SybilShield: An Agent-Aided Social Network-Based Sybil Defense among Multiple Communities

Lu Shi*, Shucheng Yu*, Wenjing Lou† and Y. Thomas Hou‡
*Department of Computer Science, University of Arkansas at Little Rock, Little Rock, AR 72204
Email: lxshi@ualr.edu, sxyu1@ualr.edu
†Department of Computer Science, Virginia Tech, Falls Church, VA 22043
Email: wjlou@vt.edu
‡Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061
Email: thou@vt.edu

***Abstract*—Lacking trusted central authority, distributed systems have received serious security threats from Sybil attack, where an adversary forges identities of more than one node and attempts to control the system. By utilizing the real-world trust relationships between users, social network-based defense schemes have been proposed to mitigate the impact of Sybil attacks. These solutions are mostly built on the assumption that the social network graph can be partitioned into two loosely linked regions – a tightly connected non-Sybil region and a Sybil region. Although such an assumption may hold in certain settings, studies have shown that the real-world social connections tend to divide users into multiple inter-connected small worlds instead of a single uniformly connected large region. Given this fact, the applicability of existing schemes would be greatly undermined for inability to distinguish Sybil users from valid ones in the small non-Sybil regions.**

This paper addresses this problem and presents SybilShield, the first protocol that defends against Sybil attack utilizing multi-community social network structure in real world. Our scheme leverages the sociological property that the number of cutting edges between a non-Sybil community and a Sybil community, which represent human-established trust relationships, is much smaller than that among non-Sybil communities. With the help of agent nodes, SybilShield greatly reduces false positive rate of non-Sybils among multiple communities, while effectively identifying Sybil nodes. Analytical results prove the superiority of SybilShield. Our experiments on a real-world social network graph with 100,000 nodes also validate the effectiveness of SybilShield.

I. Introduction

Sybil attack [1] is a well-know attack in distributed systems, where a malicious user creates as many bogus identities (or Sybil nodes) as it wants, masquerades as different entities, and then launches attacks through these fake identities, making itself untraceable. With Sybil nodes compromising a large fraction of remaining nodes in the system, the adversary is able to take control of the system. Examples of Sybil attack have been observed in real world, including collusion behaviors in the Maze incentive-based peer-to-peer file-sharing network [2], manipulating polls by voting repeatedly with multiple identities [3], [4], promoting the rank of content or pages on YouTube [5] and Google [6], and others.

Existing solutions to Sybil attack can be generally categorized into centralized defenses and decentralized defenses. With a trusted authority, centralized defense system certifies identities by unique credentials issued to them. While previous work [1], [3], [7]–[9] has shown that it is costly and unrealistic to deploy centralized solutions in distributed systems, researchers turn to exploring the more challenging decentralized defense approaches. However, there is still no universally applicable distributed solution that completely eliminates the threats of Sybil attack. All of the current schemes focus on reducing the negative impacts of Sybil attack.

Recently, there has been an increasing interest in taking advantage of common characteristics of social networks to thwart Sybil attack [7], [9]–[12]. At the heart of these social network-based schemes lies the basic idea of partitioning nodes in the network into two regions – a non-Sybil region and a Sybil region – by weighing the trust exhibited in the social graph with the help of underlying topological properties [8]. However, [13], [14] stated that a real social network graph can actually be divided into multiple communities of different types. Therefore, it is inappropriate for prior work to build their schemes on the basis of the assumption that except the Sybil region there is only one tight-knit large community for all the honest users.

Moreover, in these social network-based Sybil defenses, Sybils are detected under the algorithmic assumption that the number of attack edges between Sybil and non-Sybil nodes is limited [15]. However, Viswanath et al. in [8] showed that it is also possible for non-Sybil sub-graphs to have such a sparse cut between each other, regardless of the indirect multi-hop edges between them, which further confirms the multi-community structure of social networks. In such a social network, it is challenging to distinguish between non-Sybil and Sybil users since each user in the distributed system has no knowledge of the topology of the entire network. As a result, the information for a user to decide the identity of another user in the network is limited. Consequently, nodes within one community may mistake non-Sybil nodes in another community for Sybils with high probability due to limited direct connectivity between the two communities, which also creates difficulty of discriminating Sybils from non-Sybils successfully. Sybil nodes can easily disguise themselves as a non-Sybil community by establishing a small number of carefully targeted links to the community containing the trust

978-1-4673-5946-7/13/$31.00 ©2013 IEEE

node, thereby confusing the Sybil detection.

Motivated by above issues, we introduce SybilShield, a new Sybil defense protocol that detects Sybils and limits the damaging effects of Sybil attack in the context of multi-community social graphs. Our design leverages the following structural properties of social network graph: nodes within the Sybil community are sparsely connected to all the non-Sybil communities due to lack of trust; honest communities are tightly inter-connected in general since honest users are the majority in the system, but their inter-connections can be multi-hop. During verification, SybilShield first performs modified random walk in the graph and utilizes intersections between such walks to limit the number of Sybil identities being accepted. If the prover is not accepted in this step, SybilShiled takes advantage of multi-hop edges by adopting agents, i.e., nodes selected from communities excluding those comprising the verifier and the prover, to confirm whether the rejected identity of the prover, who claims to be legitimate, is non-Sybil or not. The main contributions of our work include: (1) We present the first solution to Sybil attack in multi-community social networks, while other social network-based Sybil defense schemes assume improperly that the underlying social network graph is only composed of one non-Sybil region and one Sybil region; (2) We evaluate SybilShield by analysis and experiments on social network samples from MySpace, showing that our scheme is able to detect Sybils effectively. Besides, in SybilShield the false positive rate, which represents the percentage of honest nodes mislabeled as Sybils, is greatly reduced compared to applying previous solutions to multi-community social networks, which improves the validation accuracy.

