER embeds the deep characteristics of existing user behavior and social relations of users and items in an inferable user-item-review-rating representation space where the representation of a new user can be efficiently inferred by a closed-form solution and reflects the user's most probable behavior and social relations. Thus, a cold-start fraud review can be effectively detected accordingly. Our experiments show JESTER (i) performs significantly better in detecting fraud reviews on four real-life social media data sets, and (ii) effectively infers new user representation in the cold-start problem, compared to three state-of-the-art and two baseline competitors.**

I. Introduction

Social media is becoming increasingly significant and heavily affects our daily life. Unfortunately, a large proportion of social media reviews are proposed by fraudsters for strong incentives of profit and reputation [1]. For example, 25% of Yelp reviews could be fraud as reported in 2013[1]. This proportion has increased rapidly as observed in 2017[2]. As a result, effectively detecting such fraud reviews is a critical task that has great business values [2].

Current efforts on fraud review detection mainly focus on analyzing the behavior and social relations of users and/or items corresponding to reviews [3]–[6]. They assume a fraud review may be posted by a user who has anomalous behavior, e.g. posted a lot of reviews within a short period. They also assume many fraud reviews are manipulated by a group of collaborative fraudsters, which may generate abnormal social relations of the fraudsters and their manipulated items. Because anomalous behavior and abnormal social relation have a good distinguishing ability, they have shown remarkable performance in detecting fraud reviews [7].

Although recent years have seen significant progress made in fraud review detection, detecting fraud reviews with the cold-start problem, i.e. *a new user just posted a new review*, is still a very challenging task and has been rarely studied. Specifically, the new users in the cold-start problem pose the two major challenges below. (i) A new user does not have historical information [8]. However, most of existing fraud review detection methods require the historical information to analyze user behavior [3], [4]. (ii) A new user does not show any explicit social relation, invalidating the detection of potential collaborative fraud review manipulation [6], [9].

To detect fraud reviews with the above cold-start problem, review content-based methods, such as [1], [10], [11], are the major solutions. They identify spam patterns, such as outlier review length and the large percentage of capital words, in the review content. Consequently, they avoid the negative effects brought by the lack of historical information and social relations in the cold-start case. Recent efforts further embed the relations between users, items and reviews into a vector representation of review content, resulting in a significantly better performance [8], [12].

However, an *indistinguishable problem* may arise: review content-based methods may fail to distinguish fraud reviews from honest reviews when the fraud and honest reviews have the same content. For example, review content-based methods cannot identify whether a review "*the product is good*" is fraud or honest, because the review content is easy to be imitated [9]. As a result, purely review content-based methods are ineffective when dealing with real-life fraud reviews [1].

In this paper, we propose a novel Joint bEhavior and Social relaTion infERable embedding (JESTER) method for fraud review detection with the cold-start problem. To solve the indistinguishable problem, JESTER considers a reviewing activity, which contains a user, an item, a review and a rating given by the user for the item. Subsequently, reviews with the same content can be distinguished if they are posted by different users for different items with different ratings. To tackle the challenges in the cold-start problem, JESTER jointly embeds the user behavior and user/item social relations into an inferable user-item-review-rating representation space, in

[1]https://www.bbc.com/news/technology-24299742

[2]https://www.forbes.com/sites/emmawoollacott/2017/09/09/exclusive-amazons-fake-review-problem-is-now-worse-than-ever

which the representation of a new user can be inferred through a closed-form solution.

Specifically, JESTER embeds a co-occurrence based user reviewing behavior by maximizing the success rate of existing behavior under a designated measure. It further embeds the user/item social relations according to the context information generated by random walks in the user-item networks, which is constructed by the reviewing activities. In the embedding process, JESTER seamlessly integrates the user/item social relations with user reviewing behavior to form the inferable user-item-review-rating representations. For a new user, JESTER infers the user representation as the one that can maximize the behavior success rate. In this way, JESTER enables an effective cold-start fraud review detection.

Accordingly, this paper makes three major contributions:

- We propose a novel representation learning method, JESTER, catering for cold-start fraud review detection in social media. JESTER jointly embeds user reviewing behavior and user/item social relations into inferable user-item-review-rating representations, and thus, the representations are more reliable for cold-start fraud detection than features which are extracted from review content.
- We propose a novel co-occurrence based user reviewing behavior embedding method. The embedded user reviewing behavior enables an efficient closed-form solution for the inferring of a new user representation.
- We seamlessly integrate the user/item social relations with user reviewing behavior. The integrated social relations provide more comprehensive evidence for fraud review detection, especially for reviews with collaborative manipulation.

Comprehensive experiments on four large real-world data sets demonstrate the effectiveness of the proposed model compared with three state-of-the-art and two baseline methods.

II. Related Work

A. Fraud Review Detection

Fraud review detection was initially studied in [13], and has long been an attractive research topic since then. Later, more efforts were made on employing user's behavior features, e.g. the historical reviewing statics of a user and the reviewing bias of a user [10], [14]–[19]. Also, Mukherjee et al. [20] proved that user's behavioral features are more effective than linguistic features for fraud detection. Besides the behavioral features, graph-based methods have been intensively studied. These methods reveal the users/items social relation, i.e. the user-user relation, item-item relation, and user-item relation, to identify fraud reviews. The intuition behind this is reviews posted by similar users or described similar items would have similar credibility. Wang et al. [21] first introduces review graph to capture the relationships between users and items. Spotting fraudster groups were then explored by network footprints [3], community discovery with sentiment analysis [22], social interactions for sparse group [23]. In-depth, Hooi et al. [9] involved dense subgraph mining for group fraud detection, targeting on detecting camouflage or hijacked accounts who manipulate their writing to look just like honest users.

