Exploiting Trust and Distrust Information to Combat Sybil Attack in Online Social Networks

Huanhuan Zhang, Chang Xu, Jie Zhang

Exploiting Trust and Distrust Information to Combat Sybil Attack in Online Social Networks

Huanhuan Zhang, Chang Xu, and Jie Zhang

School of Computer Engineering
Nanyang Technological University, Singapore
{zhan0376}@e.ntu.edu.sg

Abstract. Due to open and anonymous nature, online social networks are particularly vulnerable to the Sybil attack, in which a malicious user can fabricate many dummy identities to attack the systems. Recently, there is a flurry of interests to leverage social network structure for Sybil defense. However, most of graph-based approaches pay little attention to the distrust information, which is an important factor for uncovering more Sybils. In this paper, we propose an unified ranking mechanism by leveraging trust and distrust in social networks against such kind of attacks based on a variant of the PageRank-like model. Specifically, we first use existing topological anti-Sybil algorithms as a subroutine to produce reliable Sybil seeds. To enhance the robustness of these approaches against target attacks, we then also introduce an effective similarity-based graph pruning technique utilizing local structure similarity. Experiments show that our approach outperforms existing competitive methods for Sybil detection in social networks.

Keywords: Sybil Attack, Social Networks, Sybil Defense, Trust and Distrust, Transitivity

1 Introduction

Online social networks (e.g. Facebook) have gained great popularity and become an indispensable part of people's life. However, due to their open and anonymous attributes, these systems are particularly vulnerable to the Sybil attack, where adversary can create an unlimited number of fake identities with the intention to subvert the targeted system. According to a report on Facebook in August 2012, there are more than 83 million illegitimate accounts in the social network out of its 955 million active accounts.[1] These undesirable accounts are fabricated for various purposes such as spreading spam or gathering more 'likes' from users to promote products. Similarly, a lot of fake Twitter followers are sold rampantly on e-markets and bought by people to increase popularity or launch underground illegal activities.[2] Besides, malicious users can manipulate Sybils to pollute a

[1] http://www.bbc.co.uk/news/technology-19093078
[2] http://www.digitaltrends.com/social-media/guess-what-twitter-is-still-teeming-with-fake-accounts/

voting mechanism for some reputation systems (e.g. YouTube, Yelp) and thereby outvote honest users [4].

Recently, there is a flurry of interests to leverage social network structure for Sybil defense. Many proposals have been developed that attempt to detect Sybil nodes by utilizing topological features of social networks [2–4, 9, 10]. The basic rationale behind is based on two assumptions: (1) strong trust relationship among nodes, which makes it difficult for Sybil nodes to establish many social connections with honest nodes, even if they can easily fabricate substantial Sybil identities and build arbitrary topology networks among themselves. As a result, Sybil region connects to the main network via relatively few links, which results in *quotient cut* between non-Sybil and Sybil regions. (2) honest region is *fast mixing*, in which random walks from a non-Sybil node can quickly reach a stationary distribution after $O(log(n))$ steps compared to Sybil nodes.

However, most of the existing graph-based anti-Sybil mechanisms are vulnerable to *target attack* [10], in which an adversary has prior knowledge about the location of *honest seeds*, which are utilized for identity authentication, and launches Sybil attack by substantially compromising these honest entities as well as their nearby nodes. As a result, many dummy nodes seem to be honest due to direct connection with honest seeds, rendering the structure-based schemes ineffective. In addition, for existing Sybil defense mechanisms to work effectively, it is required that non-Sybil nodes in real social networks are well mixed to avoid sparse internal cuts. Nevertheless, this assumption does not conform to reality, since mixing time is substantially larger than anticipated [7]. As a result, these graph-based solutions cannot produce desirable detection accuracy by only relying on the inherent trust underlying social networks and limited topological features.

To address these problems, we propose an unified ranking mechanism by leveraging trust and distrust information in social networks to combat the Sybil attack. Specifically, we propose a simple but effective method to produce reliable Sybil seeds combining with current social network-based anti-Sybil schemes. Moreover, in order to enhance those topological designs against *target attacks*, an effective graph pruning strategy is introduced by exploiting local structure similarity between neighboring nodes. Finally, a ranking mechanism based on a variant of the PageRank-like algorithm is presented to combine trust and distrust together to output trustworthiness of nodes in the social network. Nodes with less trustworthiness scores are more likely to be Sybils. Experiments on three real data sets are conducted to verify the effectiveness of our methods. The results indicate that our mechanism can outperform existing state-of-the-art anti-Sybil approaches. Our method thus shades light on exploiting trust and distrust information for building an effective Sybil defense mechanism.

2 Related Work

The Sybil attack has attracted more and more attention in the community since it was introduced in 2002 [1]. Traditional solutions to combat the Sybil attack

rely on trusted identities provided by a certify authority. However, such centralized mechanisms suffer from the challenge of finding trusted identities due to the open membership in distributed systems.