The rest of this paper is organized as follows. Section II reviews the related work. Section III presents the model and assumptions. An overview of preliminary work is given in Section IV. We describe the design of SybilShield in detail in Section V. Section VI shows the effectiveness of SybilShield by theoretical analysis and experiments, which is followed by concluding remarks in Section VII.

II. Related Work

Defending against Sybil attack has drawn continuous interest since Sybil attack was first identified by Douceur [1]. Existing Sybil defense schemes can be classified into the following categories [3]: trusted central authority certification, resource testing, and capitalizing on trust networks. According to [1], trusted certification by central authority is the most common solution to Sybil attack, and also the only approach to radically wipe out its negative influences. However, in large-scale distributed systems, it is impractical for the central authority to assign one-to-one identities to all the entities. In addition, if the central authority becomes the target for attackers and fails as a result, the whole system would suffer.

Resource testing based solutions [16], [17] address Sybil attack by assuming that an attacker can only possess finite resources and is able to create few Sybil nodes with appropriately designed systems. But [18] shows that in some systems, Sybil attack can be accomplished even with few Sybil nodes.

Recently, there has been a substantial amount of work utilizing the underlying trust in the social network graphs to improve the resistance against Sybil attack. Proposals on the basis of social networks include SybilGuard [12], SybilLimit [7], SybilInfer [10], SumUp [11], Wh$\bar{a}$nau [9], SybilDefender [19], etc. They explore the fast mixing property of social graphs, which indicates that a random walk on a graph approaches the stationary distribution quickly. The validation procedure of SybilGuard, SybilLimit and SybilDefender basically relies on random walks and their collisions between honest nodes. While SybilLimit optimizes the limits for accepting Sybil nodes in SybilGuard from $O(\sqrt{n})$ to $O(\log n)$, SybilDefender claimes its Sybil identification rate approaches the theoretical bound. However, SybilDefender assumes the administrator knows the social network topology, which indicates it is a centralized mechanism and does not apply to distributed systems. SybilInfer labels nodes in a social network as honest users or Sybils according to probabilities determined by the Bayesian inference. SumUp is a content voting system treating nodes whose votes are accepted as non-Sybils, by the technique of adaptive vote flow aggregation. Wh$\bar{a}$nau, a Sybil-proof DHT routing protocol, combines random walk with the idea of layered identifiers.

In [8], Viswanath et al. found that these social network-based Sybil defense schemes are essentially graph partitioning algorithms, treating the underlying social network as a graph. This study also shows that these Sybil defenses are sensitive to community structure regarding mixing time in social networks. However, their ability of detecting Sybils would be undermined if multiple inter-connected communities constitute the non-Sybil region in real social network graphs. Mohaisen et al. in [15] models trust as a parameter in several different forms of modified random walk, and checks corresponding impact on the performance of Sybil defenses. But how to define the trust and trust level is not specified.

III. Model and Assumptions

A. System Model

We assume that in the system there are *n* honest users representing real human beings, and each of them has exactly one honest identity, which is denoted as an honest node in the social network graph. Different from previous social network-based Sybil defenses, we assume that a social network graph comprises multiple communities of different sizes. To verify our assumption, we conducted experiments on a 100,000-node sample graph from MySpace [20] by applying Louvain Method [21] for community detection. Our experimental result shows that these 100,000 nodes can be divided into 19 communities, with smallest size of 12 and largest size of 33,877, inter-connected by ten to hundreds of edges. This result validates our assumption and is also consistent with the observation made in previous work [21], [22].

Fig. 1 depicts the social network topology wherein honest nodes compose multiple groups of different sizes, which are

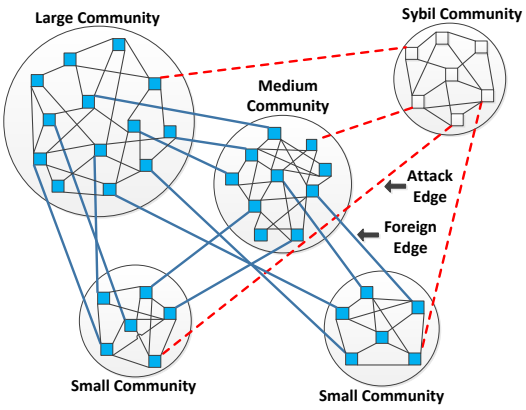

Fig. 1. The Social Network Graph

inter-connected with each other and termed *honest communities/regions*. Correspondingly, community formed by Sybil nodes is termed *Sybil community/region*. While honest nodes always comply with protocol rules, Sybil nodes may behave randomly or even in the opposite way. Additionally, we assume all the Sybil nodes can collude, which is called under the control of an adversary. Consequently, we assume all the Sybil nodes are within a single Sybil community/region. To simplify the discussion, we roughly categorize honest communities into three types according to their sizes: *small community*, *medium community*, and *large community*. If there exists an edge between two nodes located in different communities, we call it a *foreign edge*. Moreover, if one of the nodes connected by a foreign edge is a Sybil, we say this edge is an *attack edge*, through which the adversary may deceive other honest nodes.

As peers in the system, connected nodes provide and receive service from each other. Except Sybils, every node can be a verifier V or a suspect S. Since Sybil nodes mingle with honest nodes, V shall be able to determine whether or not a given S is a Sybil node. If S is a non-Sybil node, V *accepts* S, i.e., V is willing to establish a trust relationship with S for providing and receiving service between them; otherwise, V *rejects* S.

B. Design Goal

In an ideal scenario, it should be guaranteed that V accepts all the honest nodes and rejects all the Sybil nodes. But as studied in [1], only by trusted central authorization can the damaging effects of Sybils be eliminated completely. It is known to be difficult to implement a trusted central authorization in large-scale distributed systems, posing a formidable challenge of providing such a perfect guarantee. Thus, previous distributed Sybil defenses aim to reduce the negative influence of Sybil attack. Furthermore, in a social network with multiple inter-connected communities, it is hard for V to determine S's identity if S is from another community instead of where V is in. This is because a honest community with a small number of foreign edges to V resemble the Sybil community if the verification only utilizes the direct foreign edges between V and S as in previous schemes. In such a case, the false positive rate of honest nodes will be high due to the lack of ability to discriminate honest communities from Sybil ones. Therefore the main goal of SybilShield is to lower the false positive rate of honest nodes to a feasible extent, while keeping the false negative rate of Sybil nodes comparable to prior work [12].

