Minimizing Trust Leaks for Robust Sybil Detection

János Höner [1 2] Shinichi Nakajima [2 3] Alexander Bauer [2 3] Klaus-Robert Müller [2 3 4 5] Nico Görnitz [2]

Abstract

Sybil detection is a crucial task to protect online social networks (OSNs) against intruders who try to manipulate automatic services provided by OSNs to their customers. In this paper, we first discuss the robustness of graph-based Sybil detectors *SybilRank* and *Integro* and refine theoretically their security guarantees towards more realistic assumptions. After that, we formally introduce adversarial settings for the graph-based Sybil detection problem and derive a corresponding optimal attacking strategy by exploitation of trust leaks. Based on our analysis, we propose transductive Sybil ranking (*TSR*), a robust extension to *SybilRank* and *Integro* that directly minimizes trust leaks. Our empirical evaluation shows significant advantages of *TSR* over state-of-the-art competitors on a variety of attacking scenarios on artificially generated data and real-world datasets.

1. Introduction

The sheer number and variety of online social networks (OSN) today is staggering. Although the purpose and the shaping of these networks vary generously, the majority of them has one aspect in common: the value of most OSNs is in its user data and the information that one can infer from the data. This, unfortunately, results in a big incentive for culprits to intrude OSNs and manipulate their data. One popular method of intruding and attacking an OSN is referred to as *Sybil attack*, where the intruder creates a whole bunch of fake (*Sybil*) accounts that are all under the attacker's control. The intruder's influence over the OSN is multiplied by the number of accounts created, which opens possibilities of manipulation typically for gaining some monetary advantage in the end.

The term, *Sybil attack*, was coined by Douceur (2002) who showed that this kind of attack will be always possible unless a *trusted agency certifies identities*. Unfortunately, this approach is orthogonal to how OSNs grow. The threshold of registration must be as low as possible to attract as many new users as possible. On the other hand, Sybil attacks can damage the value of OSNs significantly, which has been proved by the fact that Facebook shares dropped in 2012 after the company revealed that a significant share of its network is made up by Sybil accounts (The Associated Press, 2012).

There exists a number of "classic" feature-based solutions (Stein et al., 2011; Cao et al., 2012; Stringhini et al., 2010; Yang et al., 2014). However, up until now, it remains an unsolved problem as those methods can be evaded by cleverly designed attacking schemes (Bilge et al., 2009; Boshmaf et al., 2011; Wagner et al., 2012; Lowd & Meek, 2005) and manual detection is too expensive, time consuming, and simply unfeasible in large OSNs (Cao et al., 2012).

More recent graph-based Sybil detection methods assume that *honest* (non-Sybil) nodes form a strongly connected subgraph and attackers can establish a limited amount of edges which leads to a sparse cut between the honest subraph and the Sybil nodes. The majority of the graph-based methods define *trusted nodes*, which the defender is sure to be honest, and use random walks (Yu et al., 2010; Danezis, 2009; Cao et al., 2012) or other typical graph-based algorithms like breadth-first-search (Tran et al., 2011) and belief propagation (Gong et al., 2014) to convey *trust* from the trusted nodes. A node is identified as Sybil if sufficient ammount of *trust* is not delivered to it. Among random-walk based approaches, *SybilRank* is known to be the state-of-the-art, of which the performance is theoretically guaranteed (Cao et al., 2012). However, the theory holds only under unrealistic topological assumptions of the network. In this paper, we show that the same theoretical guarantee can be obtained under more realistic situations.

We further dicuss the robustness of the random walk approach against adversarial strategies. To this end, we formally introduce adversarial settings for graph-based Sybil

[1]MathPlan, 10587 Berlin, Germany [2]Machine Learning Group, Berlin Institute of Technology, 10587 Berlin, Germany [3]Berlin Big Data Center [4]Max Planck Society [5]Korea University. Correspondence to: János Höner <janos.hoener@campus.tu-berlin.de>, Nico Görnitz <nico.goernitz@tu-berlin.de>, Klaus-Robert Müller <klaus-robert.mueller@tu-berlin.de>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

detection and derive an optimal attacking strategy that is based on the exploitation of trust leaks. Based on our analysis, we propose a transductive Sybil ranking (*TSR*), an integrated approach capable of adjusting edge weights based on sampled trust leaks. We empirically show good performance of *TSR* against the state-of-the-art baselines on a variety of attacking scenarios using artificially generated data as well as real-world Facebook data.

2. Preliminaries

We are given a graph $\mathcal{G} = (V, E)$ consisting of nodes V and pairwise edges E between nodes. We denote $\mathcal{G}_S = (V_S, E_S)$ the *Sybil* sub-graph, $\mathcal{G}_H = (V_H, E_H)$ the disjunct *honest* sub-graph, and $V_T \subseteq V_H$ our trusted (verified non-Sybil nodes) random walk seed nodes. E_A is the set of edges connecting any node in $\mathcal{G}_S$ and any node in $\mathcal{G}_H$.

