JMLR: Workshop and Conference Proceedings 45:189–204, 2015 ACML 2015

Improving Sybil Detection via Graph Pruning and Regularization Techniques

Huanhuan Zhang[1] HHZHANG@NTU.EDU.SG
Jie Zhang[1] ZHANGJ@NTU.EDU.SG
Carol Fung[2] CFUNG@VCU.EDU
Chang Xu[1] XUCH0007@NTU.EDU.SG
[1] *School of Computer Engineering, Nanyang Technological University, Singapore 639798*
[2] *Department of Computer Science, Virginia Commonwealth University, Richmond, VA*

Editor: Geoffrey Holmes and Tie-Yan Liu

Abstract

Due to their open and anonymous nature, online social networks are particularly vulnerable to Sybil attacks. In recent years, there has been a rising interest in leveraging social network topological structures to combat Sybil attacks. Unfortunately, due to their strong dependency on unrealistic assumptions, existing graph-based Sybil defense mechanisms suffer from high false detection rates. In this paper, we focus on enhancing those mechanisms by considering additional graph structural information underlying social networks. Our solutions are based on our novel understanding and interpretation of Sybil detection as the problem of partially labeled classification. Specifically, we first propose an effective graph pruning technique to enhance the robustness of existing Sybil defense mechanisms against target attacks, by utilizing the local structural similarity between neighboring nodes in a social network. Second, we design a domain-specific graph regularization method to further improve the performance of those mechanisms by exploiting the relational property of the social network. Experimental results on four popular online social network datasets demonstrate that our proposed techniques can significantly improve the detection accuracy over the original Sybil defense mechanisms.

Keywords: Sybil Attack, Social Networks, Sybil Defenses, Transitive Trust

1. Introduction

Over the last few years, social networking sites have become an indispensable part of people's lives. Social network service providers benefit from collecting a huge amount of user information. Due to their open and anonymous nature, online social networks are particularly susceptible to spamming manipulations. One of the prevalent forms is *Sybil attacks*, where an attacker creates a large number of fake identities, known as *Sybils*, to unfairly increase their power or suppress other honest users within a target community. Traditional mechanisms which rely on central trusted identities are not sufficient to defend against Sybil attacks, since it is difficult to convince users to publicize their privacy information, and electing a trusted identity is challenging in extremely large social networks.

In recent years, there is substantial growth of interest in leveraging social network structure Yu et al. (2008); Tran et al. (2011); Cao et al. (2012) to combat Sybil attacks. However, the existing structure-based Sybil defense mechanisms under-perform and suffer from high false detection rates due to their unrealistic assumptions and limited usage of topological information. Specifically, they explicitly assume that the honest region is *fast mixing*, where random walks from non-Sybil nodes can quickly reach a stationary distribution after $O(log(n))$ steps, compared to Sybil nodes. However, studies show that in real-world social networks, mixing time is much larger than that anticipated in Sybil defense mechanisms Mohaisen et al. (2010). This fact determines that structure-based solutions, which depend on the fast mixing property, cannot produce desirable accuracy. Furthermore, most of these mechanisms are vulnerable to *target attacks* Cao et al. (2012), in which an adversary has prior knowledge about the locations of *honest seeds*, and launches Sybil attacks by substantially compromising these honest seeds as well as their nearby nodes. As a result, some dummy nodes can appear to be honest due to their direct connections with these honest seeds, rendering ineffective the structure-based defense mechanisms.

Our goal in this paper is to provide effective strategies to overcome the above limitations and improve the performance of existing structure-based Sybil defense mechanisms. First, we provide a novel perspective of interpreting Sybil defense as the problem of partially labeled classification. We demonstrate that existing structure-based Sybil defense mechanisms can be seen as the processes of explicitly propagating *honest labels* across a social network, so as to partition the entire network into non-Sybil and Sybil regions, i.e., each node is declared as either *Honest* or *Sybil*.

Based on this partially labeled classification framework, we propose two effective methods– *graph pruning* and *graph regularization*, by exploring additional structural information embedded within social graphs, to improve the detection accuracy of current Sybil defense mechanisms. The *graph pruning* technique is introduced to handle the target attack problem, by exploiting local structure similarity between neighboring nodes. This strategy is performed on original social networks before Sybil detection to diminish the influence of target attacks where attack edges are established intentionally around the *honest seeds*. In addition, studies Taskar et al. (2001); Neville and Jensen (2007) show that many real-world networks such as social networks and web graphs possess relational property, implying that linked or neighboring nodes are likely to have the same class labels. This characteristic has been widely applied in many fields for classification or prediction tasks Sen et al. (2008). In this paper, a domain-specific graph regularization method is proposed based on the relational property to enhance the detection accuracy over existing Sybil defense mechanisms. To our best knowledge, none of the aforementioned structure-based Sybil defense mechanisms has taken graph regularization into account. Evaluation results demonstrate that our methods can significantly enhance the performance of existing Sybil defense mechanisms.

In all, our main contributions are three-folds: 1) we provide a novel interpretation of Sybil defense as the problem of partially labeled classification; 2) a graph pruning technique is introduced to enhance the robustness of existing Sybil defense mechanisms against target attacks; 3) we also design a specialized manifold regularizer by exploiting the relational property in social networks to further improve the accuracy of Sybil defense mechanisms. Experimental results on four popular social network datasets indicate that our regularizer can even decrease the false negative rate to nearly zero.

