A Review of Recent Advance in Online Spam Detection

Yingtong Dou YDOU5@UIC.EDU
Department of Computer Science
University of Illinois at Chicago
March 15, 2019

Abstract

Detecting spams in online social network and review platforms is a challenging task drawing attention from multidisciplinary research communities. It is a long lasting campaign between the fraudsters and platforms. In terms of data mining, spam detection is an anomaly detection task. Many algorithms which leverage the behavioral and linguistic features of the users and their posts have been proposed in this area. Graph model and deep model are gradually being adapted to this problem as well.

In this survey, I review and critique three recently published papers focusing on different aspects of the spam detection problem. They employ graph model and deep model to solve the emerging and challenging problems in spam detection. The first paper encodes the edge direction information into an optimized belief propagation model. The second one firstly tackles cold start problem in spam detection with an unsupervised embedding model. The last paper proposes a framework to detect crowdsourced spammers. After reviewing the related work, I give a brief introduction of their proposed models plus some preliminary knowledge. Then I discuss the contributions and drawbacks of their models. At last, I propose some challenging problems in spam detection along with a couple of future research directions.

1 Introduction

As the booming of the web 2.0 services like online social network and e-commerce, people spend more time online to fulfill their social need, seek information and shop goods. Meanwhile, the corresponding online businesses make a considerable profit. However, there are various kinds of suspicious behavior and spams existing in those websites and services aiming to exaggerate the reputation of businesses or even steal personal information from users. A report claims that the percentage of fake reviews on Yelp grows from 6% in 2006 to 20% in 2013 [42]. Because of the anonymity and distributive identities of the Internet, it has been a challenging task for the online service providers to tackle the growing number of fraudsters and spams. Moreover, the botnet and freelancing workers both involved in the online spam activities. It makes the spam detection task even more difficult. The evolution of cheating techniques makes the online spam detection problem become a long lasting campaign between the fraudsters and platforms. There are still numerous unsolved challenges in this area today.

In terms of data mining, such fraud/spam detection task is named as *Anomaly Detection* which is a process identifying anomaly data points from regular data points in a given dataset. The research of anomaly detection starts from the early 1970s in databases [6]. As for the online spam detection research, it develops along with the development of web applications. Spam email and short messages are the very first types of online spams which appear in the early 21^{st} century [5]. Since spam emails usually have advertisement content which is distinct to regular emails, couples of spam detectors employing semantic features have been proposed [17].

As the online social network like Facebook and Twitters thriving, new types of spam gradually emerge [34]. The fake followers and fake likes are typical types of social spam. Thanks to the abundant information of the social network, we could better model those suspicious activities. The online spam review is a unique type of online spam which draws lots of attention in recent years [27, 26, 7]. The review system of web services helps the people know more about the merchandises or services. The business owners also obtain useful feedback from users. At the same time, such mechanism fosters lots of fabricated fake reviews. Generally, those reviews are purchased by the business owners in order to boost their product ratings and increase positive reviews. In this survey, I mainly address the works that deal with spammers and spams on online social network and review platforms.

The semantic, behavioral and structural features of users are all harnessed to gauge the suspiciousness of themselves. At the early stage of spam detection research, most of the studies focus on the semantic features of reviews or posts [48, 52, 40, 27, 36, 24, 38, 15]. They assume that the spams may differ from regular reviews/posts in some handcrafted semantic features. Some high-level semantic features like sentiment are also proposed to help identify spams [23]. Some of the text features are shown in Table 1.

Besides text features, behavioral information is another essential perspective to differ the spammers from benign users. Temporal behaviors of the fraudsters like

Table 1: Behavior and text features proposed in the state-of-the-art works on spam and spammer detection. Due to the different regime of review platforms and OSNs, they share some common features and have their unique features [54].

Behavior	**Description**	**Text**	**Description**
MNR	Max. number of reviews posted in a day [46]	RL	Avg. review length [48]
PR	The ratio of positive reviews [48]	ACS	Avg. content similarity [39]
NR	The ratio of negative reviews [48]	PCW	Percentage of all capital words [36]
WRD	Weighted Rating Deviation [39]	PC	Percentage of capital letters [36]
BST	Burstiness [14]	DL_b	Description length based on bigrams [54]
RD	Rating deviation of product's avg. rating [36]	PP1	The ratio of 1st person pronouns [36]
Rank	The rank order of the review [27]	RES	The ratio of exclamation sentences [36]
ETF	The early time frame of the reviewer [46]	SW	The ratio of subjective words [36]
ISR	Is singleton? [54]	OW	The ratio of objective words [36]
DPW	Deceptive review count previous week [29]	F	The frequency of review [54]

posting frequency, number of reviews, posting burstiness are applied in the detection algorithms [14, 32, 54, 69, 37, 29]. Besides the temporal features, other behavior features like group activity [47, 4], rating distribution [39] and rating deviation of products [46] have been proposed as well.

In recent years more graph-based model and deep learning models are proposed. They leverage the hidden connections between data points in spam detection datasets. Meanwhile, some new problems especially new types of spam are emerging along with the attack-defense campaign. In this survey, I mainly review and critique three recent papers published at top-tier conferences. They investigate the novel problems in online spam detection and employ the state-of-the-art learning methods to solve the challenging problems.

The first paper proposes a graph-based algorithm called **GANG** which identifies spammers on online social networks [61]. They quantify the difference between edges with different directions and approximate the belief propagation algorithm with a matrix form. The matrix form makes belief propagation more scalable. They also provide a theoretical guarantee to the convergence of **GANG**. Traditional spam detectors usually utilize user behavior and network features for detecting spams. However, in some scenarios, spammers are new coming singletons users, where behavior and network features are not available [48]. In the second paper, Wang et al. firstly tackle such cold start problem under spam review detection settings [66]. Meanwhile, it is a typical work that adapts deep model into spam detection research. The last paper focuses on another challenging problem called crowdsourcing [28]. It is a kind of fraud activity that hires freelancing workers to write and post fake reviews. Those crowd workers are very similar to ordinary users in perspective of traditional

Table 2: Comparison of three state-of-the-art papers reviewed in this survey.

Section	Paper	Problem	Model	Target	Dataset	Venue
Section 2	Paper 1 [61]	Optimizing Alg	Graph	Social Spammer	Twitter	ICDM2017
Section 3	Paper 2 [66]	Cold Start	Deep Model	Spam Review	Yelp	ACL2017
Section 4	Paper 3 [28]	Crowdsourcing	Graph&Deep	Crowd Worker	Amazon	WSDM2018

features. Another challenge is the lack of the ground truth of those crowd workers. A detection framework called **TwoFace** is proposed in the paper. It discovers more suspicious users from local neighborhood and distant communities based on only few high-suspicious crowd workers. An overall summary of three papers is shown in Table 2.

The remaining parts of the paper are organized as follows. Section 2, 3, 4 present the reviews and critiques of the three papers mentioned above respectively. Section 5 discusses some challenges and pitfalls that exist in the current research of spam detection, and it then points out a few future research venues.

