SYBILFUSE: Combining Local Attributes with Global Structure to Perform Robust Sybil Detection

Peng Gao*, Binghui Wang†, Neil Zhenqiang Gong†, Sanjeev R. Kulkarni*, Kurt Thomas‡ and Prateek Mittal*
*Princeton University †Iowa State University ‡Google
*{pgao, kulkarni, pmittal}@princeton.edu †{binghuiw, neilgong}@iastate.edu ‡kurtthomas@google.com

***Abstract*—Sybil attacks are becoming increasingly widespread and pose a significant threat to online social systems; a single adversary can inject multiple colluding identities in the system to compromise security and privacy. Recent works have leveraged social network-based trust relationships to defend against Sybil attacks. However, existing defenses are based on oversimplified assumptions about network structure, which do not necessarily hold in real-world social networks. Recognizing these limitations, we propose SYBILFUSE, a defense-in-depth framework for Sybil detection when the oversimplified assumptions are relaxed. SYBILFUSE adopts a collective classification approach by first training local classifiers to compute local trust scores for nodes and edges, and then propagating the local scores through the global network structure via weighted random walk and loopy belief propagation mechanisms. We evaluate our framework on both synthetic and real-world network topologies, including a large-scale, labeled Twitter network comprising 20M nodes and 265M edges, and demonstrate that SYBILFUSE outperforms state-of-the-art approaches significantly. In particular, SYBILFUSE achieves 98% of Sybil coverage among top-ranked nodes.**

I. INTRODUCTION

Our systems today are vulnerable to Sybil attacks, in which an attacker injects multiple fake accounts into the system [1]. Recently, the increasing popularity of online social networks has made them attractive targets for Sybil attacks. It is estimated that tens of millions of Sybil accounts exist in popular social networks such as Facebook and Twitter [2], [3]. The attacker can leverage Sybil accounts to disrupt democratic election and influence financial market via spreading fake news [4], [5], as well as compromise system security and privacy via propagating social malware, disseminating scams, and learning users' private data [2], [3], [6]–[10].

Limitations in existing Sybil defenses: An important thread of research proposes to mitigate Sybil attacks using social network-based trust relationships. The key insight is that in a social graph where edges represent strong trust relationships, it is hard for the attacker to set up links to benign users. As a result, the number of edges between benign users and Sybil identities (called *attack edges*) is limited. Approaches such as SybilGuard [11], SybilLimit [12], SybilInfer [13], SybilRank [14], CIA [15], SybilBelief [16], and SybilSCAR [17] rely on such *strong trust assumption* and separate the benign and Sybil regions by identifying communities [18]. Íntegro [19] extends SybilRank by considering victim prediction (i.e., benign accounts that connect to Sybils).

While these approaches have pioneered the use of social network structure for Sybil defenses, the actual deployment of these ideas in real-world networks remains controversial. Real-world social networks do not necessarily have strong trusts. Yang et al. showed that RenRen, the largest social network in China, does not follow this assumption [20]. Previous work [8], [21] also showed that Sybil nodes could befriend benign users on Facebook at a large scale. Ghosh et al. [22] showed that on Twitter, a link farming phenomenon is widespread, in which certain benign accounts blindly follow back accounts who follow them. On such weak trust networks, the number of attack edges can be larger than what is typically considered in previous works, making it challenging to separate the benign region from the Sybil region.

SYBILFUSE outperforms state-of-the-art: Complex attack strategies in real-world social networks make it hard for methods that are based on single source of information to succeed. Recognizing the limitations in existing approaches, we propose SYBILFUSE, a defense-in-depth framework that leverages heterogeneous sources of information to perform robust Sybil detection. Different from existing approaches which assume strong trust networks [11]–[17], [23] or assume strong victim prediction [19], SYBILFUSE relaxes these limitations by adopting a collective classification scheme. Given social network data as input, SYBILFUSE first leverages *local attributes* to train local classifiers. Local node classifier predicts a trust score for each node, which indicates the probability of that node to be benign. Local edge classifier predicts a trust score for each edge, which indicates the probability of that edge to be a non-attack edge. These local trust scores are then combined with the *global structure* through weighted trust score propagation. Existing approaches do not leverage rich local information and treat edges equally, thus do not work well when the number of attack edges exceeds their assumption. In contrast, SYBILFUSE captures local account information in *node trust scores*, and propagates these scores through the global structure. During the score propagation, SYBILFUSE leverages *edge trust scores* to enforce unequal weights, so that attack edges will have reduced impacts on the propagation. After the propagation completes, final trust scores of accounts are used for Sybil classification and ranking.

Evaluation: We conduct comprehensive evaluations of SYBILFUSE against state-of-the-art approaches: (1) we extensively evaluate the robustness of SYBILFUSE under different network settings and observe that SYBILFUSE is robust to errors in local classifiers (Section IV-B); (2) we evaluate SYBILFUSE in a real-world, labeled Twitter network in which the assumptions that existing approaches require are not satisfied (e.g., large number of attack edges (49.5 per Sybil) and

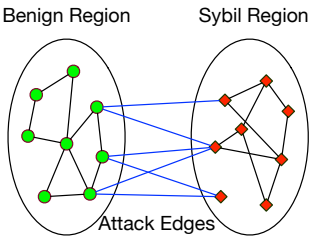


Fig. 1: Sybil attack scenario

victims (75.4% of benign nodes)). We observe that SYBILFUSE outperforms state-of-the-art approaches significantly in Sybil ranking and achieves 98% of Sybil coverage among top-ranked nodes (Section V-D); (3) we evaluate SYBILFUSE in a real-world, large-scale, labeled Twitter network comprising over 20 million nodes and 256 million edges. We observe that SYBILFUSE outperforms state-of-the-art approaches significantly in both AUC and Sybil ranking (Section VI-D).

II. BACKGROUND AND RELATED WORK

A. Sybil Attack Scenario

Consider a network topology $G = (V, E)$, comprising a set V of nodes (i.e., user accounts) and a set E of edges (i.e., friendship relationships). To model real-world social networks, graph G can be either directed or undirected. A directed graph models the follower-following topology (e.g., Twitter), in which $(u, v) \in E$ denotes that u follows v. An undirected graph models the mutual relationship topology (e.g., Facebook), in which $(u, v) \in E$ is equivalent to $(v, u) \in E$.

