Unsupervised User Behavior Representation for Fraud Review Detection with Cold-Start Problem

Qian Li[1], Qiang Wu[1], Chengzhang Zhu[2,3], Jian Zhang[1], and Wentao Zhao[3]

[1] Global Big Data Technologies Centre, University of Technology Sydney, Australia
Qian.Li-7@student.uts.edu.au
{Qiang.Wu, Jian.Zhang}@uts.edu.au
[2] Advanced Analytics Institute, University of Technology Sydney, Australia
kevin.zhu.china@gmail.com
[3] College of Computer, National University of Defense Technology, China
wtzhao@nudt.edu.cn

Abstract. Detecting fraud review is becoming extremely important in order to provide reliable information in cyberspace, in which, however, handling cold-start problem is a critical and urgent challenge since the case of cold-start fraud review rarely provides sufficient information for further assessing its authenticity. Existing work on detecting cold-start cases relies on the limited contents of the review posted by the user and a traditional classifier to make the decision. However, simply modeling review is not reliable since reviews can be easily manipulated. Also, it is hard to obtain high-quality labeled data for training the classifier. In this paper, we tackle cold-start problems by (1) using a user's behavior representation rather than review contents to measure authenticity, which further (2) consider user social relations with other existing users when posting reviews. The method is completely (3) unsupervised. Comprehensive experiments on Yelp data sets demonstrate our method significantly outperforms the state-of-the-art methods.

Keywords: Fraud review detection · Cold-start · Behavior representation · Unsupervised learning.

1 Introduction

With the increasing popularity of E-commerce, a large number of online reviews are manipulated by fraudsters, who intend to write fraud reviews driven by strong incentives of profit and reputation. Early in 2013, it has been found that around 25% of Yelp reviews could be fake[4]. This situation becomes worse than ever recently. As reported by Forbes news[5] in 2017, Amazon is seeing a lot more suspicious reviews than before. As a result, it has become a critical and urgent task to effectively detecting such fraudsters and fraud reviews.

Recent years have seen significant progress made in fraud detection. Current efforts mainly focused on extracting linguistic features (n-grams, POS, etc) and behavioral features [27, 5]. However, linguistic features are ineffective when dealing with real-life fraud reviews [19], especially when linguistic features are easy to be imitated, a.k.a. camouflage [6]. Also, extracting behavior features require a large number of samples and usually takes months to make observations. When facing the *cold-start* problem, i.e. *a new user just posted a new review*, extracting behavior features becomes even harder because none historical information is available for a new user [28].

[4] https://www.bbc.com/news/technology-24299742.
[5] https://www.forbes.com/sites/emmawoollacott/2017/09/09/exclusive-amazons-fake-review-problem-is-now-worse-than-ever/#501eccb87c0f.

Recently, the cold-start problem in fraud review detection has been first studied by [25]. This method handles the cold-start problem by considering correlations between users, items, and reviews. Later, [28] makes one step further by incorporating the relations between entities (users, items, and reviews) with their attributes from different domains. Both of the above methods feed the embedded **review** representation into a classifier for cold-start fraud detection. However, two problems may arise when adopting these methods. (1) Only using review itself is ineffective as discussed in [19], and is easy to be affected by camouflage [6]. (2) Also, high-quality labeled data are required in both methods, which is really hard to obtain in real life.

We address the above problems in current cold-start fraud detection methods by focusing on **user behavior**. The rationale is similar users may result in similar behaviors when posting a review. Specifically, in a behavior representation space (Figure 1), if a new user is closer to a group of existing fraudsters, those fraudsters will be identified as his/her similar users. Then, the new user is likely to be detected as a fraudster. Thus, the cold-start problem can be transferred to identifying the existing users who have similar behavior with a new user. Although limited information is available for a new user, existing fraudsters can be effectively detected by many methods [23, 6, 13].

Motivated by [25, 28], we first represent user according to the relations among entities. Further, we integrate the entities relations with user social relations. The intuition is that fraud reviews are always manipulated by a group of fraudsters with close social relations [6, 13]. For example, a group of users usually work together to effectively promote or demote target products. They may even know one another and copy reviews among themselves [17].

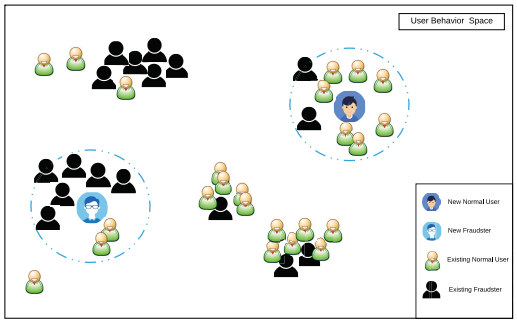


Fig. 1. Example of User Behavior Space. In this space, similar users will close to each other, i.e. a normal user will be majorly surrounded by normal users, and vice versa.