In recent years, there is a surge of interests to leverage social network structures for Sybil defense. SybilGuard [2] and SybilLimit [3] are the first two decentralized protocols to exploit topological features to detect Sybil nodes. In SybilGuard, each node performs random route of the length $\Theta(\sqrt{n}logn)$, and a suspect is accepted if its random route intersects with a verifier's. When the number of attack edges is bounded to $O(\sqrt{n}/logn)$, 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 suffer from high false rate. SybilInfer [12] adopts the Bayesian inference technique that assigns to each node its probability of being Sybil, but suffers from high computational cost. Viswanath et al. [6] explain the rationale behind graph-based anti-Sybil schemes from the perspective of *graph partitioning*. They state that existing community detection algorithms can be utilized to detect Sybils. However, it is not easy to choose a reasonable metric to achieve better detection accuracy. And such community-based algorithms are vulnerable to targeted Sybil attacks. In addition, Mohaisen et al. point out that mixing time is much larger than what is anticipated in Sybil defense schemes, implying that social networks are generally not fast mixing [7]. Such a finding renders ineffective all defense schemes that are based on the mixing property. Cao et al. [10] develop a Sybil ranking mechanism which distinguishes Sybil from non-Sybil nodes based on their relative trustworthiness. SybilRank is validated in a real social graph-Tuenti to be effective and efficient against the Sybil attack. Since it depends on the *honest seeds* to propagate trust among network, this approach also suffers from target attacks.

In addition, some proposals are developed to incorporate distrust information in social graphs to mitigate the Sybil attack. SumUp [4] is an anti-Sybil approach designed for a distributed voting system. It leverages the social network among users to limit the number of fake votes collected from Sybil identities to $O(1)$ per attack edge. This design utilizes negative feedback to further diminish the voting capability of attackers and accumulates less fake votes. SybilDefender [9] proposes a Sybil community detection algorithm to detect the Sybil group surrounding a Sybil seed. However, no theoretical or empirical analysis is provided to guarantee that such a seed is actually a Sybil node, which is one of the main concerns of our work. Another recent work using the distrust factor is presented by Chao et al. [11]. They take the insight into the topological structure of criminal accounts' social relationship on Twitter and provide an inference algorithm to detect criminal accounts by propagating malicious scores from seeds (i.e., a set of known fake accounts). But their work is unable to incorporate known honest seeds and cannot differentiate non-Sybil from Sybil nodes. The purpose of our work is to leverage trust and distrust information in social networks against the Sybil attack.

3 Problem Formulation

3.1 System and Threat Model

A social network is modeled as a graph $G = (V, E)$, where each node in V represents a user in the network and each edge in E represents trust relationship between users. We use $n = |V|$ to denote the total number of users and $m = |E|$ to denote the total number of trust edges. The degree of a node $v_i \in V$ is $deg(v_i)$.

In the attacking scenario, there may be one or more attackers in a social network. All of these participants are controlled by an adversary. To launch the Sybil attack, an adversary fabricates multiple fake identities, which disguise as real users in the system to participate in illegal activities. However, they can only establish few *attack edges* with honest nodes. We divide the whole graph into *non-Sybil* and *Sybil* regions illustrated in Fig. 1. The trusted identity and Sybil seed will be used in the unified ranking mechanism.

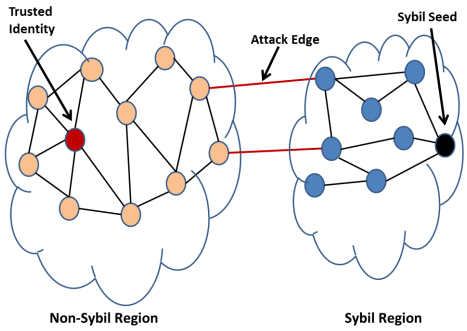

Fig. 1. Illustration of online social network under Sybil attack.

3.2 Assumptions

Our design is based on previous graph-based Sybil defense mechanisms that satisfy the following basic assumptions:

- There exist one or more known honest nodes. These nodes are utilized to break the symmetry and considered as *honest seeds* to implement identity verification.
- Honest region is *fast mixing*, in which random walks from a benign node can quickly reach a stationary distribution after $O(log(n))$ steps, compared to random walks from Sybil nodes. Although this characteristic is not so strictly satisfied in the real world, we assume nodes in non-Sybil regions are more tightly connected compared with Sybil nodes.
- There is a limited number of attack edges. For the inherent trust relationship among nodes, an adversary can create an arbitrary size of Sybil group but

establish a limited number of connections with honest nodes. Thus, it results in disproportionately *small cut* between non-Sybil and Sybil regions, which is an obvious sign for detecting Sybils.

4 Similarity-based Graph Pruning

Most of the 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 the non-Sybil and Sybil regions. However, once honest seeds are compromised by a set of disruptive nodes, these topological systems would underperform [10]. 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 attacks is called target seeding attack or simply *target attack*. To the best of our knowledge, no work has been proposed in the literature to solve this problem.

