Internet Mathematics Vol. 10: 360–420

Communities, Random Walks, and Social Sybil Defense

Lorenzo Alvisi, Allen Clement, Alessandro Epasto, Silvio Lattanzi, and Alessandro Panconesi

Abstract. Sybil attacks, in which an adversary forges a potentially unbounded number of identities, are a danger to distributed systems and online social networks. The goal of sybil defense is to accurately identify sybil identities.

This article surveys the evolution of sybil defense protocols that leverage the structural properties of the social graph underlying a distributed system to identify sybil identities. We make two main contributions. First, we clarify the deep connection between sybil defense and the theory of random walks. This leads us to identify a community detection algorithm that, for the first time, offers provable guarantees in the context of sybil defense. Second, we advocate a new goal for sybil defense that addresses the more limited, but practically useful, goal of securely white-listing a local region of the graph.

1. Introduction

The possibility that malicious users may forge an unbounded number of *sybil* identities, indistinguishable from honest ones, is a fundamental threat to

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uinm.

ISSN: 1542-7951 print

distributed systems that rely on voting [Douceur 02]. This threat is particularly acute in decentralized systems, where it may be impractical or impossible to rely on a single authority to certify which users are legitimate [Margolin and Levine 05]. The goal of sybil defense is to accurately identify sybil identities[1]—"ideally, the system should accept all legitimate identities but no counterfeit entities" [Douceur 02]—but simple techniques can be either too brittle (beating a Completely Automated Turing Test to tell Computers and Humans Apart [CAPTCHA; Von Ahn et al. 03] costs a fraction of a cent) or too blunt (IP filtering penalizes all users behind a Network Address Translation [NAT]).

Against this background, Yu et al. have put forward a radically different approach [Yu et al. 06, 08]: protecting a distributed system by leveraging the social network that connects its users. Intuitively, as long as sybil identities are unable to create too many *attack edges* connecting them to honest identities, it may be possible to separate the wheat from the chaff by analyzing the topological structure of the users' social graphs. This style of sybil defense[2] promises not only to be more surgical, but offers a mathematically precise and elegant way to characterize the robustness of a sybil defense technique in terms of the number of attack edges it can handle. The vision is to offer *universal* sybil defense to all honest nodes in the system: as long as the social graph conforms to certain assumptions, an honest node will correctly classify almost all honest nodes in the graph while rejecting all but a bounded number of sybil nodes [Yu et al. 08].

Several protocols that embrace this style of sybil defense have since been proposed [Yu et al. 06, Danezis and Mittal 09, Tran et al. 11, Wei et al. 12, Cao et al. 12], and higher-level distributed applications that rely on them are beginning to emerge [Lesniewski-Laas 10, Lesniewski-Laas and Kaashoek 10, Quercia and Hailes 10, Tran et al. 09].

The first goal of this work is to examine the promise and the fundamental limits of universal sybil defense. Indeed, as [Viswanath et al. 10] pointed out in their recent analysis of social network-based sybil defenses it is not known whether "there are fundamental limits to using only the structure of social networks to defend against Sybils."

We offer a first answer to this question by establishing both precise theoretical bounds on the resilience of several well-known social network properties that

[1] Although this goal may be more accurately characterized as sybil *detection* [Viswanath et al. 12a], we use here the term sybil *defense* originally proposed by [Yu et al. 08] and widely adopted in the literature.

[2] Henceforth, mentions of sybil defense, unless specified otherwise, refer to techniques that leverage the structure of social networks.

have been leveraged in the context of sybil defense and by evaluating in depth the validity of the *social defense* vision.

As we shall see, at the core of social sybil defense are a set of assumptions about the structure of a social graph under sybil attacks that, in essence, amount to modeling the social graph as consisting of two sparsely connected regions: one comprised of sybil nodes, and the other of honest nodes, homogeneously connected with one another. We will discuss several studies, including our own experimental results, suggesting that this representation of the world lacks essential nuance. Rather, the evidence suggests that although honest entities in social graphs do organize in tightly knit overlapping communities, those communities together form a network that, as a whole, is more vulnerable than each single community.

Our second goal for this study is to advocate a realignment of the focus of sybil defense to leverage effectively the robustness of communities to sybil infiltration. The intuition that motivates us is not new. Prior work has suggested casting sybil defense as a community detection problem [Viswanath et al. 10] and asked whether it is possible to use off-the-shelf community detection algorithms to find sybil nodes. On this front, we make two contributions. First, we show that this approach requires extreme caution, as the choice of the community detection protocol can dramatically affect whether sybil nodes are accepted as honest. Second, we identify the mathematical foundations on which the connection between sybil defense and community detection rests: we identify a well-founded theory and point to established literature to guide the development of future sybil defense protocols.