C. Assumptions

Like previous related work [7], [9]–[12], SybilShield utilizes some properties of social networks as assumptions and intuition, which can be categorized into sociological assumption and algorithmic assumption [15].

1) Sociological Assumption: In social network-based Sybil defense schemes, trust is used to rationalize assumptions about the nature of social graphs and attackers' capabilities. Since an edge reflects real human being trust relationship, it is difficult for every Sybil node to convince many honest people. This means although an attacker can create unlimited Sybil identities, he/she is not able to arbitrarily establish links to honest nodes, which sets up a barrier to penetrate the whole social graph. Therefore, compared to the honest nodes, Sybils have a tendency to be poorly connected to the rest of the network. According to this observation, similar to [12] we assume that the total number of attack edges between the Sybil region and all the honest communities is small. On the contrary, for any given honest community, its total number of foreign edges to the rest of the honest communities is relatively large even if it has a limited number of foreign edges to each individual community. This follows the idea that it is easier for honest nodes, as compared to Sbyils, to establish trust relationships with other honest nodes.

2) Algorithmic Assumption: Algorithmic properties exhibited in social networks are always used to argue for the effectiveness of applications built on top of the social network. For Sybil defense schemes, the fast-mixing property is used to support the claimed effectiveness of Sybil identities detection and yield a feasible solution. The fast-mixing property means that a random walk on the social graph converges quickly to a node following the stationary distribution of the graph [23]. With this property, the ending node of the random walk probably stays in the same community as the starting node by a certain length of walking. However, Mohaisen et al. in [24] stated that these prior Sybil defense schemes performed experiments to show their applicability to real-world social graphs, without validating the fast-mixing assumption. Thus, [24] measured the mixing time in several social graphs. They noticed that mixing patterns in social graphs are associated with the underlying social model, i.e. social networks with confined social models. Social networks with strict trust properties are slow mixing whereas those with less strict trust properties are fast mixing. Furthermore, Mohaisen et al. in [25] experimentally show that fast-mixing graphs tend to have few large cores whereas slow mixing graphs tend to have multiple smaller cores. Following [7], [10]–[12], since most of the social network do not exhibit strict trust properties, the fast-mixing property is sufficient to support our SybilShield mechanism.

IV. PRELIMINARIES

SybilGuard [12] is the first formal attempt to use social networks for defending Sybil attack in distributed systems, which took the lead in exploiting fast-mixing properties for Sybil detection. SybilGuard provides two guarantees: the number of Sybil groups to *g* is limited, where *g* represents the number of attack edges between the non-Sybil region and Sybil region; the effective size of Sybil groups are bounded below *w*, i.e. a node will accept at most $g \cdot w$ Sybil nodes.

Based on the same assumptions and similar intuition described above, SybilGuard uses verifiable random walks (called random routes) and intersections to distinguish non-Sybil nodes and Sybil nodes. It labels a suspect as non-Sybil if the random walk from the trusted node intersects with that from the suspect; otherwise the suspect is labeled as a Sybil. Notice that each node creates a persistent routing table that maps each incoming edge to an outgoing edge in a unique one-to-one mapping.

Specifically, to determine whether to accept a suspect node S as an honest user, the verifier node V creates *w*-hop random route, which is deterministic formed according to the stored routing table entries at *w* consecutive nodes. Meanwhile, S starts a similar *w*-hop random route. V accepts S if the two random routes intersect. Note that the number of Sybil nodes accepted under the SybilGuard protocol will rise with the increase of *w*. Thus, *w* must be small enough to keep random routes within the honest region with high probability. From another point of view, *w* must be adequately large to ensure that intersections are made with high probability by random routes initiated from V and S separately. That is to say, it is beneficial for *w* to be in some reasonable range, neither too large nor too small. As Yu et al. in [12] analyzed, for a social network with $O(\log n)$ mixing time, based on the generalized birthday paradox, two non-Sybil nodes with $\sqrt{n}$ samples from the non-Sybil region will have an intersection with high probability. To design the appropriate length of random routes, SybilGuard makes an estimation by taking samples from the non-Sybil region of *n* nodes using $O(\sqrt{n}\log n)$ random walks.

While providing a promising direction for Sybil defenses, SybilGuard suffers from some limitations. As mentioned above, the underlying social network structure of SybilGuard is problematic. In real social networks comprised of few giant components and numerous medium and small communities, the ability of SybilGuard is weakened. The reason is that, a verifier node V might easily mark an honest suspect node S as Sybil mistakenly due to disability to tell apart the Sybil community and non-Sybil communities, which increases the false positive rate. Viswanath et al. also stated this issue in [8] by examine existing Sybil defense schemes [7], [10]–[12] over different community structures. However, [8] did not propose any solution to this problem. By referring to SybilGuard, our protocol of SybilShield achieves higher accuracy with a relatively low false positive rate among multiple communities in the real social networks.