Sybil Rank is considered the state-of-the-art graph-based method to detect Sybil accounts as it outperformed all its contestants (Cao et al., 2012). It is also based on random walks and operates solely on the topology of the graph. *Sybil Rank* starts from the initial distribution $\{p_0^{(i)} \in [0,1]\}_{i=1}^{|V|}$ (without superscript refers to a vector containing all elements), in which "trust" is assigned to the known honest nodes V_T:

$$p_0^{(i)} = \begin{cases} \frac{1}{|V_T|} & \text{if } v_i \in V_T, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then, it "propagates" the trust via a short (k steps) random walk:

$$p_k^\top = p_{k-1}^\top Q = \cdots = p_0^\top Q^k \,, \tag{2}$$

where $Q \in \mathbb{R}^{|V| \times |V|}$ is the transition matrix through the edges with $Q_{i,j} = (\sum_{j'} \mathbf{1}[(i,j') \in E])^{-1}$, if $(i,j) \in E$, and else 0. It is known that the stationary distribution $\pi \equiv p_\infty$ is the normalized degree distribution (Behrends, 2000)

$$\pi^\top = \left(\frac{\text{deg}(v_1)}{\text{Vol}(V)}, \ldots, \frac{\text{deg}(v_{|V|})}{\text{Vol}(V)} \right) \,, \tag{3}$$

where $\text{deg}(v)$ is the degree of node v, i.e., the number of all incident edges of v, and $\text{Vol}(V) = \sum_{v \in V} \text{deg}(v)$ is the sum of the degrees for all nodes in V. SybilRank conpensates the effect of degrees, and use the degree-normalized probability

$$p^{(i)} = p_k^{(i)} / \pi^{(i)} \tag{4}$$

as the *ranking score*, where a high ranking indicates a high probability of being an honest node.

Essentially, SybilRank relies on the assumption that the total number of attacking edges is bounded. Under this assumption, only a small fraction of the trust is propagated through the sparse cut between the honest network and the Sybil nodes during the short random walk, while ”trust” go through the ”non-trusted” honest nodes through the dense connections within the the honest subgraph.

Boshmaf et al. (2016) developed *Integro* to cope with a larger number of attacking edges. To this end, *Integro* introduces weights on the edges to bias the random walk, where the weights are determined after its pre-processing step to detect *victims*. Here a victim is defined as a node that established a connection to one of the Sybil nodes. The set of all victim nodes is defined by $V_v = \{v \in V_h : \exists (v, s) \in E_A\}$. After the detection step, *Integro* lowers the weights to all incident edges to the detected victims, which prevents the trust to propagate through victim nodes. As the victims form a natural border between the honest and the Sybil graph, this reduces the overall flow of trust into the Sybil graph. Boshmaf et al. found that traditional feature-based classification methods yield good and robust detection of victims. A notable advantage against the feature-based Sybil detection is that, unlike Sybils, victims generally do not behave adversarial, as they don't have any incentive to ”hide”.

3. SybilRank's Security Guarantee Under More Realistic Assumptions

Cao et al. (2012) gave a security guarantee for *SybilRank*. Let $g := |E_A|$ be the number of attacking edges and $n := |V|$ be the number of all nodes in the graph. their theory relies on the notion of trust leaks.

Definition 1. *(Trust leaks) Let $r_{k'} = \sum_{i \in V_H} p_{k'}^{(i)}$ be the trust that remains in the honest graph after k' random walk steps. We call $l = \sum_{k'=1}^{k} (r_{k'+1} - r_{k'})$ the absolute trust leak. Assume that the attacking edges are created randomly, following a distribution $\alpha(E_A)$. We call $C_H(k') = \mathbb{E}_{\alpha(E_A)}[\frac{r_{k'+1} - r_{k'}}{r_{k'}}]$ the expected relative trust leak.*

$C_H(k')$ is actually a constant with respect to k' under reasonable assumptions on $\alpha(E_A)$. The following lemma has been proved:

Lemma 1. *(Cao et al., 2012) Assume that the graph $\mathcal{G}$ is created randomly, following the configuration model (Molloy & Reed, 1995). Then, the expected relative trust leak in each iteration is given by $C_H = \frac{g}{vol(V_H)}$.*

This leads to a theoretical guarantee of SybilRank.

Theorem 1. *(Cao et al., 2012) Assume that the graph $\mathcal{G}$ is created randomly, following the configuration model. The total number of Sybils that are ranked higher than non-Sybils by SybilRank is $\mathcal{O}(g \log n)$.*

Theorem 1 implies good performance of SybilRank, but it holds under the assumption that the attacking edges are created in the same process as the honest graph,[1] which is not realistic.

Below, we show that the same guarantee is obtained under the following more realistic assumption:

Assumption 1. *The graph $\mathcal{G}$ is constructed by the following steps:*

1. *Honest graph $\mathcal{G}_H$ construction: $\mathcal{G}_H$ is connected, non-bipartite, and scale free, i.e., the degree distribution follows the power law distribution.*
2. *Sybil graph $\mathcal{G}_S$ construction: The topology of $\mathcal{G}_S$ is arbitrary.*
3. *Attacking edges E_A generation: The attacking edges are genarated on all possible edges $E_A \subset V_S \times V_H$ between the honest and the Sybil subgraphs with equal propability.*

Under Assumption 1, evaluating the expected trust leak is less straightforward. Nevertheless, we can show that it results in the same formal security guarantee stated in Theorem 1.

To properly compute the expected trust leak, the following random variables are defined. X_v counts the number of attacking edges incident to node v, $Y_v = \pi(v)\frac{X_v}{\deg(v,\mathcal{G}_H)+X_v} = \pi(v)\frac{X_v}{\deg(v,\mathcal{G})}$ is the trust leak in node v and $Z = \sum_{v \in V_H} Y_v$ is the total trust leak. Note that here $\pi(v)$ is the current amount of trust in node v and not the stationary distribution of the random walk. This notation is used to avoid confusion with the probability mass function denoted by P. From Assumption 1 it follows that X_v is hypergeometrically distributed (Tuckwell, 1995) with the following parameters: the *population size*: $N = |V_H \times V_S|$, *successes*: $K = |\{v\} \times V_S|$, and the *draws* $n = |E_A|$. Let $g := |E_A|$ be the number of attacking edges. Moreover, let $n_H := |V_H|$ and $n_S := |V_S|$ denote the number of honest nodes and Sybil nodes, respectively.