B. Cold-Start Problem

The cold-start problem in fraud review detection has rarely been studied. Although review content-based method, i.e. extracting linguistic features to identify fraud reviews [18], [24]–[26], can alleviate the cold-start problem, they are insufficient when dealing with real life fraud reviews [1]. Recently, Wang et al. [12] embeds the relation between existing users, items, and reviews into the review representation that makes a significant progress in cold-start fraud review detection. Motivated by [12], You et al. [8] further leverage both attribute and domain knowledge information for better understanding review representation.

While the existing research on cold-start problem focus on user's review representation, we believe fraud reviews are easy to be manipulated to look like honest reviews [9] and thus may confuse existing methods. In this paper, we propose a novel model for jointly embedding user reviewing behavior and users/items social relations into inferable user-item-review-rating representations for fraud reviews detection with cold-start problem. Accordingly, the inferred new user representation and the well represented item, review, and rating information provide more comprehensive evidence for fraud review detection with cold-start problem.

III. Proposed Method

A. Representation Learning Architecture

The representation architecture of JESTER model is shown in Figure 1. Given a reviewing activity, i.e. a user u write a review t for an item d with a rating r, JESTER adopts an embedding network to represent u, t, d, r to m-dimensional vector representations $\mathbf{u}, \mathbf{t}, \mathbf{d}, \mathbf{r} \in \mathcal{R}^m$, respectively. Here m is a hyperparameter that determines the representation capacity and can be adjusted according to different social media data. JESTER further feeds these representations into a neural network for fraud review detection. In the representation learning process, JESTER simultaneously considers three tasks: *user reviewing behavior learning*, *social relation preservation*, and *fraud review detection*, corresponding to three learning loss functions: *behavior learning loss*, *social relation preservation loss*, and *fraud detection loss*. By jointly optimizing these three loss functions, JESTER learns the user-item-review-rating inferable representations for fraud review detection.

To enable an efficient inferring process of the new user representation, we constrain the $\mathbf{u}, \mathbf{t}, \mathbf{d}, \mathbf{r}$ as unit vector, i.e. $\|\mathbf{u}\|_2 = \|\mathbf{t}\|_2 = \|\mathbf{d}\|_2 = \|\mathbf{r}\|_2 = 1$. The property of this constraint will be further discussed in Sec. III-F.

B. JESTER Network Structure

To achieve the above learning objectives, the neural network structures in JESTER model are designed as follows.

The *embedding network* consists four parts: user embedding layers, item embedding layers, review embedding networks, and rating embedding layers. The embedding layers have a

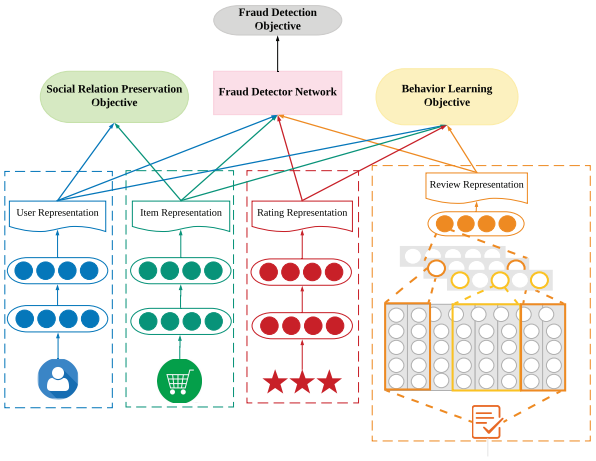


Fig. 1. The Proposed JESTER Model

two-layer structure where the first layer is a fully connected layer with m nodes and the second layer is a normalization layer. While the fully connected layer maps the input to a vector, the normalization layer normalizes the vector to its unit vector. The review embedding networks can be implemented by any text embedding network, e.g. recurrent neural network, followed by a normalization layer. Followed by [12], in this paper, we implement the text embedding network as a convolutional neural network (CNN) used in [12].

The *fraud review detection network* is implemented by a fully connected neural network with the concatenate of user-item-review-rating representation vectors as the input and the fraud label as the output. In the fully connected neural network, the $ReLU$ activation function is used in the hidden layers, and the $sigmoid$ activation function is used in the output layer. The number of hidden layers and the number of nodes in each hidden layer are two hyper-parameters that can be adjusted according to different data.

C. User Reviewing Behavior Learning Loss

Inspired by [27], we propose a co-occurrence based user reviewing behavior. The user reviewing behavior can be formally defined as follows:

Definition III.1. (User Reviewing Behavior) In the context of the reviewing activity, a *user reviewing behavior* is a relationship among a user, an item, a review, and a rating. Denoting the user, item, review, rating as u, t, d, r, respectively, the behavior b can be represented as a set $\{u, t, d, r\}$.