2. Related Work

2.1. Social Network-based Sybil Defense

Many topological solutions to counter Sybil attacks have been proposed in recent years. SybilGuard Yu et al. (2006) and SybilLimit Yu et al. (2008) are the first decentralized protocols to leverage social network structures to detect Sybil nodes. In SybilGuard, each node performs random route of length $w = \Theta(\sqrt{n}logn)$ and a suspect is accepted if its random route intersects with the verifier's. When the number of attack edges is bounded to $g = O(\sqrt{n}/logn)$, SybilGuard accepts at most $\Theta(\sqrt{n}logn)$ Sybil nodes per attack edge with a high probability. SybilLimit improves upon SybilGuard's bound by using multiple walks, which allows it to accept at most $O(logn)$ Sybil nodes per attack edge. However, both of them make unrealistic assumptions about the number of honest nodes in a network and suffer from high false negatives. SybilInfer Danezis and Mittal (2009) adopts the Bayesian inference technique that assigns to each node its probability of being Sybil, but suffers from high computational cost. Gatekeeper Tran et al. (2011) is another decentralized Sybil defense scheme that improves over the guarantees provided by SybilLimit. It heavily relies on the *expander-like* property, which is a stronger assumption than fast mixing and has not been validated in real social networks. Similarly, Mohaisen et al. (2010) point out that mixing time is much larger than that anticipated in Sybil defense schemes, implying that social networks are generally not fast mixing. Such a finding renders ineffective all defense schemes based on the mixing property. In order to eliminate negative impact of mixing time, Mohaisen et al. (2011) model different levels of trust in social networks and design modified random walks upon SybilLimit to improve its performance. More recent work such as Integro Boshmaf et al. (2015) integrates behavior analyzes with graph structure for sybil detection and also used modified random walk mechanism to compute rankings. The insight is that by incorporating trust information, the algorithmic property–*quotient cut* can be greatly shrunk and becomes clearer, which is an obvious sign for Sybil detection. Therefore, even the honest region does not strictly satisfy the fast mixing property, random walks originated from non-Sybil nodes will land on non-Sybil nodes with higher probabilities compared to Sybil nodes. However, such a trust-driven model may affect many honest nodes, leading to high false positives.

Viswanath et al. (2010) explain the rationale behind structure-based Sybil defense mechanisms from the perspective of *graph partitioning*. They state that existing community detection algorithms can be utilized to detect Sybils. However, such algorithms are vulnerable to target attacks and rarely provide provable guarantees. Cao et al. (2012) develop a Sybil ranking mechanism (called SybilRank) which distinguishes Sybil nodes from non-Sybil nodes based on their relative trustworthiness. SybilRank is validated on real social graph and is proven to be effective and efficient against Sybil attacks. Since it depends on *honest seeds* to propagate trust across the network, this approach also suffers from target attacks. We enhance the robustness of the existing topological Sybil defense mechanisms against target attacks by exploiting additional structural features in social networks.

2.2. Partially Labeled Classification

Partially labeled classification is a well-studied topic in the machine learning field. Traditional classifiers solely utilize labeled data for training. However, it is not easy to obtain sufficient labeled instances due to costly human efforts and being time consuming. Semi-supervised classification techniques well address this problem by utilizing the similarity among unlabeled data. Together with the labeled data, they form a better classifier. Zhu's recent survey Zhu (2006) summaries commonly used approaches for semi-supervised learning with various ways to operate the classification task. In general, semi-supervised classification algorithms fall into one of the three categories: self-training, feature extraction approaches, and graph-based regularization.

The key to graph-based regularization in a semi-supervised setting is the *label smoothness* or cluster assumption Seeger (2001), which states that data points in a high dense region (cluster) tend to have the same labels. Many proposals have been developed for graph-based semi-supervised classification Bousquet et al. (2003)Grandvalet et al. (2004) Zhou et al. (2003). They differ in particular choices of their objective functions and regularizers which manifest the underlying structure among unlabeled data. For example, Blum and Chawla (2001) formulate semi-supervised learning as a graph mincut (called *st*-cut) problem. Their objective is to seek for a minimum set of edges whose removal cuts the connections between sources and sinks. Thus, the nodes connecting to the sources are labeled as positive, while the other nodes are labeled as negative. Zhu et al. (2003) apply the Gaussian random fields and harmonic function methods to construct the objective function for semi-supervised classification. This representation allows a simple closed-from solution for the node marginal probabilities and has many interesting properties. Normalized Laplacian is used by Zhou et al. (2003) to build the regularizer for label smoothing among the entire graph, which exploits both the local and global consistency in the graph.

3. Understanding Sybil Defense

To identify Sybil nodes, existing structure-based Sybil defense mechanisms partition the entire network into honest and Sybil regions. The basic rationale behind is the following two core assumptions: 1) strong trust relationships exist among nodes, making it difficult for Sybil nodes to establish many social connections with non-Sybil nodes, even if they can easily recruit a large number of Sybil nodes and build an arbitrary topology network among them. As a result, Sybil region connects to the main network via a small number of attack edges. 2) honest region is *fast mixing*, where random walks from a benign node can quickly reach a stationary distribution after $O(log(n))$ steps, compared to those from Sybil nodes. However, these Sybil defense mechanisms are less effective than expected since real-world social networks do not conform to the above assumptions (see Section 2.1).

Viswanath et al. Viswanath et al. (2010) provide an interesting common insight for current Sybil defense schemes that explains them as *graph partitioning* algorithms. They demonstrate that despite their considerable differences, these topological schemes work by identifying a local community that surrounds the trusted nodes. And then, they point out that existing state-of-the-art community detection algorithms can be utilized to solve the problem of Sybil attacks. However, the *community detection* framework is confined to consider only limited topological features despite the *local connectivity* among social

networks and *fast mixing* property of the honest region. Thereby, it does not provide a clear guidance on addressing the mixing time sensitivity problem, which incurs high false positives. Furthermore, different choices of metrics, which are utilized to measure the quality of community detection, will lead to different Sybil detection results. No work has provided a reasonable metric to achieve better detection results Cai and Jermaine (2012). As also pointed out in Section 2.1, community detection algorithms are likely vulnerable to target attacks. Thus, in this paper, we take a different perspective to understand and reformulate the problem of Sybil defense.

Basically, the existing topological methods for detecting Sybil attacks assume that an attacker infiltrates the systems by creating a large amount of 'bad' nodes and then building a network of arbitrary topology among them Cai and Jermaine (2012). Due to the inherent trust assumption that prevents an attacker from establishing too many social links with benign nodes, there is an abnormal characteristic in social networks, where a large number of nodes connect to the main network via few edges. Hence, the presence of such a *small cut* is probably a good indicator of Sybil attacks. The objective of Sybil detection is to seek for a set of minimal edges (small cut) whose removal partitions the entire graph into non-Sybil and Sybil regions. Such a process is particularly similar to the principle in Blum and Chawla (2001), which copes with the partially labeled classification problem from the perspective of graph mincut (see Section 2.2).

