Improving sybil detection in online social networks

Zhang, Huanhuan

2014

Zhang, H. (2014). Improving sybil detection in online social networks. Master' s thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/61788>

<https://doi.org/10.32657/10356/61788>

IMPROVING SYBIL DETECTION IN ONLINE SOCIAL NETWORKS

A Thesis

Submitted to the School of Computer Engineering
of the Nanyang Technological University

by

Huanhuan Zhang

Supervisor: Asst. Prof. Jie Zhang

for the Confirmation for Admission
to the Degree of Master of Engineering

July 21, 2014

Abstract

Due to open and anonymous nature, online social networks are particularly vulnerable to Sybil attack, where a malicious user can fabricate many dummy identities to target systems. Recently, there are a flurry of interests to leverage social network structure for Sybil defense. The rationale behind these approaches is to partition the whole social graph into honest and Sybil regions based on two core assumptions: (1) strong trust relationship among nodes, which makes it difficult for Sybil nodes to establish substantial 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 an benign node can quickly reach a stationary distribution after $O(\log(n))$ steps, compared to random walks from Sybil nodes.

However, these anti-Sybil mechanisms are less effective than expected since real-world social networks do not conform to the above assumptions. In the thesis, the main theme is to explore additional topological features underlying social networks in designing secure and robust Sybil defense systems. In particular, this thesis is comprised of two studies.

Firstly, graph pruning and regularization techniques are proposed to enhance existing network-based Sybil defense mechanisms. In particular, a novel perspective is provided to interpret Sybil defense as the problem of partially labelled classification. Then, based on this framework, graph pruning is introduced to enhance the robustness of current anti-Sybil schemes against target attacks, by utilizing the local structural similarity between neighboring nodes in a social network. Besides, a domain-specific graph regularization method is designed to further improve the performance of those mechanisms by exploiting the relational property of social networks. Experimental results on four popular online social network datasets demonstrate that the proposed techniques can significantly improve the detection accuracy over the original Sybil defense mechanisms.

Secondly, an unified ranking mechanism based on trust and distrust is proposed for successfully combating Sybil attack. In this study, a Sybil seed selection algorithm is designed to produce reliable Sybil seeds by combining with current Sybil detectors. Then, to leverage trust and distrust, a unified ranking mechanism is designed to output an integrated trustworthiness for each node in the social graph. Nodes with lower trustworthiness value are more likely to be Sybils. Experiments show that the proposed framework outperforms existing competitive methods for Sybil detection.

To summarize, as the initiative attempt of considering different network features to address challenging problems faced by the current anti-Sybil schemes (i.e. mixing sensitivity, target attack and distrust utilization), this thesis serves to inspire the research community to build robust and secure Sybil defense mechanisms by exploring social topological features and applying them in real cases. Besides, this work mitigates the research gap between semi-supervised learning and social security fields, and thus is hoped to draw more attention towards this direction.

Acknowledgments

I would like to express my sincere thanks to my supervisor Prof. Jie Zhang, and Prof. Chang Kuiyu from School of Computer Engineering, who have continuously encouraged, inspired and supported me throughout my master research.

I would also like to thank all my friends and staffs in Center for Computational Intelligence Lab (C2I) and Data-Intensive Scalable Computing Lab (DISCO) for their technical support and enthusiastic help. Many thanks are given to the members of the research group I am working in, particularly Chang Xu, Hui Fang, Yuan Liu, Guibing Guo, Siwei Jiang. Without them, the research cannot be such an enriching and enjoyable experience.

My immense gratitude goes to my family for their great love and continuous encouragement in both my studies and life.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	vi
List of Tables	1
1 Introduction	2
1.1 Social Network-based Sybil Defense	4
1.1.1 System and Threat Model	4
1.1.2 Assumptions	4
1.1.3 Limitations of Topological Sybil Defense Approaches	5
1.2 Proposed Approaches	7
1.3 Contributions of the Thesis	9
1.4 Organization of the Thesis	10
2 Literature Review	12
2.1 General Sybil Defense Approaches	12
2.2 Social Network-based Sybil Detection Schemes	14
2.2.1 Trust-based Sybil Detection	14
2.2.2 Trust and Distrust-based Sybil Detection Schemes	17
2.3 Social Network-based Sybil Tolerance Schemes	19
2.4 Summary	20
3 Improving Sybil Detection via Graph Pruning and Regularization Techniques	21
3.1 Graph-based Partially Labeled Classification	23

3.2	Understanding Sybil Defense	24
3.3	Graph Pruning	26
3.4	Graph Regularization	28
3.4.1	Objective Function	29
3.4.2	Derivation of the Objective Function	31
3.4.3	Sybil classification	31
3.5	Experimental Evaluation	31
3.5.1	Datasets	32
3.5.2	Benchmark Approaches	32
3.5.3	Evaluation Metrics	33
3.5.4	Effectiveness against Target Attacks	33
3.5.5	Effectiveness against Random Attacks	42
3.6	Summary	45
4	Exploiting Trust and Distrust Information to Improve Sybil Detection	46
4.1	Sybil Seed Selection	47
4.2	Unified Ranking Algorithm	49
4.2.1	Propagation Phase	49
4.2.2	Integration Phase	51
4.3	Experimental Evaluation	51
4.3.1	Datasets and Attack Model	51
4.3.2	Comparative Sybil Defense Methods	52
4.3.3	Evaluation Metrics	52
4.3.4	Performance of Sybil Selection Algorithm	53
4.3.5	Evaluation of Unified Ranking Mechanism	54
4.4	Summary	57
5	Conclusions and Future Work	58
5.1	Conclusions	58
5.2	Future Work	59
	References	62

List of Figures

1.1	Illustration of online social network under Sybil attack.	5
3.1	Performance of plugging graph pruning into Gatekeeper and SybilLimit against the ER-target attack.	34
3.2	Performance of plugging graph pruning into ACL and SybilRank against the ER-target attack.	35
3.3	Impact of regularization parameters for enhancing Sybil defense mechanisms' performance.	36
3.4	Performance of plugging graph pruning and Regularization into Gatekeeper and SybilLimit against the ER-target attack.	37
3.5	Performance of plugging graph pruning and Regularization into ACL and SybilRank against the ER-target attack.	39
3.6	Performance of improved techniques upon Gatekeeper and SybilLimit against ER-random attack.	42
3.7	Performance of improved techniques upon ACL and SybilRank against ER-random attack.	44
4.1	Performance of unified ranking mechanism under different parameter a . .	54
4.2	Impact of Sybil seeds' size in detecting Sybil nodes.	56

List of Tables

3.1	Dataset of social graph used in experiments	32
3.2	Performance of improved techniques against PA-target attack.	38
3.3	Performance of improved techniques against PA-random attack.	43
4.1	Dataset of social graph used in experiments	52
4.2	The performance of Sybil seed selection algorithm.	53

Chapter 1

Introduction

Over the last few years, social networking sites have become an indispensable part of people's lives. Nowadays people are commonly engaged in many online activities, such as communicating with their friends, following real-time information on the Web, or doing online shopping, etc. With the rapid growth in popularity, social network service providers benefit from collecting a huge amount of user information. They provide platforms for manufacturers to advertise their products, and for users to amass their popularity in their personal social networks. However, 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.

Sybil attacks has become an increasing pervasive and dangerous problem as more and more people rely on online social networks for online communication and discover real-time information on the Web. For example, 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 malware and spam, or gathering many 'likes' from users to unfairly promote products. Similarly, a lot of fake Twitter followers are sold rampantly in e-markets and bought by people to increase popularity or launch underground illegal activities.² Besides, an adversary can manipulate Sybils to conduct malignant activities

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/>

in order to pollute the voting mechanism in reputation systems (e.g. YouTube, eBay, Taobao) [1, 2]. Traditional mechanisms to tackle the Sybil attack rely on a central trusted authority that issues and verifies credentials unique to a real users. Such solutions can limit the rate with which an adversary can introduce Sybil nodes into the system instead of the total number of such fake identities. However, there are several scenarios where such designs are not desirable. For example, it may be difficult to present a single trusted entity to the open membership in a distributed setting. Moreover, although such a trusted entity is established, it can easily become a single target for specific attacks, such as the denial-of-service attacks, where a malicious user can seize up the system by intentionally launching an attack towards the unique trusted node [3]. Another type of detection schemes is based on trust or reputation, such as Advocate [4], Applesed [5], iCLUB [6, 7], SybilProof [8] and many other trust models [2, 9–11]. The basic rationale behind these reputation systems is to assign a reputation value for each node by inspecting their historical behaviours and then filter out the nodes with lower trust. However, one big pitfall of such proposals is the fact that they are prone to whitewashing attacks, where the Sybil nodes can behave honestly initially and later launch a larger attack. Therefore, the Sybil attack problem is widely considered to be quite challenging.

Recently, there is substantial growth of interest to build applications for distributed systems that leverage social networks properties. In each of these applications, social networks are assumed to be trusted and well-connected. Many studies claim that these Sybil defense schemes can be utilized in decentralized systems for successfully combating Sybil attacks. However, the ability of these topological approaches for tackling Sybil attack has not been fully developed and well exploited. In this thesis, the main theme is focused on building effective and robust Sybil defense mechanisms by leveraging different social network features than existing schemes.

To motivate our work and put it in the correct context, in this chapter we first explain the insight behind social network-based Sybil defense approaches and the challenging problems of these approaches encounter in Section 1.1. Then we proceed to describe our proposed work in Section 1.2 and summarize our main contributions in Section 1.3. Finally, Section 1.4 gives a roadmap for the thesis.

1.1 Social Network-based Sybil Defense

Most of the social network-based Sybil defense mechanisms proposed in the literature address limitations in decentralized systems due to the lack of a trusted central authority. These approaches rely on the inherent trust in social networks and make use of social network characteristic (i.e. fast-mixing) to distinguish Sybil nodes from normal users, thus to partition the whole graph into non-Sybil and Sybil regions. In the following the Sybil problem formulation is firstly illustrated. Subsequently, the challenging problems for building more robust topological defense schemes are pointed out, which motivate the research work in this thesis.

1.1.1 System and Threat Model

In these Sybil defense mechanisms, 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. They assume that $n = |V|$ 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, due to the inherent trust in social networks, the adversary has limited capability to establish substantial *attack edges* with honest nodes. This leads to a disproportional small cut in the social graph in Figure 1.1. By seeking for specific algorithmic properties to distinguish honest and Sybil nodes, they divide the whole graph into non-Sybil and Sybil regions.

1.1.2 Assumptions

For these Sybil defense mechanisms to work, there are three fundamental assumptions needed:

- There exists one or more known honest nodes. These nodes are utilized to break the symmetry and considered as *honest seeds* to implement identity verification.

Figure 1.1: Illustration of online social network under Sybil attack.

- 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.
- 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.

1.1.3 Limitations of Topological Sybil Defense Approaches

As a new paradigm, social network-based Sybil defense becomes a hot topic to tackle Sybil attacks. Many proposals have been developed that attempt to detect Sybil nodes by utilizing topological features of social networks. The studies in social network-based Sybil defense have been demonstrated as effective approaches for dealing with the Sybil attack problem. However, the ability of these topological approaches has not been fully developed and well exploited. The insufficiency of existing Sybil defense schemes is briefly discussed below.