We say a behavior $b = \{u, t, d, r\}$ is *success* if the u, t, d and r consists a *reviewing activity*, i.e. the user u posted the review d to the item t with the rating r. We here introduce a measure to estimate the success rate of the user reviewing behavior from u, t, d, r. Following [27], we first represent the behavior as the sum of vector representations of user, item, review, and rating as follows,

$$\mathbf{b} = \mathbf{u} + \mathbf{t} + \mathbf{d} + \mathbf{r}, \tag{1}$$

$\mathbf{d} = f_\psi(d)$ is the review embedding calculated by a neural network f_ψ with parameters ψ. Then, we use the length of vector $\mathbf{b}$ to measure the success rate of a user reviewing behavior. Specifically, the longer vector $\mathbf{b}$ implies a larger behavior success rate. Consequently, if the vector orientation of $\mathbf{u}, \mathbf{t}, \mathbf{d}, \mathbf{r}$ are more similar in the representation space, the behavior $b = \{u, t, d, r\}$ has higher success rate, as shown in Figure 2. Considering a behavior will be either success or fail in reality, we map the success rate to a probability close to 1 or 0 by the following success probability function,

$$s(b) = 2 \cdot \frac{1}{1 + e^{-\|\mathbf{b}\|_2}} - 1. \tag{2}$$

We denote the observed behavior success probability as

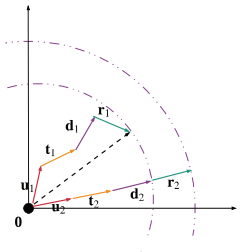


Fig. 2. The Representation Space of JESTER. In this figure, $\mathbf{u}, \mathbf{t}, \mathbf{d}, \mathbf{r}$ refer to the representations of user, item, review, and rating, respectively.

$\hat{s}(\cdot)$ where $\hat{s}(b) = 1$ if b is observed in the social media and $\hat{s}(\cdot) = 0$ otherwise. While $s(\cdot)$ describes the behavior success distribution in the representation space, $\hat{s}(\cdot)$ reflects the observed behavior success distribution in the social media. To embed the user reviewing behavior into the user-item-review-rating representations, we minimize the KL-divergence between the behavior success distribution in the representation space and the observed behavior distribution in the social media as follows,

$$\min_{\mathbf{u},\mathbf{t},\psi,\mathbf{r}} \sum_{b \in B_+} \hat{s}(b) \log \frac{\hat{s}(b)}{s(b)}, \tag{3}$$

where $B_+ = \{b_1, \cdots, b_{n_d}\}$ is the set of observed behavior in a social media data set with n_d reviews. Considering $\hat{s}(b) = 1$ if $b \in B_+$, the Eq. (3) equals to the follows,

$$\min_{\mathbf{u},\mathbf{t},\psi,\mathbf{r}} - \sum_{b \in B_+} \log s(b). \tag{4}$$

However, Eq. (4) only captures the distribution of the successful behavior, i.e. $b \in B_+$, but ignores the unsuccessful behavior, i.e. $b \notin B_+$. In practical, it is impossible to enumerate all unsuccessful behavior because of the huge behavior space. Inspired by the negative sampling used in word2vec [28], we randomly sample a set of unsuccessful behavior B_- and measure their probability in the representation space as follows,

$$s^*(b) = 2 \cdot \frac{1}{1 + e^{\|\mathbf{b}\|_2}}. \tag{5}$$

The $s^*(b)$ will be large, i.e. the behavior b is likely unsuccessful, if the vector orientation of $\mathbf{u}, \mathbf{t}, \mathbf{d}, \mathbf{r}$ is diverse in the

representation space. To capture the unsuccessful behavior, we minimize the KL-divergence between the behavior unsuccessful distribution in the representation space and the sampled negative behavior distribution as follow,

$$\min_{\mathbf{u},\mathbf{t},\psi,\mathbf{r}} - \sum_{b \in B_-} \log s^*(b). \tag{6}$$

Accordingly, we define the user reviewing behavior learning loss as follows,

$$\mathcal{L}_1 = - \sum_{b \in B_+} \log s(b) - \sum_{b \in B_-} \log s^*(b), \tag{7}$$

D. Social Relation Preservation Loss

For a social reviewing data Q with a set of users $U = \{u_1, u_2, \cdots, u_{n_u}\}$ and a set of items $T = \{t_1, t_2, \cdots, t_{n_t}\}$, we extracts a bipartite graph G from Q as $G = (U, T, E)$, where U and T are as the vertices on two sides of G, respectively, and $E \subseteq U \times T$ defines the inter-set edges. Here, the edge e_{u_i,t_j} in E carries a non-negative weight w_{u_i,t_j}, reflecting the strength between the user u_i and item t_j, and the w_{u_i,t_j} will be zero if user u_i does not review item t_j. Accordingly, the weights in the bipartite graph can be represented by a $n_u \times n_t$ matrix $\mathbf{W} = [w_{u_i,t_j}]$.

The users/items social relation in the bipartite graph reflects the cooperation of fraudsters for specific suspicious items, which is essential for collaborative fraud detection for reviews with camouflage [6], [9]. We preserve the users/items social relations by embedding them in the user-item-review-rating representations. The social relation embedding procedure captures both the explicit and implicit relations of users and items as illustrated in Figure 3.

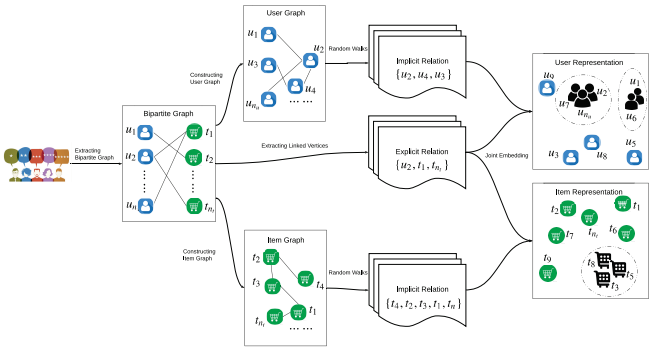


Fig. 3. The Users/Items Social Relation Embedding Workflow.