To further strength the discriminate ability of the represented behavior for fraud detection, we apply the dense subgraph mining [1, 6] to generate pseudo-fraud labels to tune the representation in an iterative way. The foundation is that an end-to-end supervised training will enable the strong task-specific discriminate ability of the generated representation, as demonstrated in most of representation learning tasks. In this process, dense subgraph mining generates high-quality labels in an unsupervised way [6]. In turn, the discriminate representation adjusts the weight of each graph link for a more precise dense subgraph mining.

Based on the above analysis, we propose a **s**ocially-aware **u**nsu**p**ervised us**e**r behavior **r**epresentation method for **cold**-start fraud detection (SUPER-COLD). Our method jointly captures entities interactions and user social relations to generate behavior representation with a strong discriminate ability for cold-start fraud detection. In summary, the main contributions of this work are as follows.

- **A user behavior representation model for cold-start fraud detection**: the represented user behavior avoids camouflage and thus is more reliable for cold-start fraud detection.
- **A socially-aware user behavior representation**: the reviewing habits and social relations of a user are jointly embedded in its behavior representation to provide comprehensive evidence for fraud detection.
- **A discriminative unsupervised representation approach for cold-start user behavior**: a dense subgraph-based approach for fraudsters detection has been involved into the unsupervised representation approach, which strengths the discriminant of the representation and tackles the problem of lacking high-quality fraud labels in real life.

Comprehensive experiments on two large real-world data sets show that: (1) SUPER-COLD effectively detects cold-start fraud reviews without manual labels (improved up to 150% in terms of F-score compared with the state-of-the-art supervised detection method); (2) SUPER-COLD enjoys a significant recall gain (up to 9.23% in terms of F-score) in general review detection tasks from incorporating entities and social relations; (3) SUPER-COLD generates a user behavior representation with a strong discriminate ability.

2 Related Work

2.1 Fraud Review Detection

Fraud review detection was initially studied in [8], and has long been an attractive research topic since then. Later, more efforts were made on exploits linguistics features [20, 11, 7], analyzing the effectiveness of n-grams, POS, ect. However, [19] found that linguistic features are insufficient when dealing with real-life fraud reviews. Therefore, researchers put more efforts in employing users' behavior features [4, 3, 10–12]. Also, [18] proved that user's behavioral features are more effective than linguistic features for fraud detection. Behavioral features were then intensively studied by introducing a set of graph-based methods. The intuition is reviews posted with similar fraud-behavior would be fraud. Wang et al. [24] first introduce review graph to capture the relationships between entities. Spotting fraudster groups were then explored by network footprints [27], community discovery with sentiment analysis [2], social interactions for sparse group [26]. In-depth, Hooi et al. [6] proposes an advanced dense subgraph mining for group fraudsters detection, targeting on detecting camouflage or hijacked accounts who manipulate their writing to look just like normal users.

2.2 Cold-Start Problem

The cold-start problem in fraud review detection was first studied in [25]. By considering the correlations between users, items, and reviews (entities relation), the review posted by a new user can be represented. Motivated by [25], the method proposed in [28] further leverages both attribute and domain knowledge for a better review representation. At last, the review representation is fed into a traditional classication model like SVM to form the fraud review classifier.

While both the above cold-start fraud review detection methods focus on user's review representation, we believe fraud reviews are easy to be manipulated to look like normal reviews [6] and thus may confuse existing methods. In this paper, we propose a novel user behavior representation model for fraud detection, where entities relations and user social relations are jointly embedded. In addition, we apply the dense subgraph mining to obtain pseudo-labels for existing users in an unsupervised way, which also avoids the difficulty to obtain high-quality labeled data.

3 Proposed Method

3.1 Behavior Representation Architecture

The behavior representation architecture of SUPER-COLD is shown in Figure 2. It consists: *entities relation embedding*, *social relation mining* and *user behavior embedding*. It also involves a dynamic link re-weighting strategy to enable a discriminative representation for fraud detection.

SUPER-COLD first embeds the relations among users, items, and reviews in their representations (entities relation embedding), and leverages the user social relation by the dense subgraph mining based on a user-item bipartite graph (social relation mining). Then, it further learns a user behavior representation by integrating the learned entities relations, which are embedded in the user representation, and the social relations, which are involved in the pseudo fraud labels generated by the dense subgraph mining, through a neural network (user behavior embedding). A dynamic link re-weighting strategy is adopted to enhance the discriminative ability of the generated behavior representation. Specifically, after user behavior representation is learned, SUPER-COLD discovers suspicious fraudsters according to the user distribution in the user behavior representation space, and assigns a higher weight to the links corresponding to these suspicious fraudsters in the user-item bipartite graph. After re-weighting, it executes the dense subgraph mining to reveal a more accurate social relation, which is further integrated with the user representation to generate a new behavior representation. SUPER-COLD repeats this process until convergence.