Symbol	Definition
n	the total number of nodes in the social network
n_a	the number of agents found by the verifier
n_w	system constant for bounding extended random routes
d_i	the degree of node i
w_i	the length of random route for node i
t_i	the threshold of acceptance ratio for node i
t	the threshold of ratio for agent voting
n_1, n_2, n_3, n_4	node number of large, medium, small and Sybil communities respectively
N_1, N_2, N_3, N_4	the number of communities of large, medium, small and Sybil communities respectively
E_{ij}	the number of foreign edge between community i and community j

TABLE I
NOTATION

V. THE SYBILSHIELD PROTOCOL

A. Overview

SybilShield defends against Sybil attack for real-world social network with multi-community structure. For a verifier node V to determine the identity of a suspect node S, both V and S first perform a modified form of random walk – random route. V accepts S if their random routes have intersections. Otherwise V refuses to admit S. Even though S is rejected by random route verification, it could not guarantee that S must be a Sybil node. To avoid mislabeling honest suspects as Sybils, SybilShield utilizes agents of V to reinspect the identity of S in the second step, which is called agent walk. In agent walk, agents are searched along all the edges of V, and carefully selected to be outside the community of V. Valid agents perform random routes to see if intersections exist with S's random routes. S gets votes from agents with intersections. A threshold *t* is set for V to make the decision of accepting S if the proportion of agents voting for S is no less than *t*. Notation of this paper is partially defined in Table I.

B. Random Route

Random route was first proposed in [12]. We refer to it and build the first step of our protocol, as shown in Algorithm 1. For random routes in honest communities, V accepts S if their random routes have at least one intersection. For each hop along a random route, its next hop is chosen strictly according to a pre-computed randomized routing table of the current hop, rather than uniformly randomly selecting one of its neighbors. Routing tables are calculated by random permutation, indicating a one-to-one mapping from incoming edges to outgoing edges. As the social network is assumed to be static, there are no nodes and edges added or deleted from the graph. Therefore, routing table of each node will keep the same and does not need to be updated.

We note that routing table is known to have the following properties: (1) Convergence property: if different random routes pass a certain node through the same incoming edge, they will share the same outgoing edges directed to the next

Algorithm 1: Initial Verification

for $i = 1$ **to** d_V **do** /* d_V: V's degree */
  V performs random route along its i^{th} edge;
  for $j = 1$ **to** d_S **do** /* d_S: S's degree */
    S performs random route along its j^{th} edge;
    Check whether an intersection exits by V's i^{th} random route and S's j^{th} random route and record the result;
  end
  if *Intersection percentage is no less than* t_S **then**
    V accepts S along its i^{th} edge;
  else
    V rejects S along its i^{th} edge;
  end
end
if *Along all V's edges, the accepting ratio is no less than* t_V **then**
  V accepts S;
else
  V finds agents for further authentication;
end

Algorithm 2: Agents Discovering

for $i = 1$ **to** d_V **do** /* d_V: V's degree */
  repeat
    V performs random route along its i^{th} edge with length w_V;
    V picks the last hop of the random route as an agent A;
    V verifies A by Algorithm 1;
    if V *accepts* A **then**
      V increases its random route length by w_V;
    else
      V accepts A as a valid agent; break;
    end
  until *random route length* $> n_w w_V$;
end
n_a valid agents are found, where $n_a \leq d_V$.

hop. (2) Back-traceable property: if two random routes have the same outgoing edges at some node, they must come to that node along the same incoming edge. But these properties will no longer be valid if entering Sybil communities. Instead, the adversary might manipulate the incoming random routes from the verifier and confine those random routes to the Sybil region. In this case, it is easier for a Sybil node to intersect with random routes initiated by the verifier and get accepted. To reduce the negative impact brought by routes extended to the Sybil region, V would not accept S unless t_V of V's random routes accept S, where t_V is a threshold shown to be $d_V/2$ to provide a good tradeoff in [12].

Lacking knowledge of the whole network, each node needs to locally decide the length of its random route w. We adopt 3-hop sampling [12] to estimate w. Specifically, a node A first performs a standard random walk with three hops, ending at a certain node B, which assures high probability that B is honest and stays in the community of A. Both A and B perform random routes along all directions to decide the lengths of the routes to reach their intersections, which are collected as samples. Finally, w is set as $2.1m$ [12], where m is the median of the samples. In the rest of the paper, we use w_i to represent the length of random route for any given node i.

Between any two communities in the social graph, inter-community connections are generally much sparser than internal connection within their own communities. Random routes with an appropriate length w initiated by nodes in one community has higher probability of staying inside their community, instead of entering another community through foreign/attack edges. As a result, verification of a suspect S from another region by V is restricted to the number of foreign/attack edges between V's and S's communities. In other words, honest nodes in one community are probably rejected and marked as Sybil by honest nodes in another community, increasing false positive rate in the system.

C. Agent Walk

Due to the limitation above, we utilize agents to make our protocol more complete and accurate. In SybilShield, to confirm the decision, V attempts to find some agents for help to check S again once S is rejected. Although inter-community connections are fewer than intra-community connections, attack edges are more sparely linked compared to foreign edges, based on the limited ability of constructing real trust relationship with honest nodes. Since the majority of communities in the system are honest, searching for agents by extending random routes to other communities has higher probability of entering honest communities than Sybil communities. Otherwise the system is compromised and controlled by the adversary. If at least a certain proportion of agents in other honest communities accept S, V treats S as an honest node.

A valid agent should be in another community other than where V is located. As Algorithm 2 shows, along each one of its edges, V starts a random route with a length of w used in Initial Verification. At the end of the route, V picks the last hop as an agent A temporarily. And then V and A initiate random routes simultaneously to see if there are intersections. If V and A are in different communities, with only a few foreign edges, their random routes are not very likely to traverse the foreign edges and enter one another's community. Thus, the probability of intersections by such routes is small. If no intersections are found, V views A as outside its community, i.e. A is a qualified agent. Otherwise, V continues the random route for w_V more hops from the last ending point in the direction that it was originally traveling, and validates the node on the tail of the extended random route. This process will be repeated until the valid agent is found in the direction of that edge. The length of extended random routes are set to be $(n_w \cdot w_V)$, where n_w is a constant. Also we suppose the node degree of V is d_V, and the number of effective agents n_a is no more than $d_V (1 \leq n_a \leq d_V)$.