2 Detecting Social Spammers with Graph Model

The users in social networks and reviewers/products on review platforms could be easily modeled with a graph model. Based on the graph homophily assumption, we believe nodes that have explicit or latent connections in a graph usually share some common attributes [51]. Thus, the graph model could help us discover more suspicious nodes beyond the feature-based model [64].

In the spam detection scenario, some of the works use the random walk algorithm to propagate the suspiciousness between each node in the graph [71, 19]. The relation between nodes could also be represented as a Markov Random Field (MRF) where nodes have joint posterior probabilities with each other. Some inference algorithms on MRF like belief propagation has been proposed for the spam detection task [54]. While in a graph with loops (most of the graphs in the real world are loopy graphs), we can only use some approximation algorithms like Loopy Belief Propagation without convergence guarantee.

In this section, I will review the paper from Wang et al. [61]. It targets the suspicious user detection problem in the online social network. The proposed **GANG** model could effectively leverage the directed edge information in the social network and calculate the user suspiciousness with theoretical guarantee efficiently. Their algorithm is based on an MRF model mentioned above. Furthermore, they give an optimized formulation of belief propagation which has convergence guarantee and scalability. In the following parts of this section, I first introduce how they model the social network with a directed graph model and how the authors incorporate edge direction information into the node suspiciousness estimation process. The next subsection includes the mechanism of belief propagation in graphical models and the techniques to optimize it. The last part is a critique of the proposed model based on the regime especially the robustness of the graph model.

2.1 Directed Graph Model

The graph model (a.k.a network model) has been proved to have strong ability in modeling the relationship between real-world entities and concepts. The research of graph problem started from the Euler's *Seven Bridge Problem* which abstracts the real world islands and bridges into a graph with nodes and edges [67]. The graph theory sheds light on the statistical machine learning research a few decades ago. The graphical model could represent the joint probabilities between variables based on the Bayes theory [49].

The social network research was first formulated by Scott et al. [57]. It models the real world human relationship and represents the unseen connections with particular nodes (person) and edges (friendship). As for the online social network, such relation are more straightforward to render. For instance, the *following* and *followed* relation between Twitter accounts can be modeled as two edges with opposite direction [45].

Generally, a graph $G = (V, E)$ includes a set of nodes $v \in V$ and a set of edges $e \in E$. The edge (u, v) and (v, u) point to the same edge in the undirected graph while they represent edges with opposite directions in the directed graph. There are three types of relationship between two accounts in the Twitter network. The bidirectional edge means two accounts follow each other; the unidirectional incoming edge means a account follows the account it points to; the unidirectional outgoing edge means the account follows its neighbor account. Figure 1 illustrates three edge types.

Definition 2.1 *(Fraudster Detection on Directed Graph Model) Given a directed social graph with labeled fraudulent nodes and normal nodes as the training set, Fraudsters detection is to predict the label of each remaining nodes in the testing set.*

Note that the directional relationship does not exist in online social networks like Facebook where friend relation is equal between two users. Wang et al. leverages such unique edge information to evaluate the suspiciousness of the accounts in microblogs like Twitter and Weibo. Their intuition is based on the homophily assumption of the graph model [51]. It suggests that the association between two accounts will

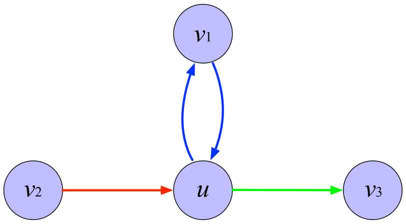


Figure 1: Different edge types in a Twitter network model. Blue is the bidirectional edge. Red is the unidirectional incoming edge. Green is the unidirectional outgoing edge [61].

reflect their inherent similarity. For instance, friends always have some common interests and preferences. The fraudulent users usually connect with each other, and the regular users seldom follow the fraudulent accounts. The formulated edge-node suspiciousness correlation is described below.

Bidirectional Edge If the account u and v follow each other, they tend to have the same property which is account suspiciousness under our problem setting. Moreover, given the labels of u's neighbors, the probability that u is fraudulent is modeled with a *sigmoid function* below:

$$\Pr(x_u = 1 | \overline{x}_{\Gamma_b(u)}) = \frac{1}{1 + \exp(-\sum_{v \in \Gamma_b(u)} J_{vu} \overline{x}_v)}, \tag{1}$$

where J_{vu} is the coupling strength of the edge vu. $J_{vu} = J_{uv}$ for the bidirectional edges and J_{vu} is set to be positive to model the homophily property. Thus, the account will be more suspicious when more of its neighbor accounts are suspicious. The sigmoid function helps us capture the homophily using a pairwise Markov Random Field which will be introduced in the next subsection.

Unidirectional Incoming Edge If v is the unidirectional incoming neighbor of the account u, it means the account v follows account u but not vice versa. Then v is not informative for u's label when v is fraudulent in such situation. Whereas if account u is followed by many benign accounts, the probability of account u being benign tend to be high. Because the fraudsters on Twitter could follow many accounts which could either be fraudulent or benign while the regular users tend to follow benign accounts. Such intuition could be modeled by slightly modifying Equation 1. By modifying the exponential factors of Equation 1, the label of node v will have difference influence on the probability of node u.

Unidirectional Outgoing Edge If v is the unidirectional outgoing neighbor of the account u, it means the account u follows account v but not vice versa. The influence of neighbors is entirely different compared to the incoming edge. If v is fraudulent, then u tends to be fraudulent since regular accounts seldom follow the fraudulent accounts. It does not influence u's suspiciousness if v is normal since both types of accounts could follow normal accounts. The formulation of this intuition is slightly different from the formulation modeling unidirectional incoming edges. Thus, the influence of the neighbors is also different from the case introduced above.

Unifying the three different formulations above could cover all types of directed edges in the directed graph model (See Figure 1). According to the related work on feature-based model, we could calculate the prior belief of a node with the node features listed in Table 1. Unifying the node prior neighbor influence, the probability of a node u to be fraudulent is:

$$\Pr\left(x_u = 1 | \overline{x}_{\Gamma(u)}\right) = \frac{1}{1 + \exp\left(-I_b(u) - I_i(u) - I_o(u) - h_u\right)}, \tag{2}$$

where $I_b(u) = \sum_{v \in \Gamma_b(u)} J_{vu} \overline{x}_v$ is the total influence of bidirectional neighbors; $\frac{1}{2} \sum_{v \in \Gamma_o(u)} J_{uv} \left(\overline{x}_v - 1\right)$ is the total influence of unidirectional incoming neighbors; $\frac{1}{2} \sum_{v \in \Gamma_o(u)} J_{uv} \left(\overline{x}_v + 1\right)$ is the total influence of unidirectional outgoing neighbors; the h_u represents the prior belief of node u.

At the next stage, the authors model such probability relations with a Markov Random Field which is introduced in the following section.