In this paper, we present a group 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 nearby honest seeds, hence evades Sybil detection. This strategy leverages local structural similarity underlying social networks. Intuitively, corresponding to the fast mixing and inherent trust assumptions, we speculate that the similarity between benign nodes and honest seeds is much higher compared to the similarity between benign nodes and Sybil nodes. Thus, by eliminating edges with low-similarity value ($w_{ij} \leq T_s$), 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 likely to be lower than that of the original network. Different structural similarity metrics in social networks have been proposed for measuring the strength of social links and predicting future interactions, such as cosine similarity, Jaccard similarity, and etc. [8]. In a social network, it is difficult for an adversary to simultaneously trick an honest node and its neighbors into trusting it. Hence, we choose the *number of common friends* as a metric to measure the structure similarity in the eliminating process.

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 satisfy the following two requirements: (1) It should minimize attack edges nearby *honest seeds*. (2) It shall also maximize the number of honest nodes because some benevolent nodes may be disconnected from the entire graph. 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 the experiments.

For those disconnected identities during pruning process, we initially mark them as Sybil accounts. These nodes will be re-visited in the following ranking phase. The detailed pruning process is described in Algorithm 1.

Algorithm 1: Graph Pruning()

Input: G : Graph $G = (V, E)$; H_0 : Set of honest seeds;
T_s : Similarity threshold; T_p : Pruned diameter
Output: G_{prune}: Pruned graph

Consider all the edge weight in graph G to be 1 ;
V_{ST_p} is the set of nodes within the distance from honest seeds less than $(T_p + 1)$;
Initially, $V_{ST_p} = \{H_0\}$;
for *all vertex* $v \in V$ **do**
    if $Distance(v, H_0) < (T_p + 1)$ **then**
        Add v to the V_{ST_p} ;
$E_{ST_p} = \{(u, v) \mid u \in V_{ST_p}$ or $v \in V_{ST_p}\}$, set of edges connecting nodes in V_{ST_p} ;
Let $G_{ST_p} = (V_{ST_p}, E_{ST_p})$;
$G_{static} = G - G_{ST_p}$, the undesired pruned graph;
$G_{static} = (V_{static}, E_{static})$, where $V_{static} = V - V_{ST_p}$ and $E_{static} = E - E_{ST_p}$;
Define W as the new weight matrix of graph G_{ST_p};
for *each pair vertice* $(u^{'}, v^{'}) \in G_{ST_p}$ **do**
    Count their number of common friends $numf$ and set $W_{u^{'},v^{'}} = numf$;
Let $G^{''} = G_{ST_p}$ and $V_{disconnet} = \emptyset$;
for *each pair vertice* $(u^{'}, v^{'}) \in G_{ST_p}$ **do**
    if $W_{u^{'},v^{'}} \leq T_s$ **then**
        Delete edge $(u^{'}, v^{'})$ from $G^{''}$;
        if $u^{'}$ *or* $v^{'}$ *is isolated* **then**
            Delete the node from $G^{''}$;
            Add the node to $V_{disconnet}$;
Finally, $G_{prune} = G_{static} \cup G^{''}$;
return G_{prune}.

5 Unified Ranking Mechanism

Our unified ranking mechanism attempts to detect Sybil nodes by taking the following three steps: (1) producing a set of well-connected Sybil seeds by the Sybil seed selection algorithm; (2) propagating trust and distrust scores from a set of known honest and Sybil seeds among the entire social network according to the closeness of *social relationships*. (3) integrating the trust and distrust scores into an unified trustworthiness for each node, ranking nodes according to their trustworthiness and filtering out Sybil nodes based on the ranked list.

The detailed and formal description as well as the insight of the unified ranking mechanism are given in the subsequent sections.

5.1 Sybil Seed Selection Algorithm

Most of graph-based Sybil defense mechanisms are developed only relying on the inherent trust underlying social networks, while ignore the distrust information. Studies conducted on Twitter reveal that criminal accounts, even those hidden deeply within complicated structure, can be detected by propagating malicious scores from a set of known criminal accounts, indicating that distrust plays an important role in unveiling malicious nodes [11]. However, few work is provided to leverage trust and distrust information to combat Sybil attacks. SybilDefender [9] introduces a Sybil community detection algorithm to identify Sybil groups from the perspective of a given Sybil seed. Such seed is randomly selected from those nodes marked as Sybils in their identification algorithm. However, this selection strategy suffers from some drawbacks. First, no theoretical or empirical analysis is provided to guarantee that each identified Sybil node is actually Sybil. Second, if the Sybil seed connects with honest users via *attack edges*, the Sybil community detection algorithm will mistakenly classify many benign nodes as Sybils. In this paper, we present a Sybil seed selection algorithm to produce reliable Sybil seeds, which can be utilized in our ranking mechanism to effectively distinguish non-Sybil from Sybil nodes.

Our method focuses on looking for connected Sybil nodes by exploiting the link dependency property among social networks. Such property indicates linked or neighboring nodes tend to have the same class labels and can be used to improve the detection accuracy. Intuitively, corresponding to the basic assumptions-*fast mixing* and *small cut*, we observe that honest users are more likely to connect with honest nodes rather than Sybils. Similarly, most Sybil nodes mainly establish social connections with their colluding entities. For well-performed Sybil detectors, most of nodes can be accurately marked despite those ambiguous nodes either located on the border between non-Sybil and Sybil regions or sparsely connected to the main network. Thus, there exists different size of clusters in which each node has the same label. Based on this insight, we can start from a Sybil seed and expand it by adding its neighboring nodes which are also identified as Sybils.