Firstly, the performance of social network-based Sybil defense mechanisms is highly associated with the quality of algorithmic properties underlying social networks [3, 12–15]. To defend against Sybil attacks, social networks are assumed to be fast mixing and

trusted. Among them, the fast mixing characteristic emphasizes a high quality of connectivity of these networks. In addition, some defense approaches are developed based on other properties, such as expander-like, well-balance and betweenness, which are stronger assumptions than fast mixing. Mohaisen *et al.* [16] conduct comprehensive experiments to measure the real-world social networks' properties including mixing characteristic, betweenness and expansion. Nevertheless, their results indicate that mixing time in real social graphs is larger than that anticipated in previous Sybil defense designs. Similarly, many studies related to the analysis of topological structures in large-scale real world social networks claim that such networks constitute of multiple communities with different types and sizes [17–24]. Moreover, they have a considerable fraction of nodes which have low degree (few social links) on the periphery. These nodes are often organized into small but tightly-knit clusters that connect to the rest of the networks via small cuts. These findings confine current topological Sybil defense schemes that strongly rely on the fast mixing property, to perform well in real cases. On the other hand, most topological Sybil defense mechanisms rely on a basic assumption that one or more honest nodes are known in advance. Such benign nodes (also known as *honest seeds*) are utilized for identity verification in order to partition 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 [15]. 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*.

Secondly, most of Sybil defense schemes are developed by solely leveraging trust relationships among users, while ignore distrust factor [3, 12–15]. Some proposals incorporate distrust information (e.g. negative feedback, malicious users) to enhance their performance and further limit the attack capability of Sybil attacks [25, 26]. For example, SumUp [27] leverages negative feedback to penalize paths from trust source to malicious nodes in order to aggregate more trustworthy votes from honest users. Ostar [26] tries to limit number of unwanted communities based on pair-wise trust credit network and illegal history records. However, these negative informations are provided by the systems. Thus, these defense models have limited applicability and are not general to other decentralized systems. Few studies have been carried out to exploit implicit distrust information embedded within social graphs for general Sybil defense schemes.

1.2 Proposed Approaches

In view of the limitations of the existing graph-based Sybil defense mechanisms, this thesis proposes two suits of strategies, one of which is focused on enhancing existing Sybil defense schemes to perform well and the other one is to develop a unified framework by leveraging trust and distrust underlying social networks in resisting Sybil attacks. To demonstrate the effectiveness and efficiency, they are respectively applied to deal with three types of challenging problems faced by the current anti-Sybil schemes. The characteristics and challenges of these three complex problems are summarized as follows:

- **Target Attack Problem** [3, 15, 17, 28–30]. To enhance the robustness of defense schemes, multiple honest seeds are established in some proposals to implement the verification process, such as Gatekeeper [28] and SybilRank [15]. Nevertheless, these approaches aim at limiting the negative impact incurred by target attacks instead of effectively addressing this issue. Few work has been proposed in the literature to well solve this problem. Thus, it is challenging to design a robust defense model against targeted Sybil attacks.
- **Mixing Sensitivity problem** [14–17]. To compensate the limitation incurred by mixing time, Cao *et al.* [15] propose a seed selection strategy for Sybil defense schemes to flexibly cope with multiple communities issues in large-scale social networks. The rationale behind is that before implementing the Sybil detection phase, they firstly partition the whole social graph into multiple small communities by using an efficient community detection algorithm [31] and distribute seeds among different communities. Such a strategy facilitates computational speed of Sybil detection and achieves higher detection accuracy. But it contributes less in dealing with the mixing sensitivity problem that also exists in small social graphs as discussed in [16]. Additionally, in order to eliminate negative impact of mixing time, Mohaisen *et al.* [14] model different levels of trust in social networks and design modified random walks upon SybilLimit to improve its performance. The insight is that by incorporating differential trust information, the small cut between non-Sybil and Sybil regions can be greatly shrunk and thus become obvious for Sybil

detection. Therefore, even the honest region does not strictly satisfy the mixing property, random walks originated from non-Sybil nodes will land on non-Sybil nodes with higher probabilities compared to Sybil nodes. But such a trust-driven model may affect many honest nodes, leading to high false positive.

- **Utilization of Distrust in Social Networks** [25–27, 32–36][37–41]. Wei *et al.* [42] develop a Sybil community detection algorithm starting from a Sybil seed. Sybil seed is randomly chosen from the rejected nodes in their verification algorithm. The difficulty stems from the fact that no theoretical or empirical analysis is provided to guarantee such a seed is actually a Sybil node, since it may either be honest node that is mistakenly classified or a single suspicious node that solely launches attack but does not belong to any malicious group [43]. As a result, the performance of the algorithm potentially suffers from high false positive rate. Thus, it is worthwhile to design an effective algorithm for choosing reliable Sybil seeds.

The proposed algorithms for respectively dealing with the above three complex problems are briefly discussed below.

Firstly, two effective strategies, graph pruning and graph regularization, are provided in Chapter 3 for coping with the target attack and mixing sensitivity problems as aforementioned. In this design, a novel perspective is provided to interpret Sybil defense as the problem of partially labelled classification. It is demonstrated that in spite of distinctions between such approaches, existing graph-based Sybil defense mechanisms can be seen as the processes of explicitly propagating honest labels among networks, so as to partition the whole 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, graph pruning and graph regularization, are devised by exploring additional structural information embedded within social graphs to improve the detection accuracy of current Sybil defense schemes. More specifically, 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 [44–46] 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 [46]. In this chapter, a novel domain-specific graph regularization method is then proposed based on the relational property to enhance the detection accuracy over existing Sybil defense approaches.

Secondly, to leverage trust and distrust underlying social networks, a unified ranking mechanism is designed in Chapter 4. In this design, a simple but effective Sybil seed selection algorithm is presented to produce reliable Sybil seeds by treating current popular anti-Sybil schemes as subroutine. Additionally, existing graph-based Sybil defense mechanisms are particularly vulnerable to target Sybil attacks. In order to guarantee the quality of the Sybil seeds, the seeding strategy is performed on the detection results returned by the Sybil detectors together with the graph pruning technique introduced in Chapter 3. Subsequently, a novel ranking mechanism is put forward citing the idea of power iteration to integrate trust and distrust together and output a unified trustworthiness score for each node in the social network. Nodes with smaller trustworthiness scores are likely to be Sybils. Comprehensive experiments are carried out in three real social networks. The results show that the ranking mechanism is more robust than the state-of-art topological anti-Sybil models.

1.3 Contributions of the Thesis

In conclusion, the research work in the thesis has three major contributions:

- As the initiative attempt of interpreting Sybil defense as the problem of partially labelled classification, a common insight into explaining how existing Sybil defense schemes work is provided. Based on this framework, it is beneficial to formulate the Sybil defense as a convex optimization problem. By solving an optimal classification function, the whole graph can be divided into non-Sybil and Sybil regions. Furthermore, this novel perspective also points to an opportunity to leverage the substantial amount of prior work on semi-supervised learning algorithms that can well address partially labelled classification, in order to defend against Sybils.

- Graph pruning and regularization techniques can effectively cope with target attack and mixing sensitivity problems. In graph pruning, the attack capability can be heavily diminished by trimming the edges with low-similarity value within the local region around honest seeds. Evaluations are conducted on four real social graphs, and the results indicate this technique can successfully aid the Sybil detectors against target attack. The graph regularization offsets the limitation of fast mixing property, which is not strictly held in real world and renders Sybil defense schemes ineffective. A domain-specific regularizer is designed that consists of two smoothness constraints, i.e. hard and soft regularized terms. Experimental results verify that this regularization technique can significantly boost the performance of existing topological anti-Sybil approaches. Moreover, they also demonstrate that the performance of the regularizer is largely governed by the soft regularization term.
- The unified ranking framework addresses the insufficiency of bootstrapping from only honest nodes for current Sybil defense schemes. It first produces reliable Sybil seeds based on the initial detection result and link dependency property in social networks. Then, it well combines trust and distrust together through propagation and integration phases, and then outputs integrated trustworthiness for each node in the social graph. Nodes with less trustworthiness are likely to be Sybils. Experimental results demonstrate that the unified approach can consistently outperform state-of-art anti-Sybil schemes. In addition, the results indicate that such a model is flexible to cope with many types of Sybil attacks including target attack.

1.4 Organization of the Thesis

The rest of the chapters in the thesis are summarized as follows.

Chapter 2. Literature Review: This chapter firstly provides an overview of relate research on Sybil defense. Then, the studies of social network-based Sybil defense mechanisms are divided into two categories, which are Sybil detection and Sybil tolerant schemes, and meanwhile the limitations of the existing network-based Sybil defense mechanisms are clearly pointed out, which motivate the research in this thesis.

Chapter 3. Improving Sybil Detection via Graph Pruning and Regularization Techniques: In this chapter, graph pruning and regularization techniques based on social network structures are proposed to improve the detection accuracy of current Sybil defense schemes. At first, an overview of related work on graph-based partially labelled classification is given. Secondly, a novel perspective for Sybil defense problem is presented, which interprets it as the problem of partially labelled classification. Then, based on this partially labelled classification framework, graph pruning and regularization are designed to handle the target attack and mixing sensitivity problems, by exploiting local similarity structure and relational property of social networks. Experimental results on four real social networks demonstrate that the proposed techniques significantly improve the performance of the state-of-art Sybil detectors.

Chapter 4. Exploiting Trust and Distrust Information to Improve Sybil Detection: In this chapter, a unified ranking mechanism by leveraging trust and distrust in social networks is proposed to combat Sybil attacks. At first, a novel Sybil seed selection algorithm is designed to seek for a set of reliable Sybil seeds, by combining with current popular social network-based Sybil detectors. Then, a ranking mechanism, citing the idea of personalized PageRank algorithm, is provided to combine trust and distrust together to assign an integrated trustworthiness value for each node in social graphs. Nodes with smaller trustworthiness values are more likely to be Sybils. Experimental results on three real social networks confirm that the proposed ranking approach is more robust than SybilRank and ACL in resisting Sybil attacks.

Chapter 5. Conclusions and Future Work: Based on the current studies, two possible directions for future work are identified. 1) For the first work, based on the current partially labelled classification framework for the Sybil defense problem, we will continue to explore and exploit additional structural information embedded within social graphs to enhance the detection accuracy over existing topological Sybil defense approaches. 2) For the second study, we will design new algorithms to seek for reliable Sybil seeds which may not rely on existing Sybil detectors; and develop novel unified mechanisms to leverage trust and distrust in social networks for defending against Sybil attacks.

Chapter 2

Literature Review

In this chapter, an overview of related research on Sybil defense is firstly provided to justify the direction and importance of the research work in Section 2.1. Section 2.2 describes a number of traditional and newly-proposed social network-based Sybil detection schemes and points out the major problems they suffer from. Subsequently, Section 2.3 illustrates some major social network-based Sybil tolerance schemes and concludes their shortcomings for combating Sybil attacks in decentralized systems. At last, the summary of this chapter is given in Section 2.4.

2.1 General Sybil Defense Approaches

The problem of Sybil attacks has attracted a lot of attention since it was introduced in 2002 [47]. Many solutions have been proposed to defend Sybil attacks in recent years. Generally, they can be classified into three broad categories: Sybil prevention approaches, social network-based Sybil detection approaches and social network-based Sybil tolerance approaches. Sybil prevention approaches attempt to limit the rate of fake identities created, to ensure that the fraction of disruptive nodes remains under a lower threshold [4, 8]. The idea behind such models relies on a trusted central authority that issues and verifies credentials unique to a real user. Typical examples include challenge-response mechanisms Cyworld system [8], etc. Sybil prevention schemes have advantages in limiting the rate at which abusive accounts are created. However, it is hard to present a unified trusted entity in decentralized systems due to their open and anonymous nature. Additionally, an attacker can easily seize up the system through launching denial-of-service

attacks to target the central authority [3]. Another type of detection schemes is based on trust or reputation, such as Advocate [4], Appressed [5] and SybilProof [8]. The basic rationale behind these reputation systems is to calculate the reputation/trust for each node by inspecting historical behaviours of nodes and suspend those identities with lower trust. One heavy pitfall of such proposals is the fact that they are prone to whitewashing attacks, where the Sybil nodes can behave honestly initially and later launch a larger attack.