After determining all the valid agents, every agent verifies S by random routes following Algorithm 3. Similarly, assuming the node degree of the agent A is d_A, for each of its random routes, if the route has at least one intersection with the random route from S, then that route accepts S. The agent A would not accepts S unless at t_a of A's random routes accepts S. There are n_a agents like A in total for V. Therefore, if more than $t \cdot n_a (0 \leq t \leq 1)$ agents accepts S, V accepts S finally. Otherwise, V treats S as a Sybil node and refuses to establish any trust relationship with S.

As we state above, honest communities are in the majority

Algorithm 3: Agent-Aid Verification

```
for i = 1 to n_a do
    for j = 1 to d_A do    /* d_A: current agent's degree */
        i^th agent performs random route along its j^th edge;
        i^th agent verifies S by Algorithm 1 and records its
        accept/reject decision;
    end
end
if Among n_a agents, the accepting ratio is no less than t then
    V accepts S;
else
    V rejects S;
end
```

in the system, and foreign edges between any two honest communities are denser than attack edges. Therefore, it can be guaranteed that the ratio of foreign edges to all the foreign edges and attack edges in the system is at least $\frac{1}{2}$ or the system has been taken control by the adversary. In our protocol, we set both thresholds t_a and t as $\frac{1}{2}$.

Note that a Sybil agent might be selected during the search. In this case, the adversary would control the random route of the Sybil agent depending on the identity of the suspect since the adversary has the knowledge of the entire network and it only does things in its favor. Therefore, if the suspect is an honest node, in order to put the suspect at risk of being rejected, the adversary will let the Sybil agent report a rejection; if the suspect is also a Sybil node, the Sybil agent reports an acceptance immediately.

VI. Evaluation

This section presents numerical analysis and experiment results on MySpace social network.

A. Numerical Analysis

1. Performance Comparison: For analysis, we set the total number of honest nodes in the social network as n. To simplify the analysis, communities are categorized into four types according to their sizes and properties, i.e., type 1, 2, 3, 4 corresponding to large, medium, small and Sybil. Their node number is averaged and denoted by n_1, n_2, n_3,and n_4, respectively. Correspondingly, the number of communities of each type of community is N_1, N_2, N_3, and N_4. The number of edges between every two communities is set to be E_{ij}, where both i and j are an integer, representing types of communities in which end point of the edge is located. Note that $E_{ij} = E_{ji}$ because of the reciprocity property. These edges include foreign edges, and attack edges as well. We assume the number of attack edges between a Sybil region and any honest community must be less than the number of foreign edges between any two of the honest regions. Due to the structure of the social network graph, following different cases are considered separately in the verification process regarding the identities of V and S.

(1) V and S are from different communities of distinct size;
(2) V and S are from different communities of same size;
(3) V is from honest communities while S is a Sybil.

Note that V must be an honest user. If V and S are from the same community, we assume V accepts S with the probability close to 1 according to previous work [12] and do not discuss this scenario in detail in this paper. Besides, since the links between nodes are assumed to be reciprocal, S verifying V is identical but reversed validation process of V verifying S.

For SybilGuard, based on the expected suspect acceptance rates in the above cases, if we randomly select a pair of V and S, the **expected acceptance rate of a honest S** P_{Honest} is:

$$P_{Honest} = \frac{1}{n(n-1)}\Bigg(\sum_{i=1}^{3} N_i n_i (n_i - 1) + 2\sum_{i=1}^{3}\binom{N_i}{2} n_i^2 \cdot P_{ii} + \sum_{i\neq j}^{3} N_i n_i \cdot N_j n_j \cdot P_{ij}\Bigg) \quad (1)$$

If V and S belong to the same region, the acceptance rate is set to be 1. P_{ij} is the probability that an honest V accepts S corresponding to the Scenario (1). And P_{ii} represents the probability that V accepts S which belongs to a different community of the same size, as Scenario (2) shown above.

The **expected acceptance rate of a Sybil S** P_{Sybil} in SybilGuard is:

$$P_{Sybil} = \frac{1}{n}\sum_{i=1}^{3} N_i n_i \cdot P_{i4} \quad (2)$$

P_{i4} is the probability that an honest V accepts a Sybil S corresponding to the Scenario (3).

For SybilShield, agents are introduced to recheck the identity of S if S is refused by V during the Initial Verification. A valid agent must be located in a different community other than the one V is in. The probability of randomly picking a valid agent depends on the total number and distribution of foreign edges connected to V's community. Note that Sybil agents might also be included through attack edges between V's community and Sybil region. Therefore, the probability of **choosing an agent from different communities of same size** can be calculated as:

$$Pa_{ik} = \frac{(N_i - 1)E_{ik}}{(N_i - 1)E_{ik} + \sum_{l\neq i, 1\leq l\leq 4} N_l E_{il}}, i = k \quad (3)$$

Similarly, **an agent from different communities of distinct size** can be obtained with the following probability:

$$Pa_{ik} = \frac{N_k E_{ik}}{(N_i - 1)E_{ii} + \sum_{l\neq i, 1\leq l\leq 4} N_l E_{il}}, i \neq k \quad (4)$$

where i and k represent the community type of V and the agent. Honest agents will authenticate suspects rejected by the verifier as shown in Algorithm 3, regardless of the real identities of these suspects. But for Sybil agents, they will definitely deviate from the normal verification protocol, refusing honest suspects and only accepting Sybil suspects. Statistically speaking, the probability of selecting an agent of a certain type is highly related to the number of foreign edges and/or attack edges of the community in which V is located. Since the adversary has limited resources to establish

real human trust relationship with honest nodes, its attack edges would be much less than the normal foreign edges between honest users. Therefore, with higher probabilities of selecting honest agents, the expected suspect acceptance rate is increased by admitting mislabeled honest nodes. Given *eq.(3)* and *eq.(4)*, we can draw the **acceptance probability of S by a single agent** Pas_{ij} in accordance with the three scenarios enumerated earlier. For V of type i, S of type j, and selected agent of type k, $1 \leq i, j \leq 4$, if S is an honest suspect:

$$Pas_{ij} = \sum_{k \neq j, 1 \leq k \leq 3} Pa_{ik} P_{kj} + Pa_{ij}(P_{jj} \frac{N_j - 1}{N_j} + \frac{1}{N_j}) \quad (5)$$