Fig. 1 shows the *Sybil attack* scenario, in which every node $v \in V$ is associated with a label that indicates its identity to be *benign* or *Sybil*. We denote the subnetwork containing all benign nodes to be the benign region, and denote the subnetwork containing all Sybil nodes to be the Sybil region. The edges that connect the two regions are called *attack edges*. Following the established convention in the literature, we do not impose any constraints on the size or the shape of the Sybil region. The attacker can create an unlimited number of Sybil nodes and set up edges between them arbitrarily. The main goal of Sybil defense research is to design a mechanism to detect as many Sybil nodes as possible, while minimizing the number of benign nodes that are misdetected.

B. Limitations in State-of-the-art Sybil Defenses

Local attribute-based approaches: Local attribute-based approaches seek to detect Sybil accounts by analyzing local account attributes (e.g., posts, status updates, connections). These approaches span a large category of schemes, including blacklisting [24], whitelisting [25], URL filtering [7], as well as various machine learning methods, such as Bayesian reasoning, Support Vector Machines, and clustering [26]–[28]. A fundamental limitation in these approaches is that Sybils can easily evade the detection by mimicking the local behaviors of benign users via manipulating their profiles and connections.

Global structure-based approaches: Recognizing the limitations in local attribute-based approaches, global structure-based approaches seek to exploit the global graph-theoretic differences between the benign region and the Sybil region. Most global structure-based approaches leverage either *random walks* or *loopy belief propagation*. For instance, random walk based approaches include SybilGuard [11], SybilLimit [12], SybilInfer [13], SybilRank [14], CIA [15], and SybilWalk [23]. Íntegro [19] extends SybilRank by incorporating victim prediction (i.e., benign accounts that connect to Sybils) in random walks. Loopy belief propagation based approaches include SybilBelief [16] and GANG [29]. Fu et al. [30] extended SybilBelief via considering user carefulness at establishing social relationships. Wang et al. [17] proposed a local rule based framework to unify random walk based and loopy belief propagation based approaches. Based on the framework, they further proposed SybilSCAR, which can be viewed as an optimized version of SybilBelief.

Research has shown that [11]–[17], [23] assume a *strong trust network*, where the number of attack edges is limited [18]. [11], [12] further assume that the benign region is fast mixing [31], which presumes the existence of a well-connected, giant community structure of benign users. However, these assumptions oversimplify the network structure and may not hold well on certain real-world social networks. First, real-world social networks do not necessarily have strong trusts. Yang et al. showed that RenRen, the largest social network in China, does not follow this assumption [20]. Previous work [8], [21] also showed that Sybil nodes could befriend benign users on Facebook at a large scale. Ghosh et al. [22] showed that on Twitter, a link farming phenomenon is widespread, in which certain benign accounts blindly follow back accounts who follow them. On such weak trust networks, the number of attack edges can be larger than what is typically considered in previous works, making it challenging to separate the benign region from the Sybil region. Second, benign users tend to form multiple small communities [18] driven by different purposes (e.g., geographical location, education, and career), which introduces a longer mixing time than the theoretical anticipated value [31]. Third, although Íntegro does not directly require a small number of attack edges, it relies on strong assumptions that the number of victims is small and that the victims are accurately predicted, which may not hold on real-world social networks like Twitter (Section V-A).

III. THE SYBILFUSE FRAMEWORK

Complex attack strategies in real-world social networks make it hard for methods that are based on single source of information to succeed. Recognizing the limitations in existing approaches, we propose SYBILFUSE, a defense-in-depth framework that leverages heterogeneous sources of information to perform robust Sybil detection. Different from existing approaches which only leverage local attribute information [7], [24]–[28], only leverage global structure information [11]–[17], [23], or assume strong victim prediction [19], SYBILFUSE combines local attributes with global structure by adopting a collective classification scheme.

A. Framework Overview

Fig. 2 shows the SYBILFUSE framework. Given social network data as input, SYBILFUSE first samples training

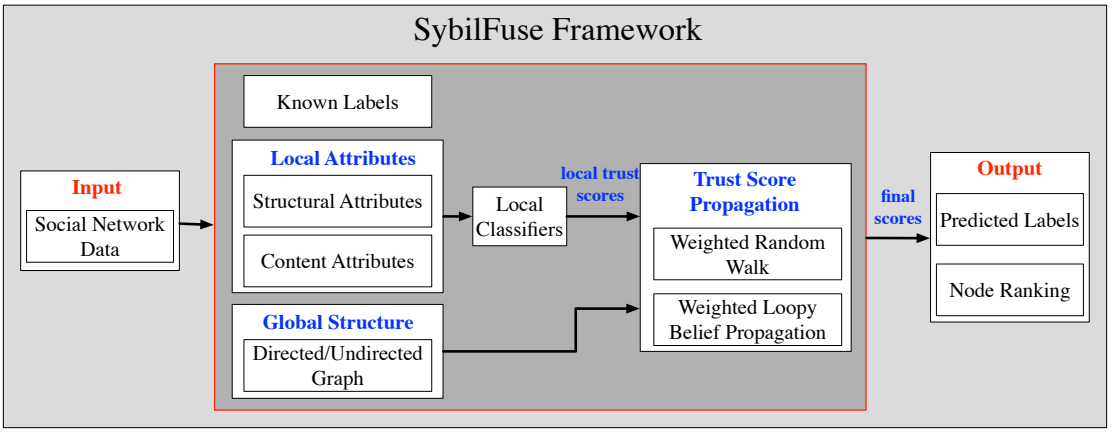


Fig. 2: The SYBILFUSE framework

data and leverages discriminative local attributes (structural or content) to train local classifiers. Local node classifier predicts a trust score for each node, which indicates the probability of that node to be benign. Local edge classifier predicts a trust score for each edge, which indicates the probability of that edge to be a non-attack edge. These local scores are then combined with the global structure through weighted trust score propagation. Existing approaches [11]–[16] do not leverage rich local information (i.e., these approaches propagate manually-set scores from a limited set of labeled seeds) and treat edges equally, thus do not work well when the number of attack edges exceeds their assumptions. In contrast, SYBILFUSE captures local account information in node trust scores, and propagates these scores through the global structure. SYBILFUSE furthermore leverages edge trust scores to enforce unequal weights for the propagation, so that attack edges will have reduced impacts. Final node scores after propagation are used for Sybil account prediction and ranking.