Intuitively, in the Sybil defense setting, there are two types of classes, i.e. *non-Sybil* and *Sybil*. To find Sybil nodes, the various mechanisms attempt to mark those unlabeled nodes by propagating *honest labels* among the network starting from some known *honest seeds*. Each node in the network is labeled as either non-Sybil or Sybil. Since initially only partial nodes are labeled as honest, the classification process proceeds by searching for specific characteristics (e.g, mixing time) that can discriminate honest nodes from Sybils. Hence, Sybil defense can be reformulated as a *partially labeled classification* problem as follows:

Given:

- An affinity graph $G = (V, E)$, where nodes in V denote identities and edges in E reflect the trust relationship between users in the social network.
- Binary class labels $Y = \{+1, -1\}$ defined on V, where $+1$ denotes honest label and -1 denotes Sybil label.
- A set of nodes H_0 with honest labels (called *honest seeds*). We have, $f(v_i) = +1, \forall v_i \in H_0$, where f is the labeling function.

Output: The mapping/labeling function $f : V \to Y$ from nodes to class labels.

As discussed in Section 2.2, the problem of *partially labeled classification* has been studied in the semi-supervised learning field Zhu (2006). Given a small portion of data points associated with class labels (called *training set*), *transductive inference* is applied to infer those unlabeled data by incorporating the intrinsic manifold structure. However, existing semi-supervised classification algorithms are not obviously applicable to detect Sybil nodes since no Sybil label information is given to *supervise* the Sybil classification problem.

Under the partially labeled classification framework, our focus in this paper is to provide effective strategies to enhance the robustness of current topological anti-Sybil designs and improve their detection accuracy. Inspired by the work of Mohaisen et al. Mohaisen et al.

(2011), which leverages trust information to improve the performance of SybilLimit, we investigate topological features embedded within social graphs to strengthen current Sybil defense mechanisms. In Section 4, we discuss how to exploit the *local structural similarity* to address the target attack problem. In Section 5, a *graph regularization* technique, based on the *relational property*, is developed to smooth the detection results of existing Sybil defense approaches.

4. Graph Pruning

Most topological Sybil defense mechanisms rely on a basic assumption that one or more honest nodes are known in advance. These nodes (also known as *honest seeds*) are utilized for identity verification and partitioning the entire network into non-Sybil and Sybil regions. However, once *honest seeds* are compromised by a set of disruptive nodes, these defense systems would under-perform Cao et al. (2012). Indeed, such attacks may be easily accomplished by an adversary through establishing as many social connections as possible with high-degree honest nodes. This type of attack is called target seeding attack or *target attack*. To the best of our knowledge, no solution has been proposed in the literature to solve this problem.

In this paper, we present a graph pruning technique that effectively reduces the impact from target attacks by enforcing that the number of attack edges around *honest seeds* is few. This avoids the situation where a large number of Sybil nodes are accepted due to their close connection to *honest seeds*, hence evade Sybil detection. This strategy leverages local structural similarity underlying social networks. Intuitively, corresponding to the fast mixing and inherent trust relationship assumptions, we speculate that the similarity between benign nodes and *honest seeds* are much higher compared to the similarity between benign nodes and Sybil nodes. Thus, by eliminating edges with low-similarity values (i.e., $w_{ij} \leq T_s$), where w_{ij} is the similarity of nodes i and j and T_s is the threshold to determine whether one edge should be trimmed, the number of attack edges is expected to be reduced. Different structural similarity metrics Mohaisen et al. (2011) in social networks have been proposed for measuring the strength of social links and predicting future interactions, such as number of common friends, cosine similarity, Jaccard similarity, etc. We choose the proximity metric of *number of common friends* to measure the local structure similarity in the graph pruning process since 1) it is simple and intuitive; 2) it well reflects the trust level between two users; 3) it is difficult for an adversary to simultaneously trick an honest node and its neighbors into trusting it.

In our method, pruning is firstly performed in local regions around *honest seeds*. Its goal is to prevent honest seeds and their nearby nodes in the network from being tricked by a set of disruptive nodes. On the other hand, pruning should not have much impact on honest users. This is partially determined by the size of the pruned region, which is denoted by T_p, the maximum diameter between *honest seeds* and the pruned nodes. The pruned network shall thus satisfy the following two requirements: 1) it should minimize attack edges nearby *honest seeds*; 2) it shall also retain as many honest nodes as possible because some benign nodes may be disconnected from the entire graph during the pruning process. We can balance the trade-off by adjusting two parameters–pruning diameter T_p and similarity threshold T_s. Specific parameter choices will be examined in our experiments.

Algorithm Graph Pruning

Require: G, graph $G = (V, E)$; H_0, set of honest seeds; T_s, similarity threshold; T_p, pruned diameter
Ensure: G_{prune}, pruned graph

```
    // Defining and Initializing Notations
 1: Initially all edges in graph G have weight 1
 2: V_{T_p}: Set of nodes within social diameter T_p from H_0
 3: Initially set V_{T_p} = {H_0}
 4: E_{T_p}: Set of edges in G connecting nodes in V_{T_p}
 5: G_{T_p}: Graph to be pruned
 6: G_{static}: Graph that will not be pruned
    // Identifying Region to be Pruned
 7: for all Node v ∈ V do
 8:     if Distance(v, H_0) < (T_p + 1) then
 9:         Add v to V_{T_p}
10:     end if
11: end for
12: E_{T_p} = {(u, v) | if u ∈ V_{T_p} or v ∈ V_{T_p}}
13: G_{T_p} = (V_{T_p}, E_{T_p})
14: G_{static} = G − G_{T_p}
    // Pruning
15: Define W as the new weight matrix of G_{T_p}
16: for all pair of connected nodes (u', v') ∈ G_{T_p} do
17:     W_{u'v'} = number of common friends of u' and v'
18: end for
19: Let G' = G_{T_p}
20: for each pair of connected nodes (u', v') ∈ G_{T_p} do
21:     if W_{u'v'} ≤ T_s then
22:         Delete edge (u', v') from G'
23:     end if
24:     if u' or v' is isolated then
25:         Delete the node from G'
26:     end if
27: end for
28: Return G_{prune} = G_{static} ∪ G';
```

For those disconnected identities during the pruning process, we initially mark them as Sybil accounts. Their class labels will be further refined in the regularization phase that will be introduced in Section 5.