If S is a Sybil:

$$Pas_{ij} = \sum_{k \neq j, 1 \leq k \leq 3} Pa_{ik} P_{kj} + Pa_{ij} \quad (6)$$

Distinct verifier nodes may have different numbers of valid agents, which are no greater than the degrees of the verifier nodes. We set a threshold *t* ($0 \leq t \leq 1$) on the percentage of agents approving the suspect for the verifier to accept that suspect on second thoughts. Taking the combinations into account, the **expected acceptance rate of S by all the agents** P'_{ij} can be obtained based on *eq.(5)* and *eq.(6)*, where n_a is a variable representing the number of valid agents.

$$P'_{ij} = \sum_{i=t \cdot n_a}^{n_a} C_i^{n_a} \cdot (Pas_{ij})^i \cdot (1 - Pas_{ij})^{n_a - i} \quad (7)$$

Therefore, for **SybilShield**, the **expected acceptance rate of a honest S** P'_{Honest} is equal to the sum of direct acceptance rate without agents' help and indirect acceptance rate provided by agents:

$$P'_{Honest} = P_{Honest} + \frac{1}{n(n-1)} \left(2 \sum_{i=1}^{3} \binom{N_i}{2} n_i^2 (1 - P_{ii}) P'_{ii} + \sum_{i \neq j}^{3} N_i n_i \cdot N_j n_j (1 - P_{ij}) P'_{ij} \right) \quad (8)$$

Similarly, the expected acceptance rate of a Sybil S P'_{Sybil} in **SybilShield** is:

$$P'_{Sybil} = P_{Sybil} + \frac{1}{n} \sum_{i=1}^{3} N_i n_i (1 - P_{i4}) P'_{i4} \quad (9)$$

Compare *eq.(1)* and *eq.(2)* for SybilGuard with *eq.(8)* and *eq.(9)* for SybilShield, we find that the difference lies in the indirect acceptance supported by agents. It is obvious that with agents, the expected suspect acceptance rate by agents P'_{ij} decides to what extent the original expected honest/Sybil nodes acceptance P_{Honest} and P_{Sybil} may increase. According to *eq.(6)*, the value of P'_{ij} is dependent on both the number of agents n_a and the expected suspect acceptance possibility by a single agent Pas_{ij}. And we can see that the calculation of Pa_{ij} is affected by the number of foreign/attack edges (E_{ij}) and the probability that an honest verifier from a certain community accepts a suspect from another community (P_{ij}). Since values of P_{ij} remain the same in both SybilGuard and SybilShield, the ratio of foreign edges to attack edges becomes the dominant factor in determining the degree of improvement of honest suspect acceptance in SybilShield.

On the other hand, as the expected honest suspect acceptance rises, the expected Sybil suspect acceptance also increase. However, previous works have shown that the resources of the adversary are too limited to build a great amount of trust relationship with honest nodes. And P'_{i4} for Sybil nodes is quite small since the attack edges are much fewer and the number of foreign edges among honest communities is greater than that of attack edges between Sybil and Honest communities. Therefore, by introducing agents, the growth of expected Sybil suspect acceptance is comparatively less significant. This property of social network assures the correctness and effectiveness of SybilShield.

Effect of System Parameters: (a) The threshold **t**. t represents the proportion of agents which vote S for V to accept S. If at least t percent ($0 \leq t \leq 1$) of all the agents believe S is honest, V admits S's identity. Among n_a agents, we assume there are n_h honest agents and n_s Sybil agents. For simplicity, the probability of S getting approved by an agent is Pas_{ij}. For Sybil agents, they will only refuse honest suspect and accept Sybil suspect. Then t can be estimated by:

$$\frac{n_s}{n_h + n_s} \leq t \leq \frac{n_h}{n_h + n_s} \quad (10)$$

Since the probability of selecting Sybil agent is small due to limited attack edges, n_h would be much greater than n_s. There would not be worse scenario than that of $n_h = n_s$ with $t = 1/2$, because the adversary would take control if most agents are Sybil. Therefore, t should be no less than $1/2$.

(b) The length of extended random route. During searching for agents in SybilShield, for efficiency the cumulative length of the extend random route shall be bounded – if no valid agent is found in the end after repeating random routes of length w by n_w times, V abandons the search on this route, turns to another edge and start the process above again. n_w can be empirically determined, e.g., the minimum number assuring $90\% \times d_V$ agents can be identified using Algorithm 2.

B. Experiments

We evaluate the effectiveness of SybilShield by experiments in terms of false positive rate of honest nodes and false negative rate of Sybil nodes. Although SybilLimit further reduced the number of accepted Sybil nodes per attack edge from $O(\sqrt{n})$ in SybilGuard to $O(\log n)$, both SybilGuard and SybilShield did not consider the issue of false positive rate of honest nodes if applied to multi-community social networks. To simplify our evaluation, the performance of SybilShield is provided and compared with SybilGuard based on the same real-world social network – MySpace. The effect of system parameters on the performance is also studied.