The probability mass function of X_v is given by $P(X_v = k) = \binom{K}{k}\binom{N-K}{n-k}/\binom{N}{n}$ and according to Tuckwell (1995), its expected value can be computed by $\mathbb{E}[X_v] = n\frac{K}{N} = |E_A|\frac{|\{v\}\times V_S|}{|V_H \times V_S|} = \frac{|E_A|}{|V_H|} = \frac{g}{n_H}$. The final goal is to compute the expected value of Z. The linearity of the expected value yields $\mathbb{E}[Z] = \sum_{v \in V_H} \mathbb{E}[Y_v]$ and for the expected value of Y_v we get

$$\begin{aligned}\mathbb{E}[Y_v] &= \tfrac{\pi(v)}{\deg(v,\mathcal{G})} \textstyle\sum_{k=0}^{\infty} kP(X_v = k) \\ &= \tfrac{\pi(v)}{\deg(v,\mathcal{G})}\mathbb{E}[X_v] = \tfrac{\pi(v)}{\deg(v,\mathcal{G})}\tfrac{g}{n_H}\ .\end{aligned}$$

Using this result, the expected value of Z becomes $\mathbb{E}[Z] = \sum_{v \in V_H} \mathbb{E}[Y_v] = \frac{g}{n_H}\sum_{v \in V_H}\frac{\pi(v)}{\deg(v,\mathcal{G})}$, where the right hand side still contains a sum that needs to be evaluated individually for each node to compute its actual value. In order to "average out" this sum, we rely on the assumption that the honest nodes $\mathcal{G}_H$ is power law-distributed (Barabási, 2009). To do this, a new random variable D_v is introduced which measures the degree of v. Then, the assumption results in the probability of a node having a degree of d being $P(D_v = d) = \frac{d^{-\gamma}}{\zeta(\gamma)}$, where ζ is the Riemann zeta function $\zeta(s) := \sum_{n=0}^{\infty} n^{-s}$ (Barabási, 2009).

With this expression, it is possible to "average out" the exact topology of the graph by computing the expected value with respect to the newly defined random variable D_v:

$$\begin{aligned}\mathbb{E}[Z] &= \tfrac{g}{n_H}\textstyle\sum_{d=1}^{\infty}\sum_{v \in V_H}\tfrac{\pi(v)}{d}P(D_v = d) \\ &= \tfrac{g}{n_H}\textstyle\sum_{v \in V_H}\pi(v)\sum_{d=1}^{\infty}\tfrac{1}{d}\tfrac{d^{-\gamma}}{\zeta(\gamma)} \\ &= \tfrac{g}{n_H}\textstyle\sum_{v \in V_H}\tfrac{\pi(v)}{\zeta(\gamma)}\sum_{d=1}^{\infty} d^{-(\gamma+1)} \\ &= \tfrac{g}{n_H}\tfrac{\zeta(\gamma+1)}{\zeta(\gamma)}\underbrace{\textstyle\sum_{v \in V_H}\pi(v)}_{\text{Total trust in the honest graph}}.\end{aligned}$$

This yields the following lemma.

Lemma 2. *Under Assumption 1 the expected relative trust leak in each iteration of the random walk is given by*

$$\tilde{C}_H = \tfrac{g}{n_H}\underbrace{\tfrac{\zeta(\gamma+1)}{\zeta(\gamma)}}_{=:e}$$

where $e < 1$ is a constant that depends on the parameter of the assumed power law distribution for the degree distribution.

Although Lemma 2 gives a different expected relative trust leak from Lemma 1, the fact that the maximum number of connection for each node is bounded in every OSN and therefore $\mathcal{O}(n_H) = \mathcal{O}(\text{vol}(V_H))$ leads to the same asymptotic behavior as Theorem 2:

Theorem 2. *Under Assuption 1, the total number of Sybils that are ranked higher than non-Sybils by SybilRank is $\mathcal{O}(g \log n)$.*

This result explicitly shows that, asymptotically, *SybilRank*'s security guarantee remains unchanged even under more realistic Assumption 1.

4. Adversarial Strategies

In this section, we discuss adversarial strategies against graph-based Sybil detection methods.

Attacker's Action Although attackers in general can take variety of actions, we restricts their action to adding attacking edges.

[1] This assumption is not explicitly stated in Cao et al. (2012), but apparent from their derivation.

Definition 2 (Attacking strategy)**.** *Given an honest graph $\mathcal{G}_{\mathcal{H}}$ and a Sybil graph $\mathcal{G}_{\mathcal{S}}$, an attacking strategy describes the set of attacking edges established by the intruder.*

The cost of action is measured by the number of attacking edges.

Attacker's Knowledge Generally, we focus on adversarial attacks against random walk based approaches. That is, an attacker's strategy for establishing edges from Sybil nodes to honest nodes in order to cloak an attacker's Sybil sub-network. For analysis, we assume different levels of knowledge that the attacker has on the defender's strategy and information:

A.1 Strategy only.

A.2 Strategy + topology.

A.3 Strategy + topology + trusted nodes (positively labeled nodes).

B.1 Strategy + topology + trusted nodes (positively labeled nodes) + untrusted nodes (negatively labeled nodes).

Here, we divided the level of access to inside information for the attacker into two groups. In group A (i.e., A.1, A.2, A.3) attackers are able to gather sophisticated information based on *publicly available sources*, whereas in group B (i.e., B.1) either some back-channel provides *non-public* information (e.g. defender marked Sybil nodes based on their analysis), or, the attackers are provided with all information *visible to the defenders*.