Our conclusion is that instead of aiming for universal coverage, sybil defense should settle for a more limited goal: offering honest nodes the ability to white-list a set of nodes of any given size, ranked accordingly to their trustworthiness. We believe that this is a good bargain, and not just because it results in a goal that, unlike its alternative, is attainable, but because (i) the guarantees it provides are, in practice, what nodes that engage in crowd-sourcing [Yuen et al. 11] or cooperative peer-to-peer (P2P) applications [Pouwelse et al. 05, Cox and Noble 03] need, and (ii) the computational cost of providing these guarantees depends only on the size of the desired white-listed set rather than, as in techniques that aim for universal sybil defense, on the total number of identities in the network.

As a first concrete step toward fulfilling the new goal, we propose for sybil defense, we present the first community detection algorithm that offers provable guarantees in the context of sybil defense. Perhaps surprisingly, the algorithm is based directly on an application, in a context much different from which it was originally designed, of the random walk algorithm of [Andersen et al. 07].

Despite these advances, we believe that it is important to acknowledge that, however narrowing, a nontrivial gap still exists between the assumptions necessary to support the theory behind the current state-of-the-art sybil defense and the reality of sybil attacks encountered in the wild.

For example, evidence from the RenRen social network [Yang et al. 11] shows attacks that differ from what current sybil defenses anticipate and that, despite their simplicity, can be devastating.

The final goal of this work is to suggest that a promising way to address this challenge is through *defense in depth*, where early defense layers (of which we sketch a few) are designed to catch the simple sybil subgraphs where defenses based on community detection techniques fail and, as a side effect, to "nudge" the attacker toward precisely those settings where these techniques can effectively detect sybil nodes.

1.1. Roadmap

This article proceeds as follows. Section 2 examines four structural properties of social graphs (popularity, small-world property, clustering coefficient, and conductance) that have been previously leveraged by sybil defense and asks: which can better serve as a foundation for sybil defense? The answer, we find, is conductance, a property intimately related to the concept of mixing time of a random walk. We then proceed in Section 3 to discuss protocols that exploit variations in conductance as a basis for decentralized universal sybil defense [Danezis and Mittal 09, Tran et al. 11, Wei et al. 12, Yu et al. 06, 08]. These protocols provide elegant worst-case guarantees when it comes to their vulnerability to sybil attacks, but these guarantees are critically sensitive to a set of assumptions that do not appear to hold in actual social networks [Bilge et al. 09, Leskovec et al. 08, Mohaisen et al. 10]. This motivates us to explore, beginning with Section 4, an alternative goal for sybil defense that leverages two observations: (i) social graphs have an internal structure organized around tightlyknit communities and (ii) the graph properties crucial for sybil defense are significantly more likely to hold within a community rather than in the entire social graph. Section 5 reviews recent work on the theory of random walks that provides a solid theoretical foundation to sybil defense based on community detection; we deepen our investigation of random walks in Section 6, where we show how the well-known concept of Personalized PageRank (not to be confused with PageRank itself) offers honest nodes a path toward a realistic target for sybil defense, more limited than universal coverage but nonetheless useful: a way to white-list trustworthy nodes that proves efficient and robust in both theory and practice. Section 7 greets us with a sobering result: in spite of their sophistication, state-of-the-art

sybil defense protocols seem helpless against very crude real-life sybil attacks. There is reason for hope (and future research), however: we show that sybil defense protocols based on random walks continue to be effective when used in combination with very simple checks that leverage structural properties of the social graph other than conductance. Section 8 offers our conclusions and points to directions for future research.

2. Foundations of Sybil Defense via Social Networks

Sybil defense via social networks is predicated on the assumption that it is possible to leverage the structural properties of the social graph G underlying a distributed system to differentiate the subgraph H comprising only of honest nodes from the sybil subgraph S. In this section, we ask a basic question: which structural property, if any, holds the greatest promise toward defending against sybil attacks?

We consider (and briefly review below) four well-known structural properties of a social graph: the popularity distribution among its nodes, the small-world property, the graph's clustering coefficient, and its conductance [Barabási and Albert 99, Watts and Strogatz 98, Leskovec et al. 08]. We focus on these particular properties because of their prominence in social network analysis and because they have been used to defend against sybil attacks.[3] The literature on social graphs discusses several other properties (including assortativity [Newman 03], betweenness centrality [Freeman 77], and modularity [Newman and Girvan 04]) that we do not consider: we see this paper as a first step towards a comprehensive characterization of the defensive powers of the structural properties of social graphs.

2.1. Structural Properties of Social Graphs

Popularity: The node degree distribution of social graphs is heavy-tailed, as in a power-law or lognormal distribution.

Small-world property: The diameter of a social graph—i.e., the longest distance between any two nodes in the graph—is small.

Clustering coefficient: A measure of how closelyknit social networks are. When we associate the vertex of a social network with the user that it represents,

[3]More specifically: conductance is at the heart of social network-based sybil defense [Yu et al. 06]; the clustering coefficient has been used for sybil defense in a recent work [Yang et al. 11]; node degrees are used as a feature in a recent defense technique based on machine learning [Yang et al. 13]; and the distance between nodes plays a fundamental role in other recent defense schemes [Xu et al. 10, Viswanath et al. 12b].

the clustering coefficient is the ratio between the actual number of friendships between the friends of a user and the number of all possible friendships between them.