1) Data Sets and Experiment Setup: Our experiments are implemented on a data set from one of the most popular social networks – MySpace [20], which was also used to validate

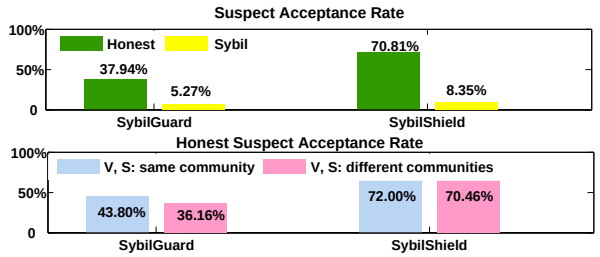

Fig. 2. Honest and Sybil suspects acceptance ratio for v and s from same/different communities

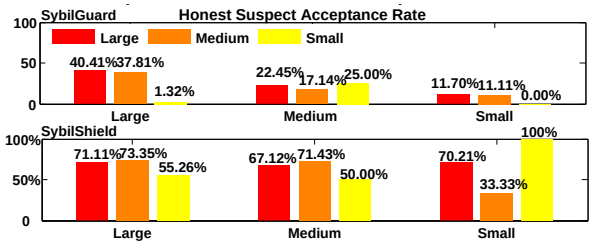

Fig. 3. Honest and Sybil suspects acceptance ratio for each type of community

our assumptions on social network structure in Section III. This data set consists of 100,000 nodes and 6,854,231 edges, with an average degree of 68. Given the multi-community property of social networks, Louvain Method [21] was applied to extract the underlying community structure. These 100,000 nodes were partitioned into 19 communities of different size. We put these 19 communities into three categories in terms of sizes – large, medium and small. Thus, in the system, there are 4 large communities with over 10,000 nodes, 3 medium communities of sizes between 1000 and 10,000 nodes, and 12 communities with no more than 1000 nodes. All the communities are connected with one another by a number of links. For large communities, the average number of foreign edges is 1,402,561, while for medium and small communities it is 404,379 and 818 respectively. Note that we do not make any modification on the MySpace data set used in this paper.

For evaluation purpose, previous work either selected random nodes from data sets as attacker nodes [7], [10], [12] or add adversarial nodes [11], [19]. Referring to the method and ratio of Sybil nodes in [11], [19], we construct a Sybil region by creating 500 Sybil nodes and randomly selecting nodes from the data set to be linked with Sybil nodes, until the number of attack edges is 50. The average node degree of Sybil region is set to be the same as in small communities in the system. In addition, we obtain the length of random route for each community by running 3-hop sampling.

2) Experiment Results: To compare the performance of SybilGuard and SybilShield, we randomly select 100,000 pairs of verifier and suspect to run the verification protocol.

First we look into the percentage of mistakenly rejected honest suspects and the percentage of accepted Sybil suspects. As shown in the first half of Fig. 2, the acceptance rate of honest suspects for SybilShield and SybilGuard is 70.81% and 37.94%. In other words, the false positive rate for SybilShield is 29.19%, while 62.06% for SybilGuard. Therefore, the false positive rate is effectively reduced by 32.87%. Obviously, the accuracy of identifying honest suspects in SybilShield is improved twice as much in SybilGuard because of the introduction of agents. However, the Sybil acceptance rates for Sybil are 8.35% and 5.29% correspondingly, with a difference of only 3.06%. Although the Sybil acceptance rate rises a little, the tradeoff between greatly reduced false positive rate and Sybil acceptance rate is acceptable.

To further study the experiment results, we divide all the test pairs into two categories according to whether both of V and S are in the same communities. Results are demonstrated in terms of honest suspects acceptance rate in the other half of Fig. 2. Whether in SybilShield or SybilGuard, the accuracy of identifying honest suspects is reduced by 1.54% and 7.64%. Since foreign edges of a community are comparatively sparser than its internal edges, it is more difficult for a verifier to reach a suspect located in another community through foreign edges by random routes. And it is clear that regardless of the relationship of V and S, with the help of agents, the honest suspects acceptance rate is enhanced from 43.8% and 36.16% in SybilGuard to 72.00% and 70.46% in SybilShiled. Furthermore, it is interesting to note that V and S belonging to the same community does not guarantee ideal 100% acceptance rate because of different underlying social network structure.

Fig.3 reveals the effect of community type on the honest suspects acceptance rate in both SybilGuard and SybilShield. We can easily see that compared to SybilGuard, the overall acceptance rate in SybilShield nearly doubles, especially for honest suspects in medium and small communities. According to our SybilShield protocol, if the communities of V and S are not only comparatively tightly inter-connected but also linked with other communities with more foreign edges, the probability of S getting accepted will be increased. And based on our observations of social networks structure, the larger the community's size is, more foreign edges the community has between itself and others. These explain why the honest nodes in small communities are harder to be accepted by a verifier. Besides, the nodes in small communities have much a lower probability to be chose due to their non-dominant percentage in the system. Because of the limited sample of test pairs from small communities, the corresponding honest suspects acceptance rate 100% in Fig. 3 does not represent the typical result of that scenario.

3) Discussion: **The Number of Agents.** As an important system parameter, the number of agents n_a of a verifier node V affects the suspect acceptance rate in our protocol. Results shows that the probability of an agent being a Sybil node is small. Compared to the average agent number of a verifier in a community of any of the three types – 113, 116, and 12, corresponding average Sybil agent number is 5, 6, and 1. Note that due to the property of this data sample, nodes in medium communities have a higher average degree than large community-nodes, so V in a medium community has the chance to obtain more agents than V in large communities.

However, there are more foreign edges in large communities rather than in medium communities. Therefore, the final number of the agents by a verifier from either of these two types of community is almost the same with very small differences. To sum up, even if Sybil agents do not obey the rule and refuse to accept honest suspects, these Sybil agents are in the minority and not able to affect the verifier node's decision.

Compared to Other Related Work. Besides SybilGuard [12], other related social network-based Sybil defense schemes also suffer from the limitation of basing their solution on non-real social networks. They all assumed the social network is comprised of two different parts, a non-Sybil region and a Sybil region. Revised from SybilGuard, SybilLimit [7] relies on tail intersections of random routes to decide whether or not the suspect should be accepted, and uses balance condition to deal with tails of the verifier entering Sybil region. Therefore, agents can be applied to SybilGuard and SybilLimit after the basic random routes. For SybilInfer [10], random walks are performed to sample nodes from non-Sybil region and determine the acceptance probability by Bayesian inference. In this case, besides sampling nodes from the region where the verifier is located, the probability of a node being non-Sybil would be more accurate if employing different agents from other regions for sampling and calculation. Our work can be easily extended to these schemes, and the performance is believed to be improved.