B. Local Trust Score Computation

Given a social graph $G = (V, E)$, we denote the trust score of node $v \in V$ by S_v, which quantifies the probability that v is benign. We denote the trust score of edge $(u, v) \in E$ by $S_{u,v}$, which quantifies the probability that node u and node v take the same label (i.e., homophily strength). To compute S_v, we leverage local node attributes (e.g., degree, local clustering coefficient, profile) and train a machine learning classifier that outputs probabilistic estimates (e.g., SVM, logistic regression). To compute $S_{u,v}$, we similarly build an edge classifier. In addition, we can measure the similarity between node u and v using various metrics (e.g., Cosine, Jaccard, Adamic-Adar [32]). The insight is that in social networks where homophily exists, connected benign nodes tend to be similar and Sybils might be different from their targeted benign nodes. As a result, attack edges tend to have low similarity scores. Note that in practice, we restrict S_v and $S_{u,v}$ to be within range $[0.1, 0.9]$ through normalization, since scoring zero would invalidate our weighted score propagation.

C. Weighted Trust Score Propagation

Score propagation is done through either weighted random walk or weighted loopy belief propagation.

Weighted random walk: Let $S^{(i)}(v)$ denote the score of node v after i-th power iteration. First, we set the initial score of every node (excluding the training data) to be the predicted local node trust score from the local node classifier:

$$S^{(0)}(v) = S_v \tag{1}$$

For nodes that belong to the training data, we set score 0.9 to the benign nodes and score 0.1 to the Sybil nodes.

Next, we set the weight of every edge to be the predicted local edge trust score from the local edge classifier, and start weighted random walk using the following update equation:

$$S^{(i)}(v) = \sum_{(u,v)\in E} S^{(i-1)}(u) \frac{S_{u,v}}{\sum_{(u,w)\in E} S_{u,w}} \tag{2}$$

Hence, u will distribute more of its round $i-1$ trust to a close friend v (i.e., $S_{u,v}$ is high) rather than an unfamiliar connection w. In this way, attack edges that have low trust scores will have reduced impacts on the propagation. After $d = O(\log n)$ steps of power iteration (i.e., early termination), where n is the number of nodes, we obtain the final score:

$$S_v^F = S^{(d)}(v) \tag{3}$$

Weighted loopy belief propagation: We model the graph G as a pairwise Markov Random Field (MRF) [33]. For each node v, we associate it with a binary random variable $X_v \in \{1, -1\}$ that represents its label ($X_v = 1$ for benign and $X_v = -1$ for Sybil). To quantify the probability of a collection of nodes jointly taking a set of labels, we introduce *node potential function* $\psi_v(X_v)$ for node v and *edge potential function* $\psi_{u,v}(X_u, X_v)$ for edge (u, v), and initialize them using our predicted trust scores from local classifiers:

$$\psi_v(X_v) = \begin{cases} S_v & \text{if } X_v = 1 \\ 1 - S_v & \text{if } X_v = -1 \end{cases} \tag{4}$$

$$\psi_{u,v}(X_u, X_v) = \begin{cases} S_{u,v} & \text{if } X_u X_v = 1 \\ 1 - S_{u,v} & \text{if } X_u X_v = -1 \end{cases} \tag{5}$$

Given a pairwise MRF (G, Ψ), we propagate local trust scores through the global structure using Loopy Belief Propagation

(LBP) [34] with the following message update function:

$$m_{u\to v}(X_v) = \sum_{X_u} \left(\psi_u(X_u)\psi_{u,v}(X_u, X_v) \prod_{s\in Neighbors(u)\setminus v} m_{s\to u}(X_s) \right) \quad (6)$$

where $m_{u\to v}(X_v)$ is initially set to 1 for all edges $u \to v$. Note that edge potentials enforce unequal contribution of edges to the propagation. After $d = 5 \sim 10$ rounds of iteration (we validated this empirically), we obtain the belief score of node v with label $X_v = x_v$:

$$bel_v(X_v = x_v) \propto \psi_v(X_v = x_v) \prod_{u\in Neighbors(v)} m_{u\to v}(X_v = x_v) \quad (7)$$

We then normalize $bel_v(X_v = 1)$ to obtain the final score:

$$S_v^F = \frac{bel_v(X_v = 1)}{bel_v(X_v = 1) + bel_v(X_v = -1)} \quad (8)$$

Computational complexity: The complexity of weighted random walk and weighted LBP is both $O(md)$, where m is the number of edges and d is the number of iterations (recall that $d = O(\log n)$ for weighted random walk and $d = 5 \sim 10$ for weighted LBP). For sparse networks, $O(md) = O(nd)$. Thus, the total computational cost is the addition of execution time of local classifier training and prediction, and that of random walk/LBP. In practice, local classifier can be trained offline and efficient implementations like *LIBSVM* [35] exist. Besides, random walk and LBP are easily parallelizable, which further speeds up the execution.

D. Sybil Account Prediction and Ranking

Sybil account prediction: For a node v, we predict its label L_v by comparing its final score S_v^F with a threshold value:

$$L_v = sign(S_v^F - threshold) \quad (9)$$

where $L_v = 1$ indicates a benign label and $L_v = -1$ indicates a Sybil label. The threshold value can be obtained by conducting cross-validation on the training data.

Sybil ranking: We can also surface Sybil accounts by ranking all nodes in an ascending order of their final scores. Sybils with low scores will be ranked upfront. Security analysts can then go through the list and surface Sybil accounts manually.

IV. Robustness Evaluation

We evaluate the robustness of SybilFuse against different network structures and different levels of local classifier errors.

A. Evaluation Setup

Network generation: We adopt a similar experimental setting as [16], in which we use Preferential Attachment [36] model to generate both the benign region and Sybil region, and randomly add attack edges between the two regions. In the basic setup, the network consists of 1,000 benign nodes and 500 Sybil nodes. The average degree of the two regions is 10. The number of attack edges is 1,000.

We simulate local trust scores from local classifiers by taking random samples in $[0.1, 0.9]$ with certain false positive rates and false negative rates (0.5 as threshold). We study the performance of SybilFuse under three factors: (1) different levels of errors in local classifiers, (2) different number of attack edges, and (3) different number of Sybil nodes. When we study one factor, we fix the other factors to be the same as the basic setup and only vary the studied one.