Formally, let f_v denote the actual number of edges between neighbors of a vertex v, i.e.,

$$f_v := |\{xy : x \in N_v, y \in N_v, xy \in E\},$$

where N_v denotes the set of neighbors of v, and let k be the maximum number of edges between neighbors of v:

$$k = \binom{\deg(v)}{2},$$

where $\deg(v)$ denotes v's degree. Then,

$$c_v := \frac{f_v}{k}.$$

The clustering coefficient of a graph is the average clustering coefficient of all its vertices, i.e.,

$$c(G) := \frac{1}{|V|} \sum_{v \in V(G)} c_v.$$

Conductance: Intuitively, the conductance $\phi(C)$ of a set C of vertices in a given network $G = (V, E)$ is the ratio between the number of edges going out from C and the number of edges inside C. More precisely, given a set of vertices C, the conductance of the set is defined as

$$\phi(C) := \frac{|\mathrm{cut}(C)|}{\mathrm{vol}(C)},$$

where the *volume* of C, $\mathrm{vol}(C)$, is defined as the sum of the degrees of the vertices in C

$$\mathrm{vol}(C) := \sum_{v \in C} \deg(v),$$

and the *cut* induced by C is the set $\mathrm{cut}(C)$ of edges with one endpoint in C and the other endpoint outside of C,

$$\mathrm{cut}(C) := \{uv \in E : u \in C, v \in V - C\}.$$

Finally, the *conductance* of a graph G is defined as

$$\phi(G) := \min_{\mathrm{vol}(C) \leq |E|} \phi(C).$$

The conductance of a graph is tightly related to its *mixing time* [Sinclair 92], a property that is at the core of many solutions developed to date for sybil defense [Yu et al. 06, 08]. Informally, the mixing time of a graph measures how fast a random walk approaches the stationary distribution. A more precise definition relies on a few important notions about random walks, which we now quickly review.

Given an undirected graph $G = (V, E)$ we define the *uniform* random walk in G as the random walk defined by the following transaction probability matrix:

$$P(u,v) = \begin{cases} \frac{1}{\deg(u)} & \text{if } uv \in E, \\ 0 & \text{otherwise.} \end{cases}$$

It is a well-known result of the theory of Markov chains (see for instance [Mitzenmacher and Upfal 05]) that any connected, non-bipartite graph has a unique stationary distribution π that depends only on the degree of the nodes:

$$\pi(v) = \frac{\deg(v)}{\mathrm{vol}(V)}.$$

Hence, if $P^t(u,v)$ is the probability of reaching node v from node u after a t-step-long random walk, we have that for all u and v

$$\lim_{t\to\infty} P^t(u,v) = \pi(v).$$

Assume now to start a random walk at a given node u and to perform t steps. The *variation distance* $\Delta_u(t)$ measures how closely the probability distribution of the endpoint approximates the stationary distribution

$$\Delta_u(t) = \frac{1}{2}\sum_v |P^t(u,v) - \pi(v)|.$$

We are finally ready to formalize the notion of mixing time.

Definition 2.1. (Mixing time). The mixing time $T(\epsilon)$ of a random walk, for any $\epsilon > 0$, is given by

$$T(\epsilon) = \max_{u\in V} \min_t \{t : \Delta_u(t) < \epsilon\}.$$

A crucial assumption underlying most of the work in social sybil defense [Mohaisen et al. 10] is that social networks are *fast mixing*, i.e., that their mixing time is $T(\epsilon) = \min(\log(n), \log(\frac{1}{\epsilon}))$, where n is the number of vertices. For $\epsilon = \Theta(\frac{1}{n})$, this implies a mixing time $T(\epsilon) = O(\log(n))$. We define τ as $T(\frac{1}{n})$.

As we have mentioned, the mixing time of a graph is intimately related to its *conductance*. Intuitively, when conductance is high, mixing time is low. In particular, it is possible to show that a class of networks is fast mixing (i.e., τ is $O(\log n)$) if and only if its conductance is asymptotically *constant* [Mitzenmacher and Upfal 05].