Clearly, it is too hard, if not impossible, to have an out-of-the-box solution for the setting described in group B and we therefore resort our analysis on the settings in group A. In the first case (A.1), no efficient adversarial strategies for graph-based random walk approaches is possible. The attackers must build up sufficient attacking edges to trusted nodes in order to absorb enough trust. In A.3 (and A.2 as a special case) on the other hand, the attacker gained enough information to guide the creation of attacking edges efficiently. This paper focuses on this most interesting situation. More specifically, we assume the following: the intruder knows defender's strategy (algorithm details), the topology of the honest graph, and the set of trusted nodes, i.e., she knows about $\mathcal{G}_H = (V_H, E_H)$ and V_T. Based on that knowledge the intruder can establish attacking edges to honest nodes of her choice with the goal to create an attacking scenario where the applied defense method fails. Although the exact topology of the Sybil graph is not specified any further, for the following results it is assumed that it is designed in a way that suits the intruder well.

Attacker's Goal Attackers can have various final goals, e.g., spamming honest users to earn money, feeding wrong information to honest nodes, stealing nonpublic information, damaging countries/companies, etc. Depending on the goal, the objective of the *optimal* strategy can differ. We assume that attacker's try to maximize their influence and hence, have an inherent need to increase the number of attacking edges.

Random-walk based approaches such as *SybilRank* and *Integro* rely on the fact that the absolute trust leak l from the honest graph to the Sybil graph is small (i.e., below the amount needed to reach the stationary distribution within the Sybil sub-graph) which ensures low trust scores for the Sybil nodes. However, if enough trust is being propagated to the Sybil graph, trust values will be close to the stationary distribution in the Sybil graph as well as in the honest graph. Consequently, the degree-normalized ranking values will be similar to the ones in the honest graph, which makes Sybil nodes indistinguishable from honest nodes and therefore disables the detector.

Definition 3 (Disabling Attacking Strategy)**.** *Let $\mathcal{G}_{\mathcal{H}}$ and $\mathcal{G}_{\mathcal{S}}$ be the honest graph and the Sybil graph. Let $l : 2^E \rightarrow \mathbb{R}$ be the absolute trust leak as a function of an attacking strategy. Then, an attacking strategy $E_A \subset V_H \times V_S$ is said to be disabling if*

$$l(E_A) \geq t_d, \tag{5}$$

where t_d is the disabling threshold, which depends on the topology of the Sybil graph and the detection algorithm.

Surely, an attacker does not aim for just any disabling strategy but for the one that comes at the lowest cost. As the cost of an attacking strategy is assumed to be increasing in the number of attacking edges, an optimal/minimal disabling strategy is given by the following definition.

Definition 4 (Optimal Disabling Strategy)**.** *An attacking strategy A_E is said to be optimal if it is the solution to the following optimization problem:*

$$\begin{aligned} \min_{E_A \subset V_H \times V_S} \quad & |E_A| \\ \textit{s. t.} \quad & l(E_A) \geq t_d. \end{aligned} \tag{6}$$

To solve this, the disabling threshold t_d and the trust leak function must be known to the attacker. Ignoring the edge weights (which are unknown to the attacker) the amount of trust needed within the Sybil graph to reach the stationary distribution of the random walk is given by $t_d = \sum_{v_i \in V_S} \pi_i = \frac{\text{vol}(V_S)}{\text{vol}(V)}$. To exactly evaluate $l(E_A)$ the entire random walk needs to be simulated which is infeasible for the attacker without knowing its exact length and the edge weights. A useful estimate is to consider only the first iteration. The computation of this value is feasible and the trust leak per attacking edge is by far the largest in the first iteration because all the trust is concentrated in the relatively small subset of trusted nodes V_T. The trust leak in the first iteration $\tilde{l}(E_A)$ is given by $l(E_A) = \sum_{v \in V_T} \frac{\kappa(v)}{\deg(v, G_H) + \kappa(v)}$, where $\kappa(v)$ is the attacking degree (i.e., the number of attacking edges) of node v. This leads to a greedy strategy where the intruder iteratively adds those attacking edges which produce the largest increase in $\tilde{l}$. In the following the term *adversarial strategy/attacker* refers to this greedy strategy.

5. Proposed Method

In this section, we propose our new method and derive its efficient solver. Our method is specifically designed to cope with a large number of attacking edges by minimizing "trust leaks", that is, minimizing a sampled trust leak by adjusting the edge weights—a missing mechanism for *SybilRank* and *Integro*.

Transductive Sybil Ranking Combining the approach of Backstrom & Leskovec (2011) and *SybilRank* (Cao et al., 2012), our proposed method, called transductive Sybil ranking (*TSR*), tries to leverage potential prior knowledge, *negative labels*, to bias a short random walk so that random walk methods work even with the existence of a large number of attacking edges.

Assume that all nodes carry attributes and $n \leq |V|$ nodes are additionally attached with label information, i.e., the defender knows a subset of nodes are honest, and another subset of nodes are sybil. More formally, the defender is given labeled nodes $\mathcal{L} := \{(x_i, y_i) \in \mathcal{X} \times \{+1, -1\}\}_{i=1}^{n}$ and unlabeled nodes $\mathcal{U} := \{x_i \in \mathcal{X}\}_{i=n+1}^{|V|}$. Since only the honest nodes can be trusted, $V_T \subseteq \{v_i \in V; y_i = +1\}$ holds.