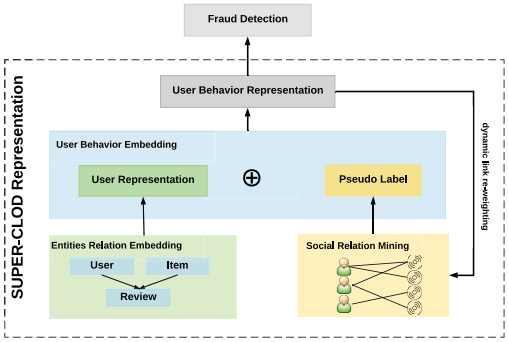


Fig. 2. The proposed SUPER-COLD Model

3.2 Entities Relation Embedding

SUPER-COLD embeds entities relation following the method in [25]. Let's note $\mathbf{v}_u$, $\mathbf{v}_o$, $\mathbf{v}_r$ as the representations of u, o, and r, where u refers to a user, o refers to an item, and r refers to a review wrote by the user u to the item o. We further denote a tuple of $< u, o, r >$ as $\nu \in S$, where S refers to an online review data set. For a given S, SUPER-COLD embeds the entities relation by the following objective function:

$$\begin{aligned} &\min_{\mathbf{V}_u, \mathbf{V}_o, \boldsymbol{\omega}} \sum_{\nu \in S} \sum_{\nu' \in S} \gamma \max\{0, 1 + \|\mathbf{v}_u + \mathbf{v}_o - \mathbf{v}_r\|^2 - \|\mathbf{v}_{u'} + \mathbf{v}_{o'} - \mathbf{v}_{r'}\|^2\}, \\ &s.t. \qquad \gamma = \begin{cases} 1 & u = u' \\ 0 & u \neq u', \end{cases} \\ &\qquad\quad \mathbf{v}_r = t_{\boldsymbol{\omega}}(r), \end{aligned} \tag{1}$$

where $\mathbf{V}_u$ and $\mathbf{V}_o$ is a set of the user and item representations, $t_{\boldsymbol{\omega}}(\cdot)$ refers to a text embedding neural network with parameters $\boldsymbol{\omega}$, and $\max\{\cdots\}$ returns the maximum in a set. SUPER-COLD implements the text embedding neural network as a hierarchical bi-directional recurrent neural network.

3.3 Social Relation Mining

SUPER-COLD models users relations by a user-item bipartite graph. Motivated by [6, 13], it adopts the dense subgraph mining to reveal the fraud behavior based on user social relation. The basic idea is greedily removing nodes in the bipartite graph which can maximize the subgraph density per a given density evaluation (Alg. 1 line 2 - 5). The final remained subgraph (Alg. 1 line 6) will have the largest density and thus can discover the users who may work together to manipulate reviews.

Given an online review data set S, SUPER-COLD constructs the user-item bipartite graph as $\mathcal{G} = (\mathcal{U} \cup \mathcal{I}, \mathcal{E})$, where $\mathcal{U}$ is a set of users, $\mathcal{I}$ is a set of items, and $\mathcal{E} = \{< u, o, r > | < u, o, r >\in S\}$ is a set of edges, i.e. links from users to items. Alg. 1 shows the process of the dense subgraph mining for SUPER-COLD. Motivated by [6], SUPER-COLD defines the density metric $g(\cdot)$ as follows,

$$g(S) = \frac{f(S)}{|S|}, \tag{2}$$

where

$$f(S) = \sum_{<u,o,r>\in\mathcal{E}} w_{u,o}, \tag{3}$$

for link weight $w_{u,o} > 0$ between user u to item o. Initially, SUPER-COLD assigns all link weights as 1. It will further adopt a dynamic re-weighting strategy to update the link weights in an iterative process.

Algorithm 1 Dense Subgraph Mining for SUPER-COLD.

Input: Bipartite graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{I}, \mathcal{E})$;
Output: The pseudo-labels set Y;
1: $\mathcal{X}_0 \leftarrow \mathcal{U} \cup \mathcal{I}$
2: **for** $t = 1, \cdots, (n_u + n_o)$ **do**
3: $i^* \leftarrow \arg\max_{i\in\mathcal{X}_{t-1}} g(\mathcal{X}_{t-1}\backslash\{i\})$
4: $\mathcal{X}_t \leftarrow \mathcal{X}_{t-1}\backslash\{i^*\}$
5: **end for**
6: $\mathcal{X}^* \leftarrow \arg\max_{\mathcal{X}_i\in\{\mathcal{X}_0,\cdots,\mathcal{X}_{n_u+n_o}\}} g(\mathcal{X}_i)$;
7: **for** $u = u_1, \cdots, u_{n_u}$ **do**
8: **if** $u \in \mathcal{X}^*$ **then**
9: $y_i = c_f$
10: **else**
11: $y_i = c_n$
12: **end if**
13: **end for**
14: **return** $Y = \{y_1, \cdots, y_{n_u}\}$

SUPER-COLD assigns pseudo-labels to existing users according to the dense subgraph mining results. Specifically, it gives the candidate fraudster label (c_f) to the users in the detected dense subgraph, and sets candidate normal user label (c_n) to other users (Alg. 1 line 7 - 14). These pseudo-labels inherit the social relations and will be used in the following behavior representation and cold-start fraud detection.