2.2. Preliminaries

Before proceeding with our analysis, we review a few important concentration results.[4]

Theorem 2.1. (Markov inequality). *For any random variable X with nonnegative values and for any $\epsilon > 0$*

$$P(X > \epsilon) \leq \frac{E[X]}{\epsilon}.$$

Theorem 2.2. (Chernoff bound). *Let $X = \sum_i^n X_i$ for $X_1, \cdots X_n$ independent random variables in $[0, 1]$. Then*

$$P(|X - E[X]| > \epsilon E[X]) \leq 2 \exp\left(-\frac{\epsilon^2}{3} E[X]\right).$$

Definition 2.2. (Lipschitz Condition). A function f satisfies the *Lipschitz condition* with respect to the random variables $X_1, \cdots, X_n$ with parameters c_j, $1 \leq j \leq n$, if for any $1 \leq j \leq n$ and a_j, a'_j.

$$\begin{aligned} |f(X_1 = a_1, \cdots, X_{j-1} = a_{j-1}, X_j = a_j, X_{j+1} = a_{j+1}, \cdots, X_n = a_n) \\ - f(X_1 = a_1, \cdots, X_{j-1} = a_{j-1}, \\ X_j = a'_j, X_{j+1} = a_{j+1}, \cdots, X_n = a_n)| \leq c_j. \end{aligned} \tag{2.1}$$

Theorem 2.3. (Bounded differences inequality). *Assume that f satisfies the Lipschitz condition with respect to the random variables $X_1, \cdots, X_n$ with parameters c_j, $j \in [n]$. Then*

$$P\left(|f - E[f]| > t\right) \leq 2 \exp\left(-\frac{t^2}{2c}\right),$$

[4] For a more comprehensive treatment, see [Mitzenmacher and Upfal 05] and [Dubhashi and Panconesi 09].

where $c = \sum_{j=1}^{n} c_j^2$.

Finally, henceforth we say than an event E occurs *with high probability* if $\lim_{n\to\infty} P(E) = 1$, where n is the number of vertices in the graph.

2.3. Which Property is More Resilient?

If an attack can add sybil identities to a social network G consisting only of honest nodes without altering a given structural property of G, then that attack will be undetectable by any sybil defense technique that leverages that property. To assess the suitability of a property to serve as a basis for sybil defense, we then compare, under a given adversarial model, the effort required to create an undetectable attack.

To this end, we assume that a graph H with n honest nodes is given and that the attack induces a graph S of sybil nodes whose topology is under total control of the adversary (unlike H, which is fixed). For each property Π, we characterize the adversary's effort as the number of edges incident to H that the adversary needs to add in order to introduce n sybil nodes in a way undetectable to Π.

To establish clear and almost-tight bounds on the number of attack edges necessary, we introduce a simple but powerful attacker that we will use for some of our bounds. To avoid detection, our adversary starts by building S so that, when looked through the filter of Π, it looks similar to H. For simplicity and only for the purpose of deriving the bounds for popularity, clustering coefficient, and diameter, we assume that the adversary builds S as a copy of H.[5]

The adversary then tries to set up $m := |E(H)|$ potential attack edges that connect H with S. The probability of a node v becoming an endpoint of an attack edge is proportional to v's degree:

$$\frac{\deg_H(v)}{2m}, \tag{2.2}$$

As we will see, this factor is crucial in leaving unaltered the properties of the social graph and in particular its degree distribution.

If the attacker is able to create arbitrarily many attack edges, no sybil defense can hope to distinguish between the two regions of the graph. Therefore, as is

[5] Although in practice it is neither necessary nor likely, this assumption, without qualitatively altering our conclusions, leads to simple bounds on the effort required to make attacks undetectable to defenses based on popularity, network diameter, and clustering coefficient. Note that neither the conductance bound nor the theorems about ACL (see Section 6) rely on this assumption.

customary in the sybil defense literature [Yu et al. 06, 08], we assume that the attacker's ability to create attack edges is limited; in particular, we postulate that tentative attack edges are accepted with probability p and rejected with probability $1 - p$. To account for the outcome of recent social engineering experiments [Bilge et al. 09], we allow p to be constant, resulting in an expected number of attack edges equal to pm. It follows easily from large deviation theory that, if m is large enough, the number of attack edges is also concentrated around pm. We denote with G the graph that results from joining H and S through the set of attack edges. We define R to be the set of *tentative* attack edges the attacker attempts to introduce. Similarly, let g be the number attack edges the attacker succeeds in establishing.

Under this simple attack model, how resilient, then, are the four structural properties of social graphs that we are considering?

2.3.1. Popularity. The intuitive motivation for popularity as a basis for social defense is that the degree distribution of nodes may be noticeably altered as a result of an adversary introducing a large number of attack edges, thereby providing evidence of an attack. Under our attack model, however, we show that the adversary can ensure that G's popularity distribution will be statistically indistinguishable from that of H even after establishing many attack edges. Intuitively, since the nodes at the endpoints of an attack edge are chosen with probability proportional to their degree, after the attack, only a few nodes will see their degree change by much; in fact, the degree of a vertex in H will increase, in expectation, by only $\frac{\deg_H(v)}{2m} p|R|$.

This intuition is formalized in the following simple proposition.