The detailed pruning process is described in Algorithm 1. First, the region G_{T_p}, which is within the pruning diameter (T_p) from the honest seeds (H_0), is identified as the graph to be pruned (Lines 1-13). The rest of the graph (G_{static}) will stay unpruned (Lines 14 and 28). Then, if the number of common friends between any two connected nodes in G_{T_p} is smaller than or equal to the similarity threshold (T_s), the edge between the two nodes is deleted (Lines 15-23). The final step is to remove all isolated nodes in G_{T_p} and label them as Sybil nodes (Lines 24-26).

5. Graph Regularization

As mentioned in Sections 2.1 and 3, the existing Sybil defense mechanisms suffer from high false detection rates due to that the fast mixing assumption does not hold in real world social networks, and criminal accounts are difficult to detect within sophisticated structures Yang et al. (2012); Ghosh et al. (2012). Many studies Taskar et al. (2001); Neville and Jensen (2007) have shown that social networks conform to the relational property. This

property is a phenomenon that linked or neighboring nodes tend to have the same class labels in a network. Also, it has been demonstrated by those studies that this property can be utilized to improve classification performance. Similarly, the key to the graph-based regularization approaches in the semi-supervised setting is the cluster assumption Zhu et al. (2003), which is consistent with the relational property. The cluster assumption refers to: 1) nearby points are likely to have the same label (local consistency); 2) nodes in the dense region are likely to have the same label (global consistency). Hence, in the semi-supervised setting, the ultimate goal is to seek for a classification function, which not only minimizes classification errors on the labeled data but also should be consistent with the intrinsic structure on unlabeled data. Inspired by this way of modeling the relational property, we develop a domain-specific graph regularization method for Sybil defense.

5.1. Objective Function

Given the initial labeled nodes (classified by the existing Sybil defense mechanisms), a set of honest seeds and an affinity graph, the key to our graph regularization method is to find out an objective function f that maps each node in the graph into the class space $\{+1, -1\}$ with the minimal classification error. The objective function consists of two parts. The first part is a *smoothness score* that measures local variations between nearby nodes, and the second part is a *fitting score* that penalizes the difference between the predicted labels and initial node labels.

Firstly, to be consistent with the intrinsic geometry of the data, i.e. the relational property, the labeling function f should not change sharply between correlated nodes. This can be well captured by the following formula:

$$\mathcal{D}_1(f) = f^T L f = \sum_{(i,j)\in E} w_{ij} \parallel f_i - f_j \parallel^2 \tag{1}$$

where $\mathcal{D}_1(f)$ denotes the *smoothness constraint*, measuring the sum of local variations, i.e., the overall changes of the labeling function between nearby points. For a *good* function f, $\mathcal{D}_1(f)$ should be small. In this representation, $L = D - W$ is the *graph Laplacian* where $W = [w_{ij}]$ is the weight matrix, and w_{ij} is the similarity value of pairwise connected nodes i and j. D is a diagonal matrix with $D_{ii} = \Sigma_j w_{ij}$. To guarantee the *convergence property*, the edge weight w_{ij} is calculated using the Gaussian kernel function with width σ Zhu (2006). Note that structure-based Sybil defense mechanisms are designed based on the *inherent trust* relationship within social networks. Hence we treat all existing edges equally. That is, if nodes u and v are connected, the edge weight for (u, v) is 1 and the weight matrix is set corresponding to the adjacency matrix $A = [a_{ij}]$ of the social graph.

Secondly, to be consistent with the initial labeling, the labeling function f should not change too much from the initial labels $\hat{C}$. So, we have:

$$\mathcal{D}_2(f) = \sum_{i=1}^{n} \parallel f_i - \hat{C}_i \parallel^2 \tag{2}$$

$\mathcal{D}_2(f)$ is the *fitting constraint*, which penalizes the deviation between predicted labels and initial labels. In our design, the fitting score covers all vertices. Note that, some *honest*

seeds are given in advance for the Sybil classification process. These specific nodes are *hard labeled* comparing to others. Thus, $\mathcal{D}_2(f)$ can be represented as the sum of the following two components:

$$\mathcal{D}_2(f) = \sum_{i\in(V-H_0)} \parallel f_i - \hat{C}_i \parallel^2 + \alpha * \sum_{i\in H_0} \parallel f_i - L_i \parallel^2 \tag{3}$$

H_0 indicates the set of honest seeds, which are *hard labeled* nodes. L_i is set to be class label $+1$. Similarly, $V - H_0$ is the set of unlabeled data before detection denoted as *soft labeled* nodes. $\hat{C}_i$ is the initially predicted label by a selected Sybil defense mechanism. Moreover, α is the parameter to measure different importance of these two terms.

Combing Equations (1) and (3), we can derive the discrete objective function for our domain-specific graph regularization method as follows:

$$J(f) = \frac{1}{2}\sum_{i=1}^{n} \sum_{j:v_i\in N(v_i)} w_{ij} \parallel f_i - f_j \parallel^2 + \frac{1}{2}\lambda_s \sum_{i\in(V-H_0)} \parallel f_i - \hat{C}_i \parallel^2 + \frac{1}{2}\lambda_h \sum_{i\in H_0} \parallel f_i - L_i \parallel^2 \tag{4}$$

where $N(v_i)$ represents the neighbouring nodes of v_i. The trade-off between the *smoothness score* and *fitting score* is captured by the positive regularization parameters λ_s and λ_h, wherein λ_s is the *soft* regularization parameter and λ_h is the *hard* regularization parameter. Obviously, $\lambda_h \geq \lambda_s$. Through the experiments in Section 6, we show that the performance of our graph regularization method is largely governed by the parameter λ_s.