Recently, there is a surge of interest in the research community to leverage social networks for combating Sybil attacks. Many topological solutions have been proposed that attempt to defend against Sybils in a social network by exploiting the social network's structure properties [12][13][42][15]. Different from Sybil prevention schemes, these approaches do not require central trusted entities, and instead rely on inherent trust relationships underlying a social network. According to their different assumptions and guarantees, social network-based Sybil defense mechanisms can be divided into two categories: Sybil detection and Sybil tolerance. Sybil detection approaches utilize global topological features of a social network to explicitly label identities as non-Sybil or Sybil nodes. The basic rationale behind is the two core assumptions: (1) Honest region is fast mixing, where random walks from a honest node can quickly reach a stationary distribution after $O(\log n)$ steps, compared to random walks from Sybil nodes. (2) Limited number of attack edges between non-Sybil and Sybil regions. Sybil tolerance schemes leverage the graph structure as well as transaction history to bound the impact an attacker can gain from fabricating multiple identities. The major issue for Sybil tolerance approaches is that they rely on application-specific information, which is not applicable to other decentralized systems.

To sum up, it is easier for social network-based Sybil defense mechanisms to defend against Sybil attacks than Sybil prevent schemes. These approaches are easy to implement and some can be operated efficiently. Although Sybil prevention schemes may largely suspend accounts that are likely to be fake in a timely manner, network-based Sybil detection is more general and easier to be implemented in real applications due to its simplicity, effectiveness and efficiency [15]. The research work in the thesis is based on the topological Sybil defense approaches and build more secure and robust mechanisms for Sybil detection by exploring additional information underlying social networks.

2.2 Social Network-based Sybil Detection Schemes

Sybil detection approaches have been developed for identity-based social systems. These proposals depend on specific assumptions and label identities in social network as either Sybil or non-Sybil. Specifically, these topological defense mechanisms are mainly designed on the basis of inherent trust and certain algorithmic structure property. Specifically, the intrinsic trust underlying a social graph confines an attacker to establish few attack edges with benign nodes, although the attacker can fabricate unlimited number of Sybil identities. And, fast mixing characteristic implies that the random walk from a non-Sybil node are more likely to lie within the non-Sybil region, compared with Sybil region.

2.2.1 Trust-based Sybil Detection

2.2.1.1 Random-based Sybil Detection

SybilGuard [3] and SybilLimit [12] are the first decentralized protocols to leverage social network structures to detect Sybil nodes. In SybilGuard, each node performs random route of length $w = \Theta(\sqrt{n \log n})$ and a suspect is accepted if its random route intersects with verifier's. When the number of attack edges is bounded to $g = O(\sqrt{n}/\log n)$, SybilGuard accepts at most $\Theta(\sqrt{n \log n})$ 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 $O(\log n)$ Sybil nodes per attack edge. Yu *et al.* claim that SybilLimit is nearly optimal in which its security guarantee is only a factor of $O(\log n)$ away from that of any optimal protocol.

SybilGuard and SybilLimit are performed in a distributed setting, where each node initially only has prior knowledge about its neighbors. In contrast, SybilInfer [13] is developed by assuming full knowledge of a social graph. This proposal adopts Bayesian inference technique that assigns the probability of being Sybil to each node. The key observation is that, if an attacker recruits a larger number of Sybils to the main network with few attack edges, the conductance of the social graph including both non-Sybil and Sybil regions becomes smaller. This leads to larger mixing time among the entire graph comparing against non-Sybil region and thereby achieves the detection of Sybil nodes. The pitfall of SybilInfer is its computational complexity since this approach is built on random walk traces with $O(n(\log n)^2)$ cost.

Gatekeeper [28] is another decentralized Sybil detection protocol that significantly improves the guarantees provided by SybilLimit. The key technique is based on a variant of the ticket distribution algorithm used in SumUp [27] and selects multiple honest identities in the graph to detect Sybil nodes. Based on a strong assumption that the social network is balanced random expander, Gatekeeper can reduce the number of Sybil nodes accepted per attack edge from $O(\log(n))$ to $O(\log(g))$, where g is the number of attack edges. Even for the worst case when $g = O(n/\log(n))$, Gatekeeper can achieve the same desirable result as SybilLimit, which accepts $O(\log(n))$ Sybil nodes per attack edge.

Recently, the fast mixing assumption in real social networks, which is crucial for building social network-based anti-Sybil schemes, is questioned by researchers. Mohaisen *et al.* [16] conduct comprehensive experiments in measuring the mixing rate of social graphs, including online social networks (e.g. Facebook,), collaborating voting system (e.g. Youtube) and reputation system (e.g. Epinion). The studies show that the mixing time in real social graphs is much larger than anticipated in Sybil defense schemes, implying that social networks are generally not fast mixing. This finding renders ineffective all defense schemes based on mixing property. Similarly, Wei *et al.* [42] find that Gatekeeper greatly suffers from high false positive and negative rates due to its assumption of random-expander which is a rather stronger characteristic than fast mixing and does not hold in real cases.

To address the mixing sensitivity problem, Mohaisen *et al.* model different level of trust in social network and design modified random walks upon SybilLimit to improve its performance [14]. Their specific design can be utilized to eliminate negative impact of mixing time to some degree. The basic rationale is that by incorporating trust information, the algorithmic property-quotient cut can be shrunk to significantly limit connections between non-Sybil and Sybil regions. Thus, despite mixing characteristic is not so strictly satisfied in honest region, random walk originating from non-Sybil nodes will land on non-Sybil nodes with higher probability compared with Sybil nodes. However, such a trust-driven model may affect many honest nodes, incurring high false positive. Motivated by their work, this work aims to cope with the mixing sensitivity problem by exploring another topological feature-relational property underlying social networks. Chapter 3 elaborates how to utilize this characteristic to solve the above problem and boost the detection accuracy of current Sybil detectors.

2.2.1.2 Community-based Sybil Detection

Viswanath *et al.* [17] conduct a comparative study on four popular Sybil defense mechanisms, including SybilGuard, SybilLimit, SybilInfer and SumUp. Through empirical analysis, they provide a novel perspective of interpreting these topological Sybil defense as local community detection algorithms. Meanwhile, they verify that existing community detection algorithms such as Mislove’s algorithm [48] can be utilized to detect Sybils. At last, they reveal two potential limitations of current social network-based Sybil defense. First, these topological approaches suffer from the mixing sensitivity issue. As discussed in Section 2.1, real social networks possess multi-community characteristic, which implying that the non-Sybil region consists of multiple, small, tightly-connected communities that are interconnected sparsely. Hence, nodes within one community might falsely classify many honest nodes within another community as Sybils due to limited connectivity between the communities. Second, these approaches are particularly vulnerable to target attacks. That is, a smart attacker may have prior knowledge about the location of honest seeds and intentionally establish attack edges with these nodes. Therefore, a large fraction of Sybil nodes are likely to evade detection due to their tight connections to the honest seeds.

Cai and Jermaine introduce a latent community detection algorithm to mitigate Sybil attacks [29]. Unlike previous work, the LC model is developed based on community structure underlying social networks instead of fast mixing property. With a hierarchical generative model for the observed social network, this approach attempts to detect Sybil communities by mapping the problem of Sybil detection into a Bayesian inference issue. Unfortunately, due to the utilization of Gibbs sampling which is an instance of Markov Chain Monte Carlo (MCMC) method, the LC model suffers from high computational cost and cannot be scalable to large social networks.

2.2.1.3 Ranking-based Sybil Detection

Cao *et al.* [15] develop a Sybil ranking mechanism to output a trustworthiness score for each node in a given social graph and distinguishes Sybil from non-Sybil nodes in terms of their relative trustworthiness. The basic design in SybilRank consists of two components, which are trust propagation and degree-normalization. Moreover, SybilRank alleviates

the influence of target attacks to some degree by adopting the multiple and flexible trust seed selection strategy. Experimental evaluation on synthetic and real world social graph (Tuenti) validate that SybilRank can be deployed to be effective and efficient against Sybil attacks. Nevertheless, the effectiveness of SybilRank will be heavily impacted if the distance between Sybils and the trust seeds decrease, thereby still suffers from target attacks.

Additionally, Alvisi *et al.* perform a series of theoretical and empirical analysis for ACL algorithm which is a variant of Personalized PageRank algorithms. And they state that this ranking mechanism can be utilized for resisting Sybil attacks [30]. ACL is developed by adopting the similar techniques as SybilRank, but solely incorporates one benign node to propagate trust among the entire network. This approach performs worse than SybilRank since the latter introduces the early-terminated and flexible seed selection strategies.

In Chapter 3, the target attack and mixing sensitivity problems faced by these social network-based Sybil defense mechanisms can be well addressed by exploring additional structural information embedded within social graphs. More specifically, 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. Then, a novel domain-specific graph regularization method is proposed based on the relational property to enhance the detection accuracy over existing Sybil defense approaches. This study is the first one to take into account the regularization technique for boosting the detection accuracy over those topological schemes.

2.2.2 Trust and Distrust-based Sybil Detection Schemes

Another direction in the literature for Sybil detection is to incorporate trust and distrust information for uncovering Sybil nodes. Comparing with trust-based approaches, the benefits of this direction are in two-fold: 1) no fundamental structure needs to be changed. Distrust information is needed to make up for the weaknesses incurred by limited topological features despite the local connectivity among social networks and fast

mixing property of the honest region. 2) it is beneficial to identify malicious groups, especially those hidden deeply within complicated structure from the perspective of distrust nodes. Some proposals are developed to incorporate distrust information in social graphs to mitigate the Sybil attack.

MobId [49] is introduced to identify Sybil nodes in the mobile setting. MobId assumes the existence of two opposite networks, which are friends network and foes network. To judge a suspect as non-Sybil or Sybil node, the benign seed should put this node into the friends and foes networks respectively, and calculate its betweenness centrality metric in these two networks. By comparing these two values, they add this suspect node to the network which possesses higher betweenness centrality with the node. However, MobID is only suitable to small to medium-size social graphs and cannot scale to large networks. SybilDefender [42] is a Sybil community detection algorithm to detect the Sybil group surrounding a Sybil seed. However, no theoretical or empirical analysis is provided to guarantee such a seed is actually a Sybil, since it may be mistakenly classified as honest node but is a single fake node that does not belong to any malicious groups and solely launches attacks. Another recent work using the distrust factor is presented by Chao *et al.* [50]. They take the insight into the topological structure of criminal accounts' social relationship on Twitter and provide an inference algorithm to detect more criminal accounts. Experimental results show that good performance is achieved by propagating malicious scores from a seed set of known criminal accounts.

In view of this, a simple but effective method is proposed in Chapter 4 to search for reliable Sybil seeds, which is the first attempt to consider the availability of distrust seeds. Subsequently, a unified ranking mechanism is designed to integrate trust and distrust together and output a unified trustworthiness value for each node in the social graph. The proposed mechanism can be utilized to effectively differentiate non-Sybil from Sybil nodes.

Not all work in detecting Sybil nodes have been proposed based upon network topology. Recently, studies conducted on the Renren, the largest online social networking website in China, have found that almost 80% of their detected Sybils do not have edges among themselves at all [50]. Thus, those structure-based approaches should not be applied to such a friendship network. Instead, some account-related statistics, like invitation frequency, outgoing request accepted ratio, clustering coefficient, *etc.*, are validated

to be effective in identifying Sybil accounts in the social graph. However, feature-based techniques are constrained to be valid in detecting those fake accounts whose behaviour slightly deviates from normal users. Some smart attackers can always adapt their behaviour or replicate the patterns of normal users, which makes it difficult to differentiate honest users from fake ones. Despite this specific social network, studies show that in many social graphs the criminal accounts are more likely to connect with other criminal nodes in order to seize up the systems. The research work in the thesis is based on previous network-based Sybil defense schemes.