2.2 pMRF and Loopy Belief Propagation

After modeling distinct indications of different edge types in the directed graph. The authors employ the pairwise Markov Random Field (pMRF) to represent the node relation and further infer node posterior probabilities (beliefs). The pMRF is a sharp tool modeling random variables in an undirected graph. If a set of random variables V are only dependent on their neighbors in the graph, such variables are in an MRF [49]. They are in pMRF if every edge (u, v) in the graph has a compatibility function $\varphi_{uv}\left(x_u, x_v\right)$.

Since the **GANG** model precisely models the different relation according to the edge types, the paper transfers Equation 2 into the probability function into pMRF below:

$$\Pr\left(x_V\right) = \frac{1}{Z} \prod_{v \in V} \phi_v\left(x_v\right) \prod_{(u,v) \in E_1 \cup E_2} \varphi_{uv}\left(x_u, x_v\right), \tag{3}$$

where Z is the sum of all probability values x_V and the $\phi_v\left(x_v\right)$ represents the node

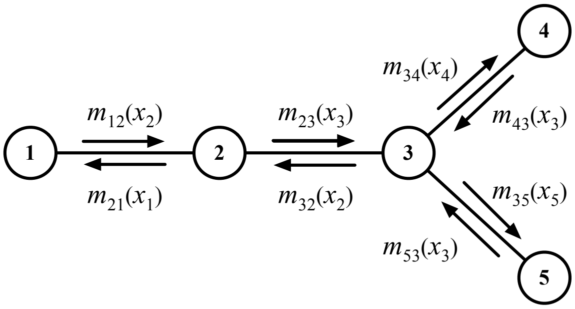


Figure 2: A illustration of message passing process in belief propagation. For instance, the message sent from node 3 to node 2 includes all the messages sent from the neighbors of node 3 and excludes $m_{23}(x_3)$ [74].

potential which is the prior belief calculated by the known information (handcrafted features). The edge compatibility $\varphi_{uv}\left(x_u, x_v\right)$ is defined based on corresponding edge type introduced in the first section above. Such modification helps adapt the original pMRF on the undirected graph to the directed graph model.

It has been proved to be an *NP-complete* problem to estimate the probability (posterior belief) of a node in pMRF. **Loopy Belief Propagation (LBP)** is a general algorithm to approximate the node posterior belief [50]. Specifically, it supposes that every node passes messages to each other iteratively. The message $m_{vu}^{(t)}\left(x_u\right)$ sent from v to u in the tth iteration is:

$$m_{vu}^{(t)}\left(x_u\right)=\sum_{x_v}\phi_v\left(x_v\right)\varphi_{vu}\left(x_v, x_u\right)\prod_{k\in\Gamma(v)/u} m_{kv}^{(t-1)}\left(x_v\right), \tag{4}$$

where ϕ_v is the node potential of node v, and $\varphi_{vu}\left(x_v, x_u\right)$ is the edge potential between node v and node u. The last term is the multiplication of the messages from all neighbor nodes excepting node u.

The LBP is a message passing algorithm which propagates the belief of a node to its neighborhood. Figure 2 illustrates how messages pass between each node. The LBP updates all the messages iteratively until it converges. The posterior belief of a node u can be calculated by the equation below:

$$\Pr^{(t)}\left(x_u\right)\propto\phi_u\left(x_u\right)\prod_{k\in\Gamma(u)} m_{ku}^{(t)}\left(x_u\right). \tag{5}$$

However, the LBP will become very slow on the large scale graph since it needs to maintain the messages on the edges at each iteration of message updating. Another pitfall of the LBP is that it lacks a theoretical guarantee on when it will converge.

The paper deals with the two issues above by eliminating the message maintenance process and approximating the belief propagation process on the graph model with a matrix form. It is motivated by recent works on linearizing LBP on undirected graphs [16]. The linearization in **GANG** model is more complicated since it models the directed graph. The optimized belief propagation formulations are:

$$\begin{cases} \mathbf{A}_i'^{(t-1)} = I\left(\mathbf{A}_i \circ \hat{\mathbf{P}}^{(t-1)^T}\right) \\ \mathbf{A}_o'^{(t-1)} = I\left(-\mathbf{A}_o \circ \hat{\mathbf{P}}^{(t-1)^T}\right) \\ \hat{\mathbf{p}}^{(t)} = \hat{\mathbf{q}} + 2 \cdot \hat{w} \cdot \left(\mathbf{A}_b + \mathbf{A}_i'^{(t-1)} + \mathbf{A}_o'^{(t-1)}\right) \cdot \hat{\mathbf{p}}^{(t-1)}, \end{cases} \tag{6}$$

where $\mathbf{A}_b \in \mathbb{R}^{V|\times|V|}$, $\mathbf{A}_i \in \mathbb{R}^{V|\times|V|}$ and $\mathbf{A}_o \in \mathbb{R}^{V|\times|V|}$ represent the bidirectional, unidirectional incoming and unidirectional outgoing adjacency matrices correspondingly. $\mathbf{p}^{(t)} = \left[p_1^{(t)}; p_2^{(t)}; \cdots; p_{|V|}^{(t)}\right]$ is the column vector represents the posterior beliefs of all nodes at the $t-$th iteration and $\hat{\mathbf{p}}^{(t)}$ is its residual vector. $\hat{\mathbf{P}}^{(t)} \in \mathbb{R}^{|V|\times|V|}$ is a matrix consisting $|V|$ times of repeats of the column vector $\hat{\mathbf{p}}^{(t)}$. Similarly, $\hat{\mathbf{q}}$ is the residual column vector of the prior beliefs of all nodes; $\hat{w}$ is the residual of the edge homophily strength. The indicator function $I(\mathbf{Y})$ means that the entry is set to be 0 if the corresponding entry of the matrix $\mathbf{Y}$ is non-negative; otherwise, it is set to be 1.

The first two lines of the Equation 6 model the influence of the unidirectional edges when we propagate the belief between nodes and their neighbors. Since the bidirectional edge indicate nothing to both sides, it is directly aggregated with A_i and A_o to yield the final adjacency matrix in the last line. The last line represents the belief propagation process updating the node posteriors directly without message maintenance. The final formulation of belief propagation is very similar to the random walk algorithm [59].

The authors further prove its efficiency comparing with former LBP and give the upper bound of $\hat{w}$ which guarantees the convergence of **GANG**. More details of algorithms and proof of them can be found in the original paper.

To summarize, the **GANG** model leverages the directed edge information and use a matrix multiplication algorithm to approximate the LBP. Further optimization is also applied to the LBP. The experiment results validate the effectiveness and scalability of the **GANG** comparing with other baselines.

2.3 Contributions & Drawbacks

From the Equation 6 of the final matrix formulation of the **GANG** model , we could find that it is a simplified random walk model. The difference is that the first two equations of 6 encode the directed edge influence with matrix entries. That is the primary contribution of the paper. The first half the paper explains how edge direction affects node suspiciousness and formulate it into a pMRF model. Eventually, the model settles with a concise matrix form. It is better than starting from a matrix form directly.