Additionally, SybilRank [10] is validated to be an effective and efficient algorithm for detecting Sybil nodes among existing anti-Sybil schemes. In this paper, we treat this algorithm as a subroutine to seek for Sybil seeds. Algorithm 2 illustrates the detailed selection procedure for SybilRank. Let Ir denote the trust vector returned by the SybilRank scheme. $N(v_i)$ is the set of neighbors for node v_i in the network. Sybil seed selection is performed as follows: first, all the nodes in the network are classified into two categories: non-Sybil (labelled as 1) and Sybil (labelled as 0) by setting a cut-off threshold η. $I(.)$ is the indicator function that takes value 1 if the trust score of node v_i is larger than η and 0 otherwise. For each Sybil node, we calculate its spamicity value according to its

neighbors' class labels. The *spamicity* metric is defined as follows:

$$SP(v_i) = \frac{\Sigma_{j \in N(v_i)} |I(j, \eta)|}{|N(v_i)|} \tag{1}$$

Then, we search for the nodes whose SP is 1. Besides, the human evaluation procedure is introduced to further filter out those misclassified honest nodes. For normalization, the human evaluation can be formalized as a binary Oracle function defined in Equation 2. Subsequently, from the $Suspend$ set, we seek for tightly connected Sybil groups as Sybil seed candidates.

$$O(v_i) = \begin{cases} 0 & \text{if } v_i \text{ is Sybil} \\ 1 & \text{if } v_i \text{ is Honest} \end{cases} \tag{2}$$

This selection process repeats until $SeedCandidate \neq \phi$. Finally, the sets $SeedCandidate$ are returned, which can be treated as Sybil seeds.

Algorithm 2: Sybil Seed Selection()

Input: **G**: Social Network; Ir: Trust Vector outputted by SybilRank.
Output: $SeedCandidate$: set of Sybil Seeds

1 $[\mathbf{r}_\nabla, Index] = SORT(Ir)$;
2 $\theta = 0.01 * k, \quad k = 1$;
3 $\eta = \mathbf{r}_\nabla(n * \theta)$;
4 $I(v_i, \eta) = \begin{cases} 1 & \texttt{if } \mathbf{r}_\nabla(v_i) > \eta \\ 0 & \texttt{if } \mathbf{r}_\nabla(v_i) \leq \eta \end{cases}$;
5 $m = \Sigma_{v_i}\{v_i | I(v_i, \eta) == 0\}$;
6 **for** $i \leftarrow 1$ ***to*** m **do**
7 $\quad Source$=Index(i);
8 $\quad$ Calculate SP for each node using Equation 1;
9 $Suspend^* = |\{v_i | SP(v_i) == 1\}|$;
10 $Suspend = \{v_i | O(v_i) == 0, v_i \in Suspend^*\}$;
11 $s = |Suspend|$;
12 $SeedCandidate = \phi$;
13 **for** $k \leftarrow 1$ ***to*** s **do**
14 $\quad$ add $Suspend(k)$ to $SeedCandidate$;
15 $\quad$ **for** $p \leftarrow 1$ ***to*** s **do**
16 $\quad\quad$ if $Suspend(p) \in N(SeedCandidate)$;
17 $\quad\quad$ add $Suspend(p)$ to $SeedCandidate$;
18 **if** $SeedCandidate == \phi$ **then**
19 $\quad k = k + 1$;
20 $\quad \theta = 0.01 * k$;
21 $\quad$ repeat step 3-17;
22 Return $SeedCandidate$.

5.2 Unified Ranking Algorithm

To leverage trust and distrust in social networks, we present our unified ranking mechanism based on a variant of the PageRank-like model–*Personalized PageRank* algorithm, which is an essential technique for ranking and prediction [14]. Our ranking algorithm consists of two main components. The first component is to respectively propagate benign and malicious scores from a seed set of known honest and Sybil seeds among the entire network and the second component is to integrate the trust and distrust values into an unified trustworthiness for each node, which can be used to effectively discriminate non-Sybil from Sybil nodes.

Propagation Phase Given the topological structure of the social network and a set of labeled nodes, we can propagate trust/distrust scores from these seeds to their neighboring nodes according to their closeness of social relationships. The propagation process can be modelled in the following formula:

$$\mathbf{r}(v_i) = \alpha * \frac{\Sigma_{j \in N(v_i)} \mathbf{r}(j)}{|N(j)|} + (1 - \alpha) * d(v_i) \tag{3}$$

where $\mathbf{r}(v_i)$ denotes the score value of node v_i. α is the jump probability. Generally, $\alpha = 0.85$ [14]. d is the normalized score vector for the seed set. After trust and distrust propagation, two opposite scores are obtained for each node. In order to distinguish them, we *negatively* bias the initial scores towards the Sybil seeds. Thus, each node is assigned a negative value after distrust propagation. And the corresponding initial vector d is defined in Equation 4, where SS denotes the set of Sybil seeds.