To summarize, Sybil detection mechanisms aim to distinguish Sybil from non-Sybil relying on the inherent trust and specific graph features of social networks. The objective of Sybil detection is to seek for a set of minimal edges (small cut), which is probably a good indicator of Sybil attacks. The direction of Sybil defense is encouraged for its flexibility and generality in dealing with the Sybil attack problem.

2.3 Social Network-based Sybil Tolerance Schemes

Another category of social network-based Sybil defense mechanisms-Sybil tolerance schemes, which are developed relying on the graph structure features and application-specific information in the system [25–27, 32]. Instead of explicitly marking each identity as Sybil or non-Sybil node, Sybil tolerance approaches aim at reducing bad influence of Sybil attacks, regardless of the number of fake accounts being fabricated.

SumUp [27] is an anti-Sybil approach designed for a distributed voting system. It utilizes a ticket distribution scheme that leverages the social network among users to limit the number of fake votes collected from Sybil identities to $O(1)$ per attack edge. SumUp chooses a vote collector and distributes tickets on the links inside a voting envelope. To cast a vote, each voter must find a path to the vote collector with sufficient credit. If no such path can be searched, the vote is considered to be discarded. This design utilizes negative feedback to further diminish the voting capability of attackers and accumulates less fake votes.

Ostra [26] is designed to bound unwanted communication or spam sent by users with multiple identities in the context of online communication media like instant messaging.

Similar to SumUp, this scheme also uses a social network and assigns the credit values on the social links between users. When a message is sent, Ostra searches a path with available token from the sender to the receiver. If such a path is located, credit is consumed from each user the next along the path. Otherwise, the message is blocked.

In summary, Sybil tolerance schemes aim to bound influence of Sybil attack as much as possible rather than explicitly filtering out fake nodes. The rationale behind is to build a credit mechanism for determining the action whether a user can initiate a transaction with another user. The major drawback of these approaches is that they are application specific and not suitable to other decentralized environments.

2.4 Summary

The problem of Sybil attack has been extensively studied in recent years. Many approaches have been proposed to defend against Sybil attack. In this chapter, a general overview about the major categories of methodologies in the literature is presented. Specifically, these approaches can be broadly classified into three types: Sybil prevention, Sybil detection and Sybil tolerance. For each category, a number of representative methods are summarized and their advantages and disadvantages are listed. Moreover, this chapter explains and justifies the direction and importance of our research work.

Although many social network-based anti-Sybil approaches have been proposed so far, it is not far that the research community achieved and better ways to defend against the Sybil attack are expected, as additional network features exploration is believed to have the potential of leading to better performance. Existing structure-based approaches suffer from target attack and mixing sensitivity problems and perform worse in real world. Incorrect classification may be produced and cause high false positive and false negative rate. New approaches should be based on exploiting more information than existing schemes, allowing Sybil defense to be effective where now it is not.

Chapter 3

Improving Sybil Detection via Graph Pruning and Regularization Techniques

As mentioned in the previous chapters, the performance of social network-based Sybil defense mechanisms is highly associated with two basic 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. Unfortunately, existing structure-based Sybil defense mechanisms are under-performed and suffer from high false rate due to their unrealistic assumptions and limited usage of topological information. Specifically, studies show that in real-world social networks mixing time is much larger than anticipated in defense mechanisms [16]. This fact determines that graph-based solutions, which are sensitive to fast mixing property, cannot produce desirable accuracy. Furthermore, most of these mechanisms are vulnerable to *target attacks* [15], in which an adversary has prior knowledge about the locations of *honest seeds*, which are utilized for identity authentication, and launch Sybil attacks by substantially compromising these honest seeds as well as their nearby nodes. As a result, many dummy nodes seem to be honest due to their direct connections with these honest seeds, rendering ineffective structure-based defense schemes.

To address the above issues, this chapter provides two effective strategies to improve the performance of Sybil detection. Firstly, a novel way of interpreting existing graph-based Sybil defense mechanisms is provided as semi-supervised learning problems. Specifically, it is demonstrated that in spite of distinctions between such approaches, those topological algorithms can be seen as the processes of explicitly propagating *honest labels* among networks, so as to partition the whole network into non-Sybil and Sybil regions, i.e. each node is declared as either *Honest* or *Sybil*.

In the proposed partially labeled classification framework, two effective methods. i.e. *graph pruning* and *graph regularization*, are proposed by exploring additional structural information embedded within social graphs to improve the detection accuracy of current Sybil defense schemes. More specifically, 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 [44–46] 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 [46]. In this chapter, a novel domain-specific graph regularization method is then proposed based on the relational property to enhance the detection accuracy over existing Sybil defense approaches. None of the aforementioned topological Sybil defense mechanisms has taken graph regularization into account.

In all, three major contributions are made as follows: 1) a novel interpretation of Sybil defense is provided 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) a specialized manifold regularizer is designed by exploiting the relational property in social networks to further improve the accuracy of Sybil defense mechanisms.

The rest of chapter is organized as follows. Section 3.1 illustrates several major semi-supervised classification approaches for partially label classification problem. Section 3.2 presents a novel perspective for existing graph-based Sybil defense mechanisms and the

problems to solve in the chapter. Section 3.3 details the process of graph pruning. And the graph regularization framework is presented in Section 3.4. The experimental results are described in Section 3.5, followed by the conclusion in Section 3.6.

3.1 Graph-based 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. Chapelle’s recent survey [51] 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 [52], 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 [53][54] [55] [56]. They differ in particular choices of their objective functions and regularizers which manifests the underlying structure among unlabeled data. For example, Blum and Chawla [57] 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 labelled as positive, while the other nodes are labeled as negative. Zhu et al. [58] apply the Gaussian random fields and harmonic function methods to construct the objective function for semi-supervised classification. This representation allows a simple closed-form solution for the node marginal probabilities and has many interesting properties. Normalized Laplacian is used by Zhou *et al.* [56] to build the regularizer for label smoothing among the entire graph, which exploits both the local and global consistency in the graph. Belkin *et al.* [54] propose a manifold regularization framework that employs two regularization terms to seek for the optimal classifier.

The work in this chapter is mostly related to the graph-based regularization approaches, but formulates the regularizer particularly for the domain of Sybil defense.

3.2 Understanding Sybil Defense

Viswanath *et al.* [17] 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 [29]. In addition, community detection algorithms are likely vulnerable to target attacks. Thus, this chapter provides 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 [29]. 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 [57], which copes with the partially labeled classification problem from the perspective of graph mincut.

Intuitively, in the Sybil defense setting, there are two types of classes, i.e. *non-Sybil* and *Sybil*. To find Sybil nodes, the various schemes 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 (such as mixing time) that can discriminate honest nodes from Sybil nodes. 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*). $f(v_i) = +1, \forall v_i \in H_0$, where f is the labeling function.

Output: A mapping function $f : V \rightarrow Y$ from nodes to class labels.

As discussed in Section 3.1, the problem of *partially labeled classification* has been studied in the semi-supervised learning field [51]. 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, the goal is to provide strategies to enhance the robustness of current topological anti-Sybil designs and improve their detection accuracy. Inspired by the work of Mohaisen *et al.* [14], which leverages trust information to improve the performance of SybilLimit, the effective approach is to investigate topological features embedded within social graphs to strengthen current Sybil defense mechanisms. Section 3.3 discusses how to exploit the *local structural similarity* to address the target attack problem. Then in Section 3.4, a graph regularization technique, based on the *relational property*, is developed to smooth the detection results of those existing Sybil defense approaches.

```

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

// 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 foreach node  $v \in V$  do
8   if  $Distance(v, H_0) < (T_p + 1)$  then
9     └ Add  $v$  to  $V_{T_p}$ ;
10  $E_{T_p} = \{(u, v) \mid \text{if } u \in V_{T_p} \text{ or } v \in V_{T_p}\}$ ;
11  $G_{T_p} = (V_{T_p}, E_{T_p})$ ;
12  $G_{static} = G - G_{T_p}$ ;
// Pruning
13 Define  $W$  as the new weight matrix of  $G_{T_p}$ ;
14 foreach pair of connected nodes  $(u', v') \in G_{T_p}$  do
15    $W_{u'v'} = \text{number of common friends of } u' \text{ and } v'$ ;
16 Let  $G' = G_{T_p}$ ;
17 for each pair of connected nodes  $(u', v') \in G_{T_p}$  do
18   if  $W_{u'v'} \leq T_s$  then
19     Delete edge  $(u', v')$  from  $G'$ ;
20     if  $u'$  or  $v'$  is isolated then
21       └ Delete the node from  $G'$ ;
22 return  $G_{prune} = G_{static} \cup G'$ ;

```

Algorithm 1: Graph Pruning

3.3 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 [15]. 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 knowledge, no solution has been proposed in the literature to solve this problem.

In this section, a target pruning technique is presented to effectively reduce 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, it can be speculated 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 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 expected to be reduced. Different structural similarity metrics [14] 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. Here, the *number of common friends* is chosen as proximity metric 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 this 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. It can balance the trade-off by adjusting two parameters, i.e. pruning diameter T_p and similarity threshold T_s . Specific parameter choices will be examined in the following experiments. For those disconnected identities during the pruning process, they are initially marked as Sybil accounts. Their class labels will be further refined in the regularization phase.

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 1-11). The rest of the graph (G_{static}) will stay unpruned (Lines 12 and 22). 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 13-19). The final step is to remove all isolated nodes in G_{T_p} and label them as Sybil nodes (Lines 20-21).

3.4 Graph Regularization

As aforementioned, 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 [50, 59]. Many studies [44, 45] 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 [58], 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, a domain-specific graph regularization method is developed for Sybil defense.

Note that in the semi-supervised classification setting, typically, both positive and negative labels are partially known. In contrast, in the Sybil defense setting, no Sybil label information is given to *supervise* the Sybil classification problem. Thus, different from those graph-based regularization approaches in the semi-supervised setting, the design of graph regularization method is domain specific and it is employed after the existing Sybil defense mechanisms are performed, to further improve their performance, which is detailed in the following subsections.

3.4.1 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 the 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:

$$\mathcal{D}_1(f) = f^T L f = \sum_{(i,j) \in E} w_{ij} \|f_i - f_j\|^2 \quad (3.1)$$

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 $D_{ii} = \sum_j w_{ij}$. To guarantee the *convergence property*, the edge weight w_{ij} is calculated using the Gaussian kernel function with width σ [55]. Note that graph-based Sybil defense mechanisms are designed based on the *inherent trust* relationship within social networks. Hence all existing edges are treated 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 $\hat{C}$, which can be captured in the following formula:

$$\mathcal{D}_2(f) = \sum_{i=1}^n \|f_i - \hat{C}_i\|^2 \quad (3.2)$$

$\mathcal{D}_2(f)$ is the *fitting constraint*, which penalizes the deviation between predicted labels and initial labels. In the 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, as follows:

$$\mathcal{D}_2(f) = \sum_{i \in (V-H_0)} \|f_i - \hat{C}_i\|^2 + \alpha * \sum_{i \in H_0} \|f_i - L_i\|^2 \quad (3.3)$$

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 (3.1) and (3.3), it can derive the discrete objective function for the domain-specific graph regularization method as follows:

$$\begin{aligned} J(f) = & \frac{1}{2} \sum_{i=1}^n \sum_{j: v_i \in N(v_i)} w_{ij} \|f_i - f_j\|^2 \\ & + \frac{1}{2} \lambda_s \sum_{i \in (V-H_0)} \|f_i - \hat{C}_i\|^2 + \frac{1}{2} \lambda_h \sum_{i \in H_0} \|f_i - L_i\|^2 \end{aligned} \quad (3.4)$$

where $N(v_i)$ represents the neighbour nodes of v_i . The trade-off between the *smoothness score* and *fitting score* is captured by the positive regularization parameters λ_s, 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 4.3, it can be showed that the performance of graph regularization method is largely governed by the soft regularization parameter λ_s .