Though the paper has a strong theoretical guarantee, it still has some pitfalls for the approximated belief propagation in Equation 6. General random walk algorithm needs normalization to avoid the aggregation effect during node prior information propagating [60]. For the **GANG** model, the node suspicious score is correlated with the node degree since it does not normalize the matrix entries according to the node degree.

The unnormalized **GANG** model will tend to assign higher suspicious scores to nodes with higher node degree since the node will aggregate more information from its neighbors. The outstanding performance of the unnormalized model illustrates that the model could better fit the dataset. However, the unnormalized model is vulnerable to attacks. When the spammer manipulates the node degree, the node could be easily evaded from the detector. For instance, the spammer could control whom to follow and how many accounts to follow. It will change the node degree and further lower down the suspiciousness of the node under the **GANG** model.

Therefore, designing a more robust node classification algorithm against various kinds of attacks is a promising research direction. More details are introduced in Section 5.

3 Handling Cold Start Problem using Review Embedding

Different from the social network model in the first paper, users and products reviewing relationship could be modeled as a bipartite graph on review platforms. In such bipartite graphs, nodes denote users and products, and edges represent the reviews connecting users and products [1]. It is a unique type of graph model where the users and products do not have edges inside themselves. Like the social network model, the LBP could be applied to the review graph model [54].

However, the feature model and graph model are unable to encode the latent connections between the reviewers, reviews and products. When they meet a new coming singleton review, they could not capture useful features from it. Recently, as the development of deep learning, some works apply the deep learning methods to the spam detection problem. Previous work demonstrates that the semantic features could not discriminate the spams effectively [48]. While a recent work employs Convolutional Neural Network to capture the document level and high dimensional features of the review text [55], they prove the deep model could better identify benign and fabricated reviews than handcrafted features.

Besides the deep semantic model, other works have tried various types of deep architecture with semi-supervised or unsupervised fashions in spam detection research [65, 20, 66, 78, 75, 77]. Recent work also leverages the deep generation model to generate spam reviews and improve the robustness of spam classifiers via adversarial training [73]. Combining deep learning and graph model, the graph neural network based fraud detection becomes a trending research topic nowadays [41].

Detecting spams from the new coming reviews with limited information is called cold start problem. The second paper I reviewed in this section is the first work to deal with the cold start problem. They leverage the strong learning capability of deep learning to learn a general embedding of the reviews. At first, I will introduce the cold start problem in details. Then I will show their framework followed by the critique review of the deep learning approach in spam detection.

3.1 Cold Start Problem

In data mining and machine learning, the classifiers capture the features of giving data and then predict their labels. The quality of the features of a data item is essential to the performance of the classifier. It reflects how much information the classifier could get from the data.

The cold start problem describes such a challenging situation where some data items are lack of useful features. More specifically, the new data items which have little historical records produce the cold start problem .

Some researchers in recommender system study first posed the cold start problem [56]. The collaborative filtering algorithm which relies on users' preferences cannot deal with the new users. For instance, when we want to recommend movies to a user, the collaborative filtering algorithm first finds the users sharing similar movie preferences to the user, and it then recommends the movies watched by those similar users to the current user. Therefore, the algorithm will work better for users with more movie preference history and cannot have an accurate prediction for the new users with few movie preferences [12].

Similarly, the cold start problem is crucial in online spam detection. According to previous studies [48, 31], most spammers in Yelp are singletons which means they are new users with little historical data. In another perspective, we want to detect the spam review as soon as possible to lessen its influence. The task can be defined as follows:

Definition 3.1 *(Cold Start Problem in Spam Detection) Given a labeled dataset, we want to detect the spam reviews among all the new singleton reviews posted after the training data with the well-trained model on training data.*

It is different from traditional spam detection tasks which do not consider the time factor and divide the training-testing dataset evenly. Such a model has been proved to have worse performance on detecting new spams. Despite the dataset splitting reason, another point is that the traditional handcrafted features (see Table 2) rely too much on the behavior data. Moreover, the handcrafted feature capture limited information of the training data.

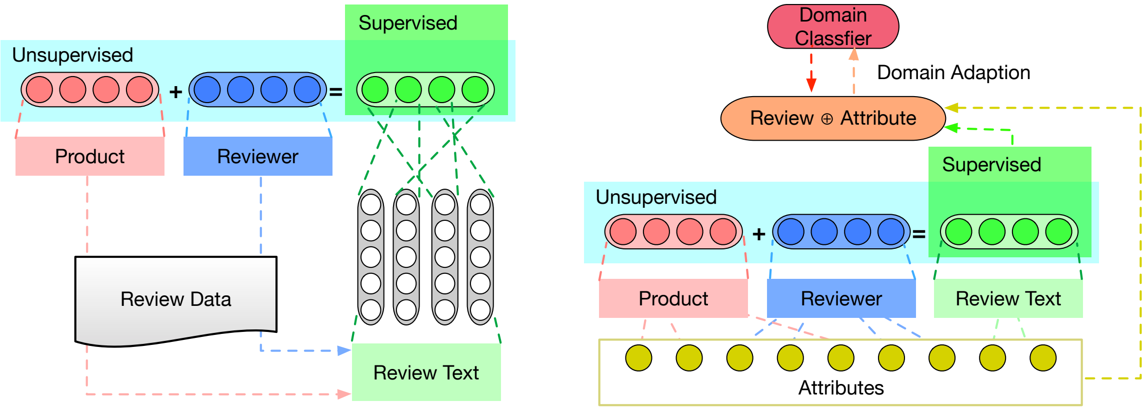
In the era of deep learning, an intuitive way to solve such problems is to employ the deep model for learning the general representation (embedding) of review which could capture more high-dimensional information [68]. In the next subsection, I will introduce the model detail of Wang et al. and why it can tackle the cold start problem in spam detection.

3.2 Text and Behavior Embedding

Based on the experiment result and previous observations [48], Wang et al. states that the traditional linguistic features in not effective in spam detection. The traditional behavior features only works well with sufficient behavioral information. More behavior information will lead to better performance.

To learn more behavior information, the authors adapt the TransE model to the spam detection problem [3]. The TransE model is an unsupervised deep model originally designed to learn the embedding of the relations in knowledge graph. Specifically, the model includes a triplet (h, ℓ, t) where h represents the head entity, t is the tail entity. l is the label vector related to h. It models the relation named *label* between two entities in the knowledge graph. The embedding of $h + l$ should be very close to the embedding of the t. The loss function is defined as follows:

$$\sum_{(h,\ell,t),(h',\ell,t'))\in T_{\text{ batch}}} \nabla \left[\gamma + d(h+\ell,t) - d\left(h'+\ell,t'\right)\right]_{+} \tag{7}$$



(a) Wang et al. Model

(b) You et al. Model

Figure 3: Two deep embedding models which tackle the cold start problem in spam review detection.