$$d(v_i) = \begin{cases} \dfrac{-1}{|SS|} & \text{if } v_i \in SS \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Integration Phase In propagation phase, each node is assigned two scores, namely trust value and distrust value. The following questions are: can they solely be used to differentiate non-Sybil from Sybil nodes? If not, how can we combine them together such that the integrated value can identify Sybil nodes with lower false rate? In this paper, we utilize a simple but effective weighted scheme to obtain the final trustworthiness shown in Equation 5. Empirical analysis in the following section demonstrates that such combination model can greatly filter out most of Sybil nodes from rankings.

$$Total(v_i) = a * TR(v_i) + (1 - a) * DTR(v_i) \tag{5}$$

where $TR(v_i)$ and $DTR(v_i)$ respectively denote trust and distrust scores for node v_i. The parameter a is used to measure the weights of trust and distrust values for the overall trustworthiness.

6 Experimental Analysis

6.1 Experimental Design

Datasets and Attack Model We use three data sets from popular online social networks to stimulate the honest region. Table 1 summarizes the properties of these datasets. These social graphs have been commonly utilized to evaluate existing anti-Sybil schemes [3].

In addition, two kinds of topological structures, *random graph* (ER model) and *scale-free* (PA model), are used to simulate attack regions. For each type of attack, we first generate m nodes to be *Sybil supporters*, which serve for compromising honest region by establishing social connections with them. Then these dummy supporters introduce ψ additional Sybil nodes to form *ER* or *scale-free* topology among themselves with average degree of 10. The number of attack edges connecting non-Sybil and Sybil regions is g. In our simulations, we have $m = 100$, $g = 200$. The experiment is repeated 100 times with different attack scenarios. In addition, 50 honest nodes are picked from the top 500 non-Sybil nodes that have the highest degree to perform as verifiers or trust sources.

Table 1. Dataset of social graph used in experiments

OSN	Node	Edge	Average Degree	CC
Facebook	4,039	88,234	19.88	0.221
AstroPh	18,772	396,160	22	0.3158
HepTh	9,877	51,971	5.67	0.2734

Evaluation Metrics Three metrics are used to exhibit the effectiveness of our proposed techniques: number of accepted Sybil nodes (*false negative*), number of rejected benign nodes (*false positive*) and AUC curve. AUC represents the area under the Receive Operating Characteristic (ROC) curve and is a widely used metric for evaluating the quality of ranking within networks [10]. The AUC ranges between 0 and 1, with larger numbers indicating that a randomly selected honest node is ranked higher than a random Sybil node.

Comparative Sybil Defense Methods Two most recent and effective graph-based Sybil detection mechanisms are evaluated, namely SybilRank and ACL. SybilRank [10] is a ranking mechanism that sorts nodes in a network according to their trustworthiness. Nodes with low trust values are likely to be Sybils. ACL [14] is originally proposed to detect a local community in a social graph and it is based on the normalized version of Personalized PageRank algorithm. Alvisi *et al.* proved that such an approach can be utilized to detect Sybil nodes. Both SybilRank and ACL employ the power iteration technique, but SybilRank terminates the iteration process after only $O(log(n))$ steps. In addition, we choose the SybilRank algorithm to seek for Sybil seeds due to its better performance.

[3] http://snap.stanford.edu/data/

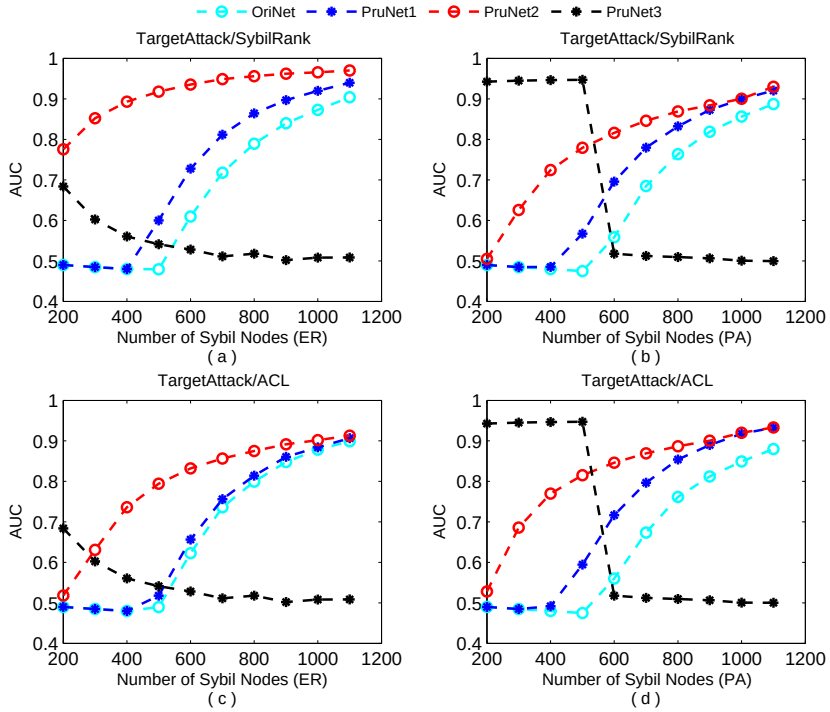

Fig. 2. The performance comparison of SybilRank and ACL methods when graph pruning technique is applied with respect to the size of Sybil region under target attacks. OriNet denotes the original network, PrunNet1, PrunNet2, PrunNet3 correspond to pruned graphs by setting $T_p = 1$, $T_p = 2$, $T_p = 3$ respectively.