1) Explicit Relations Embedding: The explicit relations refer to the direct links between users and items through edges in a bipartite graph G, which reflect the preference of users to items. We assume the representations of a user and an item should be similar if the user prefers the item. Under this assumption, actually, the explicit relations have already preserved in the user reviewing behavior learning. Specifically, if a user u_i and an item t_j is directly linked in graph G, i.e. the user u_i have posted a review to the item t_j, they must be involved in a success behavior $b \in B_+$. By optimizing the loss function (7), $\mathbf{u}_i$ and $\mathbf{t}_j$ will have similar vector orientations in the representation space as shown in Figure 2. Because both $\mathbf{u}_i$ and $\mathbf{t}_j$ are unit vectors, similar vector orientations indicate similar vector representations. As a result, the explicit relations embedding has been already achieved by the user reviewing behavior learning.

2) Implicit Relations Embedding: The implicit relations refer to the relations between users/items that are not directly connected by edges. In a graph, two users/items may have an implicit relation if exists a path between them where a path is a sequence of a limited number of edges with shared vertices. The users/items implicit relations reveal the potential similarity between users/items. Similar to [29], we reconstruct the bipartite graph G into two graphs $G^{(u)}$ and $G^{(t)}$ to discover the implicit relations where $G^{(u)}$ only contains user vertices U and $G^{(t)}$ only contains item vertices T. In $G^{(u)}$, u_i and u_j will have an edge e_{u_i,u_j} if exists a t_k that $e_{u_i,t_k} \in E$ and $e_{u_j,t_k} \in E$ where E is the edge set of G. In $G^{(t)}$, t_i and t_j will have an edge e_{t_i,t_j} if exists a u_k that $e_{u_k,t_i} \in E$ and $e_{u_k,t_j} \in E$ where E is the edge set of G. Similar to [30], we calculate the weights of e_{u_i,u_j} and e_{t_i,t_j} as follows,

$$w_{u_i,u_j} = \sum_{e_{u_i,t_k}, e_{u_j,t_k} \in E} w_{u_i,t_k} \cdot w_{u_j,t_k}, \tag{8}$$

or

$$w_{t_i,t_j} = \sum_{e_{u_k,t_i}, e_{u_k,t_j} \in E} w_{u_k,t_i} \cdot w_{u_k,t_j}. \tag{9}$$

To embed the implicit relation, we need to discover the paths in the graph $G^{(u)}$ and $G^{(t)}$. However, counting all paths in $G^{(u)}$ and $G^{(t)}$ has a great high complexity, which is impracticable for social media data. Inspired by DeepWalk [31], we perform a truncated random walks on a graph from each node where the weight of an edge is proportional to the walking probability on the edge. Subsequently, we adopt the walked edges as the paths to reveal the implicit relation. In other words, two vertices are treated having an implicit relation if they are in a random walk path. The random walk paths generation procedure generates a set of random walk paths $D^{(u)}$ of U and a set of random walk paths $D^{(t)}$ of T.

We assume two users or items in the same path have implicit relation that they will have or be affected by a similar user reviewing behavior. To seamlessly integrate the implicit relations with user reviewing behavior, we maximize the similarity of the orientations of vector of users/items in the representation space if they are in the same path. In this way, these users/items will have similar user reviewing behavior success rate according to Eq. (2). The similarity of the orientations of vectors can be measured by cosine function. Since all representations are unit vector, it equals to an inner dot of two vectors. Accordingly, we can calculate the probability of two vertices (can either be user or item) in a path from their vector representations as follows,

$$p(v_i, v_j) = 2 \cdot \frac{1}{1 + e^{-\mathbf{v}_i^\top \mathbf{v}_j}} - 1, \tag{10}$$

where $\mathbf{v}$ is the vector representation of v. For the vertices that are not in a path, their probability can be calculated as,

$$p^*(v_i, v_j) = 2 \cdot \frac{1}{1 + e^{\mathbf{v}_i^\top \mathbf{v}_j}}. \tag{11}$$

Similar to Eq. 7, we minimize the KL-divergence between the distribution of vertices in a path in the representation space and the observed distribution of vertices in a path in social media data. Consequently, we get the social relation preservation loss as follows,

$$\begin{aligned}\mathcal{L}_2 = &- \sum_{u_i \in P \wedge P \in D^{(u)}} \sum_{u_j \in C_P(u_i)} \log p(u_i, u_j) \\ &- \sum_{u_i \in P \wedge P \in D^{(u)}} \sum_{u_j \in C_-(u_i)} \log p(u_i, u_j) \\ &- \sum_{t_i \in P \wedge P \in D^{(t)}} \sum_{t_j \in C_P(t_i)} \log p^*(t_i, t_j) \\ &- \sum_{t_i \in P \wedge P \in D^{(t)}} \sum_{t_j \in C_-(t_i)} \log p^*(t_i, t_j),\end{aligned} \tag{12}$$

where P refers to a path in $D^{(\cdot)}$, $C_P(\cdot_i)$ refers the other vertices of in the path P instead of u_i or t_i, and $C_-(\cdot_i)$ refers to the negative sampled vertices that do not in any path contained u_i or t_i.