The $(h' + \ell, t')$ is called the corrupted triplet. The h' and t' are randomly sampled from the complete original triplet (h, ℓ, t). d is a distance function which is generally set to be Euclidean distance function. $\gamma > 0$ is a hyperparameter. The loss function manifests that the distance of entity embeddings inside the same triplet should be very different from the distance of entity embeddings from different triplets. For instance, in book knowledge graph, h represents the book title, t is the author name, l represents the relation *written by*. If we change h and t to other two different book title and author name, then such *written by* may not be established.

To train the model, we need to get a set of randomly sampled corrupted triplets for each entity. Then we update the embedding of each entity according to the loss function. The training process is totally in an unsupervised fashion which could learn the latent and high dimensional features between data items.

Since the *reviewer-product-review* is a typical triplet in the review platforms where a review has a strong connection with its corresponding reviewer and product. Wang et al. adapts the TransE model to spam review detection problem to learn the embeddings of review texts based on relations above. Notably, the new triplet is modified to (β, α, τ) where β is the product set, α is the reviewer set, and τ is the review set. The input of the deep network is the id vectors of reviewers and products plus the review text embedding learning by Convolutional Neural Network [55]. The embedding training process is similar to TransE. After the unsupervised embedding process, the learned review embeddings encode the review relation among the complete dataset. The embeddings are finally sent to the SVM (Support Vector

Machine) to finish the training and testing part of the cold start spam detection problem.

The traditional handcrafted semantic features have been proved to have limited generation ability in spam detection task. Motivated by [55], Wang et al. employs the Convolutional Neural Network (CNN) model for learning an embedding of review text as the feature vector of a review. The CNN model has been proved to be able to better model the aspect information of reviews than the RNN (Recurrent Neural Network).

To enrich the model, the authors introduce product rating as the input since the rating is a unique factor in review platforms. Because product rating is the clearest indicator of the quality of the product, it becomes a primary attack target being manipulated by spammers. Moreover, the rating score (usually from 1 star to 5 stars) can indicate the sentiment polarity of the corresponding review. Based on this observation, the authors add the rating as the constraint of review during training the review embedding. It is similar to the TransE model which aims to obtain a rating-review embedding where the review is highly related to its corresponding rating and different to other rating scores. A toy example of the architecture of the model is shown in Figure 3.

At the experiment stage, the authors validate the effectiveness of their model in detecting new coming (cold start) spams comparing with few baselines. Two baselines are designed based on an intuition way. The behavior feature of a cold start spammer is obtained from the most similar old spammer. The similarity is calculated by the similarity of edit similarity review texts or the similarity of word2vec transformation of review texts [44]. The experiment states that the deep model could beat such intuitive models. Furthermore, the added rating constraint could help improve the performance of the original deep embedding model.

3.3 Contributions & Drawbacks

Wang et al. firstly explores the cold start problem in review spam detection. They employ the unsupervised deep embedding model for encoding the behavior information other than handcrafted features. They prove that the deep model could work better than the previous feature-based models. However, there are still some empirical and theoretical issues regarding the proposed model.

The TransE model is initially proposed on the knowledge graph. It aims to learn a low-dimension representation of entity and relation. The *head* and *tail* are two entities; the translate vector l is the relation between two entities. In the user-review-product triplet, the *head*, l and *tail* should be product, review and user comparing with the original TransE model. Wang et al. switches the position of user and review in their model without further explanation.

Moreover, there are many approaches that learn low-dimension embeddings of node relations in a graph from the traditional matrix decomposition algorithms to deep learning frameworks. The paper only compares the performance between the proposed model and some naive approaches. That leaves a question of why TransE model is good at encoding global behavior information in review graph.

As for the embedding of the review text, the authors harnesses a recent model using CNN [55]. They give a brief explanation that CNN could better model the aspect information of the review. In fact, other models like Long-Short-Term-Memory (LSTM) and auto-encoder have been proved to learn a better sentence embedding in spam detection task [25, 43]. Besides the issues above, the lack of theoretical guarantee is the fundamental shortcoming of those deep models, the embedding model proposed in this paper leaves many unsolved questions which need further study.

Following the idea of Wang's paper, another paper targeting cold start in spam detection problem is proposed recently (See Figure 3) [77]. Based on the framework of Wang's model, You at al. apply some incremental improvement of the deep embedding model. Firstly, they propose an attribute aiding factor which indicates the common features shared by different products. For instance, all restaurants share the attributes about the price and location. Similarly, users and reviews also have their attributes. Those attribute would help the model encode more latent information in order to solve the cold start problem. The second improvement is adding a domain classifier at the top of the embedding model. It borrows the idea of transfer learning. For two different domains restaurant and hotel, domain adaption makes the review embedding closer to the category the review belongs to and far away from the review embeddings in other domains.

The two strategies above introduced in [77] both leverage the characteristics of the review dataset and aim to extract more information from global data. Following such a clue, more information encoding algorithms can be developed. Though such deep models are lack of explanations, they are able to capture the latent relations of the review data which is essential in the cold start scenario. Traditional explainable features and models could not fit the data better than the proposed deep models. Since spam detection is an empirical area, detectors with stellar performance are preferable by the decision makers. To this end, the deep model has its strong points.

Table 3: Some of the crowdsourcing tasks released on an underground crowdsourcing website.

Task	Reward
Vote a YouTube video	$0.10
Sign up insurance form	$1.50
Upload five photos to a website	$0.40
Follow me on Twitter	$0.12
Sign up online game	$0.20
Review a product on Amazon	$1.00

4 Identifying Crowd Workers via User Similarity

In Section 3, I introduced the weakness of the graph based model in modeling new coming reviews/reviewers. Besides this issue, the graph model is unable to pass information between two cliques independent cliques in the graph. In the deep learning era, node embedding could solve the problem. The paper reviewed in this section targets a scenario which has a limited amount of ground truth data points. They employ both graph model (random walk) and deep model (node embedding) to discover similar data points. At first, I will introduce the crowdsourcing attack scenario. Then, I will show the details of their proposed model which named **TwoFace**. At last, some critiques of the model will be listed.

4.1 Crowdsourcing Attack

Crowdsourcing means finishing specific tasks with a large number of people on the Internet. It leverages the mass intelligence of the billions of people on the Internet. People can release personalized tasks like Photoshop photos, filling surveys on crowdsourcing websites. The crowd workers will be rewarded by money after finishing the task [22]. Usually, it is an economical and efficient way to hire people via crowdsourcing platforms other than purchasing services from specialized firms. For instance, Amazon Mechanical Turk is a typical crowding sourcing marketplace hosted by Amazon [2].

However, crowdsourcing is a double-sided sword where the crowd workers may be hired to work on malicious tasks [63]. Table 3 shows some malicious tasks released on a crowdsourcing website in the black market. Those tasks are legal but would bring fake data to the targeted platforms and further distain the online environment.