6.2 Performance of Similarity-based Graph Pruning Technique

Based on the three real-world datasets including Facebook, AstroPh and HepTh described in Table 1, we conduct experiments to investigate the performance of our graph pruning strategy against target attacks. To infiltrate into the entire graph, we let Sybil supporters intentionally connect to the 1000 benign nodes which are the closest to the honest seeds. The number of additional Sybil nodes ψ varies from 100 to 1000. Then, SybilRank and ACL are implemented separately for Sybil classification on original graph and pruned graphs. Fig. 2 depicts the improved results on Facebook graph. The performance of pruning strategy implemented on other social graphs yield similar results. Fig. 2(a) and (b) show the detection results of SybilRank for ER and PA attack models, and (c) and (d) correspond evaluation result returned by ACL. Specifically, we have $T_s = 1$ for our experiments. This appears to be reasonable since it is hard for an adversary to fool a real user and his/her friends together.

As illustrated, both SybilRank and ACL schemes can be enhanced through graph pruning against target attacks, especially when the threshold $T_p = 2$. As we expected, no benign node is disconnected from the network in this case. However, when increasing T_p to 3, the AUC curve for Sybil defense exhibits

instability and becomes even worse than the original graph. By checking the false positive metric, we find that for both attack scenarios at least 800 benign nodes are isolated from the social graph. Furthermore, as the size of ER Sybil region increases, the AUC curves of both detection schemes monotonous decrease. We speculate the reason behind this is that although attack capacity is reduced due to elimination, many Sybil nodes can take priority to be accepted over disconnected honest nodes. But for PA Sybil region the curve keeps higher and falls sharply when number of Sybil nodes is 600. This phenomenon is attributed to the underlying Sybil structure. Since a large fraction of nodes in scale-free model have low degree which constitute the *heavy-tail* in the power-law node degree distribution, the pruning process will heavily affect these Sybil nodes to be isolated for larger T_p. Hence, despite those isolated benign nodes, most of honest nodes can be accurately classified. In the following experiments, we set threshold $T_p = 2$.

Table 2. The Sybil seeds selected in Algorithm 2 for different compromised network, where θ denotes the cut-off threshold to classify all the nodes to non-Sybil and Sybil categories.

`Num.Sybil`	`Threshold` θ	`Sybil Seed Sets`
$No.Sybil = 200$	0.02	0
	0.03	2 two-seeds
	0.04	3 two-seeds, 2 three-seeds, `4-seeds cluster`
	0.05	2 two-seeds, 3 three-seeds, `40-seeds cluster`
$No.Sybil = 300$	0.02	0
	0.03	2 two-seeds, 1 three-seeds, `7-seeds cluster`
	0.04	`78-seeds cluster`
	0.05	`97-seeds cluster`
$No.Sybil = 400$	0.02	2 two-seeds
	0.03	2 two-seeds
	0.04	1 two-seeds, `102-seeds cluster`
	0.05	`209-seeds cluster`
$No.Sybil = 500$	0.02	0
	0.03	0
	0.04	4 two-seeds, 2 three-seeds, `10-seeds cluster,16-seeds cluster`
	0.05	1 two-seeds, `210-seeds cluster`
$No.Sybil = 600$	0.02	0
	0.03	0
	0.04	5 two-seeds
	0.05	`230-seeds cluster`

6.3 Performance of Sybil Selection Algorithm

The experimental results illustrated in Fig. 2 have validated the effectiveness of our pruning strategies against target attacks. In this experiment, we combine

the SybilRank algorithm with graph pruning technique to seek for reliable Sybil seeds. We treat the trust vector output by SybilRank as an input value for Sybil seed selection algorithm. By adjusting the threshold θ to be used in partitioning the whole graph into non-Sybil and Sybil regions, we obtain the following Sybil seed selection results for different attack scenarios shown in Table 2. From all these results, we can see that our method can catch tightly connected Sybil seed, whereas the size is very small by setting the cut-off threshold θ to be a lower value. With the increment of θ, it is more likely to catch relatively large Sybil clusters which occupy large coverage of Sybil community. However, larger θ implies more nodes should be manually inspected which is not applicable in real case. Since we are attempting to cope with the Sybil attack problem, the performance of using these Sybil seeds to detect Sybils is our major concern. In the following experiment, we verify that the factor of Sybil seeds' size has a smaller impact on the defense performance.