Furthermore, in order to reduce the degree bias which may impact false positives from low-degree benign nodes and false negatives from high-degree Sybil nodes [15], the first term of f can be modified by dividing the degree for each node, which is represented as follows:

$$\begin{aligned} J(f) = & \frac{1}{2} \sum_{i=1}^n \sum_{j: v_i \in N(v_i)} w_{ij} \left\| \frac{f_i}{D_{ii}} - \frac{f_j}{D_{jj}} \right\|^2 \\ & + \frac{1}{2} \lambda_s \sum_{i \in (V-H_0)} \|f_i - \hat{C}_i\|^2 + \frac{1}{2} \lambda_h \sum_{i \in H_0} \|f_i - L_i\|^2 \end{aligned} \quad (3.5)$$

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

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

$$f^* = \operatorname{argmin}_f J(f) \quad (3.6)$$

3.4.2 Derivation of the Objective Function

For simplicity, Equation (3.5) can be expressed as the following matrix form:

$$J(f) = \frac{1}{2}f^T Lf + \frac{1}{2}(f - f_0)^T \Lambda(f - f_0) \quad (3.7)$$

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:

$$\Lambda(i, i) = \begin{cases} \lambda_s & \text{if } i \in V - H_0 \\ \lambda_h & \text{if } i \in H_0 \end{cases} \quad (3.8)$$

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 , it shows

$$\frac{\partial J}{\partial f}|_{f=f^*} = Lf^* + \Lambda(f^* - f_0) = 0 \quad (3.9)$$

which derives a closed-form solution:

$$f^* = (L + \Lambda)^{-1} \Lambda f_0 \quad (3.10)$$

3.4.3 Sybil classification

Since f^* obtained in Equation (3.10) is a real-value function, the final class label C_v^* for a vertex $v \in V$ is given by the following formula.

$$C_v^* = \begin{cases} +1 & \text{if } f_v^* > 0 \\ -1 & \text{if } f_v^* \leq 0 \end{cases} \quad (3.11)$$

3.5 Experimental Evaluation

Two sets of experiments are carried out to evaluate the effectiveness of the graph pruning and regularization techniques by verifying whether they can be used to enhance the detecting accuracy of existing Sybil defense mechanisms against both target attacks and random attacks.

Table 3.1: 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

3.5.1 Datasets

All these experiments are conducted on four datasets from popular online social networks, representing the honest regions. Table 3.1 summarizes the properties of those datasets. Among them the Facebook graph [60] is a connected component sampled using the similar sampling strategy in [15]. The rest of the social graphs have been commonly utilized to evaluate existing Sybil defense mechanisms¹.

In the experiments, two types of topological structures are considered to simulate attack regions, the random graph (the ER model) [61] and the scale-free graph (the PA model) [62]. For either type of attack, it can 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 $d = 10$. 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.

3.5.2 Benchmark Approaches

Four representative graph-based Sybil defense approaches, Gatekeeper [28], SybilLimit [12], ACL [30] and SybilRank [15], are chosen to validate the effectiveness of the proposed 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

¹<http://snap.stanford.edu/data/>

adopts an early-termination technique during the propagation process while the former implements its trust propagation process iteratively until convergence.

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

3.5.3 Evaluation Metrics

Three metrics are used to measure the performance of the 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. More specifically, 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.

3.5.4 Effectiveness against Target Attacks

This section presents the experimental results on the performance of these two proposed methods upon the four different Sybil defense mechanisms. The effectiveness of the regularization model with different parameters are also evaluated on the Facebook dataset.

3.5.4.1 Graph Pruning against Target Attacks

To emulate the target attack, Sybil supporters intentionally connect to the top 1000 benign nodes which are the closest to the honest seeds. The number of attack edges is set to be 200 and let the size of additional Sybil nodes ψ vary from 100 to 1000. Figure 3.1 and 3.2 summarize the performance comparison of the four Sybil defense mechanisms, i.e. Gatekeeper, SybilLimit, ACL and SybilRank, after graph pruning in terms of false positive, false negative and AUC against the ER-target attack on the Facebook dataset, where OriNet denotes original network and PrunNet1, PrunNet2, PrunNet3 correspond

Figure 3.1: Performance of plugging graph pruning into Gatekeeper and SybilLimit against the ER-target attack.

to pruned graphs by setting $T_p = 1$, $T_p = 2$, $T_p = 3$ respectively. Similar results are obtained on the other three datasets. Specifically, T_s is set to be 1 for this experiments as it is difficult for an adversary to fool both a real user and his/her friends.

First, it can be observed from Figure 3.2 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 [15]. In addition, as shown in Figure 3.2 (a-f), 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, more than 900 benign nodes are isolated

Figure 3.2: Performance of plugging graph pruning into ACL and SybilRank against the ER-target attack.

from the non-Sybil region when $T_p = 3$. In contrast, with the increment of additional Sybil nodes, the false negative curves monotonously increase. It can be speculated 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 stronger 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. It is worth noting that the

Figure 3.3: Impact of regularization parameters for enhancing Sybil defense mechanisms' performance.

performance for Gatekeeper and SybilLimit is significantly improved through the pruning technique. As illustrated in Figure 3.1 (a-f), 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 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, the pruning diameter is chosen such that both

Figure 3.4: Performance of plugging graph pruning and Regularization into Gatekeeper and SybilLimit against the ER-target attack.

the preservation of original network and detection accuracy are high. In the following experiments, the pruning diameter T_p is set to 2 for the Facebook dataset. Furthermore, it can be seen that all Sybil defense mechanisms show the similar trend that the detection performance on the original graph improve 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. Thus, adversaries gain no advantage to launch large Sybil attacks by solely fabricating more Sybil identities.

3.5.4.2 Graph Regularization against Target Attacks

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

Table 3.2: Performance of the graph pruning and regularization techniques in different social graphs against the PA-target attack, where GP and GR represent graph pruning and regularization techniques respectively, and GR is performed by setting $\lambda_h = 0.10, \lambda_s = 0.05$.

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

Figure 3.5: Performance of plugging graph pruning and Regularization into ACL and SybilRank against the ER-target attack.

An important factor for graph regularization is to determine the regularized parameters (λ_h, λ_s). Figure 3.3 illustrates the impact of regularization parameters for enhancing the performance of Sybil defense mechanisms, by incorporating the manifold regularizer with SybilRank and Gatekeeper against ER-target attack where an attacker fabricates 800 Sybil nodes². By comparing Figure 3.3 with Figure 3.1 and 3.2, it can be seen that this manifold regularizer can significantly decrease the false detection rate of SybilRank and Gatekeeper in terms of false positive and false negative, implying that the relational property can inherently improve the detection accuracy. In addition, the impact from the hard parameters λ_h on the results is rather trivial, since the four curves in Figure 3.3 (a-d) appear to overlap in most situations. This demonstrates that the regularizer is merely influenced by the hard parameter (λ_h). In contrast, as λ_s increases from 0.005 to 0.1, it is clear that a monotonic increase in false positive, but drastic degradation in

²Results on SybilLimit and ACL exhibit the similar trends.

false negative, especially in the case when λ_s is very low (e.g., less than 0.01). Hence, a reasonable soft parameter should be determined to meet the two conditions that both false positive and false negative are low. However, it is difficult to fix an optimal value for λ_s , since it may depend on different social graphs and attack strategies. Nevertheless, as Figure 3.3 shows, graph regularization can achieve best results when the soft parameter λ_s is within the range $[0.01, 0.1]$.

Next, the performance of graph regularization under different parameter settings and attack scenarios is investigated. More Specifically, four different combinations of parameters (λ_h, λ_s) are included to exhibit their regularization ability. Figure 3.4 and 3.5 show the overall performance of the four Sybil defense mechanisms with graph pruning and regularization against ER-target attack on the Facebook dataset. The observations are in the following:

First, it can be seen that all the four Sybil defense mechanisms achieve more consistent and better performance in both false positive and false negative compared to their original designs after the pruning and regularization processes. Especially, the performance of 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 graph-based Sybil defense mechanisms.

Second, from all the results, it can be seen that Sybil classification performs worst when $\lambda_h = 0.1, \lambda_s = 0$. 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 , λ_s is increased 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. It is

speculated that 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. However, there is not much difference when varying λ_h from 0.05 to 0.1 and keep $\lambda_s = 0.05$. These results imply that the improvement of Sybil classification heavily depends on the soft regularization parameter λ_s .

Furthermore, the effectiveness of the methods is evaluated upon Sybil defense mechanisms on the four datasets. Table 3.2 presents the representative results against PA-target attack. For a fair comparison, the fraction of mistakenly classified non-Sybil and Sybil nodes is taken to be false positive and false negative respectively due to the different sizes of these four datasets. As aforementioned, a suitable pruning threshold should be determined 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 the pruning and regularization techniques under PA-target attack and obtain relatively high detection accuracy. For WikiVote, it can be found 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, the 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 3.2. 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.

Figure 3.6: Performance of improved techniques upon Gatekeeper and SybilLimit against ER-random attack.

These results confirm the effectiveness of the proposed strategies in enhancing the Sybil detectors.

3.5.5 Effectiveness against Random Attacks

The above experimental results have confirmed that the proposed graph pruning and regularization techniques can significantly improve Sybil detection accuracy under target attacks. In the following experiments the effectiveness of the techniques is investigated against random attacks.

h

Table 3.3: Performance of graph pruning and regularization techniques in different social graphs against the PA-random attack, where GP and GR represent the graph pruning and regularization techniques respectively, and GR is performed by setting $\lambda_h = 0.10, \lambda_s = 0.05$.

Dataset	WikiVote			HelpTh		
	FP	FN	AUC	FP	FN	AUC
GK(Ori)	0.075	0.316	0.805	0.264	0.405	0.665
GK(Ori+GP)	0.082	0.317	0.800	0.328	0.285	0.693
GK(Ori+GP+GR)	0	0.016	0.922	0.136	0.071	0.896
SL(Ori)	0.021	0.094	0.943	0.071	0.573	0.678
SL(Ori+GP)	0.025	0.100	0.937	0.081	0.247	0.736
SL(Ori+GP+GR)	0	0	1	0.063	0.043	0.942
ACL(Ori)	0.001	0.003	0.998	0.067	0.377	0.778
ACL(Ori+GP)	0.006	0.039	0.978	0.069	0.297	0.797
ACL(Ori+GP+GR)	0	0	1	0.052	0.049	0.932
SR(Ori)	0	0	1	0.063	0.342	0.798
SR(Ori+GP)	0.006	0.039	0.978	0.058	0.298	0.822
SR(Ori+GP+GR)	0	0	1	0.058	0.098	0.922
Dataset	Facebook			AstroPh		
	FP	FN	AUC	FP	FN	AUC
GK(Ori)	0.208	0.357	0.717	0.372	0.217	0.705
GK(Ori+GP)	0.244	0.181	0.788	0.380	0.045	0.788
GK(Ori+GP+GR)	0.073	0	0.963	0.150	0	0.925
SL(Ori)	0.083	0.346	0.785	0.028	0.528	0.722
SL(Ori+GP)	0.178	0.126	0.848	0.043	0.284	0.837
SL(Ori+GP+GR)	0.041	0	0.979	0.009	0.013	0.989
ACL(Ori)	0.020	0.195	0.893	0.021	0.382	0.798
ACL(Ori+GP)	0.019	0.193	0.894	0.022	0.385	0.797
ACL(Ori+GP+GR)	0.002	0	0.999	0.006	0.123	0.935
SR(Ori)	0.016	0.156	0.914	0.018	0.324	0.829
SR(Ori+GP)	0.015	0.153	0.916	0.017	0.310	0.836
SR(Ori+GP+GR)	0.003	0.001	0.998	0.006	0.117	0.938

Figure 3.7: Performance of improved techniques upon ACL and SybilRank against ER-random attack.