E. Joint Loss Function

Denoting the fraud detector network as f_ω, the predicted fraud label l to a review in a successful behavior b equals to:

$$l(b) = f_\omega(\mathbf{u}, \mathbf{t}, \mathbf{d}, \mathbf{r}), \tag{13}$$

where ω refers to the parameters of the fraud detector network, and u, t, d, and r consist b. In the learning process, we adopt cross entropy to evaluate the loss of the fraud detector network. Denoting the supervised fraud review label of b as $\hat{l}(b)$, the loss of the fraud detector network can be calculated as follows,

$$\mathcal{L}_3 = \sum_{b \in B} -(\hat{l}(b) \log l(b) + (1 - \hat{l}(b)) \log(1 - l(b))). \tag{14}$$

JESTER jointly optimize the user reviewing behavior learning loss, the social relation preservation loss, and the fraud detector network loss to learn the inferable user-item-review-rating representations. The joint loss function is as follows,

$$\mathcal{L} = \alpha_1 \mathcal{L}_1 + \alpha_2 \mathcal{L}_2 + \alpha_3 \mathcal{L}_3 \tag{15}$$

where α_1, α_2, and α_3 are hyper-parameters that control the affects of three $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$.

F. User Representation Inferring in Cold-start Problem

For a review proposed by a new user, JESTER first represents the item, review, and rating into their representations $\{\mathbf{t}, \mathbf{d}, \mathbf{r}\}$. It then infers the user representation by optimizing the following objective function:

$$\max_{\mathbf{u}} 2 \cdot \frac{1}{1 + e^{-\|\mathbf{u}+\mathbf{t}+\mathbf{d}+\mathbf{r}\|_2}} - 1, \tag{16}$$

which aims to maximize the behavior success rate according to the item, review and rating representations. Finally, it feeds the four representations $\mathbf{u}, \mathbf{t}, \mathbf{d}, \mathbf{r}$ into the learned fraud detector network f_ω to predict the fraud label.

The objective function Eq. (16) can be efficiently solved by a closed-form solution. Maximizing Eq. (16) equals to maximizing the $\|\mathbf{u}+\mathbf{t}+\mathbf{d}+\mathbf{r}\|_2$. Given $\mathbf{t}, \mathbf{d}$, and $\mathbf{r}$, as shown in Figure 4, the $\mathbf{u}$ that can maximize $\|\mathbf{u}+\mathbf{t}+\mathbf{d}+\mathbf{r}\|_2$ must in the direction of $\mathbf{t} + \mathbf{d} + \mathbf{r}$. Considering the unit vector constraint of the representations, the representation of a new user can calculated as follows,

$$\mathbf{u} = \frac{\mathbf{t} + \mathbf{d} + \mathbf{r}}{\|\mathbf{t} + \mathbf{d} + \mathbf{r}\|_2}. \tag{17}$$

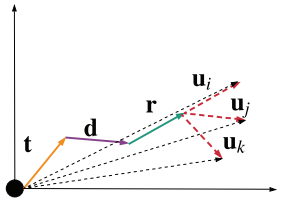


Fig. 4. The New User Representation Inferring.

IV. Experiments

A. Data Sets

Following the literature [8], [12] about cold-start fraud detection, our experiments are carried on four real-life data sets, including Yelp-hotel, Yelp-restaurant, Yelp-NYC, and Yelp-Zip, which are also commonly used in previous fraud detection researches [1], [4], [32]. Table I displays the statistics of the data sets.

We split the original data sets into several subsets according to the time period to evaluate the fraud review detection performance in a stable way. We further split each subset into two parts by setting a time point. The first part includes the reviews posted before the time point, while the second part contains the rest reviews. From the second part, we pick up the reviews which are posted by new users at first time as cold-start reviews. We train the fraud detection methods on the first part and evaluate them on the second part.

TABLE I
Data Characteristics

Name	Training Data		Testing Data				
	Time Period	#R	Time Period	#F	#FC	#N	#NC
Zip_1	24/10/08 – 24/03/09	10530	25/03/09 – 25/06/10	6267	4848	43744	15952
Zip_2	24/03/09 – 24/08/09	13252	25/08/09 – 25/12/09	1396	1075	10220	3820
NYC_1	24/10/08 – 24/03/09	6780	25/03/09 – 25/06/10	3183	2539	27974	11313
NYC_2	24/03/09 – 24/08/09	8243	25/08/09 – 25/12/09	748	594	6664	2754

In this table, #R refers to the number of reviews; #F and #FC refer to the number of fraud reviews and cold-start fraud reviews, respectively; and #N and #NC refer to the number of honest reviews and cold-start honest reviews, respectively.

B. Evaluation Metrics

We evaluate the fraud review detection performance of each method by three metrics, including *precision*, *recall*, and *F-score*. Here, the precision evaluates the ratio of correct detected

TABLE II
COLD-START FRAUD DETECTION PERFORMANCE OF DIFFERENT METHODS

Data Info.		JESTER			JETB			Behavior			Bigram			Improvement		
Name	Category	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Zip_1	Normal	**0.81**	0.90	0.85	0.77	**1.00**	0.87	0.77	0.99	0.87	**0.78**	0.96	**0.86**	**0.03**	-0.10	-0.02
	Fraud	0.37	**0.21**	**0.27**	0.24	0.00	0.00	**0.54**	0.05	0.09	0.42	0.10	0.16	-0.17	**0.11**	**0.11**
Zip_2	Normal	**0.82**	0.84	0.83	0.78	**1.00**	0.88	0.79	0.99	0.88	0.80	0.92	**0.85**	**0.02**	-0.16	-0.05
	Fraud	0.33	**0.30**	**0.31**	0.45	0.01	0.02	**0.54**	0.06	0.11	0.37	0.17	0.23	-0.21	**0.13**	**0.08**
NYC_1	Normal	**0.84**	0.84	0.84	0.82	1.00	**0.90**	0.82	1.00	**0.90**	0.82	0.96	0.89	**0.02**	-0.16	-0.06
	Fraud	0.26	**0.26**	**0.26**	0.00	0.00	0.00	0.38	0.00	0.00	0.31	0.08	0.13	-0.12	**0.18**	**0.13**
NYC_2	Normal	**0.85**	0.90	0.87	0.82	**1.00**	0.90	0.82	**1.00**	0.90	0.83	0.94	0.88	**0.02**	-0.10	-0.03
	Fraud	**0.31**	**0.23**	**0.27**	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.12	0.17	**0.02**	**0.11**	**0.10**