Evaluation metrics: We measure the accuracy and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) [37] of SybilFuse with weighted random walk (*SF-RW*) and SybilFuse with weighted LBP (*SF-LBP*). Given the ranking of scores of all nodes from the smallest to the largest, AUC measures the the probability that a randomly selected Sybil nodes ranks higher than a randomly selected benign node. A higher AUC indicates better Sybil ranking performance. For *SF-LBP*, we use a natural threshold 0.5 to compute accuracy. For *SF-RW*, we do not compute accuracy since the final scores of random walk-based approaches (e.g., [14], [15]) are typically very low due to the nature of score distribution and finding a proper threshold is hard.

B. Evaluation Results

Fig. 3 shows the accuracy and AUC of SybilFuse with local node trust scores. We set the edge trust scores to be the default value 0.9 to model homophily. For (a), we vary the FPR and FNR of local node classifier from 0 to 0.4 (i.e., from perfect local classifier to better than random guess). For (b) and (c), we set FPR and FNR to be 0.3 (i.e., noisy local classifier) and vary the number of attack edges & the number of Sybil nodes. Fig. 4 shows the accuracy and AUC of SybilFuse with local edge trust scores. We randomly sample 1 benign node and 1 Sybil node as trusted seeds, and set their node trust scores to be 0.9 and 0.1, correspondingly. For the rest of nodes, we set their scores to be the default value 0.5.

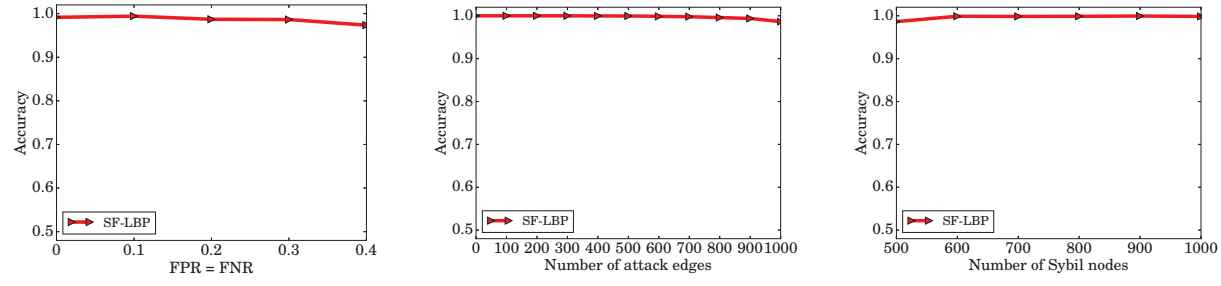
We observe that: (1) when $FPR = FNR \leq 0.3$, *SF-RW* and *SF-LBP* achieve > 0.98 accuracy and AUC under all evaluated network settings given local node trust scores, and achieve > 0.92 accuracy and AUC given local edge trust scores; (2) the performance of both *SF-RW* and *SF-LBP* improves under more accurate local classifiers, less number of attack edges, and more number of Sybil nodes; (3) *SF-LBP* performs better than *SF-RW* in all settings.

V. Labeled Twitter Evaluation

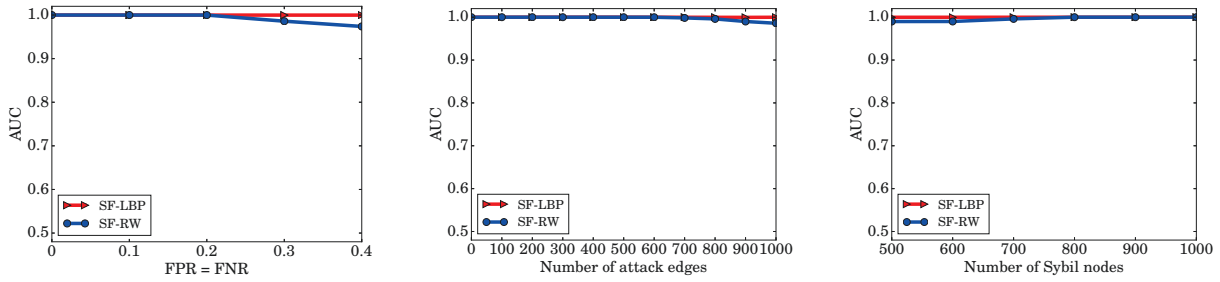
We evaluate SybilFuse against state-of-the-art approaches in a real-world, labeled Twitter network. We demonstrate that by combining local attributes with global structure, SybilFuse outperforms state-of-the-art significantly.

A. Network Preprocessing and Measurement

The original directed network was obtained from [15]. Since it is easy for the attacker to manipulate one-way directed edges, we followed the established convention [14]–[17], [19] and preprocessed it to an undirected network by retaining an undirected edge if both directions exist.

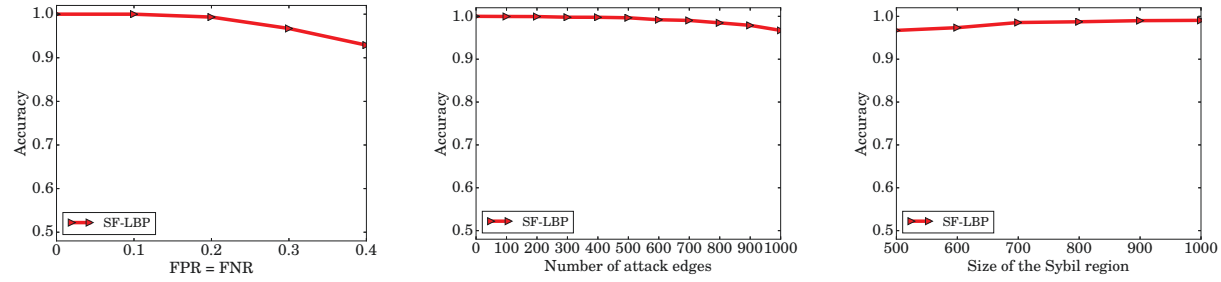


(a) Vary FPR=FNR (b) Vary the number of attack edges (c) Vary the number of Sybil nodes