We define an edge feature function $\psi_{u,v}$ between nodes u and v as $\psi_{u,v} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Y}$. A corresponding parameterized, non-negative scoring function $f_w : \mathcal{Y} \rightarrow \mathbb{R}^{+}$ is learned during training and applied as edge weight $a_{u,v} = f_w(\psi_{u,v})$ in the weighted adjacency matrix $Q \in \mathbb{R}^{|V| \times |V|}$:

$$Q_{u,v} = \begin{cases} \frac{a_{u,v}}{\sum_x a_{u,x}} & \text{if } (u,v) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Throughout our experiments, we restrict ourselves to the following differentiable edge feature function:

$$f_w(\psi_{u,v}) = (1 - \exp(-w^\top \psi_{u,v}))^{-1}. \tag{8}$$

Once the transition matrix is fixed, The remaining procedure is the same as SybilRank. Namely, starting form the initial distribution (1), k-steps random walk (2) is applied with the transition matrix (7). After that, the degree-normalized ranking probability (4) is used for classification. However, we are also given negatively labeled nodes, which are used to train the parameter w of the edge feature function (8), so that $p^{(i)} < p^{(j)},\ \forall\, i, j \in \{1, \ldots, n\}$ with $y_i = -1$ and $y_j = +1$. In the spirit of regularized risk minimization (Vapnik, 1999), this problem is formalized as follows:

Definition 5 (*TSR* optimization problem)**.** TSR *solves a quadratically regularized, non-convex optimization problem with generic loss-functions* $h : [0,1] \times \{+1, -1\} \rightarrow \mathbb{R}$*:*

$$\underset{w}{\text{minimize}}\, F(w) = \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} h(p^{(i)}(w), y_i)\,. \tag{9}$$

Using the notion of $p^{(i)}(w)$ visually indicates that node ranking probabilities p are (non-linearly) dependent on the parameter vector w. As for the choice of loss-functions, we examine the following:

- **Wilcoxon-Mann-Whitney (WMW) loss** (Yan et al., 2003). WMW maximizes the area under the ROC curve:

$$h(p,y) = \sum_{j=1}^{n} \mathbf{1}[y = +1 \wedge y_j = -1] \left(1 + \exp^{-\frac{p - p_j}{b}}\right)^{-1}.$$

- **Smooth hinge-loss variant** A smooth variant of the classical support vector machine hinge-loss with two additional parameters: a decision boundary $b \in \mathbb{R}$ and a scaling parameter $a \in \mathbb{R}$:

$$h(p,y) = \begin{cases} \frac{1}{2} - y(ap - b) & \text{if } y(ap-b) \leq 0, \\ \frac{1}{2}(1 - y(ap-b))^2 & \text{if } 0 < y(ap-b) \leq 1, \\ 0 & \text{if } 1 < y(ap-b). \end{cases}$$

In this work, we focus on smooth, differentiable loss-functions only, ensuring fast convergence to local optima via gradient-based methods, i.e., fast second-order methods (BFGS). A pivotal point is hence, to assess the gradient w.r.t. w.

Gradient Computation The remaining of this section is dedicated to the derivation of the gradient:

$$\frac{\partial F(w)}{\partial w} = \frac{\partial \lambda \|w\|^2}{\partial 2w} + \sum_i^n \frac{\partial h(p^{(i)}(w), y_i)}{\partial w}\,,$$

where the loss-function h can be further split into $\frac{\partial h(p^{(i)}(w), y_i)}{\partial w} = \frac{\partial h(p^{(i)}(w), y_i)}{\partial p^{(i)}(w)} \frac{\partial p^{(i)}(w)}{\partial w}$. Since we constrained ourselves to differentiable loss-function $h(p, y)$, the partial derivative w.r.t. p can be calculated rather straightforward. More complicated is the evaluation of

$$\frac{\partial p^{(i)}}{\partial w} = \frac{\partial}{\partial w} \frac{p_k^{(i)}}{\pi^{(i)}} = \left(\frac{\partial p_k^{(i)}}{\partial w} \pi^{(i)} - p_k^{(i)} \frac{\partial \pi^{(i)}}{\partial w} \right) {\pi^{(i)}}^{-2}\,. \tag{10}$$

The derivative of the i-th component of π is given by:

$$\frac{\partial\pi^{(i)}}{\partial w} = \left(\frac{\partial \mathrm{deg}^*(v_i)}{\partial w}\mathrm{vol}(V) - \frac{\partial \mathrm{vol}(V)}{\partial w}\mathrm{deg}^*(v_i)\right)\mathrm{vol}(V)^{-2},$$

where $\frac{\partial \mathrm{deg}^*(v_i)}{\partial w} = \sum_{\substack{e\in E\\ v_i\in e}} \frac{\partial a_e}{\partial w} = \sum_{\substack{e\in E\\ v_i\in e}} \frac{\partial f_w(\psi_e)}{\partial w}$ and $\frac{\partial \mathrm{vol}(V)}{\partial w} = 2\sum_{e\in E}\frac{\partial a_e}{\partial w} = 2\sum_{e\in E}\frac{\partial f_w(\psi_e)}{\partial w}$. As f_w is said to be differentiable the only part of Eq. (10) that remains is the Jacobian $\partial p_k/\partial w$.

Theorem 3. *The derivative $\partial p_k/\partial w$ for $k \geq 1$ is given by:*

$$\frac{\partial p_k}{\partial w} = \left(\sum_{l=0}^{k-1} p_l Q^{k-1-l}\right)\frac{\partial Q}{\partial w}. \tag{11}$$

(the proof is given in Appendix A). The derivative of Q, defined in Eq. (7), is given by

$$\frac{\partial Q_{uv}}{\partial w} = \begin{cases} \frac{\frac{\partial a_{uv}}{\partial w}\sum_x a_{ux} - a_{uv}\sum_x \frac{\partial a_{ux}}{\partial w}}{\left(\sum_x a_{ux}\right)^2} & \text{if } (u,v)\in E, \\ 0 & \text{otherwise.} \end{cases}$$

This completes the computation of the gradient and enables the application of gradient-based methods, i.e., BFGS to find a (locally) optimal estimate $\hat{w}$. By using this estimate, *TSR* weights the whole graph, with which a short random walk is performed to obtain the final ranking p.