Precision (P), Recall (R) and F-score (F) are reported per normal and fraud reviews. The best results are highlighted in bold.

reviews over all detected reviews, recall reflects the ratio of undetected reviews over all relevant reviews, and the F-score indicates an average of precision and recall. We use all of them because the fraud detection in an imbalance classification problem [33], i.e. the number of fraud reviews are much less than honest reviews, that cannot be considered only from either precision or recall perspective. We report these three metrics per ground-truth honest and fraud classes to illustrate the performance for different categories, and further average them to show the overall performance. Higher precision, recall, and F-score indicate a better performance.

We follow the literature [4], [12] to use the results of the Yelp commercial fake review filter as the ground-truth for performance evaluation. Although its filtered (fraud reviews) and unfiltered reviews (honest reviews) are likely to be the closest to real fraud and honest reviews [1], they are not absolutely accurate [17]. The inaccuracy exists because it is hard for the commercial filter to have the same psychological state of mind as that of the users of real fraud reviews who have real businesses to promote or to demote, especially for cold-start problems.

C. Parameters Settings

In our experiments, we use a CNN network to embed reviews following [12]. The CNN network adopts 100 filters with size 3×100 on the pre-trained 100-dimensional word embedding by GloVe algorithm [34] [3]. We embed the user, item and rating into a 100-dimension vector representation. We implement the fraud detector network by a 3-layer fully connected neural network with 100 nodes in the hidden layers and use ReLU as the activation function of all hidden nodes. We train our model by Adam [35] and batch size 32. For the parameters in the compared methods, we take their recommended settings.

D. Effectiveness on Cold-start Fraud Detection

1) Experimental Settings: JESTER is compared with the state-of-the-art method JETB [12]. This method handling cold-start problem by considering entities (user, item and review) relations to embed reviews. When a new user posted a new review, this review can be represented by the trained network and classified by the classifier. In the literature [12], support vector machine (SVM) is used as the fraud classifier based on the JETB generated review features. However, SVM is with a time complexity $O(n^3)$, where n is the number of training samples. It is not suitable for the problem with large amount of data. To make JETB practicable, we use a 3-layer fully connected neural network instead of SVM as the fraud classifier of JETB.

We further compared with two review content-based fraud detection methods used in [12] as baseline competitors. Both of them extract features from review content, and feed these features into a classifier for fraud review detection. Specifically, the first method (denoted as Bigram) uses the bigram feature. The second method (denoted as Behavior) uses (i) the bigram feature, (ii) the length of review, (iii) the absolute rating diversity of a review compared with other reviews of the same item, and (iv) the similarity of a review to its most similar reviews of the same item under the cosine similarity. We also use 3-layer fully connected neural network as their fraud classifier.

2) Findings - JESTER Significantly Outperforming the State-of-the-art Cold-start Fraud Detection Method: Table II illustrates the cold-start fraud detection performance of JESTER compared with JETB, Behavior, Bigram on four time period of Yelp-Zip and Yelp-NYC data sets. JESTER gains largely improvement for cold-start fraud review detection, i.e. 0.11, 0.08, 0.13, and 0.10 F-score increase on Zip_1, Zip_2, NYC_1, and NYC_2, respectively. This averaged performance improvement is mainly contributed by the increased recall for the fraud reviews (corresponding recall increase values are 0.11, 0.13, 0.18, and 0.11). As shown in the results, JESTER slightly "decreases" the performance of honest reviews detection. This "decreased" may be caused by the *noising ground-truth* of the cold-start fraud reviews that do not be detected by the Yelp commercial filter.

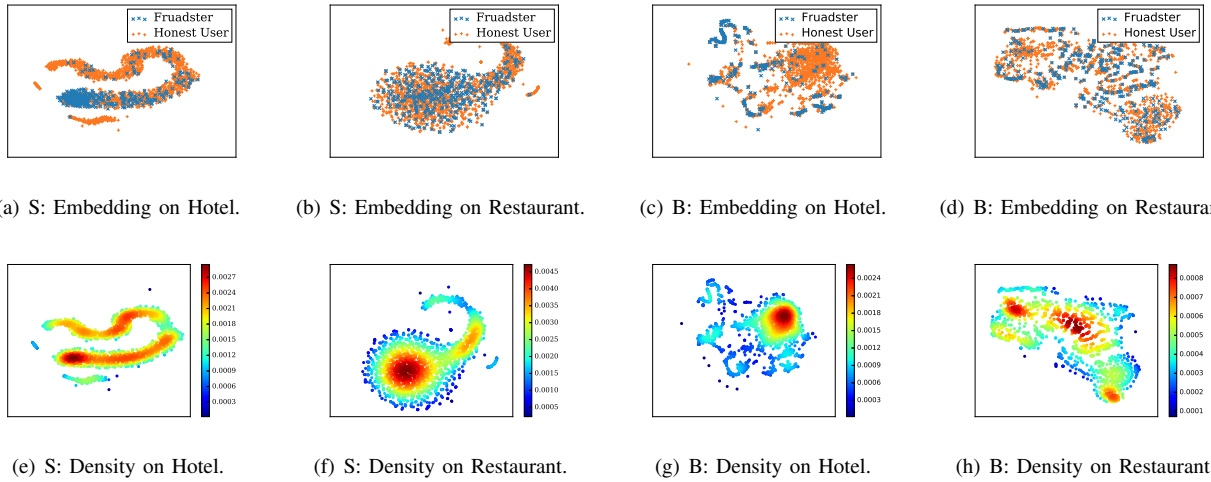
In addition to review, JESTER further leverages information from user, item, and rating, which are guaranteed by the inferable representation and enable JESTER to effectively capture more fraud evidence from multiple views. As a result, JESTER can achieve significant performance improvement in cold-start fraud detection.