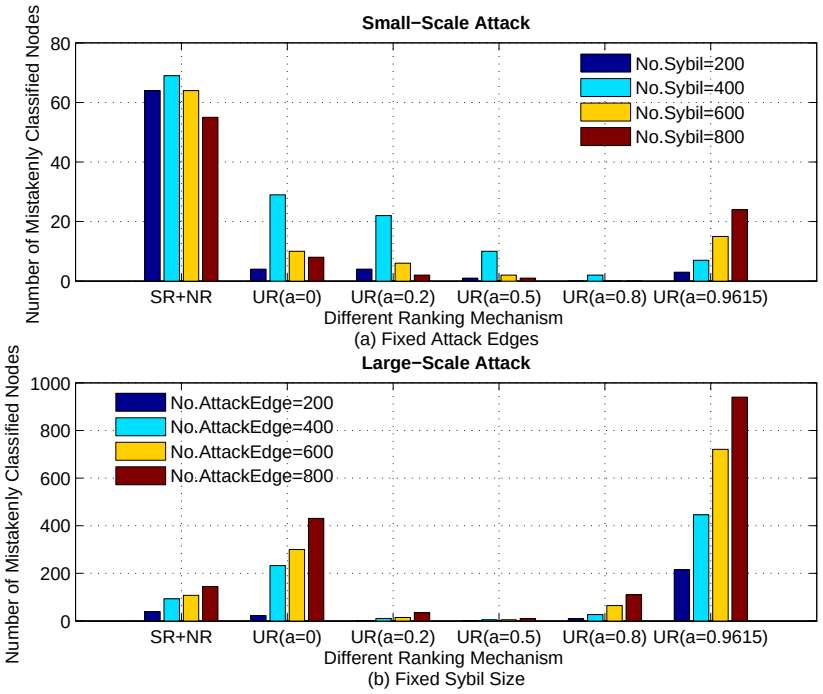

Fig. 3. The performance of unified ranking mechanism by varying weighting parameter a. $SR+NR$ means the SybilRank scheme without pruning step and UR is the unified ranking scheme.

6.4 Evaluation of Unified Ranking Mechanism

In this section, we investigate the effects of two components in our unified ranking mechanism, namely weighting parameter a and the size of Sybil seeds. In

addition, to have a fair comparison, we simulate another type of attack scenarios. To simulate the Sybil region, we let Sybil supporters connect to non-Sybil region starting from 200 attack edges. Meanwhile, Sybil supporters introduce 5000 additional Sybil nodes and establish an ER topology amongst themselves. Then we gradually increase the number of attack edges to a larger number 800. This attack type is called *large-scale attack*. Correspondingly, the attack type utilized in Sections 6.2 and 6.3 refers to *small-scale attack*.

We first check the effectiveness of our unified ranking mechanism by varying the weighting parameter a. By performing the Sybil seed selection algorithm, we obtain two Sybil seeds for each attack scenario. The detection results are illustrated in Fig. 3, where the value 0.9615 denotes the ratio between size of Sybil and benign seeds. It can be obviously seen that our unified mechanism can possess strong defense ability against small-scale attack regardless of the choice of parameter a. Even when the parameter $a = 0$, implying that the unified model solely relies on distrust information, this algorithm can still effectively differentiate non-Sybil from Sybil nodes.

However, the results are not so promising for large-scale attack compared with small-scale attack. It can be seen that the unified model performs worse for defending large-scale attack by choosing larger weighting parameter a. This might be due to the fact that the attack region is comparatively large such that malicious scores are assigned sparsely to each Sybil node, especially those honest nodes near the Sybil region. Thus the distinction of distrust values between non-Sybil nodes and Sybil nodes is not so clear. Instead, better performance can be achieved when the parameter a lies in the interval $[0.2, 0.8]$. Hence, we can conclude that the strength of weighting parameter's impact on the unified ranking mechanism depends on the size of Sybil region. Moreover, as shown in Fig. 3, the SybilRank approach also achieves good detection result for combating large-scale attack. This phenomenon is attributed to the fundamental assumption-*small cut*. As the Sybil region becomes larger, the small cut becomes increasingly narrow and distinct, which makes Sybil detection more effective. Finally, we can observe that the resilient unified model can always be derived by treating the trust and distrust information uniformly, that is to set $a = 0.5$.

Next, we examine whether the size of Sybil seeds plays an important role in uncovering more Sybil nodes. To explore the effect of this factor, we increase the number of seeds from 2 to a larger value. Additionally, to have a fair comparison, we randomly select another two Sybil nodes in order to verify the usefulness of our selected Sybil seeds. By setting the parameter $a = 0.5$, we obtain the following detection results using the unified mechanism shown in Fig. 4.

First, we observe that the unified mechanism can achieve higher detection accuracy by incorporating large Sybil seed cluster. Despite this case, the detection accuracy does not appear to heavily fluctuate with the increment of number of Sybil seeds. We speculate the reason is also due to the *small cut* assumption, which is the basis for designing anti-Sybil mechanisms. That is, due to the limited number of *attack edges* connecting non-Sybil and Sybil regions, the Sybil community surrounding Sybil seeds will accumulate a large fraction of

malicious scores regardless how many malicious nodes propagate distrust value initially. During the distrust propagation process, most of Sybil nodes can be penalized and assigned more malicious scores than honest users. It indicates that the performance of the unified model is not so sensitive to the size of Sybil seeds. Second, the model performs worse when incorporating randomly chosen Sybil nodes, which demonstrates that the Sybil seeds selected in Algorithm 2 are much reliable and useful in uncovering more Sybil nodes.