Robustness of *TSR* against Attacks By using the negative label information, our *TSR*, in principle, monitors "trust leak" through random walk, and adjusts the edge weights so that the leak is minimized. As a result, the weights tend to be lower on the attacking edges (to reduce the propagation), and higher on the Sybil edges (to boost the stationary distribution). Thus, we can expect that our *TSR*, which is an advanced integrated method, is more robust against attacks than the SybilRank and the two-step Integro approach.

6. Empirical Evaluation on Synthetic Data

To assess the robustness of the proposed method and the baseline methods, we generate artificially network topology and edge and node attributes in order to have full control of the underlying ground truth. We separately create two graphs, the honest and the Sybil graph. Both use the generation method proposed by Holme & Kim (2002) for scale free networks. Node features are generated randomly and correlated through dependency injection. The edge features function $\psi_{u,v}$ simply stacks node features of the two adjacent nodes x_u and x_v (see Appendix B for more details). Connections between Sybil and honest graphs are established according to a *random attacking strategy* that iteratively adds attacking edges randomly, i.e., equally distributed on the set of all possible attacking edges $V_H \times V_S$ or a *adversarial attacking strategy* that solves Problem (6) for optimal attacks. This strategy only chooses an honest node to be attacked next and the corresponding Sybil node is chosen randomly (equally distributed on the set of all Sybil nodes V_S). We test our method, *TSR*, using the proposed loss functions and compare against the state-of-the-art methods *SybilRank* and *Integro*. As *Integro* depends on a preceding victim prediction, we simulated one that achieves highest possible rankings (ROC-AUC close to 1.0).[2]

Random Attacking Strategy We generated a sample network ($|V_H| = 200$ and $|V_S| = 30$) and select 15 honest nodes and 8 Sybil nodes randomly, which will be used as labeled examples for our *TSR*. The labeled honest nodes are also used as the set V_T of trusted seeding nodes for the random walks in all methods. We evaluate the performance in terms of ROC-AUC-values for the computed ranking. This procedure was repeated 20 times for varying number of attacking edges (10-200 edges). Figure 1 shows ROC-AUC curves for all methods under the random attacking setting. We can obsreve that our *TSR*, regardless of the choice of loss function, performs superior to the other methods. *Integro*'s accuracy deteriorates fast but still has an edge over *SybilRank* up to the point where the ROC-AUC-value reaches 0.5. After that *SybilRank* and *Integro* essentially perform similar.

Adversarial Attacking Strategy For the adversarial setting, we ran the same benchmarks but this time attacking edges were added according to the adversarial attacking strategy. Due to the much more aggressive setting, we varied the number of attacking edges from 1-40 and repeated this procedure 20 times to report averaged ROC-AUC accuracies. The results are depicted in Figure 2. All choices of loss functions outperform *SybilRank* and *Integro* clearly. The results confirm our considerations that *SybilRank*'s performance drops fast and steep as soon as a certain amount of attacking edges is established. *Integro* behaves more robust than *SybilRank*, but, ultimately, must resign after a few more attacking edges. Again, our *TSR* is significantly more robust against adversarial attacks and can withstand higher number of attacking edges until its performance finally deteriorates.

7. Empirical Evaluation on Real-world Data

We also evaluated our method on a sample of the Facebook graph Leskovec & Mcauley collected from survey participants using the Facebook app. The dataset includes the topology ($|V| = 4039$ users and $|E| = 88234$ friend-

[2] *SybiRank*, *Integro*, and *TSR* rely on different information, and therefore, the fairness of comparison is not trivial. We discuss this issue in Appendix C.

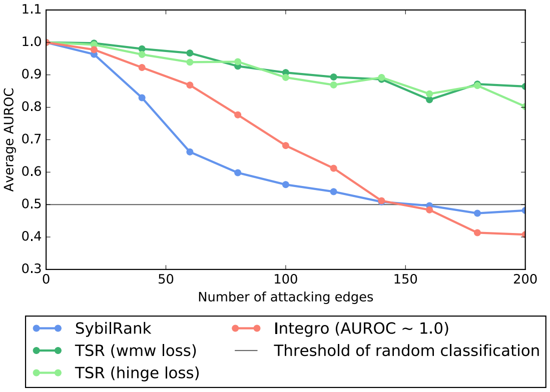

Figure 1. **Results for the random attacking setting.** Accuracy in terms of ROC-AUC for all methods on the generated graph ($|V_H| = 200$, $|V_S| = 30$) with 20 repetitions and varying number of attacking edges.

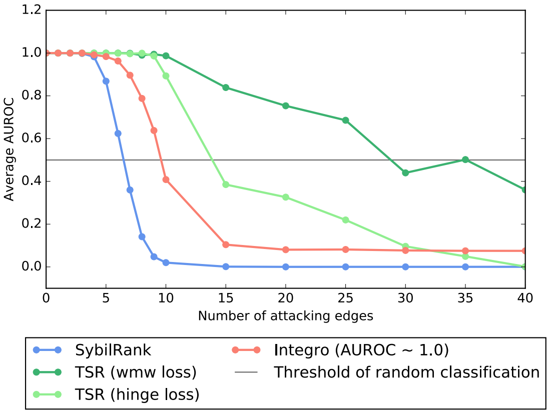

Figure 2. **Results for the adversarial setting.**

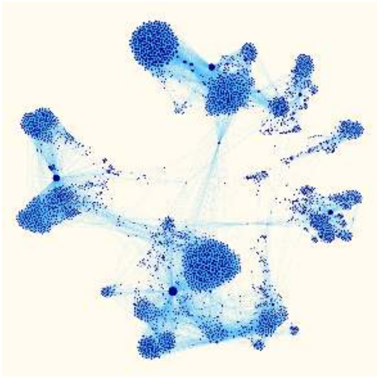

Figure 3. **Visualization of the Facebook graph.** The size of a node represents its degree.