Proposition 2.1. *Let H be the input graph, S be the attack graph, R the set of tentative attack edges between S and H and p the probability that each attempt to add an attack edge succeeds. Finally, let G be the resulting (random) graph. Then, for each $v \in G$,*

$$E[\deg_G(v)] = \deg_H(v)\left(1 + |R|\frac{p}{2m}\right),$$

where m is the number of edges in the honest region. Furthermore, if $\deg_H(v) > \log^2 n$ in H, then with high probability the final degree of honest nodes is concentrated, i.e.,

$$\deg_H(v)\left(1 + |R|\frac{p}{2m} - o(1)\right) \leq \deg_G(v) \leq \deg_H(v)\left(1 + |R|\frac{p}{2m} + o(1)\right).$$

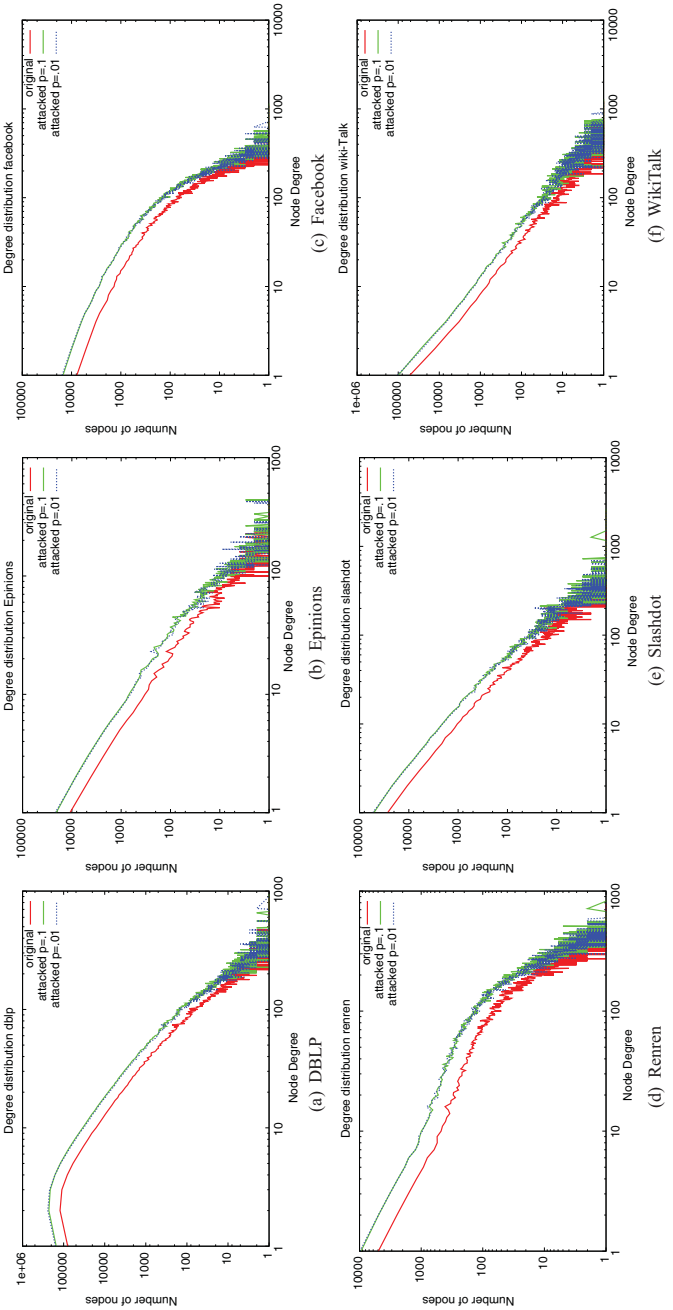


Figure 1. Degree distribution of the graphs before and after attack. The attack shifts the distribution up (because it doubles the size of the graph) and to the right (proportionally to the number of attack edges), but does not change the shape of the curves.

Proof. Let v be a node in H (for nodes in S the same analysis applies). Then, because all attack edges are added with probability proportional to the original degree of the node in H, $E[\deg_G(v)] = \deg_H(v) + |R|\frac{\deg_H(v)}{2m} = \deg_H(v)\left(1 + |R|\frac{p}{2m}\right)$. Furthermore, if v's degree is larger than $\log^2 n$ in H, then by the Chernoff bound (Theorem 2.2) it follows that:

$$P(|\deg_G(v) - E[\deg_G(v)]| > 6\log^{1.5} n) \leq e^{-\frac{1}{3}6\log n} \in O(n^{-2}).$$

Thus, with high probability, the post-attack degree of a node v with degree larger than $\log^2 n$ will be

$$\deg_H(v)\left(1 + |R|\frac{p}{2m} - o(1)\right) \leq \deg_G(v) \leq \deg_H(v)\left(1 + |R|\frac{p}{2m} + o(1)\right).$$

□

Our experiments with real-life social networks confirm the above conclusion. Figure 1 shows the degree distribution of snapshots of several social networks before and after two attacks in which attack edges are inserted with probability $p = 0.01$ and $p = 0.1$, respectively: the curves before and after the attacks have the same shape. We conclude that popularity is ill-suited as a foundation for sybil defense.