In the scope of spam detection problem, the crowdsourced spams have been a big challenge for the online platforms. Since the crowdsourced reviews are written by the regular users instead of bots, it is hard to track and model the crowdsourced reviews from the backend system log [11]. Checking the features listed in Table 1, most of the features could not differ the crowd workers from benign users. That is why many business owners choose crowdsourcing to cheat the platform.

In the perspective of defense, acquiring the ground truth of spam reviews is the primary work to train and test a spam classifier. However, the crowdsourced reviews are usually mixed with a bunch of reviews written by regular users. Finding *real* spams is difficult. Tackling such crowd workers and crowd reviews is a practical problem, and few works have addressed it due to the lack of ground truth [10].

To acquire the ground truth, setting up a honeypot is a general approach borrowed from the security research [34]. We need to set up a honeypot then purchase fake reviews from crowd workers. It usually costs much money to get enough amount of training/testing samples. The paper from Kaghazgaran et al. proposes a cross-domain method which maps the products releasing crowdsourcing tasks with their reviews on Amazon. This approach could acquire ground truth with strong belief without expenses. The authors form a **TwoFace** crowd spam detection framework along with other techniques. More details are introduced in the following part.

4.2 TwoFace Framework

The **TwoFace** is composed of three parts. The first part is to identify suspicious seed users at the review platforms (i.e., Amazon). The second part is to propagate the suspiciousness of users a via random walk over a suspiciousness graph to find more suspicious users. The last part is to discover similar distant suspicious users via mapping users into a low-dimensional embedding space. The last two parts discover different kinds of suspicious users while the distant suspicious users share similar local structures with users discovered in the second part.

4.2.1 Identifying Suspicious Users

Despite setting honeypot, previous feature-based ground truth acquiring methods has the pitfall to missample regular users as the suspicious one. For example, the suspicious users sampled according to their review posting frequency in a short period may include some regular users have similar behavior.

The authors propose a cross-domain mapping method composed of two crawlers. The first crawler crawls the information of products that have published the review written tasks on a crowdsourcing website. Then the second one crawls all reviewers who have reviewed those products on Amazon plus their reviews. It also crawls the reviews to other products written by the suspicious reviewers above. Those review data compose the suspicious review dataset.

The set of non-fraudulent users is composed by the users and reviews sampled from a public Amazon review dataset except the reviewers has reviewed the crowdsourcing targeted products above.

The authors analyze some classic handcrafted features on the two curated datasets above. The result indicates that previous features like burstiness and review length could not identify the fraudulent and non-fraudulent reviews efficiently since the suspicious set may exist legitimate reviews.

4.2.2 Propagating Suspiciousness

To better target the suspicious reviewers, the authors propose a random walk algorithm to propagate the suspiciousness of seed reviewers to similar users [59].

The authors design three approaches to select seed users. The best choice approach is to select the users that have reviewed ten targeted products and all his purchases are unverified. Such seed users are highly suspicious users. The second approach is a random choice which selects seed users from suspicious set randomly. The last approach is a worst choice which selects the users only have few reviews on targeted crowdsourcing products.

Given the seed set, the authors build a co-review graph to run the random walk. The nodes of the co-review graph represent the reviewers; there is an edge between two reviewers if they reviewed the same product. The weight of the edge represents the number of common reviewed products. The suspicious scores of seed users are initialized as 1. After several rounds random walk, the suspicious score is propagated across the complete co-review graph.

The intuition of the local suspicious propagation algorithm is that the users who reviewed the same product tend to have similar behavior. The review buyers also purchase many workers to post multiple reviews for their products. The experiment result on the random walk with different seed selection approach shows that the best choice reaches best precision@k. The performance of random choice is between the other two approaches. The experiment result also reflects the precision@k of random walk algorithm decays very fast when the k increase. It indicates the random walk could only capture the local homophily which only finds a limited amount of suspicious nodes.

4.2.3 Uncovering Distant Users

To discover more suspicious users through the seed nodes, the authors leverage the recent advance in network embedding. The goal is to learn a low-dimensional representation of the node's local structure and find the nodes with similar structures. The intuition is that the nodes played similar roles in the graph may not connect with each other.

The authors employ the *node2vec* model for learning an embedding of each node [18]. The optimization objective function is defined below:

$$\max_{f} \sum_{u \in V} \log p_r(N(u)|f(u)) \tag{8}$$

, where $f(u)$ is the node feature vector (embedding) of node u, $N(u)$ is the neighborhood nodes of node u. The learning object is to maximize the probability of its neighborhood given the embedding a node. Suppose each node has common effects on the feature space. The conditional probability of neighbor node n_i given node u could be represented as a *softmax* function below:

$$\operatorname{pr}\left(n_i|f(u)\right) = \frac{\exp\left(f(u) \cdot f\left(n_i\right)\right)}{\sum_{v \in V} \exp(f(u) \cdot f(v))}. \tag{9}$$

According to the [18], the neighborhood of a node is defined by interpolation of BFS (Breadth-First Sampling) and DFS (Depth-First Sampling) methods. After running the algorithm above, the learned node embeddings could encode the neighbourhood information of each node. Combining the seed node selection and random walk suspiciousness propagation algorithms above, the **TwoFace** model could discover local and distant suspicious users given the seed users. Figure 4 shows the entire framework of the model.

The following experiment part validates the effectiveness of the proposed model with more than 90% recall on testing data. The learned user node embeddings are feed as the input of traditional classifiers like SVM, Naive Bayes and tree-based classifiers. The authors also compare the feature based and dense-block based algorithms which all hit a very poor precision below 50%. It indicates that traditional approaches misclassify many legitimate reviews as suspicious ones. The proposed **TwoFace** beats those algorithms with more than 25% precision.

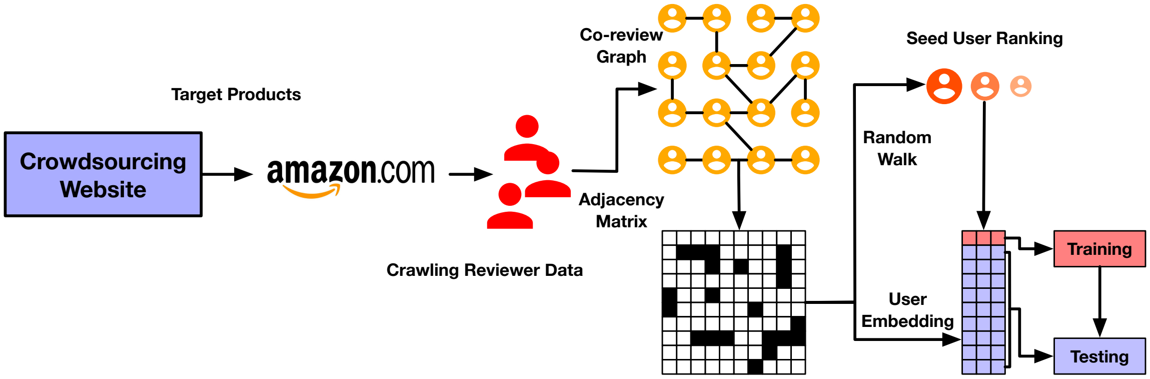


Figure 4: **TwoFace**: a crowdsourcing review spam detection framework.