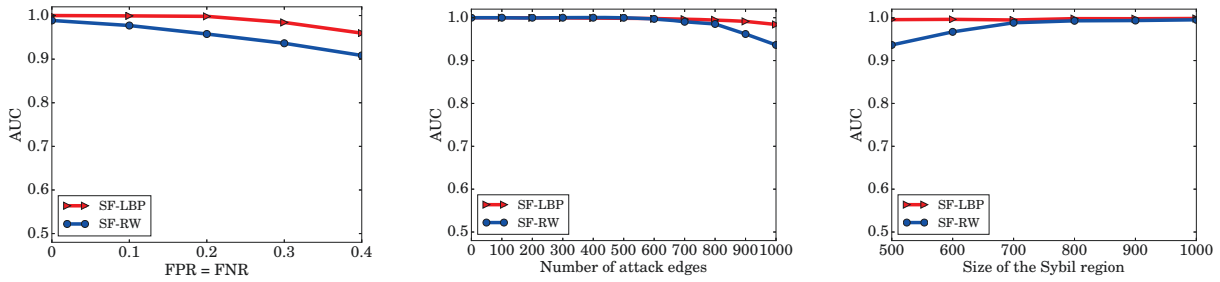


(d) Vary FPR=FNR (e) Vary the number of attack edges (f) Vary the number of Sybil nodes

Fig. 3: Accuracy and AUC of SYBILFUSE with weighted random walk (*SF-RW*) and weighted loopy belief propagation (*SF-LBP*) given local **node** trust scores. For (a), we vary FPR=FNR of local node classifiers. For (b) and (c), we fix FPR=FNR=0.3.



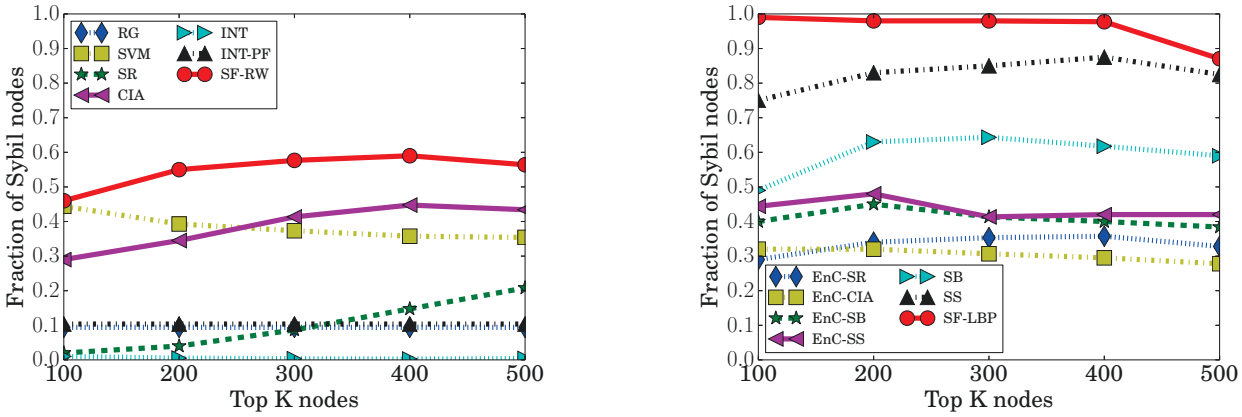
(a) Vary FPR=FNR (b) Vary the number of attack edges (c) Vary the number of Sybil nodes



(d) Vary FPR=FNR (e) Vary the number of attack edges (f) Vary the number of Sybil nodes

Fig. 4: Accuracy and AUC of SYBILFUSE with weighted random walk (*SF-RW*) and weighted loopy belief propagation (*SF-LBP*) given local **edge** trust scores. For (a), we vary FPR=FNR of local edge classifiers. For (b) and (c), we fix FPR=FNR=0.3.

After preprocessing, the network consists of 8,167 nodes and 54,146 edges, with verified 7,358 benign nodes and 809 Sybil nodes. We observe that: (1) Sybil nodes emit a large number of attack edges (40,010), with 49.5 attack edges on average per Sybil; (2) 53.4% of Sybils are isolated (i.e., they only connect to benign nodes). These isolated Sybils emit 37.0% of attack edges. In addition, 41.2% of Sybils form a largest connected component (emitting 42.8% of attack edges), and 5.4% of Sybils form several small connected groups (emitting 20.2% of attack edges); (3) the number of victims is large (5,546; 75.4% of benign nodes). Thus, the benign region and the Sybil region can hardly be viewed as separate communities, and the assumptions that existing approaches require (Section II-B) are not satisfied. As we will show in Section V-D, existing approaches [14]–[17], [19] have limited performance on this network.



(a) Random walk-based approaches

(b) LBP-based approaches and ensemble methods

Fig. 5: Fraction of Sybils among top K nodes of all evaluated approaches. We observe that *SF-LBP* has the best ranking performance among all approaches and achieves $> 98\%$ Sybil ranking up to top 400 nodes.

B. Local Trust Score Computation

Feature extraction: We extract three discriminative node features from the original directed network, and map these features to the corresponding nodes in the undirected network. Since extracting discriminative edge features from this dataset is difficult, we set local edge scores to be 0.9 by default to model homophily.

- Incoming requests accepted ratio: $Req_{in}(v) = \frac{|In(v) \cap Out(v)|}{|In(v)|}$, where $In(v)$ ($Out(v)$) represents the set of all incoming (outgoing) edges of v. The insight is that Sybils are more likely to accept incoming requests than benign users in order to quickly propagate spam, resulting in a higher Req_{in}. Since the dataset only contains structural information, we use in-degree & out-degree to model the ratio.
- Outgoing requests accepted ratio: $Req_{out}(v) = \frac{|In(v) \cap Out(v)|}{|Out(v)|}$. The insight is that Sybils are less reliable than benign users and hence their outgoing friend requests are less likely to be accepted, resulting in a lower Req_{out}.
- Local clustering coefficient: $CC(v) = \frac{|\{(i,j): i,j \in Nei(v), (i,j) \in E\}|}{|Nei(v)|(|Nei(v)|-1)}$), where $Nei(v)$ represents the set of neighbors of v. The insight is that benign users often have well-connected social cliques, and users in such cliques are often friends, resulting in to a high CC.

Training a SVM classifier: We randomly sample 50 benign nodes and 50 Sybil nodes as the training set, and train a SVM classifier with RBF kernel using *LIBSVM* [35]. We then output probabilistic estimates as local node scores. Note that we extract these features and adopt the SVM classifier in particular for this Twitter network. The system administrator is free to explore other features and classifiers under the overall SYBILFUSE framework.