2.3.2. Small-World Property. The small-world property does not fare much better than popularity, since the adversary can easily keep the diameter of G from growing suspiciously. First, it is easy for the adversary to bound the relative growth of the diameter of G with respect to that of H: if $S = H$ and the adversary succeeds in inserting just one attack edge, the diameter can at most double. Our experimental evaluation of several real-life social networks shows (see 90% diameter column of Table 1) that the 90%-effective diameter [Leskovec et al. 05], which measures the maximum distance between 90% of the pairs of nodes, is indeed barely affected under attack.

2.3.3. Clustering Coefficient. Leveraging the clustering coefficient appears promising because attack edges reduce its value. Unfortunately, although the clustering coefficient of social networks is typically high, its value varies significantly from network to network [Leskovec et al. 08], from 0.79 in the actor collaboration network of IMDB, down to 0.35 for Live Journal and to a mere 0.09 for the social network of Yahoo! Messenger chat exchanges. Thus, if an attack modifies the clustering coefficient by a small multiplicative factor, the change is difficult to detect. This intuition is captured in the following result.

Lemma 2.2. *Let H be the input graph, S be the attack graph obtained by copying H and p the probability that a tentative attack edge succeeds in attaching to a node*

in H. Then, if $c(H)$ is the clustering coefficient of H and $c(H) \in O(1)$, with high probability, $c(G) \geq \alpha^{-1} c(H)$, where $\alpha := 14(1 + \frac{1}{2}p)^2$.

Proof. We show that the insertion of attack edges does not increase, for most nodes, their degree by much. This implies a lower bound for the final clustering coefficient of the graph under attack.

First, note that, by definition, all nodes of degree 1 in H have a clustering coefficient of 0. So, in the following we consider only nodes with degree greater than 1. After the attack, the expected degree of a node v in G is equal to $\deg_H(v)(1 + \frac{1}{2}p)$. By the Markov inequality (Theorem 2.1), it follows that the final degree of v is at least $\frac{3}{2}\deg_H(v)(1 + \frac{1}{2}p)$ with probability less than $\frac{2}{3}$. So each node v has $\deg_G(v) < \frac{3}{2}\deg_H(v)\left(1 + \frac{1}{2}p\right)$ with probability at least $\frac{1}{3}$ and thus

$$\deg_G^2(v) < \frac{9}{4}\deg_H^2(v)\left(1 + \frac{1}{2}p\right)^2.$$

$$\deg_G(v)(\deg_G(v) - 1) < \frac{9}{4}\left(1 + \frac{1}{2}p\right)^2 \deg_H(v)$$

$$\times(\deg_H(v) - 1) + \frac{9}{4}\left(1 + \frac{1}{2}p\right)^2 \deg_H(v).$$

As $\deg_H(v) > 1$ we have

$$\deg_G(v)(\deg_G(v) - 1) < \frac{9}{2}\left(1 + \frac{1}{2}p\right)^2 \deg_H(v)(\deg_H(v) - 1).$$

It follows that the clustering coefficient of each node decreases by, at most, a factor of $\frac{27}{2}(1 + \frac{1}{2}p)^2$ and that, by linearity of expectation, the clustering coefficient of G decreases by, at most, a factor of $\frac{27}{2}(1 + \frac{1}{2}p)^2$.

Consider the sum of the clustering coefficients of the nodes in the graph H (the same argument applies also to nodes in S). By assumption, we know that this sum is $\Theta(|V_H|)$. Now, each node $v \in H$ contributes to this sum by at most 1 and, with probability at least $\frac{1}{3}$, by $\frac{c_H(v)}{\frac{9}{2}(1+\frac{1}{2}p)^2}$, where $c_H(v)$ is the initial value of v's clustering coefficient.

By linearity of expectation, the expected sum of the clustering coefficients after inserting the attack edges is also in $\Theta(|V_H|)$. To prove that the bound promised by the lemma holds with high probability, we then apply the Bounded Difference Inequality (Theorem 2.3) with a Lipschitz-condition coefficient of $c_j = 1$ for each