Furthermore, to reduce the degree bias which may impact false positives from low-degree benign nodes and false negatives from high-degree Sybil nodes Cao et al. (2012), we modify the first term of f by dividing the degree for each node, which is represented as follows:

$$J(f) = \frac{1}{2}\sum_{i=1}^{n} \sum_{j:v_i\in N(v_i)} w_{ij} \parallel \frac{f_i}{D_{ii}} - \frac{f_j}{D_{jj}} \parallel^2 + \frac{1}{2}\lambda_s \sum_{i\in(V-H_0)} \parallel f_i - \hat{C}_i \parallel^2 + \frac{1}{2}\lambda_h \sum_{i\in H_0} \parallel f_i - L_i \parallel^2 \tag{5}$$

where D_{ii} denotes node i's degree, and the same for D_{jj}.

The optimal classification function f^* can be obtained by minimizing the objective function $J(f)$:

$$f^* = argmin_f J(f) \tag{6}$$

5.2. Derivation of the Objective Function

For simplicity, Equation (5) can be expressed as the following matrix form:

$$J(f) = \frac{1}{2} f^T L f + \frac{1}{2}(f - f_0)^T \Lambda (f - f_0) \tag{7}$$

L is the normalized Laplacian matrix $I - D^{-1/2}AD^{-1/2}$, where I is the identity matrix and A is the adjacency matrix of the social graph. Recall that if nodes i and j are connected, their edge weight is 1, D is the diagonal matrix, and D_{ii} denotes the node i's degree. f_0

denotes the *initial class label* combining both the hard and soft labels. Λ is a diagonal matrix and can be represented as:

$$\Lambda(i,i) = \begin{cases} \lambda_s & \texttt{if } i \in V - H_0 \\ \lambda_h & \texttt{if } i \in H_0 \end{cases} \tag{8}$$

To find the optimal classifier, the objective function J should be minimized by explicitly taking its derivatives with respect to the f's and setting them to zero. Differentiating $J(f)$ with respect to f, we have

$$\frac{\partial J}{\partial f}|_{f=f^*} = Lf^* + \Lambda(f^* - f_0) = 0 \tag{9}$$

which derives a closed-form solution:

$$f^* = (L + \Lambda)^{-1}\Lambda f_0 \tag{10}$$

5.3. Sybil Classification

Since f^* obtained in Equation (10) is a real-value function, the final class label C_v^* for a vertex $v \in V$ is given by the following formula:

$$C_v^* = \begin{cases} +1 & \texttt{if } f_v^* > 0 \\ -1 & \texttt{if } f_v^* \leq 0 \end{cases} \tag{11}$$

6. Experimental Evaluation

We perform two sets of experiments to evaluate the effectiveness of our graph pruning and regularization techniques by verifying whether they can enhance the detection accuracy on existing Sybil defense mechanisms under both target attacks and random attacks.

6.1. Datasets

All our experiments are conducted on four datasets from popular online social networks, representing the honest regions. Table 1 summarizes the properties of those datasets. Among them the Facebook graph Gjoka et al. (2009) is a connected component sampled using the similar sampling strategy in Cao et al. (2012). The rest of the social graphs have been commonly utilized to evaluate existing Sybil defense mechanisms[1].

Table 1: Dataset of social graph used in experiments

OSN	Node	Edge	Average Degree	CC
Facebook	9,943	60,870	19.88	0.221
AstroPh	18,772	396,160	22	0.3158
HepTh	9,877	51,971	5.67	0.2734
WikiVote	7,115	103,689	3	0.1250

It is widely acknowledged that obtaining an annotated Sybil attack dataset is extremely difficult. Thus, following the common practice in the literature Danezis and Mittal (2009); Tran et al. (2011); Cao et al. (2012), we also simulate attack regions. With additional

1. http://snap.stanford.edu/data/

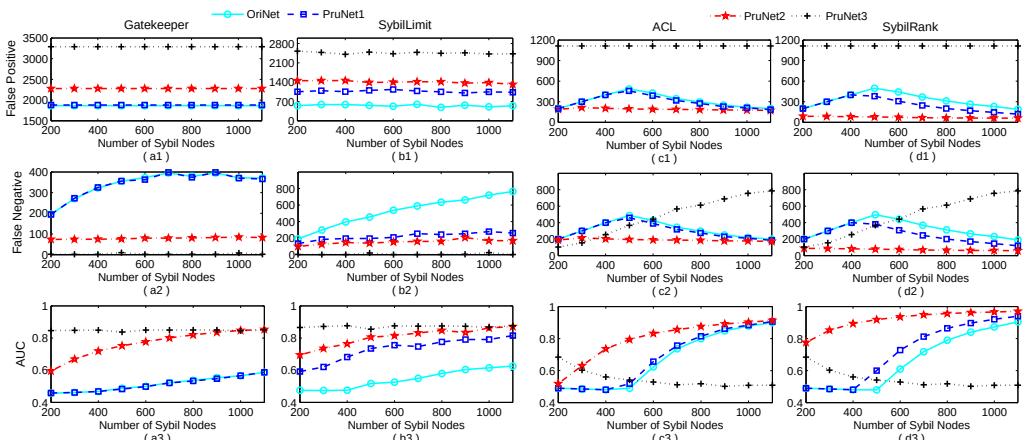

Figure 1: Performance comparison of the four Sybil defense schemes after applying graph pruning against the ER-target attack where OriNet denotes original network, and PrunNet1, PrunNet2, PrunNet3 correspond to pruned graphs by setting $T_p = 1$, $T_p = 2$ and $T_p = 3$ respectively.

efforts, we consider two types of attacks: random attacks and target attacks. For each type of attacks, we also consider two types of topological strictures for the attack regions, the random graph (the ER model) and the scale-free graph (the PA model). Specifically, for each attack type, we first generate m nodes as *Sybil supporters*, to establish social connections with nodes in the honest region. Then these dummy supporters introduce ψ additional Sybil nodes to form ER or PA topology among themselves with the average degree of $d = 10$. Note that this setting has been widely adopted in the literature Danezis and Mittal (2009); Tran et al. (2011); Cao et al. (2012). The social links between non-Sybil and Sybil regions are called *attack edges*. Each experiment is repeated 100 times and the average is computed to obtain statistically significant results.