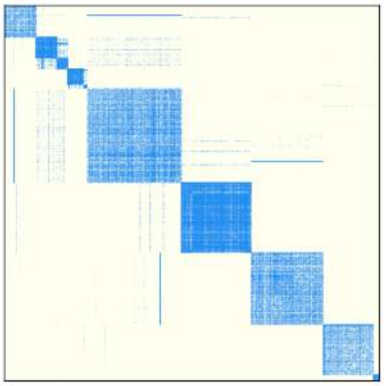

Figure 4. **Adjacency matrix of the Facebook graph.** Nodes have been grouped together.

ships) as well as node features for every node (see Table 1 for summary), Figure 3). Node features are comprised of obfuscated categorical features of users profiles including education, work, hometown, language, last name, etc. As with most of real world social graphs, the data exhibits strong multi-cluster structure, as seen in Figure 3 and Figure 4. These clusters pose additional challenges to the application of random walk-based methods as the trust propagation between two loosely inter-connected clusters is low (Cao et al., 2012; Boshmaf et al., 2016). Hence, trust seeds should be distributed among all clusters. Following *SybilRank* and *Integro* (Cao et al., 2012), we employ the *Louvian clustering method* (Blondel et al., 2008) first.

As common, the Sybil graph needs to be generated. For this purpose, a (small) subgraph was copied and declared as Sybil. The attacking edges were created to link the honest and the Sybil graph following one of the attacking strategies (random or adversarial). It was made sure that no Sybil node attacked one of the direct neighbors of its origin which is reasonable for most social graphs. Edge features for *TSR* are as follows: the number of shared features (in total), the number of shared friends, and the number of shared features within specific categories. The other experimental setup is the same as the previous section.

Random Attacks The trusted nodes $|V_T| = 50$ were randomly distributed among all clusters and a small subset of Sybils $|V_D| = 30$ was chosen as known Sybil nodes. Attacking edges E_A were established following the random attacking strategy ranging from $|E_A| = 1$ to $|E_A| = 1400$. Experiments were repeated 10 times. *Integro* was run with

Table 1. Topological properties of the Facebook sample graph

Property	Value
Number of nodes	4039
Number of edegs	88234
Strongly connected	True
Weighted (edges)	False
Avg. clustering coefficient	0.6055
Diameter	8

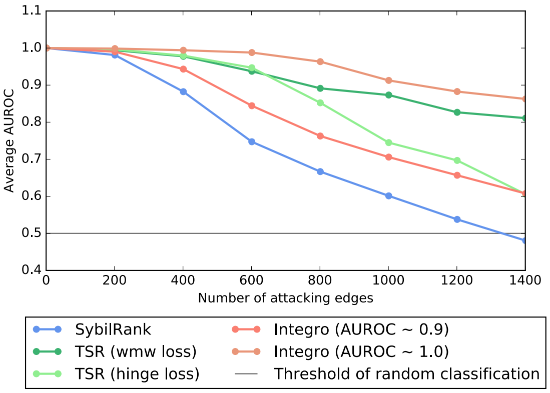

Figure 5. **Comparison of the detection methods in a random attacking scenario on the Facebook graph.** Accuracy in terms of average AUROC for all evaluated methods.

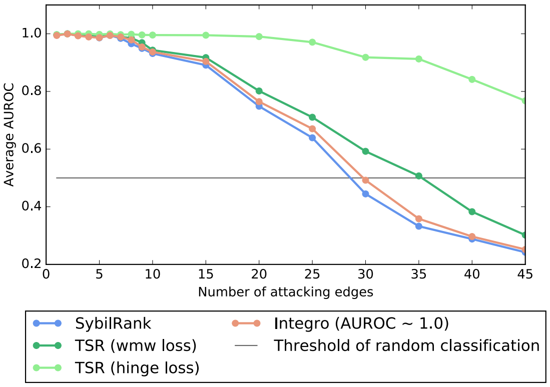

Figure 6. **Comparison of the detection methods in an adversarial attacking scenario on the Facebook graph.**

two levels of accuracy in simulated victim detection, i.e., perfect (AUROC $= 1$) and almost perfect (AUROC $= 0.9$). Figure 5 shows the AUROC-values. The detection performance of *SybilRank* is the lowest and drops soon as attacking edges increase. *Integro* with the *perfect* victim detection outperforms the other methods, but with just a slight reduction in the victim detection accuracy (AUROC $= 0.9$), its performance drops significantly. All versions of *TSR* perform almost on par with perfect version of *Integro* in the lower range of attacking edges (1—800). In the higher range (800—1400), the hinge loss drop fast to end up with a performance similar to *Integro* with the almost perfect victim detection. However, the variant that uses the WMW-loss does not show this performance drop and stays close to the upper-bound of *Integro*.