[3]The pre-trained word embedding can be downloaded from: http://nlp.stanford.edu/data/glove.6B.zip

TABLE III
GENERAL FRAUD DETECTION PERFORMANCE OF DIFFERENT METHODS

Data Info.		JESTER			JETB			FRAUDER			HoloScope			Improvement		
Name	Category	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Zip_1	Normal	**0.89**	0.92	0.91	0.87	**1.00**	0.93	0.87	0.95	0.91	0.86	0.86	0.86	**0.02**	-0.08	-0.02
	Fraud	**0.23**	**0.17**	0.19	0.18	0.00	0.01	0.01	0.00	0.00	0.03	0.03	0.03	**0.05**	**0.14**	**0.16**
Zip_2	Normal	**0.90**	0.87	0.88	0.78	**1.00**	0.88	0.88	0.95	0.91	0.87	0.88	0.88	**0.02**	-0.13	-0.03
	Fraud	0.22	**0.29**	0.25	**0.45**	0.01	0.02	0.04	0.02	0.02	0.04	0.04	0.04	-0.23	**0.25**	**0.21**
NYC_1	Normal	**0.91**	0.88	0.90	0.90	**1.00**	0.95	0.88	0.86	0.87	0.88	0.86	0.87	**0.01**	-0.12	-0.05
	Fraud	**0.18**	**0.25**	0.21	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	**0.17**	**0.24**	**0.02**
NYC_2	Normal	**0.92**	0.92	0.92	0.90	**1.00**	0.95	0.88	0.82	0.85	0.87	0.69	0.77	**0.02**	-0.08	-0.03
	Fraud	**0.24**	**0.22**	0.23	0.00	0.00	0.00	0.01	0.02	0.02	0.02	0.06	0.03	**0.22**	**0.16**	**0.20**

Precision (P), Recall (R) and F-score (F) are reported per normal and fraud reviews. The best results are highlighted in bold.



(a) S: Embedding on Hotel. (b) S: Embedding on Restaurant. (c) B: Embedding on Hotel. (d) B: Embedding on Restaurant.

(e) S: Density on Hotel. (f) S: Density on Restaurant. (g) B: Density on Hotel. (h) B: Density on Restaurant.

Fig. 5. User Representation with Density of Different Methods on Yelp-Hotel and Yelp-Restaurant. The sub-figures (a), (b), (c), (d) contain the user representation information with the ground-truth labels, and the sub-figures (e), (f), (g), (h) show the density in the representation space. S refers to the social relation embedding-based method, and B refers to the behavior embedding-based method.

E. Effectiveness on General Fraud Detection

1) Experimental Settings: JESTER is further compared with JETB [12] and two state-of-the-art competitors: FRAUDER [9] and HoloScope [6] in detecting *general fraud reviews*, i.e. all the reviews contained in the testing data set. Different from JETB which is review content-based method, FRAUDER and HoloScopre are two social relation-based fraud review detection methods. Specifically, FRAUDER models the social relation as a graph and detects fraud reviews by dense subgraph mining. HoloScope also adopts graph to model social relation but detects fraud reviews by jointly considering the graph topology and review temporal spikes.

2) Findings - JESTER Significantly Improving General Fraud Detection Performance: The precision, recall and F-score of JESTER, JETB, FRAUDER, and HoloScope are reported in Table III. Overall, JESTER significantly outperforms the competitors in fraud review detection. It improves 0.16, 0.21, 0.20, and 0.20 compared with the best-performing method in terms of F-score on four data sets for fraud review detection.

The dramatic performance improvement of JESTER is mainly contributed by jointly embedding user reviewing behavior and user/item social relations in its user-item-review-rating representations: (1) compared to FRAUDER and HoloScope that capture the social relations, JESTER further considers the user reviewing behavior to effectively detect personalized fraud; and (2) compared to JETB, JESTER seamlessly integrates user/item social relations to avoid camouflage. Consequently, JESTER obtains up to 0.24 recall improvement compared with the competitors.

F. Evaluating the Effectiveness of User Reviewing Behavior and User/Item Social Relations for Fraud Review Detection

1) Experimental Settings: We visualize the user representation in a two-dimensional space trough TSNE [36], and plot the ground-truth labels of each user at their positions in the representation space. The user representation learned according to user reviewing behavior learning loss function Eq. (7) is compared with that learned according to social relation preservation loss function Eq. (12) on Yelp-Hotel and Yelp-Restaurant data sets.