Graph	Nodes	Edges	Attack Edges	Diameter	90% Diameter	Clust. Coeff	Est. Conductance
AstroPh	17903	196972	0	14	4.99	0.67	0.010
... $p = 0.01\%$	35806	395924.2 (±22.2)	1980.2 (±22.2)	14.1 (±0.2)	5.63 (±0.006)	0.66 (±0.0001)	0.005 (±0.0001)
... $p = 0.1\%$	35806	413663.3 (±54.4)	19719.3 (±54.4)	11.8 (±0.4)	4.90 (±0.004)	0.61 (±0.0003)	0.036 (±0.0039)
DBLP	718115	2786906	0	20	7.43	0.73	0.016
... $p = 0.01\%$	1436230	5601647.3 (±85.7)	27835.3 (±85.7)	19.4 (±0.4)	7.95 (±0.004)	0.72 (±0.0000)	0.004 (±0.0003)
... $p = 0.1\%$	1436230	5852543.5 (±224.4)	278731.5 (±224.4)	17.0 (±0.3)	7.01 (±0.010)	0.67 (±0.0001)	0.013 (±0.0012)
Enron	33696	180811	0	12	4.83	0.70	0.005
... $p = 0.01\%$	67392	363426.6 (±24.6)	1804.6 (±24.6)	12.4 (±0.3)	5.12 (±0.013)	0.70 (±0.0001)	0.005 (±0.0002)
... $p = 0.1\%$	67392	379691.2 (±71.8)	18069.2 (±71.8)	10.9 (±0.3)	4.67 (±0.004)	0.64 (±0.0004)	0.022 (±0.0015)
Epinions	26588	100120	0	16	5.98	0.23	0.020
... $p = 0.01\%$	53176	201240.1 (±20.0)	1000.1 (±20.0)	16.4 (±0.2)	6.73 (±0.009)	0.23 (±0.0001)	0.005 (±0.0001)
... $p = 0.1\%$	53176	210213.5 (±36.3)	9973.5 (±36.3)	14.6 (±0.3)	5.97 (±0.005)	0.21 (±0.0002)	0.030 (±0.0026)
EuAll	32430	54397	0	9	4.57	0.52	0.031
... $p = 0.01\%$	64860	109337.5 (±13.8)	543.5 (±13.8)	9.9 (±0.1)	5.06 (±0.024)	0.51 (±0.0004)	0.005 (±0.0001)
... $p = 0.1\%$	64860	114245.0 (±33.2)	5451.0 (±33.2)	8.6 (±0.2)	4.70 (±0.002)	0.42 (±0.0008)	0.051 (±0.0057)
Facebook	63392	816886	0	12	5.15	0.25	0.020
... $p = 0.01\%$	126784	1641941.2 (±46.1)	8169.2 (±46.1)	14.2 (±0.2)	5.79 (±0.002)	0.25 (±0.0000)	0.005 (±0.0000)
... $p = 0.1\%$	126784	1715443.9 (±121.8)	81671.9 (±121.8)	13.2 (±0.2)	5.24 (±0.005)	0.23 (±0.0001)	0.031 (±0.0042)
RenRen	33294	705248	0	11	4.29	0.23	0.032
... $p = 0.01\%$	66588	1417543.1 (±48.4)	7047.1 (±48.4)	12.9 (±0.1)	4.82 (±0.002)	0.23 (±0.0000)	0.005 (±0.0000)
... $p = 0.1\%$	66588	1481107.0 (±68.3)	70611.0 (±68.3)	11.6 (±0.2)	4.44 (±0.002)	0.21 (±0.0001)	0.060 (±0.0040)

(Continued on next page)

Slashdot	70999	365572	0	11	4.84	0.10	0.023
... $p = 0.01\%$	141998	734795.4 (±26.5)	3651.4 (±26.5)	12.0 (±0.0)	5.49 (±0.005)	0.10 (±0.0000)	0.005 (±0.0000)
... $p = 0.1\%$	141998	767694.4 (±85.1)	36550.4 (±85.1)	11.1 (±0.1)	4.92 (±0.002)	0.09 (±0.0001)	0.036 (±0.0042)
WikiTalk	92117	360767	0	9	4.63	0.14	0.047
... $p = 0.01\%$	184234	725141.2 (±26.0)	3607.2 (±26.0)	10.1 (±0.1)	5.01 (±0.005)	0.13 (±0.0000)	0.005 (±0.0000)
... $p = 0.1\%$	184234	757628.1 (±79.5)	36094.1 (±79.5)	10.0 (±0.0)	4.76 (±0.001)	0.12 (±0.0001)	0.048 (±0.0007)

Table 1. Statistical properties of the largest connected component in a collection of real world data sets.

The values reported reflect the properties of the data set before and after the attack specified in Section 2.3. The results for sybil graphs are averaged over 20 attack instances and the 95% confidence intervals, obtained by the t-student distribution [Walpole et al. 93], are reported between parenthesis. In directed graphs, we removed edge direction to obtain an undirected network. The AstroPh [Leskovec et al. 07] is a co-authorship graph from 2003; the DBLP [dblp 11] graph is a snapshot of the DBLP coauthor graph from 2011; the Enron [Klimt and Yang 04, Leskovec et al. 09] graph is an e-mail communication network from 2009; the Epinions [Richardson et al. 03] graph is a dataset from the Epinions product review site obtained in 2003; the EuAll [Leskovec et al. 07] graph is an e-mail communication network of a European research institution from 2005; the HE Physics [Leskovec et al. 05] graph is a citation network of high-energy physics from 2003; the Facebook [Viswanath et al. 09] graph is a crawl of the Facebook-New Orleans community in 2007; the RenRen [Jiang et al. 10] graph is snapshot of the RenRen social network from 2009; the Slashdot [Leskovec et al. 09] graph is a crawl of the website social network from 2008; the WikiTalk [Leskovec et al. 10] graph is derived from the Wikipedia page edit history as of January 2008.

of the random variables corresponding to the clustering coefficient of the nodes in H. □

Note that the constants in the theorem are large only to make the statement hold with high probability. In practice, one can expect much smaller variations, as shown in Table 1.