6.2. Benchmark Approaches

Four representative structure-based Sybil defense mechanisms, Gatekeeper Tran et al. (2011), SybilLimit Yu et al. (2008), ACL Alvisi et al. (2013) and SybilRank Cao et al. (2012), are chosen to validate the effectiveness of our techniques. Gatekeeper and SybilLimit adopt the random walk approach to directly partition the social graph into non-Sybil and Sybil regions, while ACL and SybilRank utilize power iteration and degree-normalization techniques, and output a ranked list according to the trustworthiness of each node. Nodes with the lowest trustworthiness are highly likely to be Sybils. The key difference between ACL and SybilRank is that the latter adopts an early-termination technique during the propagation process while the former implements its trust propagation process iteratively until convergence.

For honest seeds selection, we use the same honest seeds for all the Sybil defense mechanisms in the experiments. For SybilLimit that uses only one honest seed, we randomly choose one node from the top-50 benign nodes with the highest degree. For other mechanisms which require multiple seeds, we choose all the 50 benign nodes, including the same node used in SybilLimit.

6.3. Evaluation Metrics

We use three metrics to measure the performance of our proposed techniques. One is the false detection rate including *false positive* and *false negative*, which correspond to the

misclassified number of benign and Sybil nodes respectively. To better assess the quality of node ranking, the area under the Receiver Operation Characteristic (e.g. AUC) is used as the evaluation metric. The AUC curve exhibits the probability that a random non-Sybil node is ranked higher than a random Sybil node. The AUC value of 1 represents a perfect classification results; 0.5 represents a random guess; and -1 represents the worst results.

6.4. Effectiveness against Target Attacks

In this section, we present the experimental results on the performance of our proposed methods upon four different Sybil defense mechanisms against target attacks, and the effectiveness of our regularization model with different parameters on the Facebook dataset.

6.4.1. Performance of Graph Pruning against Target Attacks

To emulate the target attack, we let Sybil supporters intentionally connect to the top 1000 benign nodes which are the closest to the honest seeds. We set the number of attack edges to be 200 and let the size of additional Sybil nodes ψ vary from 100 to 1000. Figure 1 summarizes the performance comparison of the four Sybil defense mechanisms after graph pruning in terms of false positive, false negative and AUC against the ER-target attack on the Facebook dataset. Similar results are obtained on the other three datasets. Specifically, we choose $T_s = 1$ for our experiments as it is difficult for an adversary to fool both a real user and his/her friends.

First, we can observe that SybilRank performs the best against the ER-target attack before pruning, followed by ACL, SybilLimit, and Gatekeeper, which is consistent with those illustrated in Cao et al. (2012). In addition, as shown in Figure 1 (c1-c3) and (d1-d3), both ACL and SybilRank achieve improved performance after the appropriate pruning process, and the best performance is reached when the pruning threshold $T_p = 2$. In this case, few benign nodes are disconnected from the network, which largely reserves the original network structure. However, when T_p is increased to 3, the AUC curves for both ranking methods exhibit instability and become even worse than before pruning. By examining the false positive and false negative, we found that by setting $T_p = 3$, more than 900 benign nodes are isolated from the non-Sybil region. In contrast, with the increment of additional Sybil nodes, the false negative curves monotonously increase. We speculate the reason is that although attack capacity is largely reduced due to the pruning procedure, many Sybil nodes can take priority to be accepted over those disconnected benign nodes. In addition, SybilRank outperforms all other approaches in terms of resistance to target attacks. Although ACL is also designed relying on trust propagation, SybilRank achieves better detection accuracy due to its early-termination strategy.

Gatekeeper performs the worst on defending against target attacks, because it relies on a strong assumption–*expander-like*, which requires tight connectivity among the non-Sybil region so that a breadth-first search starting from a benign node will highly likely stop at a non-Sybil node after $O(log(n))$ steps. However, such assumption is not always true in real-world social networks. SybilLimit performs slightly better than Gatekeeper, but still suffers from high false positive and false negative. As illustrated in Figure 1 (a1-a3) and (b1-b3), when varying the pruning threshold T_p from 1 to 3, the false positive increases by a moderate percentage but the false negative decreases drastically. Hence, the overall

Table 2: Performance of graph pruning and regularization in different social graphs against PA-target attack, where GP and GR represent graph pruning and regularization respectively, and GR is performed by setting $\lambda_h = 0.1$, $\lambda_s = 0.05$.

Dataset	WikiVote			HelpTh			Facebook			AstroPh		
	FP	FN	AUC	FP	FN	AUC	FP	FN	AUC	FP	FN	AUC
GK(Ori)	0.086	0.738	0.588	0.262	0.506	0.615	0.208	0.266	0.763	0.372	0.146	0.741
GK(Ori+GP)	0.095	0.388	0.779	0.329	0.094	0.789	0.244	0.056	0.850	0.380	0.028	0.796
GK(Ori+GP+GR)	**0.052**	**0.138**	**0.905**	**0.140**	**0**	**0.930**	**0.073**	**0**	**0.967**	**0.149**	**0**	**0.926**
SL(Ori)	0.021	0.102	0.939	0.077	0.736	0.594	0.089	0.514	0.699	0.030	0.663	0.654
SL(Ori+GP)	0.025	0.105	0.935	0.130	0.403	0.734	0.177	0.072	0.875	0.044	0.121	0.918
SL(Ori+GP+GR)	**0**	**0**	**1**	**0.065**	**0.387**	**0.904**	**0.038**	**0**	**0.982**	**0.010**	**0**	**0.995**
ACL(Ori)	0.003	0.017	0.992	0.086	0.738	0.588	0.045	0.444	0.756	0.039	0.699	0.631
ACL(Ori+GP)	0.006	0.039	0.978	0.055	0.388	0.779	0.016	0.163	0.910	0.017	0.307	0.838
ACL(Ori+GP+GR)	**0.001**	**0**	**1**	**0.052**	**0.138**	**0.905**	**0.015**	**0**	**0.992**	**0.017**	**0**	**0.991**
SR(Ori)	0	0	1	0.080	0.680	0.618	0.034	0.337	0.815	0.032	0.570	0.699
SR(Ori+GP)	0.006	0.039	0.978	0.040	0.344	0.808	0.007	0.074	0.959	0.006	0.102	0.946
SR(Ori+GP+GR)	**0.006**	**0**	**0.997**	**0.030**	**0.043**	**0.964**	**0.003**	**0.007**	**0.995**	**0.006**	**0.006**	**0.994**