3.4 Integrating Entities and Social Relation for Behavior Representation

SUPER-COLD further integrates the entities and social relation for behavior representation. Specifically, it adopts a fully connected neural network, $Dense_{\mathbf{p}}(\cdot)$, to transform a user representation $\mathbf{v}_u$ to a user behavior representation $\mathbf{v}_u^*$, and minimizes the pseudo-labels prediction loss (defined as the cross-entropy between the predicted labels and pseudo-labels) based on $\mathbf{v}_u^*$ by updating the parameters $\mathbf{p}$ of $Dense_{\mathbf{p}}(\cdot)$ using a softmax function. The objective function can be formalized as follows,

$$\begin{aligned} \min_{\mathbf{p},\mathbf{w},b} \quad & \sum_{i=1}^{n_u} \sum_{y=\{c_f,c_n\}} \mathbf{1}[y_i = y] \log q_i \\ s.t. \quad & q_i = \text{softmax}(\mathbf{w} \cdot Dense_{\mathbf{p}}(\mathbf{v}_{u_i}) + b), \end{aligned} \tag{4}$$

where y_i is the pseudo-label of the i-th user assigned by Alg. 1, n_u is the number of existing users, $\mathbf{w}$ and b are the parameters of the softmax function.

3.5 Dynamic Link Re-weighting Strategy

The target of the user behavior representation is detecting fraudsters. To this end, the discriminative ability of the behavior representation should be strong. In SUPER-COLD, this discriminative ability is mainly obtained from the pseudo-labels generated by the dense subgraph mining. However, social relation may not comprehensively indicate all kinds of fraudsters [23]. As a result, the discriminative ability of behavior representation may not be good if only learning from the social relation.

To enhance the discriminative ability, SUPER-COLD reinforces the focusing of the dense subgraph mining on the suspicious users discovered in the behavior representation space, which reflects both the entities relation and the social relation. Specifically, SUPER-COLD clusters a set of users into two categories according to their behavior representations. It then re-weights the link of each user by the reciprocal of the number of its assigned categories. Formally, the links weight of a user u is assigned as,

$$w_{u,\cdot} = \frac{1}{|C_u|}, \tag{5}$$

where C_u refers to a set of users with the same category as user u, and $|\cdot|$ returns the size of the set. The assumption behind this re-weighting is that a user with less similar users are more suspicious as a fraudster. After re-weighting the links, SUPER-COLD conducts the dense subgraph mining again to generates new pseudo-labels, which are further integrated with the entities relation for the behavior representation. SUPER-COLD repeats this dynamic re-weighting strategy until convergence. With the dynamic link re-weighting strategy, the SUPER-COLD behavior representation procedure is summarized in Algorithm 2.

3.6 SUPER-COLD Fraud Review Detection

SUPER-COLD detects fraud reviews according to the behavior representation in an unsupervised way. The SUPER-COLD fraud review detection procedure is shown in Alg. 3. For a new review tuple $< u^*, o^*, r^* >$, the user behavior representation cannot be directly got from the existing model because u^* never appears and thus are not in the learned embedding layer. Alternatively, SUPER-COLD deduces the new user behavior representation by using the entities relation and the social relation. SUPER-COLD first looks up the learned item representation $\mathbf{v}_{o^*}$ and generates the review representation $\mathbf{v}_{r^*} = t_{\boldsymbol{\omega}}(r^*)$ by the learned text embedding neural network. Then, it calculates an approximate representation of the user as $\mathbf{v}_{u^*} = \mathbf{v}_{r*} - \mathbf{v}_{o^*}$. After

Algorithm 2 SUPER-COLD User Behavior Representation.