Here, Sybil supporters randomly connect to the non-Sybil region starting from 200 attack edges and add additional 1000 fake nodes into the network. Then the number of attack edges is gradually increased to a large number 1000, so that the ability of the Sybil defense mechanisms degrades significantly. Figure 3.6 and 3.7 show the performance comparison of Gatekeeper, SybilLimit, ACL and SybilRank under ER-random attacks by incorporating the 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. Table 3.3 illustrates representative results by performing Sybil detectors together with the proposed methods against PA-random attack with 400 attack edges. It can be seen that the proposed methods perform consistently better upon those four Sybil defense mechanisms. For SybilRank, the detection accuracy slightly increases after pruning on the HelpTh, Facebook and AstroPh datasets. This is mainly because the pruning technique leverages the *structure similarity* property in social networks and can eliminate some attack edges nearby honest seeds. As a result, more honest nodes can benefit from gaining more trustworthiness during the trust propagation process. But the detection results on the other schemes such as Gatekeeper, SybilLimit and ACL appear a little worse on pruned graphs compared to the original graphs. Nevertheless, the graph regularization technique can cope with such a problem and assist those defense schemes to accurately identify Sybil nodes. As it is difficult to know whether there exist target attacks in a particular real-world online social network, these two methods work as a single pack to effectively enhance the performance of Sybil defense.

3.6 Summary

In this chapter, two effective techniques based on social network features are proposed to enhance the performance of existing graph-based Sybil defense mechanisms. First, a novel insight is provided that interprets Sybil defense as the problem of partially labeled classification. Then, based on this understanding, graph pruning is 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 is also introduced to enhance Sybil classification results based on the relational property underlying social networks. Extensive experimentation on four popular online social network datasets confirms that the proposed techniques can significantly improve the detection accuracy over the four representative Sybil defense mechanisms.

Chapter 4

Exploiting Trust and Distrust Information to Improve Sybil Detection

As mentioned in chapters 1-2, most of the existing topological Sybil detection mechanisms are designed only considering inherent trust underlying social networks, while ignore the distrust factor. However, many studies reveal that distrust plays an important role in uncovering illegitimate accounts [37, 41, 63, 64]. For example, Yang *et al.* [50] perform an empirical analysis of characteristics of criminal accounts' social relationships on Twitter and find that the cyber criminal ecosystem in real world is more sophisticated than anticipated, which constitutes of multi-level organizations. They also verify that the state-of-art graph-based Sybil defense approaches such as SybilGuard, SybilInfer, etc, are ineffective to identify these criminal accounts. Instead, they propose a Criminal Inference Algorithm (CIA) by exploiting the social relationships between criminal nodes and propagating malicious scores in the graph from a set of known *criminal seeds*. Experimental result demonstrates that this algorithm can accurately identify most of the criminal accounts. Nevertheless, one pitfall with CIA is the fact that it only focuses on distrust information and does not give the false rate for mistakenly classifying legitimate accounts. Another recent work using the distrust factor (i.e. SybilDefender) is proposed by Wei *et al.* [42]. In their work, they explore Sybil community by performing partial random walks starting from an unveiled node. The unveiled node is selected as Sybil seed if it is rejected in their first verification process. However, no theoretical or empirical analysis is provided to guarantee that such a seed is actually a Sybil, since it may

be mistakenly classified as a honest node. Moreover, it may be a single fake node that does not belong to any malicious groups and solely launches attacks. Thus, the challenging problem for building a secure and robust Sybil detection mechanism by leveraging trust and distrust information is to explore the distrust information implicitly embedded within social networks [65].

In this chapter, a simple but effective Sybil seed selection algorithm is presented to produce reliable Sybil seeds by treating current popular anti-Sybil schemes as subroutine. Additionally, existing graph-based Sybil defense mechanisms are particularly vulnerable to target Sybil attacks as discussed in the previous chapters. In order to guarantee the quality of the Sybil seeds, the seeding strategy is performed on the detection results returned by the Sybil detectors together with the graph pruning technique introduced in chapter 3. Subsequently, a novel ranking mechanism that cites the idea of power iteration is put forward to integrate trust and distrust together and output a unified trustworthiness score for each node in the social network. Nodes with the smaller trustworthiness scores are likely to be Sybils. The comprehensive experiments are carried out in three real social networks. Experimental results show that the ranking mechanism is more robust than the state-of-art topological anti-Sybil models.

The rest of this chapter is organized as follows. Section 4.1 introduces the Sybil seed selection algorithm. Section 4.2 presents the unified ranking mechanism for combating Sybil attacks. The experimental results are illustrated in Section 4.3. Finally, Section 4.4 summarizes this work.

4.1 Sybil Seed Selection

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 [50]. However, few work is provided to leverage trust and distrust information to combat Sybil attacks. SybilDefender [42] 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 chapter, a Sybil seed selection algorithm is presented to produce reliable Sybil seeds, which can be utilized in the ranking mechanism to effectively distinguish non-Sybil from Sybil nodes.

This 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 thus can be used to improve the detection accuracy. Intuitively, corresponding to the basic assumptions—*fast mixing* and *small cut*, it can be observed 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, it can start from a Sybil seed and expand it by adding its neighboring nodes which are also identified as Sybils.

Additionally, SybilRank [15] is validated to be an effective and efficient algorithm for detecting Sybil nodes among existing anti-Sybil schemes. Hence, this algorithm is treated as a subroutine to seek for Sybil seeds. Algorithm 2 illustrates the detailed selection procedure for SybilRank. Let I_r 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 . $I(\cdot)$ 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, its spamicity value can be calculated according to its neighbors' class labels. The *spamicity* metric is defined as follows:

$$SP(v_i) = \frac{\sum_{j \in N(v_i)} |I(j, \eta)|}{|N(v_i)|} \quad (4.1)$$

Then, those nodes with $SP = 1$ are searched. 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 (4.2). Subsequently, from the *Suspend* set, tightly connected Sybil groups are searched as Sybil seed candidates.

$$O(v_i) = \begin{cases} 0 & \text{if } v_i \text{ is Sybil} \\ 1 & \text{if } v_i \text{ is Honest} \end{cases} \quad (4.2)$$

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

4.2 Unified Ranking Algorithm

The 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 seed 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 a 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..

To leverage trust and distrust in social networks, the unified ranking mechanism is presented based on a variant of the PageRank-like model, i.e. Personalized PageRank algorithm, which is an essential technique for ranking and prediction [37, 41, 63, 64, 66]. The 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 a unified trustworthiness for each node, which can be used to effectively discriminate non-Sybil from Sybil nodes.

4.2.1 Propagation Phase

Given the topological structure of a social network and a set of labeled nodes (honest and Sybil seeds), it can propagate trust/distrust scores from these seeds in the entire

```

Input:  $\mathbf{G}$  : 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   $I(v_i, \eta) = \begin{cases} 1 & \text{if } \mathbf{r}_\nabla(v_i) > \eta \\ 0 & \text{if } \mathbf{r}_\nabla(v_i) \leq \eta \end{cases}$  ;
5   $m = \sum_{v_i} \{v_i | I(v_i, \eta) == 0\}$ ;
6  for  $i \leftarrow 1$  to  $m$  do
7  |    $Source = Index(i)$ ;
8  |   Calculate  $SP$  for each node:  $SP(Source) = \frac{\sum_{j \in N(Source)} |I(j, \eta)|}{|N(Source)|}$ ;
9  |    $Suspend^* = |\{v_i | SP(v_i) == 1\}|$ ;
10 |    $Suspend = \{v_i | O(v_i) == 0, v_i \in Suspend^*\}$ ;
11 |    $s = |Suspend|$ ;
12 |   for  $k \leftarrow 1$  to  $s$  do
13 | |    $SeedCandidate = \phi$ ;
14 | |   add  $Suspend(k)$  to  $SeedCandidate$  ;
15 | |   for  $p \leftarrow 1$  to  $s$  do
16 | | |   if  $Suspend(p) \in N(SeedCandidate)$ ;
17 | | |   add  $Suspend(p)$  to  $SeedCandidate$ ;
18 |   if  $SeedCandidate == \phi$  then
19 | |    $k = k + 1$  ;
20 | |    $\theta = 0.01 * k$  ;
21 | |   repeat step 3-17;
22 Return  $SeedCandidate$ .
    
```

Algorithm 2: Sybil Seed Selection

graph according to the closeness of social relationships between neighboring nodes. The propagation process can be modelled in the following formula:

$$\mathbf{r}(v_i) = \alpha * \frac{\sum_{j \in N(v_i)} \mathbf{r}(j)}{|N(j)|} + (1 - \alpha) * d(v_i) \quad (4.3)$$

where $\mathbf{r}(v_i)$ denotes the score value of node v_i . α is the jump probability. Generally, $\alpha = 0.85$ [66]. 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, the initial scores towards the Sybil seeds are *negatively* biased. Thus, each node is

assigned a negative value after distrust propagation. And the corresponding initial vector d is defined in Equation (4.4) where SS denotes the set of Sybil seeds.

$$d(v_i) = \begin{cases} \frac{-1}{|SS|} & \text{if } v_i \in SS \\ 0 & \text{otherwise} \end{cases} \quad (4.4)$$

4.2.2 Integration Phase

In the 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 to combine them together such that the integrated value can identify Sybil nodes with lowest false rate? To address these problems, a simple but effective weighted scheme is utilized to obtain the final trustworthiness shown in Equation (4.5). Empirical analysis in the following section demonstrates that such combination model can correctly filter out most of Sybil nodes.

$$Total(v_i) = a * TR(v_i) + (1 - a) * DTR(v_i) \quad (4.5)$$

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 weights of trust and distrust values for the overall trustworthiness.

4.3 Experimental Evaluation

4.3.1 Datasets and Attack Model

Three data sets from popular online social networks are used to stimulate the honest region. Table 4.1 summarizes the properties of these datasets. Among them the Facebook graph is the same one as used in chapter 3. These rest of the social graphs have been commonly utilized to evaluate existing anti-Sybil schemes¹.

As illustrated in chapter 3, the two types of topological structures, i.e. *random graph* (ER model) [61] and *scale-free* (PA model) [62], are used to simulate attack regions. For each type of attack, 100 nodes are firstly generated to be *Sybil supporters*, which serve

¹<http://snap.stanford.edu/data/>

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 the average degree of 10. The number of additional Sybil nodes ψ varies from 100 to 1000. And, the number of attack edges connecting non-Sybil and Sybil regions is set to be 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 honest seeds.

Table 4.1: Dataset of social graph used in experiments

OSN	Node	Edge	Average Degree	CC
Facebook	9943	60,870	19.88	0.221
AstroPh	17,903	197,000	20	0.3040
HepTh	8,604	24,386	5.60	0.2434

4.3.2 Comparative Sybil Defense Methods

Two most recent and effective graph-based Sybil detection mechanisms are evaluated, namely SybilRank and ACL. SybilRank [15] 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 [30] 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, SybilRank algorithm is chosen to seek for Sybil seeds due to its better performance.

4.3.3 Evaluation Metrics

Three metrics are used to exhibit the effectiveness of the 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 in networks [15]. 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.