C. Evaluated Approaches

For SYBILFUSE, we propagate local scores through weighted random walk (*SF-RW*) and weighted loopy belief propagation (*SF-LBP*). We also evaluate the following existing approaches: (1) local node classification: *SVM*; (2) global structure-based approaches: SybilRank (*SR*), CIA (*CIA*), SybilBelief (*SB*), SybilSCAR (*SS*), Íntegro (*INT*), Íntegro with perfect victim prediction (*INT-PF*). We use the same training data as propagation seeds. For *INT*, we further sample 50 victims and 50 non-victims to learn a victim predictor based on Random Forest algorithm [19]. For *INT-PF*, we model a perfect victim predictor by setting the probability score of each victim to be 1 and the score of each non-victim to be 0; (3) ensemble approaches: *EnC-SR*, *EnC-CIA*, *EnC-SB*, *EnC-SS*. We combine the scores from SVM classifier with the scores from structure-based approaches in a standard voting scheme; (4) random guess: *RG*.

D. Evaluation Results

Sybil ranking performance: Following the established convention [14], [16], [17], [19], we evaluate the *ranking performance* of the compared approaches by ranking all nodes in an ascending order of their final scores. A more effective approach will rank more Sybil nodes upfront.

Fig. 5 shows the fraction of Sybils among top K nodes. We observe that: (1) *SF-RW* has the best performance among all random walk-based approaches. Specifically, the average improvement of *SF-RW* w.r.t. *SR*, *CIA*, and *INT* is 44.8%, 16.2%, and 54.3%, respectively (the improvement of A w.r.t. B is computed as $A-B$, where A and B represent the fraction of Sybils); (2) *SF-LBP* has the best performance among all evaluated approaches. Specifically, the average improvement of *SF-LBP* w.r.t. *RG*, *SVM*, *SR*, *CIA*, *INT*, *SB*, and *SS* is 85.7%, 57.5%, 85.9%, 57.4%, 95.5%, 36.5%, and 13.4%, respectively. Besides, *SF-LBP* achieves $> 98\%$ Sybil ranking up to top 400 nodes; (3) *SF-RW* and *SF-LBP* outperform *INT*, *INT-PF*, and ensemble approaches significantly. Surprisingly, even under perfect victim prediction, the performance of *INT-PF* is just slightly better than *RG*. In fact, Íntegro assigns low weights to all edges originated from victims according to the formula $w(v_i, v_j) = \min\{1, \beta \cdot (1 - \max\{p(v_i), p(v_j)\})\}$ [19]. Hence, the propagation of scores from trusted seeds to victims and from victims to other benign nodes is inhibited. When the number of victims is large (75.4% in our Twitter network), even under perfect victim prediction, a large number of benign

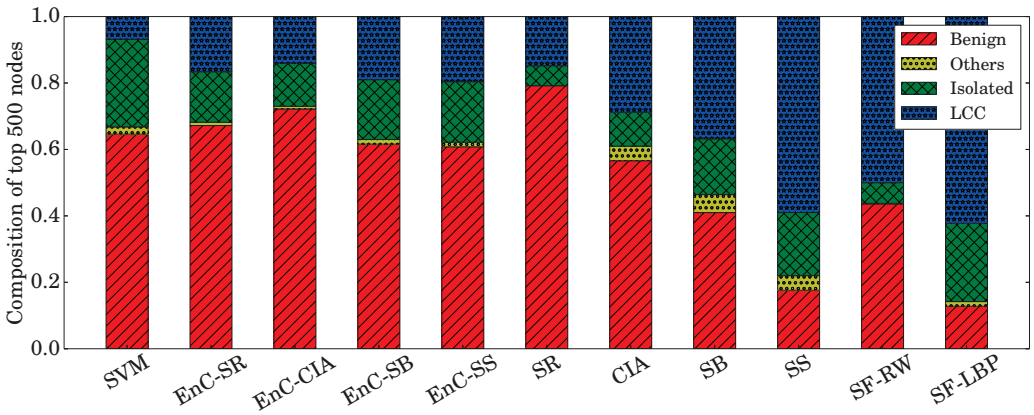


Fig. 6: Decomposition of top 500 nodes. We observe that *SF-LBP* has the best performance in ranking both isolated Sybils and Sybils in the largest connected component (LCC).

nodes will receive low final scores, including the victims and the benign nodes that are separated from propagation seeds by victims. This problem becomes worse in practice since it is impossible to have perfect victim prediction. For *INT*, the victim predictor has 18.3% precision and 19.6% recall, and we can see that *INT* has nearly 0 Sybil ranking capability. In short, *INT* has very limited performance when the number of victims is large or the victim predictor is inaccurate.

Measuring the composition of top-ranked Sybils: We study the power of each approach in ranking different types of Sybils: isolated Sybils (*Isolated*), Sybils in the largest connected component (*LCC*), and Sybils in small, connected groups (*Others*). Fig. 6 shows the decomposition of top 500 nodes. We omit *INT* and *INT-PF* due to their very limited performance. We observe that: (1) local node classification approach (*SVM*) is more powerful in ranking isolated Sybils (74.6% of all Sybils ranked by *SVM*); (2) global structure-based approaches are more powerful in ranking Sybils in the LCC (71.2%, 66.4%, 62.4%, and 71.6% of all Sybils ranked by *SR*, *CIA*, *SB*, and *SS*, respectively); (3) *SF-LBP* has the best performance in ranking both isolated Sybils and Sybils in the LCC.

Summary: In summary, SYBILFUSE methods (*SF-RW*, *SF-LBP*) significantly outperform existing approaches in terms of Sybil ranking, and the most effective approach is to propagate local trust scores from local classifiers through weighted loopy belief propagation (*SF-LBP*).

VI. LARGE-SCALE LABELED TWITTER EVALUATION

We further evaluate SYBILFUSE against state-of-the-art approaches in a real-world, large-scale, labeled Twitter network comprising over 20 million nodes and 256 million edges.

A. Ground Truth Collection

We obtained a snapshot of the Twitter follower network from [38]. Similar to Section V-A, we preprocessed the original directed network to an undirected network by only retaining bi-directional edges. After preprocessing, the network consists of 21,297,772 nodes and 265,025,545 edges. To collect ground truth, we re-crawled every account using Twitter's API, which told us the account status. We found that 145,156 (0.7%) nodes were suspended by Twitter, 1,911,482 (9.0%) nodes were deleted, and the rest were still active. We treated the suspended accounts as Sybil nodes and the active accounts as benign nodes. For the deleted accounts, since we were not sure whether they were deleted by users or by Twitter, we did not include them in the training and evaluation.