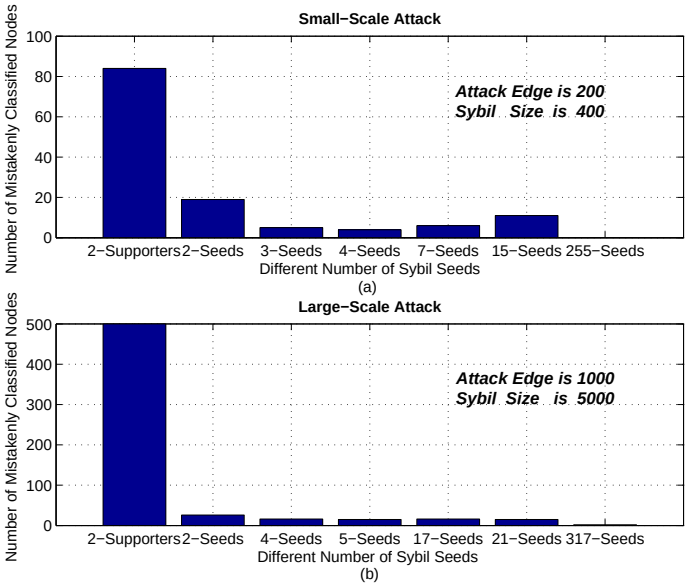

Fig. 4. The performance of unified ranking mechanism by varying the size of Sybil seeds. 2-Supporters are the randomly selected Sybil seeds. k-Seeds represents a Sybil cluster consisting of k connected Sybil nodes.

7 Conclusion

In this paper, we focus on leveraging both trust and distrust information to defend against Sybil attacks in social networks. First, a graph pruning strategy is introduced to diminish the attack ability near honest seeds by utilizing local structure similarity, leading to the improved robustness of Sybil defense mechanisms against target attacks. Moreover, we provide a Sybil seed selection algorithm to produce reliable Sybil seeds combining with current anti-Sybil schemes. Then, an unified ranking mechanism based on a variant of PageRank-like algorithm is proposed to combine trust and distrust information together to output integrated trustworthiness for nodes in a network. These trustworthiness values can be utilized to effectively distinguish Sybil from non-Sybil nodes. Ex-

perimental results demonstrate that our unified ranking mechanism can achieve better performance and outperform state-of-the-art Sybil defense approaches.

8 Acknowledgement

This work is supported by the MoE AcRF Tier 2 Grant M4020110.020 and the ACI Seed Funding M4080962.C90 awarded to Dr. Jie Zhang.

References

1. Douceur, John R: The sybil attack. Peer-to-peer Systems. pp. 251-260. Springer Berlin Heidelberg (2002)
2. Yu Haifeng, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman: Sybilguard: defending against sybil attacks via social networks. ACM SIGCOMM Computer Communication Review 36, pp. 267-278. (2006)
3. Yu Haifeng, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao: Sybillimit: A near-optimal social network defense against sybil attacks. In Security and Privacy. (2008)
4. Tran, Dinh Nguyen, Bonan Min, Jinyang Li, and Lakshminarayanan Subramanian: Sybil-Resilient Online Content Voting. In NSDI, vol. 9, pp. 15-28. (2009)
5. MLA Tran, Nguyen, Lakshminarayanan Subramanian, and Sherman SM Chow: Optimal sybil-resilient node admission control. In INFOCOM. (2011)
6. Viswanath, Bimal, Ansley Post, Krishna P. Gummadi, and Alan Mislove: An analysis of social network-based sybil defenses. ACM SIGCOMM Computer Communication Review, vol. 41(4), pp. 363-374. (2011)
7. Mohaisen, Abedelaziz, Aaram Yun, and Yongdae Kim: Measuring the mixing time of social graphs. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM, (2010)
8. Mohaisen, Abedelaziz, Nicholas Hopper, and Yongdae Kim: Keep your friends close: Incorporating trust into social network-based sybil defenses. In INFOCOM. (2011)
9. Wei, Wei, Fengyuan Xu, Chiu Chiang Tan, and Qun Li: Sybildefender: Defend against sybil attacks in large social networks. In INFOCOM, (2012)
10. Cao Qiang, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro: Aiding the detection of fake accounts in large scale social online services. In NSDI. (2012)
11. Yang Chao, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In Proceedings of the 21st international conference on World Wide Web, pp. 71-80. ACM, (2012)
12. Danezis, George, and Prateek Mittal: SybilInfer: Detecting Sybil Nodes using Social Networks. In NDSS. (2009)
13. Ghosh, Saptarshi, 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 Proceedings of the 21st international conference on World Wide Web. ACM, (2012)
14. Berkhin, Pavel: A survey on pagerank computing. Internet Mathematics. Vol. 2(1), pp. 73-120. (2005)