Table 4.2: The Sybil seeds selected in Algorithm 2 for different compromised networks, where θ denotes the cut-off threshold that is utilized to classify all the nodes into non-Sybil and Sybil categories.

Num.Sybil	Threshold θ	Sybil Seed Sets
<i>No.Sybil</i> = 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
<i>No.Sybil</i> = 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
<i>No.Sybil</i> = 400	0.02	2 two-seeds
	0.03	2 two-seeds
	0.04	1 two-seeds, 102-seeds cluster
	0.05	209-seeds cluster
<i>No.Sybil</i> = 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
<i>No.Sybil</i> = 600	0.02	0
	0.03	0
	0.04	5 two-seeds
	0.05	230-seeds cluster

4.3.4 Performance of Sybil Selection Algorithm

Based on the three real-world datasets including Facebook, AstroPh and HepTh described in Table 4.1, the experiments are conducted to evaluate the performance of the Sybil seed selection algorithm in seeking for reliable Sybil seeds. In particular, the state-of-art anti-Sybil approach, i.e. SybilRank, is chosen as a subroutine to give the initial detection result. In order to enhance this approach against target attacks, the graph pruning technique that is introduced in chapter 3 is incorporated into the SybilRank scheme. Subsequently, the Sybil seed selection algorithm is utilized to output the Sybil seeds based on the initial detection result and social relationships among nodes in the graph.

By adjusting the threshold θ to be used in partitioning the whole graph into non-Sybil and Sybil regions, the Sybil seed selection results are output for different attack scenarios

Figure 4.1: 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.

shown in Table 4.2. From all these results, it can be seen that this method can catch small but tightly connected Sybil seed 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 that more nodes should be manually inspected which is not applicable in a real case. Since the goal is to cope with the Sybil attack problem, the performance of using these Sybil seeds to detect Sybils is the major concern. In the following experiment, it can be verified that the factor of Sybil seeds' size has small impact on the defense performance.

4.3.5 Evaluation of Unified Ranking Mechanism

In this section, the effectiveness of the unified ranking mechanism is investigated, by varying the weighting parameter a and the size of Sybil seeds. In addition, to have a fair

comparison, another type of attack scenarios is simulated. To simulate the Sybil region, 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 the number of attack edges is gradually increased to a larger number 800. This attack type is called *large-scale attack*. Correspondingly, the attack type utilized in the previous experiments refers to *small-scale attack*.

The effectiveness of the unified ranking mechanism is firstly checked by varying the weighting parameter a . By performing the Sybil seed selection algorithm, two Sybil seeds are obtained for each attack scenario. The detection results are illustrated in Figure 4.1, where the value 0.9615 denotes the ratio between the size of Sybil and benign seeds. It can be obviously seen that the unified mechanism can possess strong defense ability against small-scale attack regardless of the choice of parameter a . Even though 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. 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, it can be concluded that the strength of weighting parameter's impact on the unified ranking mechanism depends on the size of Sybil region. Moreover, as discussed in [15], the SybilRank approach also achieves good detection results for combating large-scale attack. This phenomenon is attributed to the small cut assumption. As the Sybil region becomes larger, the small cut becomes increasingly narrow and distinct, which makes Sybil detection more effective. Finally, it is observed that the unified mechanism is robust against both large-scale and small-scale attacks by setting the weight parameter a to be 0.5.

Next, the factor of Sybil seeds' size is examined whether it plays an important role in uncovering more Sybil nodes. To explore the effect of this factor, the number of seeds is increased from 2 to a larger value. Additionally, to have a fair comparison, another two

Figure 4.2: The performance of the unified ranking mechanism by varying the size of Sybil seeds. 2-Supporters is the randomly selected Sybil seeds. k-Seeds represents k connected Sybil seeds output by Algorithm 2.

Sybil nodes are randomly selected in order to verify the usefulness of the selected Sybil seeds. By setting the parameter $a = 0.5$, the following detection results are obtained using the unified mechanism shown in Figure 4.2.

First, it is observed that the unified mechanism can achieve higher detection accuracy by incorporating a large Sybil cluster. Despite this case, the detection accuracy does not appear to heavily fluctuate with the increment of the number of Sybil seeds. It is speculated that 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 of 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 defending against Sybil attacks.

4.4 Summary

In this chapter, a unified ranking mechanism based on trust and distrust is proposed for resisting Sybil attacks. In this design, a Sybil seed selection algorithm is firstly presented to produce reliable Sybil seeds, based on the current anti-Sybil schemes. Then, honest and Sybil seeds are incorporated into a novel ranking model to output an integrated trustworthiness score for each node in the social network. These trustworthiness values can be utilized to effectively distinguish Sybil from non-Sybil nodes. Experiments on three real data sets demonstrate that the proposed ranking mechanism can achieve better performance and outperform the state-of-art Sybil defense approaches. This mechanism thus shades light on exploiting trust and distrust information for building secure and robust Sybil defense mechanisms.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This section concludes the completed work so far in the research topic of social network-based Sybil defense. As a new paradigm emerging in recent years, leveraging social network structures for defending against Sybil attacks draws more and more attention. This thesis proposes two suits of strategies, one of which is utilized to enhance the detection accuracy over existing topological Sybil defense approaches and the other is to build a unified ranking mechanism by leveraging trust and distrust for improving Sybil detection. These approaches respectively try to address three challenging problems faced by current anti-Sybil schemes: target attack, mixing sensitivity and distrust utilization.

More specifically, two effective approaches based on social topological structures are proposed in Chapter 3 to improve the detection accuracy for current Sybil defense mechanisms. The purpose is to address the target attack and mixing sensitivity problems, which render those topological anti-Sybil schemes ineffective. On the other hand, the research community lacks of a common insight to interpret these defense schemes. In this study, a novel perspective is provided to interpret Sybil defense as the problem of partially labelled classification. Then, based on this partially labelled classification framework, the graph pruning technique is introduced to handle the target attack problem, by exploiting local structure similarity between neighboring nodes. Moreover, in order to address mixing sensitivity issue and boost the detection accuracy over these Sybil defense schemes, a domain-specific graph regularization method is designed based on the relational property of social networks. Experimental results on four real social networks

demonstrate that the proposed techniques significantly improve the performance of four representative Sybil defense mechanisms.

In Chapter 4, a new unified ranking mechanism by leveraging trust and distrust in social networks is designed to defend against Sybil attacks. The goal is to resolve the insufficiency of bootstrapping from only honest nodes, which limits the detection accuracy over existing Sybil defense approaches. In addition, in many real world application (e.g. Digg), distrust information (e.g. negative feedback, malicious users) are almost provided or suspended by the systems. In this work, a Sybil seed selection algorithm is introduced to seek for a set of reliable Sybil seeds, based on current Sybil detectors. Then, to leverage trust and distrust, a unified ranking mechanism, which is a variant of the personalized PageRank algorithm, is designed to output an integrated trustworthiness score for each node in the social graph. Experimental results on real social graphs show that the proposed ranking approach can achieve high detection accuracy and outperform the state-of-art Sybil defense approaches.

In summary, as the initiative attempt of considering different network features to address the challenging problems faced by the current anti-Sybil schemes (i.e. mixing sensitivity, target attack and distrust utilization), this thesis serves to inspire the research community to build robust and secure Sybil defense mechanisms by exploring social topological features and applying them in real cases. Besides, this work mitigates the research gap between semi-supervised learning and social security fields, and thus is hoped to draw more attention towards this direction.

5.2 Future Work

Based on the current studies, the future research will be conducted in the following two aspects.

Extensions for the Graph Pruning and Regularization Techniques. As stated in Chapter 3, graph pruning and regularization techniques are effective to improve the detection accuracy of the current Sybil defense schemes. These two approaches are developed by exploiting local structure similarity and relational property in social networks. In future, the current work will be extended in the following three possible directions.

- **Utilization of other similarity metrics in graph pruning.** The pruning technique considers the number of common friends as the similarity metric to eliminate edges with low similarity values for diminishing attack capability around honest seeds. In future, it will be interesting to investigate the impact of other similarity metrics, e.g. cosine similarity, Jaccard similarity, for resisting target attacks.
- **Incorporation of pairwise features in graph regularization.** In the current work, all existing edges are treated uniformly due to the inherent trust assumption underlying social graphs. Local pairwise features are not yet involved. In general, pairwise features are considered as effective components for measuring the connectivity of neighboring nodes. Inspired by studies in [67–71], pairwise features will be incorporated into the graph regularization framework, representing a promising direction for enhancing the smoothing ability of regularized approaches.
- **Other complex problems and real applications.** The performance of the improved techniques in this work is conducted in simulated attack models. To more accurately measure the competency of these techniques and parameters, it is worthwhile to applying the Sybil detectors together with the improved techniques into practice, such as detecting criminal accounts in Twitter [50, 59].

Extensions for the Trust- and Distrust-based Unified Framework. In Chapter 4, the novel unified ranking mechanism is proposed by leveraging trust and distrust for combating Sybil attacks. This algorithm outputs an integrated trustworthiness score for each node in the social graph, which can effectively distinguish Sybil from non-Sybil nodes. In future, the current work will be extended in the following three potential directions.

- **Sybil seed selection without initial input.** Existing Sybil detectors play an important role in this work for outputting an initial detection result. An interesting direction is to seek for Sybil seeds only based on the topological structure of the social network without basing on those detectors in the first stage. Then, Sybil seeds are incorporated into the unified mechanism to output detection results. Finally, based on the detection results and social network, a set of reliable Sybil seeds can be searched.

- **Designing unified mechanisms based on trust and distrust.** As stated in Chapter 4, this study adopts a linear weighted scheme to combine trust and distrust scores together for deriving the final integrated trustworthiness. In future, it is interesting to develop other unified approaches by leveraging trust and distrust for defending against Sybil attack.
- **Coping with complicated attack models and real application.** Inspired by the study in [50], an interesting direction is to conduct empirical analysis on Sybil nodes' behaviours in real social networks and then to simulate these attack models. The robustness of unified mechanisms will be investigated under those attack models.

References

- [1] C. Xu, J. Zhang, K. Chang, and C. Long, “Uncovering collusive spammers in chinese review websites,” in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 979–988, ACM, 2013.
- [2] S. Jiang, J. Zhang, and Y.-S. Ong, “An evolutionary model for constructing robust trust networks,” in *Proceedings of the International Conference on Autonomous Agents and multi Agent Systems*, pp. 813–820, International Foundation for Autonomous Agents and Multiagent Systems, 2013.
- [3] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, “Sybilguard: defending against sybil attacks via social networks,” *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 4, pp. 267–278, 2006.
- [4] R. Levien and A. Aiken, “Attack-resistant trust metrics for public key certification,” in *Proceedings of the 7th USENIX Security Symposium*, pp. 229–242, San Antonio, Texas, 1998.
- [5] H. Zhang, A. Goel, R. Govindan, K. Mason, and B. Van Roy, “Making eigenvector-based reputation systems robust to collusion,” in *Proceedings of the International Conference on Algorithms and Models for the Web Graph*, pp. 92–104, Springer, 2004.
- [6] S. Liu, J. Zhang, C. Miao, Y.-L. Theng, and A. C. Kot, “An integrated clustering-based approach to filtering unfair multi-nominal testimonies,” *Computational Intelligence*, 2012.