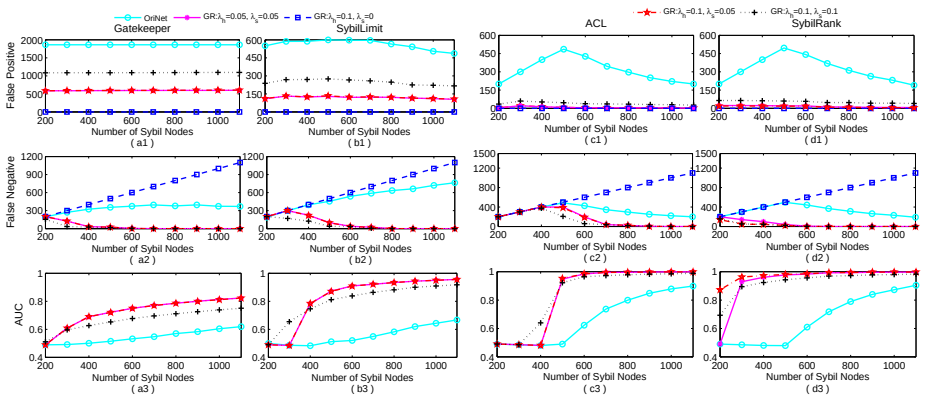

Figure 2: Performance after applying both graph pruning and regularization under the ER-target Attack.

quality is enhanced, which is represented by the AUC curve. Interestingly, it seems that Gatekeeper and SybilLimit perform the best under the target attacks by setting $T_p = 3$ in terms of AUC curve. However, in this case, the original network structure is damaged greatly since a large fraction of honest nodes are disconnected from the social graph. Thus, we choose the pruning diameter such that both the preservation of original network and detection accuracy are high. In the following experiments, we set the pruning diameter T_p to 2 for the Facebook dataset. Furthermore, we can see that all Sybil defense mechanisms show the similar trend that the detection performance on the original graph improves as the number of Sybil nodes increases. The reason is that the small cut between non-Sybil and Sybil regions becomes increasingly narrow and distinct as the Sybil group size gets larger, which makes the Sybil group more distinguishable from the non-Sybil region.

6.4.2. PERFORMANCE OF GRAPH REGULARIZATION AGAINST TARGET ATTACKS

In the following experiments, we investigate the impact of graph regularization on the performance of the Sybil defense mechanisms. Note that graph regularization is always employed after performing Sybil detection on pruned graphs.

An important factor for graph regularization is to determine the regularized parameters (λ_h, λ_s). Figure 2 shows the overall performance of the four Sybil defense mechanisms with graph pruning and regularization against ER-target attack on the Facebook dataset. We have the following observations. First, we can see that all the four Sybil defense mechanisms

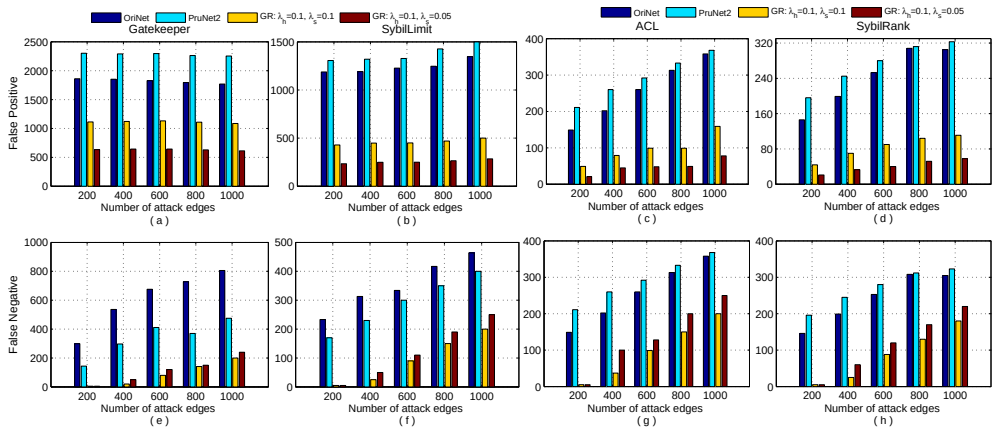

Figure 3: Performance after applying both graph pruning and regularization under the ER-random attack.

achieve more consistent and better performance in both false positive and false negative after the pruning and regularization processes, compared to their original designs. Especially, SybilRank, after graph pruning and regularization, performs the best and its AUC value is close to 1 in all scenarios, implying that such a framework can be seen as an ideal choice for Sybil nodes detection. Besides, the detection rates of Gatekeeper and SybilLimit are also significantly improved, where the false positive is decreased to 500 in each scenario and the false negative is close to 0 when the Sybil size is larger than 500. It confirms that utilizing the *relational property* of social topologies can significantly improve the detection accuracy for existing structure-based Sybil defense mechanisms.

Second, from all the results, it can be seen that Sybil classification performs worst when $\lambda_h = 0.1$ and $\lambda_s = 0$. In this case, due to the absence of *fitting constraint* on soft labeled data, the Sybil classification task tends to degenerate case where all the Sybils are assigned with honest label $+1$. Thus, the AUC is meaningless here since this metric is a graphical approach for displaying the tradeoff between true positive rate and false positive rate of a classifier. With fixed λ_h, we increase λ_s to 0.05. It can be seen that the numbers of mis-classified honest and Sybil nodes decrease greatly, demonstrating that *graph regularization* can significantly improve the detection accuracy. When λ_s increases to 0.1, the false negative continuously decreases but the false positive increases. We speculate the reason is that pure Sybil defense mechanisms depend on the *fast mixing* assumption. In addition, although the defense ability against target attacks can be improved through graph pruning, some honest nodes which are loosely connected to the rest of graph may be mis-classified. Both of them cause mis-classification. A higher value of λ_s indicates that the labeling function f strongly relies on the predictive results which include many spurious labels. Therefore, to minimize the cost function, nodes which are located in the periphery of graph and organized into small but tightly-connected clusters tend to be labeled as Sybil.