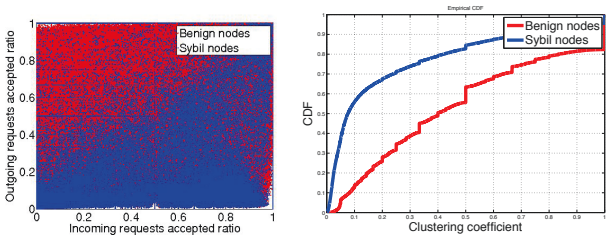
B. Network Structure Measurement

We adopt *modularity* [39], ranging from -0.5 to 1, to quantify if a network partition can be viewed as two communities. Clauset et al. [40] concluded that modularity > 0.3 indicates a significant community structure. However, we find that the partition consisting of the benign and Sybil regions only has modularity 0.0042. Thus, the two regions cannot be viewed as separate communities. Next, we show two reasons:

(1) Half of Sybils are isolated: In total, we find 77,917 connected components in the Sybil region. Among them, 50% of Sybils are isolated, 45% of Sybils form a largest connected component, and the rest of Sybils form small connected components whose sizes are less than 20. Furthermore, we find that the modularity of the partition consisting of the benign region and the largest Sybil connected component is still only 0.0046, which means that even this largest connected component cannot be viewed as a community.

(2) Large number of attack edges: In total, there are 18,414,469 attack edges, which means each Sybil node successfully attacks around 127 benign nodes on average. Furthermore, we find that the number of neighbors of both benign and Sybil nodes follow long-tail distributions, which is also widely observed in other OSNs such as LiveJournal [41] and Google+ [42]. Thus, we speculate that Sybils are imitating the benign region to evade automatic detection. Besides, 90% of attack edges concentrate on 3% of benign nodes. Thus, we speculate that such nodes are celebrities that tend to follow back any account that follows them.

Note that these structural properties of Sybil nodes in our Twitter network match those in another large-scale Twitter network [22] and those in the RenRen network [20], which indicates the representativeness of our observations. As a



(a) Scatter plot (b) CDF

Fig. 7: Analysis of local node features

TABLE I: AUC in the large-scale Twitter network

SR	*CIA*	*INT*	*INT-PF*	*SB*	*SS*	*SF-RW*	*SF-LBP*
0.57	0.80	0.48	0.54	0.74	0.74	0.81	0.85

result, the assumptions that existing approaches require are not satisfied, leading to their limited performance (Section VI-D)

C. Local Trust Score Computation

We extract the same set of features as Section V-B. Fig. 7 shows the scatter plot of Req_{out} vs. Req_{in} and the CDF plot of CC. As expected, Sybil nodes tend to have a higher Req_{in}, a lower Req_{out}, and a lower CC. We randomly sample 3,000 benign nodes and 3,000 Sybil nodes, and train a SVM classifier using *LIBSVM* [35]. We then output probabilistic estimates as local node scores. Since extracting discriminative edge features from this dataset is difficult, we set local edge scores to be 0.9 by default to model homophily.

D. Evaluation Results

Following the notations in Section V-C and the established convention [14], [16], [17], [19], we evaluate the *ranking performance* of SYBILFUSE approaches (*SF-RW*, *SF-LBP*) against state-of-the-art Sybil defense approaches *SR*, *CIA*, *INT*, *INT-PF*, *SB*, and *SS* (we follow the same notation as Sec. V-C). We measure two metrics: (1) AUC; (2) fraction of Sybils among top K ranked nodes.

Table I shows the AUC of the evaluated approaches and Fig. 8 shows the fraction of Sybils among top K nodes. We observe that: (1) *SF-RW* has the best AUC among all random walk-based approaches. Specifically, the improvement of *SF-RW* w.r.t. *SR*, *CIA*, *INT*, and *INT-PF* is 0.24, 0.01, 0.33, and 0.27, respectively (measured as $A - B$); (2) *SF-LBP* has the best AUC among all evaluated approaches. Specifically, the average improvement of *SF-LBP* w.r.t. *SR*, *CIA*, *INT*, *INT-PF*, *SB*, and *SS* is 0.28, 0.05, 0.37, 0.31, 0.11, and 0.11, respectively; (3) *SF-LBP* outperforms state-of-the-art Sybil defenses significantly in terms of Sybil ranking. Specifically, among the top 1K nodes, the Sybil ranking improvement of *SF-LBP* w.r.t. *SR*, *CIA*, *INT*, *INT-PF*, *SB*, and *SS* is 54.8%, 46.5%, 55.1%, 55.1%, 55.1%, 53.9%, and 55.1%, respectively.

E. Limitations in Twitter's Sybil Detection Policy

Recall that we obtained our ground truth based on whether the account was active or suspended by Twitter. Thus, it is possible that some accounts are actually Sybil but evade Twitter's detection policy. To test this, we manually examined the top 100 accounts, of which 71 were suspended and 29

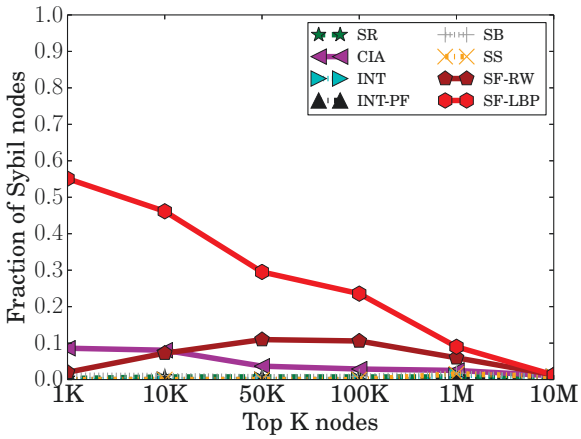


Fig. 8: Fraction of Sybils in the large-scale Twitter network

were active. Among the 29 active accounts, we found: (1) 3 benign accounts, with a long timeline and diverse tweets; (2) 24 Sybil accounts, with common low-quality profile pictures and spam tweets; (3) 2 suspicious accounts with protected information. *These 24 (24/29=82.8%) active Sybil accounts evaded Twitter's detection policy but were successfully captured by* SYBILFUSE*'s ranking mechanism.*

VII. DISCUSSION

Resilience against social churn: According to [43], existing Sybil defenses such as SybilInfer [13] and SybilRank [14] are vulnerable to the churn in social graphs. To attack these systems, [43] considers churn in attack edges by gradually moving the attack edges closer to the trusted seeds, so that the seed-based score propagation will eventually fail. The success of this temporal attack requires two assumptions: (1) the attacker knows the location of the trusted seeds in the social graph; (2) the location of the trusted seeds will not change drastically for a certain period of time, so that the attacker has enough time to perform the attack. However, we propose that the system administrator can frequently change the trusted seeds and rerun the detection program. Furthermore, SYBILFUSE does not start propagation from the trusted seeds. Instead, SYBILFUSE computes local scores for all nodes and propagates these scores. As a result, the attacker does not have a direction for moving the attack edges gradually.