Input: Online review set S, convergence threshold ϵ;
Output: User behavior representation $\mathbf{V}_u^*$;
1: Entities relation embedding by Eq. (1)
2: Generating pseudo-label set Y by Alg. 1
3: Generating behavior representation $\mathbf{V}_u^*$ by Eq. (4)
4: Initializing $\Delta = +\infty$
5: **while** $\Delta > \epsilon$ **do**
6: $Y' \leftarrow Y$
7: Clustering $\mathbf{V}_u^*$ into two categories
8: Re-weighting user-item graph links by Eq. (5)
9: Generating pseudo-label set Y by Alg. 1
10: Generating behavior representation $\mathbf{V}_u^*$ by Eq. (4)
11: $\Delta = 1 - \frac{\sum_{y_i \in Y, y_i' \in Y'} 1[y_i = y_i']}{|Y|}$
12: **end while**
13: **return** $\mathbf{V}_u^*$

that, it uses the learned fully connected neural network to generate behavior representation of the user as $\mathbf{v}_{u^*}^* = Dense_{\mathbf{p}}(\mathbf{v}_{u^*})$, which integrates the social relation with the approximate representation. Finally, SUPER-COLD retrieves the k-nearest existing users of the detecting user, and uses the majority voting to ensemble the retrieved users' pseudo-labels as the label assigned to the u^* (Alg. 3 line 5). If most of the nearest users are candidate fraudsters, u^* will be treated as a fraudster and the r^* will be assigned as a fraud review (Alg. 3 line 6-11).

Algorithm 3 SUPER-COLD Fraud Review Detection.

Input: An online review tuple $< u^*, o^*, r^* >$, the number of nearest users k;
Output: The fraud detection label y^*;
1: Looking up $\mathbf{v}_{o^*}$
2: Generating $\mathbf{v}_{r^*} = t_{\boldsymbol{\omega}}(r^*)$
3: Calculating $\mathbf{v}_{u^*} = \mathbf{v}_{r^*} - \mathbf{v}_{o^*}$
4: Calculating $\mathbf{v}_{u^*}^* = Dense_{\mathbf{p}}(\mathbf{v}_{u^*})$
5: Retrieving $U = \arg\min_{U} \sum_{u \in U} \|\mathbf{v}_u - \mathbf{v}_u^*\|^2$ with $|U| = k$
6: Looking up pseudo-label set $|Y|$ of $|U|$
7: **if** $|\{y|y = c_f, y \in Y\}| > |\{y|y = c_n, y \in Y\}|$ **then**
8: $y^* = Fraud$
9: **else**
10: $y^* = Normal$
11: **end if**
12: **return** y^*

4 Experiments

4.1 Data Sets

Following the literature [25, 28] about cold-start fraud detection, our experiments are carried on two Yelp data sets including Yelp-hotel and Yelp-restaurant, which are also commonly used in previous fraud detection researches [19, 22, 16]. Table 1 and 2 display the statistics of the data sets.

We split original data sets into two parts for cold-start fraud detection performance evaluation. The first part includes 90% earliest posted reviews. The users who posted these reviews are treated as existing users. The second part is the 10% latest posted reviews. From the second part, we pick up the reviews which wrote by new users at the first time as cold-start reviews. Besides, we use the whole data sets to evaluate the general fraud detection performance and do the ablation study.

4.2 Evaluation Metrics

We evaluate their performance by three metrics - *precision*, *recall*, and *F-score*. While precision evaluates the fraction of relevant review among detected reviews, recall reflects the fraction of relevant reviews that have been detected over the total amount of relevant reviews. The precision and recall should be jointly considered since fraud detection is an imbalance problem [14], i.e. fraud reviews are much less than normal reviews. Thus, we use F-score, which balances the precision and recall, as an averaged indicator. Higher F-score indicates a better performance for a fraud detection method. We report these three metrics per ground-truth normal and fraud classes to illustrate the performance for different categories. We further average them to show an overall performance.

We follow the literature [25] 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 (normal reviews) are likely to be the closest to real fraud and normal reviews[19], they are not absolutely accurate [10]. The inaccuracy exists because it is hard for the commercial filer 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 the cold-start problem.

4.3 Parameters Settings

In our experiments, we use a hierarchical bi-directional GRU structure with 100 nodes to embed reviews, in which the pre-trained word embedding by GloVe algorithm [21] is used[6]. We train the user/item/review embedding by Adam [9]. We set the word embedding dimension as 100 and training batch size as 32. To integrate the entities relation and the social relation, we adopt a 3-layer fully connected neural network with 100 nodes in each hidden layer. We train this fully connected neural network 10 epochs by Adam with batch size 32. All hidden nodes of the neural network used in the experiments use ReLU as their activation function. We choose k-means as the clustering method in SUPER-COLD. When inferring the behavior of a new user, we retrieve the 5 closest existing users of the user according to the distance in the embedding space. For the parameters of the compared methods, we take their recommended settings.

4.4 Effectiveness on Cold-start Fraud Detection

Experimental Settings. SUPER-COLD is compared with the state-of-the-art method JETB [25]. This method handling the cold-start problem by considering entities (users, items and reviews) relations to represent reviews. When a new user posts a new review, this review can be represented by the trained network and classified by the classifier. JETB is also the first work that exploits the cold-start problem in review fraud detection.

In [25], 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 a large amount of data. In this experiments, the training data contains $619,496$ and $709,623$ reviews on Yelp-Hotel and Yelp-Restaurant data sets, respectively. To make JETB practicable, we use a 5-layer fully connected neural network instead of SVM as the fraud classifier of JETB.