4.3 Contributions & Drawbacks

To summarize the TwoFace framework, it discovers the similar users in two perspectives, from the neighborhood and the structure-similar users. Traditional graph models like the random walk and belief propagation are unable to pass the information from a node to some distant nodes with no connections. The deep model has its advantage in modeling the node embeddings which could match users with similar neighborhood structure.

Since the paper only focuses on the structural features but without behavior and semantic features. If we include such features as the node prior knowledge aiding the embedding model, the performance is expected to be better. In the classification phase, the authors only investigated some linear and tree-based classifiers. Feeding the embeddings into deep classifiers may improve the performance furthermore. Similarly, the embedding approaches like SDNE and DeepWalk are another two effective node embedding models that learn to encode the node's local structure [62, 53]. Comparing their performance with *node2vec* will unveil more intrinsic relations among the crowd workers.

The **TwoFace** model has captured some crowd spammers with high confidence. Analyzing the behavioral and linguistic features of the spammers would help us better model their characteristics and intentions behind cheating behavior. In the future, adapting the proposed model to other platforms like Yelp and App Store will further validate its generation ability.

5 Discussion and Future Work

5.1 Discussion

The three paper introduced in this survey reflect the trending topics in online spam detection research. As some works focus on optimizing previous algorithm (Paper 1), the deep learning methods are introduced to encode the behavioral and semantic information of spammers and reviews/posts (Paper 2 & 3). As the evolution of online services, more sophistic spammers with advanced cheating tactics appear. Crowdsourcing and cold start are emerging problems which need more effort to handle (Paper 2 & 3).

Comparing the three papers, the first one investigates the social spammers with traditional Bayes based framework which has a strong theory guarantee. The second one applies the deep learning framework to encode the review data into a low dimension representation. The last paper targets a novel problem and leverages both the traditional model and deep model. Together with other papers published on top-tier conferences recently [41, 35, 75, 78, 72], we could find that deep learning becomes a trending tool in online spam detection research.

As I discussed in Section 3.3, the great success of deep learning in recent years brings deep learning into almost every computer science research areas even the biology and social science study. In the perspective of achieving better learning target, deep learning is a primary choice. However, it still has the threat to be attacked by various kinds of noises. The reproducibility of deep learning models is still under doubt now. The lack of theoretical guarantees brings out many shortcomings of deep models. Recently, some researchers begin to address the interpretability of the deep learning model also the node embedding models [33, 9]. Believing the interpretable deep learning will lead the spam detection research into a new era.

Since spam detection is an application-oriented research, the empirical performance of the detectors is the primary criteria to evaluate its performance. The belief of the ground truth is significant to the study on new datasets (like Paper 3). If the classifier were trained and optimized on a dataset with noise, though it achieves reasonable detecting performance, it would have no help to the spam detection in practice and even damage the business. More practically, we need to know the working regime of spammers/fraudsters as much as possible when designing the defense strategies. For instance, through investigation of the black market that trade reviews will help us know the behavior patterns of spammers even acquire some strong ground truth [70]. As the evolution of the spammers, the spam detection framework needs to be kept up-to-date too.

5.2 Future Work

Deep learning is a promising future research venue for spam detection. Some directions like evaluating different learning frameworks on behavior modeling and semantic information encoding has been mentioned at Section 3.3. The graph neural network (GNN) and graph convolutional network (GCN) feed the graph structure (generally adjacency matrix) into the deep neural network. They exploit the learning ability of deep network and adapt it on the graph-structured data. It has been proved very effective under node classification tasks [30]. Its adaption to the recommender systems also demonstrates its strong learning ability [76]. Currently, there is no paper adapting the GCN to the spam detection problem. Moreover, adapting the GCN to a bipartite graph. Along with the advance of deep learning in network embedding, natural language processing and time series analyzing, many domains of **deep spam detection** research remain unexplored.

The adversarial attack on graph-structured data is another hot topic recently [79, 8]. The spam detection is a typical adversarial attacking and defending scenario in the real world without any high-level modeling and assumptions. Though few works have addressed the adversarial attacks under spam detection settings [58, 21], there are still some attacking settings remained unstudied. For instance, we could imitate the attack tactic of the spammers to choose appropriate controlled accounts to post fake reviews with respect to some constraints. Investigating the robustness of the state-of-the-art spam detectors would help us eliminate potential threatens to them. Besides the plans above , most of the research venues under adversarial machine learning are applicable to the spam detection setting.

The cold start problem in spam detection introduced by Paper 2 brings up a new research direction. Excepting two static embedding model above, some dynamic embedding approaches could be developed to deal with the new coming singleton reviews/reviewers.

According to Paper 3, the crowdsourcing spam is still a challenging problem in today's online social network and review platforms. Figuring out their working mechanism and further proposes plans to build up better information sharing schemes would be a promising research direction. It is a multidisciplinary research area with knowledge from computer science and social science.

From a recent news report, some new types of spam reviews appear on Amazon. Few Amazon sellers purchase fake reviews to the products of their rivals on Amazon intentionally [13]. They mislead the spam detector of Amazon to block their rivals. Mining such *professional* spammers (also crowdsourced) is a new challenge problem.

Due to the limited amounts of available benchmark datasets and their limited information, most of the spam detectors only focus on raising the metrics like AUC and recall. In order to deal with the spammers nowadays, a new benchmark is expected to be published. Some new tasks like identifying spammers with different intentions, differing bots from crowd workers and mining the relationship between the text and product metadata could only be studied under a new dataset.

From a broader perspective, online false information like fake news attracts both researchers and mass media in these years. Some spam detection frameworks should be useful for fake news detection problem. Some of the fake news detection algorithms like combining context and social information would shed light on the spam detection problems especially dealing with the emerging spammer types.

References

[1] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion Fraud Detection in Online Reviews by Network Effects. *ICWSM*, 2013.

[2] Amazon. Amazon mechanical turk. https://www.mturk.com. Accessed: 2019-02-25.

[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[4] Q. Cao, X. Yang, J. Yu, and C. Palow. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 477–488. ACM, 2014.

[5] X. Carreras and L. Marquez. Boosting trees for anti-spam email filtering. *arXiv preprint cs/0109015*, 2001.

[6] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[7] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada. Survey of review spam detection using machine learning techniques. *J. Big Data*, 2(1):1, 2015.

[8] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song. Adversarial Attack on Graph Structured Data. *ICML*, 2018.

[9] A. Dalmia, M. Gupta, et al. Towards interpretation of node embeddings. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 945–952. International World Wide Web Conferences Steering Committee, 2018.

[10] V. Dave, S. Guha, and Y. Zhang. Measuring and fingerprinting click-spam in ad networks. In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, pages 175–186. ACM, 2012.

[11] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, pages 26–30, 2012.

[12] Y. Dou, H. Yang, and X. Deng. A survey of collaborative filtering algorithms for social recommender systems. In *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 40–46. IEEE, 2016.