2) Findings - Behavior-embedded Representation contributes to Personalized Fraud Review Detection and Social Relation-embedded Representation Contributes to Collaborative Fraud Review Detection: The behavior-embedded and social relation-embedded user representations are visualized

in Figure 5. As shown in Figure 5, users have more diverse representations in the behavior-embedded representation space compared with social relation-embedded representation space. This indicates more personalized information is captured by the behavior-embedded representation, which is important to identify personalized fraud reviews. However, in the behavior-embedded representation space, the users with large density are not consistent with the ground-truth fraudster label. In contrast, the density of social relation-embedded representation is consistent with the ground-truth fraudsters distribution. As evidenced by [9], the collaborative manipulation of reviews will generate density connection between users. Accordingly, the results demonstrate our embedded social relation is essential for collaborative fraud review detection. A high quality user representation will enable a dense distribution for fraudsters because of the collaborative manipulation [9]. This qualitative illustrates that the social relation of users is essential for collaborative fraudsters detection.

V. Conclusion

This paper introduces a novel joint behavior and social relation inferable embedding method, JESTER, for fraud review detection with cold-start problem. JESTER jointly embeds user reviewing behavior and user/item social relations into the inferable representations of user, item, review and rating, which provides more comprehensive information for fraud review detection. For cold-start problem, JESTER efficiently infers the most probable representation of a new user in a closed-form solution according to the embedded user reviewing behavior. Two large real-word social media data sets demonstrate the performance of JESTER is substantially better than the state-of-the-art competitors.

References

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

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

[3] J. Ye and L. Akoglu, "Discovering opinion spammer groups by network footprints," in *ECML*. Springer, 2015, pp. 267–282.

[4] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *ACM SIGKDD*. ACM, 2015, pp. 985–994.

[5] B. Hooi, K. Shin, H. A. Song, A. Beutel, N. Shah, and C. Faloutsos, "Graph-based fraud detection in the face of camouflage," *TKDD*, vol. 11, no. 4, p. 44, 2017.

[6] S. Liu, B. Hooi, and C. Faloutsos, "Holoscope: Topology-and-spike aware fraud detection," in *CIKM*. ACM, 2017, pp. 1539–1548.

[7] S. Rayana and L. Akoglu, "Collective opinion spam detection using active inference," in *ICDM*. SIAM, 2016, pp. 630–638.

[8] Z. You, T. Qian, and B. Liu, "An attribute enhanced domain adaptive model for cold-start spam review detection," in *COLING*, 2018, pp. 1884–1895.

[9] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, "Fraudar: Bounding graph fraud in the face of camouflage," in *ACM SIGKDD*. ACM, 2016, pp. 895–904.

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

[11] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *IJCAI*, vol. 22, no. 3, 2011, p. 2488.

[12] X. Wang, K. Liu, and J. Zhao, "Handling cold-start problem in review spam detection by jointly embedding texts and behaviors," in *ACL*, vol. 1, 2017, pp. 366–376.

[13] N. Jindal and B. Liu, "Opinion spam and analysis," in *WSDM*. ACM, 2008, pp. 219–230.

[14] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *WWW*. ACM, 2012, pp. 191–200.

[15] S. Feng, L. Xing, A. Gogar, and Y. Choi, "Distributional footprints of deceptive product reviews." *ICWSM*, vol. 12, pp. 98–105, 2012.

[16] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection." *ICWSM*, vol. 13, pp. 175–184, 2013.

[17] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in *ICDM*. IEEE, 2014, pp. 899–904.

[18] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, "Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns." in *ICWSM*, 2015, pp. 634–637.

[19] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao, "Modeling review spam using temporal patterns and co-bursting behaviors," *arXiv preprint arXiv:1611.06625*, 2016.

[20] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "Fake review detection: Classification and analysis of real and pseudo reviews," *Technical Report UIC-CS-2013–03, University of Illinois at Chicago, Tech. Rep.*, 2013.

[21] G. Wang, S. Xie, B. Liu, and S. Y. Philip, "Review graph based online store review spammer detection," in *ICDM*. IEEE, 2011, pp. 1242–1247.

[22] E. Choo, T. Yu, and M. Chi, "Detecting opinion spammer groups through community discovery and sentiment analysis," in *IFIP*. Springer, 2015, pp. 170–187.

[23] L. Wu, X. Hu, F. Morstatter, and H. Liu, "Adaptive spammer detection with sparse group modeling." in *ICWSM*, 2017, pp. 319–326.

[24] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *ACL HLT*. Association for Computational Linguistics, 2011, pp. 309–319.

[25] C. Xu, J. Zhang, K. Chang, and C. Long, "Uncovering collusive spammers in chinese review websites," in *CIKM*. ACM, 2013, pp. 979–988.

[26] D. Hovy, "The enemy in your own camp: How well can we detect statistically-generated fake reviews–an adversarial study," in *ACL*, vol. 2, 2016, pp. 351–356.

[27] D. Wang, M. Jiang, Q. Zeng, Z. Eberhart, and N. V. Chawla, "Multi-type itemset embedding for learning behavior success," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2397–2406.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[29] M. Gao, L. Chen, X. He, and A. Zhou, "Bine: Bipartite network embedding," in *SIGIR*, 2018.

[30] H. Deng, M. R. Lyu, and I. King, "A generalized co-hits algorithm and its application to bipartite graphs," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 239–248.

[31] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.

[32] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," in *ACM SIGKDD*. ACM, 2013, pp. 632–640.

[33] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.

[34] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, no. Nov, pp. 2579–2605, 2008.