Furthermore, we evaluate the effectiveness of our methods upon Sybil defense mechanisms on the four datasets. Table 2 presents the representative results against PA-target attack. For a fair comparison, we take the fraction of mistakenly classified non-Sybil and Sybil nodes to be false positive and false negative respectively due to the different sizes of these four datasets. As aforementioned, we intend to choose a suitable pruning threshold such that both the coverage of original network and detection accuracy are high. Hence, T_p is set to 1 for WikiVote, and 2 for other social graphs. It is worth noting that on these datasets, current popular Sybil defense mechanisms can also be improved by our pruning

and regularization techniques under PA-target attack and obtain relatively high detection accuracy. For WikiVote, we find that graph pruning does not work since both the false detection rates and AUC get worse than being only performed on the original graph. This result is due to the extremely sparsity property underlying the WikiVote topology, reducing capability of target attacks. Thus, Sybil detectors can perform well to combat target attacks on this social graph. On another hand, graph pruning will damage the network structure somehow, leading to the disconnection of a fraction of honest nodes with lower degree. Nevertheless, our graph regularization method can effectively address such a problem incurred by pruning and allow these Sybil defense mechanisms to consistently perform well, which can be observed from the AUC metrics in Table 2. Besides, Gatekeeper and SybilLimit suffer from high false positive and false negative on the original social networks. After the pruning and regularization processes, they are able to achieve consistently much better results similar to SybilRank and ACL. These results confirm the effectiveness of our proposed strategies in enhancing the Sybil detectors.

6.5. Effectiveness against Random Attacks

We have shown that our proposed graph pruning and regularization techniques can significantly improve Sybil detection accuracy under target attacks. In the following experiments we investigate the effectiveness of the techniques against random attacks.

We let Sybil supporters randomly connect to the non-Sybil region starting from 200 attack edges and add additional 1000 fake nodes into the network. Then we gradually increase the number of attack edges to a large number 1000, so that the ability of the Sybil defense mechanisms degrades significantly. Figure 3 shows the performance comparison of Gatekeeper, SybilLimit, ACL and SybilRank under ER-random attacks by incorporating our pruning and regularization methods, respectively. It can be seen that the detection accuracy decreases slightly when incorporating the graph pruning process. However, with the graph regularization technique, the performance consistently outperforms the pure defense mechanisms. The similar results are also obtained on the other three data sets, but are not presented due to the space limitation. to know whether there exist target attacks in a particular real-world online social network, our two methods work as one package that can effectively enhance the performance of Sybil defense regardless of attack types.

7. Conclusion

In this paper, we focused on enhancing the performance of existing structure-based Sybil defense mechanisms. First, we provided a novel insight that Sybil defense can be modeled as a *partially labeled classification* problem. Then, based on this understanding, graph pruning was proposed to reduce attacking capacity of target attacks by exploiting the local structural similarity among nodes, leading to the improved robustness of Sybil detection mechanisms. A domain-specific graph regularization technique was also proposed to enhance Sybil classification results based on the relational property in social networks. Experimentation on popular online social network datasets confirms that our techniques can significantly improve the detection accuracy over the four representative Sybil defense mechanisms.

References

Lorenzo Alvisi, Allen Clement, Alessandro Epasto, Silvio Lattanzi, and Alessandro Panconesi. Sok: The evolution of sybil defense via social networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 382–396, 2013.

Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. In *ICML*, 2001.

Yazan Boshmaf, Dionysios Logothetis, Georgos Siganos, Jorge Lería, Jose Lorenzo, Matei Ripeanu, and Konstantin Beznosov. Íntegro: Leveraging victim prediction for robust fake account detection in osns. In *Proc. of NDSS*, 2015.

Olivier Bousquet, Olivier Chapelle, and Matthias Hein. Measure based regularization. In *NIPS*, 2003.

Zhuhua Cai and Christopher Jermaine. The latent community model for detecting sybil attacks in social networks. In *NDSS*, 2012.

Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *NSDI*, 2012.

George Danezis and Prateek Mittal. Sybilinfer: Detecting sybil nodes using social networks. In *NDSS*, 2009.

Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. Understanding and combating link farming in the twitter social network. In *WWW*, 2012.

Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. A walk in facebook: Uniform sampling of users in online social networks. *arXiv preprint arXiv:0906.0060*, 2009.

Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. In *NIPS*, 2004.

Abedelaziz Mohaisen, Aaram Yun, and Yongdae Kim. Measuring the mixing time of social graphs. In *IMC*, 2010.

Abedelaziz Mohaisen, Nicholas Hopper, and Yongdae Kim. Keep your friends close: Incorporating trust into social network-based sybil defenses. In *INFOCOM*, 2011.

Jennifer Neville and David Jensen. Relational dependency networks. *Journal of Machine Learning Research*, 8: 653–692, 2007.

Matthias Seeger. Learning with labeled and unlabeled data. Technical report, University of Edinburgh, 2001.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93, 2008.

Benjamin Taskar, Eran Segal, and Daphne Koller. Probabilistic classification and clustering in relational data. In *IJCAI*, volume 17, pages 870–878, 2001.

Nguyen Tran, Jinyang Li, Lakshminarayanan Subramanian, and Sherman SM Chow. Optimal sybil-resilient node admission control. In *INFOCOM*, pages 3218–3226. IEEE, 2011.

Bimal Viswanath, Ansley Post, Krishna P Gummadi, and Alan Mislove. An analysis of social network-based sybil defenses. In *ACM SIGCOMM*, 2010.

Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. Analyzing spammers' social networks for fun and profit. In *WWW*, pages 16–20, 2012.

Haifeng Yu, Michael Kaminsky, Phillip B Gibbons, and Abraham Flaxman. Sybilguard: defending against sybil attacks via social networks. In *ACM SIGCOMM*, pages 267–278, 2006.

Haifeng Yu, Phillip B Gibbons, Michael Kaminsky, and Feng Xiao. Sybillimit: A near-optimal social network defense against sybil attacks. In *IEEE Symposium on Security and Privacy (SP)*, pages 3–17, 2008.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NIPS*, volume 16, pages 321–328, 2003.

Xiaojin Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison, 2006.

Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.