Strategic adversaries: Strategic adversaries who are aware of the features used in SYBILFUSE's local classifiers may attempt to evade detection by changing attack strategies. SYBILFUSE's multi-layer protection restricts such possibilities by combining heterogeneous information sources. Specifically, if the attacker aims to evade the detection of local node classifier by mimicking the local features of benign users (i.e., lower Req_{in}, higher Req_{out}, and higher CC), he needs to establish more connections between Sybil identities. Consequently, Sybils will be much more densely connected, and the trust score propagation module (Fig. 2) will be more effective to detect them. Machine learning in these adversarial scenarios remains a challenging problem for the research community. In general, we advocate periodically retraining the local classifiers to respond to the dynamics of attack behaviors.

Broader applicability: Our framework that combines local attributes with global structure has broad applicabilities for network security. For example, the area of botnet detection can benefit from similar techniques that combine host-level information with network structure-based information.

VIII. Conclusion and Future Work

In this work, we proposed SybilFuse, a defense-in-depth framework that novelly combines local attributes with global structure to perform robust Sybil detection. SybilFuse overcomes the limitations in existing approaches by leveraging local attributes to train local classifiers and propagating the local classifier scores through global structure via weighted score propagation. Experimental results in synthetic topologies and real-world topologies demonstrate that SybilFuse outperforms state-of-the-art approaches significantly. In future, we plan to generalize SybilFuse to detect Sybils in *directed* social graphs [29] and apply SybilFuse to other security applications such as botnet detection and reputation systems.

Acknowledgement

This work was supported in part by DARPA RADICS under contract FA-8750-16-C-0054, and the National Science Foundation under grants CNS-1553437, CNS-1409415, and CNS-1750198.

References

[1] J. R. Douceur, "The Sybil attack," in *IPTPS*, 2002.

[2] Malicious/fake accounts in Facebook. (2012, August). [Online]. Available: https://goo.gl/BC8zjt

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *CEAS*, 2010.

[4] Hacking Election. (2016, May). [Online]. Available: http://goo.gl/G8o9x0

[5] Hacking Financial Market. (2016, May). [Online]. Available: http://goo.gl/4AkWyt

[6] K. Thomas, C. Grier, V. Paxson, and D. Song, "Suspended accounts in retrospect: An analysis of twitter spam," in *IMC*, 2011.

[7] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time url spam filtering service," in *IEEE S & P*, 2011.

[8] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: Automated identity theft attacks on social networks," in *WWW*, 2009.

[9] P. L. Fong, "Preventing Sybil attacks by privilege attenuation: A design principle for social network systems," in *IEEE S & P*, 2011.

[10] K. Thomas, C. Grier, and V. Paxson, "Adapting social spam infrastructure for political censorship," in *LEET*, 2012.

[11] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman., "SybilGuard: Defending against Sybil attacks via social networks," in *SIGCOMM*, 2006.

[12] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "SybilLimit: A near-optimal social network defense against Sybil attacks," in *IEEE S & P*, 2008.

[13] G. Danezis and P. Mittal, "SybilInfer: Detecting Sybil nodes using social networks," in *NDSS*, 2009.

[14] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *NSDI*, 2012.

[15] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammer's social networks for fun and profit," in *WWW*, 2012.

[16] N. Z. Gong, M. Frank, and P. Mittal, "SybilBelief: A semi-supervised learning approach for structure-based sybil detection," *IEEE TIFS*, 2014.

[17] B. Wang, L. Zhang, and N. Z. Gong, "SybilSCAR: Sybil detection in online social networks via local rule based propagation," in *INFOCOM*, 2017.

[18] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An analysis of social network-based Sybil defenses," in *SIGCOMM*, 2010.

[19] Y. Boshmaf, D. Logothetis, G. Siganos, J. Leria, J. Lorenzo, M. Ripeanu, and K. Beznosov, "Integro: Leveraging victim prediction for robust fake account detection in osns," in *NDSS*, 2014.

[20] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network Sybils in the wild," in *IMC*, 2011.

[21] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in *ACSAC*, 2011.

[22] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in *WWW*, 2012.

[23] J. Jia, B. Wang, and N. Z. Gong, "Random walk based fake account detection in online social networks," in *IEEE DSN*, 2017.

[24] A. Ramachandran, N. Feamster, and S. Vempala, "Filtering spam with behavioral blacklisting," in *CCS*, 2007.

[25] G. S. S. Yardi, D. Romero and D. Boyd, "Detecting spam in a Twitter network," *First Monday*, vol. 15(1), 2010.

[26] A. H. Wang, "Don't follow me - spam detection in twitter," in *SECRYPT 2010*, 2010.

[27] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. Choudhary, "Towards online spam filtering in social networks," in *NDSS*, 2012.

[28] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *ACSAC*, 2010.

[29] B. Wang, N. Z. Gong, and H. Fu, "GANG: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs," in *IEEE ICDM*, 2017.

[30] H. Fu, X. Xie, Y. Rui, N. Z. Gong, G. Sun, and E. Chen, "Robust spammer detection in microblogs: Leveraging user carefulness," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2017.

[31] A. Mohaisen, A. Yun, and Y. Kim, "Measuring the mixing time of social graphs," in *IMC*, 2010.

[32] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Am. Soc. Inf. Sci. Technol.*, 2007.

[33] G. Cross and A. Jain, "Markov random field texture models," *IEEE Trans. PAMI*, vol. 5, 1983.

[34] K. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief-propagation for approximate inference: An empirical study," in *UAI*, 1999.

[35] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, 2011.

[36] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, 1999.

[37] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Mach. Learn.*, 2001.

[38] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW*, 2010.

[39] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, 2004.

[40] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, 2004.

[41] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *IMC*, 2007.

[42] N. Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song, "Evolution of social-attribute networks: Measurements, modeling, and implications using google+," in *IMC*, 2012.

[43] C. Liu, P. Gao, M. Wright, and P. Mittal, "Exploiting temporal dynamics in sybil defenses," in *ACM CCS*, 2015.