[13] J. Dzieza. Dirty dealing in the $175 billion amazon marketplace. https://www.theverge.com/2018/12/19/18140799/amazon-marketplace-scams-seller-court-appeal-reinstatement. Accessed: 2019-02-25.

[14] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting Burstiness in Reviews for Review Spammer Detection. *ICWSM*, 2013.

[15] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics, 2012.

[16] W. Gatterbauer, S. Günnemann, D. Koutra, and C. Faloutsos. Linearized and single-pass belief propagation. *Proceedings of the VLDB Endowment*, 8(5):581–592, 2015.

[17] K. R. Gee. Using latent semantic indexing to filter spam. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 460–464. ACM, 2003.

[18] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[19] Z. Gyongyi, H. Garciamolina, and J. Pedersen. Combating Web Spam with TrustRank. In *Proceedings 2004 VLDB Conference*, pages 576–587. Elsevier, 2004.

[20] N. Hernandez, M. Rahman, R. Recabarren, and B. Carbunar. Fraud de-anonymization for fun and profit. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 115–130. ACM, 2018.

[21] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. FRAUDAR - Bounding Graph Fraud in the Face of Camouflage. *KDD*, pages 895–904, 2016.

[22] J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[23] X. Hu, J. Tang, H. Gao, and H. Liu. Social spammer detection with sentiment information. In *2014 IEEE International Conference on Data Mining*, pages 180–189. IEEE, 2014.

[24] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in microblogging. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[25] G. Jain, M. Sharma, and B. Agarwal. Optimizing semantic lstm for spam detection. *International Journal of Information Technology*, pages 1–12, 2018.

[26] N. Jindal and B. Liu. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM, 2007.

[27] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230. ACM, 2008.

[28] P. Kaghazgaran, J. Caverlee, and A. Squicciarini. Combating crowdsourced review manipulators: A neighborhood-based approach. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 306–314. ACM, 2018.

[29] S. KC and A. Mukherjee. On the temporal dynamics of opinion spamming: Case studies on yelp. In *Proceedings of the 25th International Conference on World Wide Web*, pages 369–379. International World Wide Web Conferences Steering Committee, 2016.

[30] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[31] D. Kumar, Y. Shaalan, X. Zhang, and J. Chan. Identifying Singleton Spammers via Spammer Group Detection. In *Advances in Knowledge Discovery and Data Mining*, pages 656–667. Springer, Cham, 2018.

[32] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. S. Subrahmanian. FairJudge - Trustworthy User Prediction in Rating Platforms. *CoRR*, cs.SI, 2017.

[33] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

[34] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[35] C. Li, S. Wang, L. He, S. Y. Philip, Y. Liang, and Z. Li. Ssdmv: Semi-supervised deep social spammer detection by multi-view data fusion. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 247–256. IEEE, 2018.

[36] F. H. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[37] 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 *ninth international AAAI conference on web and social Media*, 2015.

[38] J. Li, M. Ott, C. Cardie, and E. Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1566–1576, 2014.

[39] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948. ACM, 2010.

[40] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[41] Z. Liu, C. Chen, X. Yang, J. Zhou, X. Li, and L. Song. Heterogeneous graph neural networks for malicious account detection. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2077–2085. ACM, 2018.

[42] M. Luca and G. Zervas. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science*, 62(12):3412–3427, Jan. 2016.

[43] G. Mi, Y. Gao, and Y. Tan. Apply stacked auto-encoder to spam detection. In *International Conference in Swarm Intelligence*, pages 3–15. Springer, 2015.

[44] 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*, pages 3111–3119, 2013.

[45] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[46] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh. Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 632–640. ACM, 2013.

[47] A. Mukherjee, B. Liu, and N. Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200. ACM, 2012.

[48] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance. What Yelp Fake Review Filter Might Be Doing? *ICWSM*, 2013.

[49] K. P. Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[50] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

[51] M. Newman. *Networks*. Oxford university press, 2018.

[52] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

[53] 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*, pages 701–710. ACM, 2014.

[54] S. Rayana and L. Akoglu. *Collective Opinion Spam Detection: Bridging Review Networks and Metadata*. Bridging Review Networks and Metadata. ACM, New York, New York, USA, Aug. 2015.

[55] Y. Ren and Y. Zhang. Deceptive opinion spam detection using neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 140–150, 2016.

[56] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260. ACM, 2002.

[57] J. Scott. Social network analysis. *Sociology*, 22(1):109–127, 1988.

[58] N. Shah, A. Beutel, B. Gallagher, and C. Faloutsos. Spotting Suspicious Link Behavior with fBox - An Adversarial Perspective. *ICDM*, pages 959–964, 2014.

[59] F. Spitzer. *Principles of random walk*, volume 34. Springer Science & Business Media, 2013.

[60] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 613–622. IEEE, 2006.

[61] B. Wang, N. Z. Gong, and H. Fu. Gang: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 465–474. IEEE, 2017.

[62] D. Wang, P. Cui, and W. Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2016.

[63] G. Wang, T. Wang, H. Zheng, and B. Y. Zhao. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 239–254, 2014.

[64] G. Wang, S. Xie, B. Liu, and S. Y. Philip. Review graph based online store review spammer detection. In *2011 IEEE 11th International Conference on Data Mining*, pages 1242–1247. IEEE, 2011.

[65] X. Wang, K. Liu, and J. Zhao. Detecting deceptive review spam via attention-based neural networks. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 866–876. Springer, 2017.

[66] X. Wang, K. Liu, and J. Zhao. Handling Cold-Start Problem in Review Spam Detection by Jointly Embedding Texts and Behaviors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 366–376, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics.

[67] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ, 1996.

[68] J. Weston, F. Ratle, H. Mobahi, and R. Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012.

[69] S. Xie, G. Wang, S. Lin, and P. S. Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 823–831. ACM, 2012.

[70] Z. Xie and S. Zhu. Appwatcher: Unveiling the underground market of trading mobile app reviews. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, page 10. ACM, 2015.

[71] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 71–80. ACM, 2012.

[72] Z. Yang, Y. Zhang, and Y. Dai. Defending against Social Network Sybils with Interaction Graph Embedding. In *2018 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9. IEEE, 2018.

[73] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1143–1158. ACM, 2017.

[74] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

[75] C. M. Yilmaz and A. O. Durahim. Spr2ep: A semi-supervised spam review detection framework. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 306–313. IEEE, 2018.

[76] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983. ACM, 2018.

[77] Z. You, T. Qian, and B. Liu. An attribute enhanced domain adaptive model for cold-start spam review detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1884–1895, 2018.

[78] S. Yuan, X. Wu, J. Li, and A. Lu. Spectrum-based deep neural networks for fraud detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2419–2422. ACM, 2017.

[79] D. Zügner, A. Akbarnejad, and S. Günnemann. Adversarial Attacks on Neural Networks for Graph Data. *KDD*, pages 2847–2856, 2018.