VII. Conclusion

This paper presents SybilShield, a novel decentralized defense protocol against Sybil attacks in multi-community social networks, which limits the negative influences of accepting Sybils mistakenly and mislabeling honest nodes. SybilShield is based on underlying properties of real-world social networks that the non-Sybil regions are fast mixing and the number of attack edges created by an adversary is relatively less than that of foreign edges among honest communities, which are validated on the given MySpace topology data sample. Inspired by these social network properties, we introduce agents for help if the initial validation by performing random routes denies to accept the suspect node. Through the theoretical probability analysis and experiments on the MySpace data set, SybilShield is shown to greatly outperform SybilGuard, reducing the false positive rate while keeping the effectiveness of identifying Sybil nodes with an acceptable tradeoff.

For future work, we will run SybilShield on more real social network data with different structures and further improve the efficiency of our SybilShield algorithm.

Acknowledgement

This work was supported in part by the US national science foundation grants CNS-1217889 and CNS-1155988.

References

[1] J. Douceur, "The sybil attack," in *Peer-to-Peer Systems*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2002, vol. 2429, pp. 251–260.

[2] Q. Lian, Z. Zhang, M. Yang, B. Zhao, Y. Dai, and X. Li, "An empirical study of collusion behavior in the maze p2p file-sharing system," in *Distributed Computing Systems, 2007. ICDCS '07. 27th International Conference on*, Jun. 2007, p. 56.

[3] B. N. Levine, C. Shields, and N. B. Margolin, "A survey of solutions to the sybil attack," *World*, no. Technical Report 2006-052, 2006.

[4] H. Hsieh, "Doonesbury online poll hacked in favor of mit," The Tech, MIT, 2006.

[5] D. Riley. (2007) Stat gaming services come to youtube. TechCrunch.

[6] M. Bianchini, M. Gori, and F. Scarselli, "Inside pagerank," *ACM Trans. Internet Technol.*, vol. 5, pp. 92–128, February 2005.

[7] H. Yu, P. Gibbons, M. Kaminsky, and F. Xiao, "Sybillimit: A near-optimal social network defense against sybil attacks," *Networking, IEEE/ACM Transactions on*, vol. 18, no. 3, pp. 885–898, Jun. 2010.

[8] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An analysis of social network-based sybil defenses," in *Proc. ACM SIGCOMM 2010 conference on SIGCOMM*, ser. SIGCOMM '10. New York, NY, USA: ACM, 2010, pp. 363–374.

[9] C. Lesniewski-Laas and M. F. Kaashoek, "Whanau: A sybil-proof distributed hash table," in *Proc. NSDI'10*. San Jose, CA, USA: USENIX Association, Apr. 2010, pp. 111–126.

[10] G. Danezis and P. Mittal, "Sybilinfer: Detecting sybil nodes using social networks," in *Proc. NDSS'09*, San Diego, CA, USA, Feb. 2009.

[11] N. Tran, B. Min, J. Li, and L. Subramanian, "Sybil-resilient online content voting," in *Proceedings of the 6th USENIX symposium on Networked systems design and implementation*. Berkeley, CA, USA: USENIX Association, 2009, pp. 15–28.

[12] H. Yu, M. Kaminsky, P. Gibbons, and A. Flaxman, "Sybilguard: Defending against sybil attacks via social networks," *Networking, IEEE/ACM Transactions on*, vol. 16, no. 3, pp. 576–589, Jun. 2008.

[13] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 611–617.

[14] ——, "Structure and evolution of online social networks," in *Link Mining: Models, Algorithms, and Applications*. Springer New York, 2010, pp. 337–357.

[15] A. Mohaisen, N. Hopper, and Y. Kim, "Keep your friends close: Incorporating trust into social network-based sybil defenses," in *INFOCOM, 2011 Proceedings IEEE*, Apr. 2011, pp. 1943 –1951.

[16] H. Rowaihy, W. Enck, P. McDaniel, and T. La Porta, "Limiting sybil attacks in structured p2p networks," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, May 2007, pp. 2596 –2600.

[17] N. Borisov, "Computational puzzles as sybil defenses," in *Peer-to-Peer Computing, 2006. P2P 2006. Sixth IEEE International Conference on*, Sep. 2006, pp. 171 –176.

[18] M. K. Wright, M. Adler, B. N. Levine, and C. Shields, "The predecessor attack: An analysis of a threat to anonymous communications systems," *ACM Trans. Inf. Syst. Secur.*, vol. 4, pp. 489–522, 2004.

[19] W. Wei, F. Xu, C. Tan, and Q. Li, "Sybildefender: Defend against sybil attacks in large social networks," in *INFOCOM, 2012 Proceedings IEEE*, march 2012, pp. 1951 –1959.

[20] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 835–844.

[21] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, October 2008.

[22] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[23] D. Quercia and S. Hailes, "Sybil attacks against mobile users: Friends and foes to the rescue," in *INFOCOM, 2010 Proceedings IEEE*, Mar. 2010, pp. 1 –5.

[24] A. Mohaisen, A. Yun, and Y. Kim, "Measuring the mixing time of social graphs," in *Proceedings of the 10th annual conference on Internet measurement*, ser. IMC '10. New York, NY, USA: ACM, 2010, pp. 383–389.

[25] A. Mohaisen, H. Tran, N. Hopper, and Y. Kim, "Understanding social networks properties for trustworthy computing," in *SIMPLEX'11: The 3rd Annual Workshop on Simplifying Complex Networks for Practitioners (with ICDCS'11)*, Minneapolis, MN, USA, Jun. 2011.