Findings - SUPER-COLD Significantly Outperforming the State-of-the-art Cold-start Fraud Detection Method. Table 1 demonstrates the SUPER-COLD fraud detection performance, compared to JETB

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

Table 1. Cold-start Fraud Detection: Precision (P), Recall (R) and F-score (F) are reported.

Data Info.				SUPER-COLD			JETB			Improvement		
Name	Category	#Existing	#Cold-start	P	R	F	P	R	F	P	R	F
Hotel	Normal	376,671	60	**0.45**	0.15	0.23	0.32	**0.90**	0.47	40.63%	-83.33%	-51.06%
	Fruad	242,825	122	**0.69**	**0.91**	**0.78**	0.57	0.07	0.12	21.05%	1200.00%	550.00%
	Overall	619,496	182	**0.61**	**0.66**	**0.6**	0.49	0.34	0.24	24.49%	94.12%	150.00%
Restaurant	Normal	412,435	1,654	0.64	**0.84**	**0.73**	**0.68**	0.74	0.71	-5.88%	13.51%	2.32%
	Fruad	297,188	873	**0.62**	**0.68**	**0.65**	0.41	0.34	0.37	51.22%	100.00%	75.30%
	Overall	709,623	2,527	**0.63**	**0.78**	**0.70**	0.59	0.60	0.59	7.90%	30.39%	18.07%

on Yelp-Hotel and Yelp-Restaurant data sets. SUPER-COLD largely improves the fraud detection performance, i.e. 150% and 18.07% F-score increase on Yelp-Hotel and Yelp-Restaurant data sets. This averaged performance improvement is mainly contributed by the increased detection performance of fraud category (550% on Yelp-Hotel and 75.3% on Yelp-Restaurant). As shown in the results, SUPER-COLD "decreases" the performance of normal reviews detection. It reflects SUPER-COLD is more tough for fraud reviews. On one hand, this "decreased" performance of normal reviews detection does not decrease the averaged fraud detection performance. On the other hand, this "decreased" performance may be caused by the *noising ground-truth labels* of the cold-start fraud reviews that do not be detected by the Yelp commercial filter.

SUPER-COLD uses the represented user behavior instead of review features for cold-start fraud detection to avoid the camouflage in reviews. Because of the more reliable information, SUPER-COLD achieves significant performance improvement in cold-start fraud detection.

4.5 Effectiveness on General Fraud Detection

Experimental Settings. SUPER-COLD is further compared with three state-of-the-art competitors: Frauder [6], HoloScope [13], and SPEAGLE [23] in detecting *general fraud reviews* - all the reviews contained in a data set. These three competitors have different but relevant mechanisms compared with SUPER-COLD.

- *Fixed weighting dense subgraph mining-based method - FRAUDER* [6]. FRAUDER is a fraud detection method by dense subgraph mining. To detect camouflage and hijacked accounts, it adopts a fixed weighting strategy. Different from FRAUDER, the dense subgraph mining method used in SUPER-COLD is with a dynamic link weighting strategy to further fuse the entities relation with the social relation.
- *Dynamic weighting dense subsgraph mining-based method - HoloScope* [13]. HoloScope uses graph topology and temporal spikes to detect fraudsters groups, and employs a dynamic weighting approach to enable a more accurately fraud detection. However, the dynamic weighting is only conducted once according to the user temporal spikes. In contrast, SUPER-COLD interactively updates the dynamic weighting along the user behavior embedding process.
- *Metadata and social relation integration-based method - SPEAGLE* [23]. SPEAGLE proposes a unified framework to utilize metadata and the social relation in Markov Random Field for fraud detection. While SPEAGLE needs fraud labels, SUPER-COLD is a completely unsupervised method which jointly considers the entities relation and the social relations for user behavior representation.

While FRAUDER and HoloScope directly predict fraud reviews, SPEAGLE gives a probability of a review may be fake. To make a fair comparison, we only report the averaged precision of SPEAGLE but ignore its the recall and F-score.

Table 2. General Fraud Detection: Precision (P), Recall (R) and F-score (F) are reported.