Adversarial Attacks The number of adversarial attack edges ranged from $|E_A| = 1$ to $|E_A| = 45$. Figure 6 shows the recorded average AUROC-values. Again, *SybilRank*'s performance drops the fastest and steepest and *Integro* is insignificantly better in this adversarial scenario. Both variants of *TSR* performs better than the baselines. However, the WMW-loss variant performs only slightly better than *SybilRank* and *Integro*, while the hinge-loss variant keeps good performance even for a large number of attacking edges. As our future work, we will investigate which loss function should be chosen, depending on data and assumed attacker's strategy. Overall, whereas *SybilRank*'s and *Integro*'s performance drops to an average AUROC-value below 0.5 at $|E_A| = 30$, the hinge-loss variant of *TSR* still achieves an average value over 0.9 at the same amount of attacking edges.

8. Conclusion & Outlook

In this paper, we studied the problem of Sybil detection. We first refined the security guarantees of random walk approaches towards more realistic assumptions. Then, we formalized and coined the adversarial setting and introduced optimal strategies for attackers. Further, we proposed a new method, transductive Sybil ranking (*TSR*), that leverages prior information, network topology as well as node and edge features. Unlike *Integro*, it is fused in a single optimization framework and can be solved efficiently by using gradient-based optimizer. In our empirical evaluation, we showed the advantages of our method and investigated the susceptibility of our method and baseline competitors to adversarial attacks. Further research will focus on the application of our method to real-world, large-scale OSNs.

9. Acknowledgments

JH was supported by MathPlan GmbH and innoCampus, TU-Berlin. SN, AB and KRM were supported by the German Ministry for Education and Research as Berlin Big Data Center BBDC, funding mark 01IS14013A. KRM thanks for the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No.2017-0-00451). NG was supported by BMBF ALICE II grant 01IB15001B.

References

Backstrom, Lars and Leskovec, Jure. Supervised random walks. *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, pp. 635, 2011.

Barabási, Albert-Lzl. Scale-free networks: A decade and beyond. *Science*, 325(5939):412–413, 2009.

Behrends, Ehrhard. *Introduction to Markov Chains—With*

Special Emphasis on Rapid Mixing. Vieweg+Teubner Verlag, 2000.

Bilge, Leyla, Strufe, Thorsten, Balzarotti, Davide, Kirda, Engin, and Antipolis, Sophia. All your contacts are belong to us: Automated identity theft attacks on social networks. *WWW 2009*, pp. 551–560, 2009.

Blondel, Vincent D., Guillaume, Jean-Loup, Lambiotte, Renaud, and Lefebvre, Etienne. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10008(10):6, 2008.

Boshmaf, Yazan, Muslukhov, Ildar, Beznosov, Konstantin, and Ripeanu, Matei. The socialbot network: when bots socialize for fame and money. *Acm*, pp. 93, 2011.

Boshmaf, Yazan, Logothetis, Dionysios, Siganos, Georgos, Lería, Jorge, Lorenzo, Jose, Ripeanu, Matei, Beznosov, Konstantin, and Halawa, Hassan. Íntegro: Leveraging victim prediction for robust fake account detection in large scale OSNs. *Computers and Security*, 61 (February):142–168, 2016.

Cao, Qiang, Sirivianos, Michael, Yang, Xiaowei, and Pregueiro, Tiago. Aiding the detection of fake accounts in large scale social online services. *NSDI'12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pp. 15, 2012.

Danezis, George. sybilinfer: Detecting Sybil nodes using social networks. *Network & Distributed System Security Symposium(NDSS)*, 2009.

Douceur, John R. The Sybil attack. *Peer-to-peer Systems*, pp. 1–6, 2002. ISSN 00278424. doi: 10.1016/S0140-6736(07)60784-3.

Gong, Neil Zhenqiang, Frank, Mario, and Mittal, Prateek. Sybilbelief: A semi-supervised learning approach for structure-based sybil detection. *IEEE Transactions on Information Forensics and Security*, 9(6):976–987, 2014.

Holme, Petter and Kim, Beom Jun. Growing scale-free networks with tunable clustering. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 65(2):2–5, 2002.

Leskovec, Jure and Mcauley, Julian J. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems 25*, pp. 539–547, 2012.

Lowd, Daniel and Meek, Christopher. Adversarial learning. *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 641–647, 2005.

Molloy, Michael and Reed, Bruce. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161180, 1995.

Stein, Tao, Chen, Erdong, and Mangla, Karan. Facebook immune system. *Proceedings of the 4th Workshop on Social Network Systems*, m(5):1–8, 2011.

Stringhini, Gianluca, Kruegel, Christopher, and Vigna, Giovanni. Detecting spammers on social networks. *Proceedings of the 26th Annual Computer Security Applications Conference*, 159:1, 2010.

The Associated Press. Facebook shares drop on news of fake accounts, 2012.

Tran, Nguyen, Li, Jinyang, Subramanian, Lakshminarayanan, and Chow, Sherman S M. Optimal Sybil-resilient node admission control. *Proceedings - IEEE INFOCOM*, pp. 3218–3226, 2011.

Tuckwell, Henry C. *Elementary applications of probability theory, second edition*. Chapman and Hall/CRC, 1995.

Vapnik, V N. An overview of statistical learning theory, 1999.

Wagner, Claudia, Mitter, Silvia, Körner, Christian, and Strohmaier, Markus. When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts (# MSM2012)*, pp. 2, 2012.

Yan, Lian, Dodier, Robert, Mozer, Michael C, and Wolniewicz, Richard. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. *ICML 03*, 2003.

Yang, Z, Smola, AJ, Song, L, and Wilson, AG. A la carte-learning fast kernels. *arXiv preprint arXiv:1412.6493*, 2014.

Yu, Haifeng, Gibbons, Phillip B., Kaminsky, Michael, and Xiao, Feng. SybilLimit: A near-optimal social network defense against Sybil attacks. *IEEE/ACM Transactions on Networking*, 18(3):885–898, 2010.