REFERENCES

- [7] S. Liu, J. Zhang, C. Miao, Y.-L. Theng, and A. C. Kot, “iCLUB: An integrated clustering-based approach to improve the robustness of reputation systems,” in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1151–1152, International Foundation for Autonomous Agents and Multiagent Systems, 2011.
- [8] A. Cheng and E. Friedman, “Sybilproof reputation mechanisms,” in *Proceedings of the ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems*, pp. 128–132, ACM, 2005.
- [9] A. A. Irissappane, F. A. Oliehoek, and J. Zhang, “A pomdp based approach to optimally select sellers in electronic marketplaces,” in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 1329–1336, International Foundation for Autonomous Agents and Multiagent Systems, 2014.
- [10] A. A. Irissappane, S. Jiang, and J. Zhang, “A framework to choose trust models for different e-marketplace environments,” in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 213–219, AAAI Press, 2013.
- [11] H. Fang, J. Zhang, and N. M. Thalmann, “A trust model stemmed from the diffusion theory for opinion evaluation,” in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 805–812, International Foundation for Autonomous Agents and Multiagent Systems, 2013.
- [12] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, “SybilLimit: A near-optimal social network defense against sybil attacks,” in *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 3–17, IEEE, 2008.
- [13] G. Danezis and P. Mittal, “SybilInfer: Detecting sybil nodes using social networks,” in *Proceedings of the International Conference on Network and Distributed System Security Symposium*, 2009.
- [14] A. Mohaisen, N. Hopper, and Y. Kim, “Keep your friends close: Incorporating trust into social network-based sybil defenses,” in *Proceedings of the IEEE International Conference on Computer Communications*, pp. 1943–1951, IEEE, 2011.

REFERENCES

- [15] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, “Aiding the detection of fake accounts in large scale social online services,” in *Proceedings of the International Conference on Networked Systems Design and Implementation*, 2012.
- [16] A. Mohaisen, A. Yun, and Y. Kim, “Measuring the mixing time of social graphs,” in *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*, pp. 383–389, ACM, 2010.
- [17] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, “An analysis of social network-based sybil defenses,” *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 363–374, 2011.
- [18] S. Wasserman, *Social network analysis: Methods and applications*, vol. 8. Cambridge University Press, 1994.
- [19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and analysis of online social networks,” in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pp. 29–42, ACM, 2007.
- [20] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, “Analysis of topological characteristics of huge online social networking services,” in *Proceedings of the 16th International Conference on World Wide Web*, pp. 835–844, ACM, 2007.
- [21] V. Gómez, A. Kaltenbrunner, and V. López, “Statistical analysis of the social network and discussion threads in slashdot,” in *Proceedings of the 17th International Conference on World Wide Web*, pp. 645–654, ACM, 2008.
- [22] M. Cha, A. Mislove, and K. P. Gummadi, “A measurement-driven analysis of information propagation in the flickr social network,” in *Proceedings of the International Conference on World Wide Web*, pp. 721–730, ACM, 2009.
- [23] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, “Characterizing user behavior in online social networks,” in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, pp. 49–62, ACM, 2009.

REFERENCES

- [24] S. Gonzalez-Bailon, A. Kaltenbrunner, and R. E. Banchs, “The structure of political discussion networks: a model for the analysis of online deliberation,” *Journal of Information Technology*, vol. 25, no. 2, pp. 230–243, 2010.
- [25] N. Chiluka, N. Andrade, J. Pouwelse, and H. Sips, “Leveraging trust and distrust for sybil-tolerant voting in online social media,” in *Proceedings of the World Wide Web Workshop on Privacy and Security in Online Social Media*, p. 1, ACM, 2012.
- [26] A. Mislove, A. Post, P. Druschel, and P. K. Gummadi, “Ostra: Leveraging trust to thwart unwanted communication,” in *Proceedings of the International Conference on Networked Systems Design and Implementation*, vol. 8, pp. 15–30, 2008.
- [27] N. Tran, B. Min, J. Li, and L. Subramanian, “Sybil-resilient online content voting,” in *Proceedings of the International Conference on Networked Systems Design and Implementation*, pp. 15–28, USENIX Association, 2009.
- [28] N. Tran, J. Li, L. Subramanian, and S. S. Chow, “Optimal sybil-resilient node admission control,” in *Proceedings of the IEEE International Conference on Computer Communications*, pp. 3218–3226, IEEE, 2011.
- [29] Z. Cai and C. Jermaine, “The latent community model for detecting sybil attacks in social networks,” in *Proceedings of the International Conference on Network and Distributed System Security Symposium*, 2012.
- [30] L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi, “SoK: The evolution of sybil defense via social networks,” in *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 382–396, IEEE, 2013.
- [31] Y. Zhang, J. Wang, Y. Wang, and L. Zhou, “Parallel community detection on large networks with propinquity dynamics,” in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 997–1006, ACM, 2009.
- [32] B. Viswanath, M. Mondal, K. P. Gummadi, A. Mislove, and A. Post, “Canal: Scaling social network-based sybil tolerance schemes,” in *Proceedings of the European Conference on Computer Systems*, pp. 309–322, ACM, 2012.

REFERENCES

- [33] C.-N. Ziegler and G. Lausen, “Spreading activation models for trust propagation,” in *Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service*, pp. 83–97, IEEE, 2004.
- [34] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, “Propagation of trust and distrust,” in *Proceedings of the 13th International Conference on World Wide Web*, pp. 403–412, ACM, 2004.
- [35] M. De Cock and P. P. Da Silva, “A many valued representation and propagation of trust and distrust,” in *Proceedings of the International Workshop on Fuzzy Logic and Applications*, pp. 114–120, Springer, 2006.
- [36] T. DuBois, J. Golbeck, and A. Srinivasan, “Predicting trust and distrust in social networks,” in *Proceedings of the 3th IEEE International Conference on Privacy, Security, Risk and Social Computing*, pp. 418–424, IEEE, 2011.
- [37] C.-N. Ziegler and G. Lausen, “Propagation models for trust and distrust in social networks,” *Information Systems Frontiers*, vol. 7, no. 4-5, pp. 337–358, 2005.
- [38] P. Metaxas, “Using propagation of distrust to find untrustworthy web neighborhoods,” in *Proceedings of the International Conference on Internet and Web Applications and Services*, pp. 516–521, IEEE, 2009.
- [39] P. Victor, C. Cornelis, M. De Cock, and P. Pinheiro da Silva, “Gradual trust and distrust in recommender systems,” *Fuzzy Sets and Systems*, vol. 160, no. 10, pp. 1367–1382, 2009.
- [40] F. J. Ortega, J. A. Troyano, F. L. Cruz, C. G. Vallejo, and F. Enríquez, “Propagation of trust and distrust for the detection of trolls in a social network,” *Computer Networks*, vol. 56, no. 12, pp. 2884–2895, 2012.
- [41] F. Chung, A. Tsiatas, and W. Xu, “Dirichlet pagerank and ranking algorithms based on trust and distrust,” *Internet Mathematics*, vol. 9, no. 1, pp. 113–134, 2013.
- [42] W. Wei, F. Xu, C. C. Tan, and Q. Li, “SybilDefender: Defend against sybil attacks in large social networks,” in *Proceedings of the IEEE International Conference on Computer Communications*, pp. 1951–1959, IEEE, 2012.

REFERENCES

- [43] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, “Uncovering social network sybils in the wild,” in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pp. 259–268, ACM, 2011.
- [44] B. Taskar, E. Segal, and D. Koller, “Probabilistic classification and clustering in relational data,” in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 17, pp. 870–878, 2001.
- [45] J. Neville and D. Jensen, “Relational dependency networks,” *Journal of Machine Learning Research*, vol. 8, pp. 653–692, 2007.
- [46] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, “Collective classification in network data,” *Artificial Intelligence Magazine*, vol. 29, no. 3, p. 93, 2008.
- [47] J. R. Douceur, “The sybil attack,” in *Proceedings of the International Workshop on Peer-to-peer Systems*, pp. 251–260, Springer, 2002.
- [48] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, “You are who you know: inferring user profiles in online social networks,” in *Proceedings of the ACM International Conference on Web Search and Data Mining*, pp. 251–260, ACM, 2010.
- [49] D. Quercia and S. Hailes, “Sybil attacks against mobile users: friends and foes to the rescue,” in *Proceedings of the IEEE International Conference on Computer Communications*, pp. 1–5, IEEE, 2010.
- [50] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, “Analyzing spammers’ social networks for fun and profit: a case study of cyber criminal ecosystem on twitter,” in *Proceedings of the 21st International Conference on World Wide Web*, pp. 71–80, ACM, 2012.
- [51] O. Chapelle, B. Schölkopf, A. Zien, *et al.*, *Semi-supervised learning*, vol. 2. MIT Press Cambridge, 2006.
- [52] M. Seeger, “Learning with labeled and unlabeled data,” tech. rep., University of Edinburgh, 2001.

REFERENCES

- [53] O. Bousquet, O. Chapelle, and M. Hein, “Measure based regularization.,” in *Proceedings of the International Conference on Neural Information Processing Systems*, 2003.
- [54] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.
- [55] Y. Grandvalet and Y. Bengio, “Semi-supervised learning by entropy minimization.,” in *Proceedings of the International Conference on Neural Information Processing Systems*, 2004.
- [56] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, “Learning with local and global consistency.,” in *Proceedings of the International Conference on Neural Information Processing Systems*, vol. 16, pp. 321–328, 2003.
- [57] A. Blum and S. Chawla, “Learning from labeled and unlabeled data using graph mincuts,” in *Proceedings of the International Conference on Machine Learning*, 2001.
- [58] X. Zhu, Z. Ghahramani, and J. Lafferty, “Semi-supervised learning using gaussian fields and harmonic functions,” in *Proceedings of the International Conference on Machine Learning*, 2003.
- [59] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, “Understanding and combating link farming in the twitter social network,” in *Proceedings of the International Conference on World Wide Web*, pp. 61–70, ACM, 2012.
- [60] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, “Walking in facebook: A case study of unbiased sampling of osns,” in *Proceedings of the IEEE International Conference on Computer Communications*, pp. 1–9, IEEE, 2010.
- [61] W. Aiello, F. Chung, and L. Lu, “A random graph model for massive graphs,” in *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 171–180, Acm, 2000.

REFERENCES

- [62] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [63] B. Wu, V. Goel, and B. D. Davison, “Propagating trust and distrust to demote web spam,” *MTW*, vol. 190, 2006.
- [64] B. Wu and B. D. Davison, “Identifying link farm spam pages,” in *Proceedings of the International Conference on World Wide Web*, pp. 820–829, ACM, 2005.
- [65] H. Zhang, C. Xu, and J. Zhang, “Exploiting trust and distrust information to combat sybil attack in online social networks,” in *Proceedings of the 8th IFIP WG 11.11 International Conference on Trust Management*, pp. 77–92, Springer, 2014.
- [66] P. Berkhin, “A survey on pagerank computing,” *Internet Mathematics*, vol. 2, no. 1, pp. 73–120, 2005.
- [67] S. Oyama and C. D. Manning, “Using feature conjunctions across examples for learning pairwise classifiers,” in *Proceedings of the European Conference on Machine Learning*, pp. 322–333, Springer, 2004.
- [68] E. Pekalska, A. Harol, C. Lai, and R. P. Duin, “Pairwise selection of features and prototypes,” in *Proceedings of the International Conference on Computer Recognition Systems*, pp. 271–278, Springer, 2005.
- [69] A.-P. Ta, C. Wolf, G. Lavoue, A. Baskurt, and J.-M. Jolion, “Pairwise features for human action recognition,” in *Proceedings of the International Conference on Pattern Recognition*, pp. 3224–3227, IEEE, 2010.
- [70] P. Matikainen, M. Hebert, and R. Sukthankar, “Representing pairwise spatial and temporal relations for action recognition,” in *Proceedings of the European Conference on Computer Vision*, pp. 508–521, Springer, 2010.
- [71] C. Kang, X. Wang, J. Chen, C. Liao, Y. Chang, B. Tseng, and Z. Zheng, “Learning to re-rank web search results with multiple pairwise features,” in *Proceedings of the ACM International Conference on Web Search and Data Mining*, pp. 735–744, ACM, 2011.