Data Info.			SUPER-COLD			HoloScope			FRUADER			SPEAGLE			Improvement		
Name	Category	#Review	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Hotel	Normal	420,785	**0.69**	0.96	**0.8**	0.64	0.6	0.62	0.64	**0.98**	0.77	0.53	-	-	7.81%	-2.04%	3.90%
	Fruad	267,544	**0.82**	0.31	**0.45**	0.42	**0.46**	0.44	**0.82**	0.11	0.31	0.72	-	-	0.00%	-32.61%	2.27%
	Overall	888,329	**0.74**	**0.71**	**0.66**	0.55	0.55	0.55	0.71	0.65	0.55	0.60	-	-	4.23%	9.23%	20.00%
Restaurant	Normal	461,190	**0.67**	0.9	**0.77**	0.51	**0.95**	0.66	0.63	**0.95**	0.76	0.42	-	-	6.35%	-5.26%	1.32%
	Fruad	326,981	0.73	**0.37**	**0.49**	**0.74**	0.12	0.21	**0.74**	0.21	0.33	0.58	-	-	-1.35%	76.19%	48.48%
	Overall	788,741	**0.69**	**0.68**	**0.65**	0.63	0.52	0.43	0.68	0.64	0.58	0.49	-	-	1.47%	6.25%	12.07%

Findings - SUPER-COLD Significantly Improving General Fraud Detection Performance, Especially Recall. The precision, recall, and F-score of SUPER-COLD, Frauder, HoloScope, and SPEAGLE are reported in Table 2. Overall, SUPER-COLD significantly outperforms the competitors. It improves 20% and 12.07% compared with the best-performing method in terms of F-score on two data sets.

Unlike FRAUDER and HoloScope that ignore the entities relation when they perform dense subgraph mining based on the social relation, SUPER-COLD couples these two independent relations to iteratively refine their performance by the dynamic link weighting. This enables SUPER-COLD to avoid camouflage by the social relation and effectively detect personalized fraud by the entities relation. Consequently, SUPER-COLD obtains up to 76.19% recall improvement compared with the competitors.

4.6 Quality of Behavior Representation

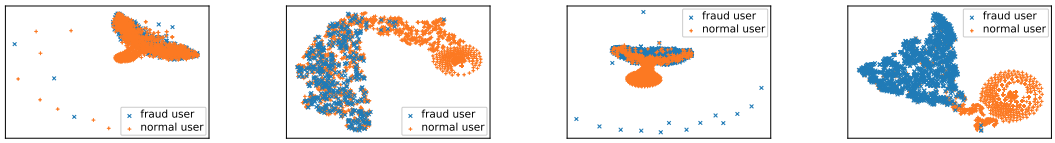
Experimental Settings. We visualize the behavior representation in a two-dimensional space trough TSNE [15]. To evaluate the representation quality, we plot pseudo-labels of each user according to the dense subgraph mining-based fraud detection results. A high-quality behavior representation will enable a separate location of users with different pseudo-labels. The behavior representation generated by SUPER-COLD is compared with that generated by JETB.

Findings - SUPER-COLD Generated Behavior Representation is with Strong Discriminate Ability. The behavior representations generated by SUPER-COLD and JETB are visualized in Figure 3. In the JETB generated representation space, there is a large overlap between users with different pseudo-labels, especially on Yelp-Hotel data set. In contrast, the SUPER-COLD generated representation is with a stronger discriminative ability that separates users with pseudo-labels well. These qualitative illustrations are consistent with the quantitative results in Table 2 that the improvement brought by SUPER-COLD on Yelp-Hotel data set is much larger than that on Yelp-Restaurant data set.

Based on the JETB generated user behavior representation, SUPER-COLD moves one step further. It adopts the pseudo-labels generated by the dense subgraph mining-based fraud detection to fine-tune the representation learning from the entities relation. Since the dense subgraph mining-based fraud detection involves much more effective patterns (e.g. group manipulation) about fraudsters, the fine-tuned representation thus has a stronger discriminative ability for fraud and normal users.

4.7 Ablation Study

Experimental Settings. We further study the contribution from each SUPER-COLD components: entities relation learning, social relation learning, dynamic graph link re-weighting, and behavior-based cold-start



(a) JETB: Hotel. (b) SUPER-COLD: Hotel. (c) JETB: Restaurant. (d) SUPER-COLD: Restaurant.

Fig. 3. User Behavior Embedding of Different Methods.

fraud detection. This contribution can be analyzed from the Table 1, 2 and Figure 3. Here, we assume FRAUDER reflects the performance of social relation learning, HoloScope demonstrates the performance of dynamic weighting, SPEAGLE stands for the performance of combining the entities and social relations, and JETB implies the performance of review-based cold-start fraud detection.

Findings - SUPER-COLD is Contributed by learning Entities and Social-Relation and Dynamically Re-weighting Graph Links, Especially by Social Relation. As shown in Table 1, SUPER-COLD outperforms FRAUDER at least 12.07% in terms of F-score. Meanwhile, SPAEGLE also achieves much better performance compared with FRAUDER. This demonstrates that incorporating entities relation with social-relation gains a large performance improvement, which reflects the contribution of entities relation.

Social relation also makes a significant contribution, which is much greater than that made by the entities relation. Compared with JETB (with entities relation but without social relation), SUPER-COLD (with entities and social relation) gains 150% F-score improvement on Yelp-Hotel data set because of integrating social relation. In contrast, on that data set, the improvement brought by entities relation is only 20% according to the comparison of SUPER-COLD (with entities and social relation) and FRAUDER (with social relation but without entities relation) shown in Table 2.

Dynamically re-weighting graph links make the dynamic weighting strategy more reliable. As shown in Table 2, the performance of HoloScope (with dynamic weighting) is similar to FRAUDER (without dynamic weighting) on Yelp-Hotel data set but slightly lower than FRAUDER on Yelp-Restaurant data set. However, SUPER-COLD (with dynamic re-weighting) consistently outperforms FRAUDER on both data set. The reason may lie in the re-weighting mechanism that iteratively enhances weighting quality.

5 Conclusion

This paper proposes a socially-aware unsupervised user behavior representation method to tackle the cold-start problem in fraud review detection. The proposed unsupervised method integrates both entities and social relations for user behavior representation, and further strengths the discriminative ability of the behavior representation by a dynamic link re-weighting strategy. It can effectively detect fraud reviews with the cold-start problem as demonstrated by comprehensive experiments.

References

1. Chen, J., Saad, Y.: Dense subgraph extraction with application to community detection. TKDE **24**(7), 1216–1230 (2012)

2. Choo, E., Yu, T., Chi, M.: Detecting opinion spammer groups through community discovery and sentiment analysis. In: IFIP. pp. 170–187. Springer (2015)
3. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. ICWSM **13**, 175–184 (2013)
4. Feng, S., Xing, L., Gogar, A., Choi, Y.: Distributional footprints of deceptive product reviews. ICWSM **12**, 98–105 (2012)
5. Hooi, B., Shin, K., Song, H.A., Beutel, A., Shah, N., Faloutsos, C.: Graph-based fraud detection in the face of camouflage. TKDD **11**(4), 44 (2017)
6. Hooi, B., Song, H.A., Beutel, A., Shah, N., Shin, K., Faloutsos, C.: Fraudar: Bounding graph fraud in the face of camouflage. In: ACM SIGKDD. pp. 895–904. ACM (2016)
7. Hovy, D.: The enemy in your own camp: How well can we detect statistically-generated fake reviews–an adversarial study. In: ACL. vol. 2, pp. 351–356 (2016)
8. Jindal, N., Liu, B.: Opinion spam and analysis. In: WSDM. pp. 219–230. ACM (2008)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Li, H., Chen, Z., Liu, B., Wei, X., Shao, J.: Spotting fake reviews via collective positive-unlabeled learning. In: ICDM. pp. 899–904. IEEE (2014)
11. Li, H., Chen, Z., Mukherjee, A., Liu, B., Shao, J.: Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In: ICWSM. pp. 634–637 (2015)
12. Li, H., Fei, G., Wang, S., Liu, B., Shao, W., Mukherjee, A., Shao, J.: Modeling review spam using temporal patterns and co-bursting behaviors. arXiv preprint arXiv:1611.06625 (2016)
13. Liu, S., Hooi, B., Faloutsos, C.: Holoscope: Topology-and-spike aware fraud detection. In: CIKM. pp. 1539–1548. ACM (2017)
14. Luca, M., Zervas, G.: Fake it till you make it: Reputation, competition, and yelp review fraud. Management Science **62**(12), 3412–3427 (2016)
15. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. JMLR **9**(Nov), 2579–2605 (2008)
16. Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R.: Spotting opinion spammers using behavioral footprints. In: ACM SIGKDD. pp. 632–640. ACM (2013)
17. Mukherjee, A., Liu, B., Wang, J., Glance, N., Jindal, N.: Detecting group review spam. In: WWW. pp. 93–94. ACM (2011)
18. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: 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)
19. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.S.: What yelp fake review filter might be doing? In: ICWSM (2013)
20. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: ACL HLT. pp. 309–319. Association for Computational Linguistics (2011)
21. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
22. Rayana, S., Akoglu, L.: Collective opinion spam detection: Bridging review networks and metadata. In: ACM SIGKDD. pp. 985–994. ACM (2015)
23. Rayana, S., Akoglu, L.: Collective opinion spam detection using active inference. In: ICDM. pp. 630–638. SIAM (2016)
24. Wang, G., Xie, S., Liu, B., Philip, S.Y.: Review graph based online store review spammer detection. In: ICDM. pp. 1242–1247. IEEE (2011)
25. Wang, X., Liu, K., Zhao, J.: Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In: ACL. vol. 1, pp. 366–376 (2017)
26. Wu, L., Hu, X., Morstatter, F., Liu, H.: Adaptive spammer detection with sparse group modeling. In: ICWSM. pp. 319–326 (2017)
27. Ye, J., Akoglu, L.: Discovering opinion spammer groups by network footprints. In: ECML. pp. 267–282. Springer (2015)
28. You, Z., Qian, T., Liu, B.: An attribute enhanced domain adaptive model for cold-start spam review detection. In: COLING. pp. 1884–1895 (2018)