The implications of this lemma are disappointingly clear: the clustering coefficient is not a sound basis for sybil defense, because, even after the attack, its value cannot drop by too much. The *Clustering Coeff* column of Table 1 confirms the theorem's predictions.

Note that even though the theorem applies to the clustering coefficient of the graph, a similar observation holds for the clustering coefficient of each single node, as the degrees of almost any node change by a tiny multiplicative factor. Thus, sybil defense techniques that rely solely on analyzing the clustering coefficient of each node [Yang et al. 11] can be easily circumvented by a capable attacker.

2.3.4. Conductance. [Yu et al. 08] proved that if H belongs to a class of graphs whose conductance is asymptotically constant, an adversary that can introduce $O(n)$ attack edges to build a graph G whose conductance is indistinguishable from that of H. In the following, we generalize this result to graphs H of arbitrary conductance.

We begin with two preliminary observations. First, because, by definition, the conductance of a graph is the minimum of $\phi(C)$ on any subset C of the graph's vertices, an adversary can always enforce $\phi(G) \in O(\phi(H))$ by introducing a suitable cut in the sybil region, whose topology is under his complete control.

Second, an adversary who wants to introduce n sybil nodes needs to add at least $n\Omega(\phi(H))$ edges, lest the cut between the sybil and honest part of G become too sparse, making it easy to use changes in conductance to detect the attack.

We now show that, by adding just a few more edges, an adversary, as defined earlier, can ensure that $\phi(G) \in \Omega(\phi(H))$.

Theorem 2.4. *Let H denote a network of n honest nodes with conductance ϕ such that $\phi \operatorname{vol}(H) \in \Omega(\log n)$ and $\phi \leq e^{-1}$, and let S be a copy of H. Suppose that the adversary is able to establish between S and H $\phi \log(\phi^{-1})\operatorname{vol}(H)$ attack edges, whose endpoints are selected with probability proportional to the degrees of the nodes. Let G be the resulting graph. Then, with high probability, $\phi(G) \in \Omega(\phi)$.*

Theorem 2.4 is actually a direct consequence of the following, more general result.

Theorem 2.5. *Let $H = (V, E)$ be a connected simple graph such that $\phi(H)\mathrm{vol}(V) \in \Omega(\log n)$, $\phi(H) \leq \frac{1}{e}$ and let $S = (V', E')$ be another connected simple graph with $\phi(S) \geq \phi(H)$. Suppose further that*

$$\phi(H)\mathrm{vol}(V) \leq \mathrm{vol}(V') \leq \mathrm{vol}(V).$$

Let $G_F = (V_F, E_F)$ be the union of S with H and let g be the number of random attack edges between H and S, whose endpoints are selected with probability proportional to the degrees of the nodes. Then, if

$$\log \phi(H)^{-1} \cdot \phi(H) \cdot \mathrm{vol}(V) \leq g \leq \mathrm{vol}(V'),$$

we have that, with high probability, $\phi(G_F) \in \Omega(\phi(H))$.

Note that the assumption that $\phi(H) \leq \frac{1}{e}$, which restricts somewhat the generality of the result, holds in real networks.

In order to avoid disrupting the flow of the article, we defer to the Appendix the rather long proof of Theorem 2.5. Its fundamental implication, however, is clear: if the adversary is able to introduce at least $\phi(H)\mathrm{vol}(V)\log\frac{1}{\phi(H)}$ attack edges (or $O(\mathrm{vol}(V))$ when the mixing time is $O(\log n)$), then the conductance of the graph will remain, with high probability, very nearly the same as that of H. This in turn implies that the mixing time of the network does not change after the attack, and so it is hard to detect such an attack using this property.

Theorem 2.5 then allows us to draw mixed conclusions about the suitability of conductance for the sybil defense. On the one hand, it proves that detection techniques based on changes in global conductance can in principle be circumvented; on the other, it shows that the effort required to do so is much higher for conductance than for any of the other properties we have considered.

Table 1 confirms the theorem's message. As expected, conductance drops significantly under a weak attack ($p = 0.01$), providing leverage for sybil detection. Under a strong attack ($p = 0.1$), however, conductance may actually *increase* because, by adding random attack edges, the adversary enlarges every cut with some probability, including the cut with minimum conductance that defines the conductance of the entire final graph.[6]

Note that computing a graph's conductance is NP-hard. The conductance values that we report are computed by a widely used technique proposed in recent social network literature [Leskovec et al. 08].

[6]Note that any hope of using an increase in conductance as an indication of a possible attack is futile, because the adversary can always insure that conductance is below a threshold by creating a sparse cut in S.

Property	Number of edges to circumvent it
Degree distribution	$g \geq 0$
Diameter	$g \geq 1$
Clustering coefficient	$0 \leq g \leq m$
Conductance	$\phi(G)m \log \phi(G)^{-1} \leq g \leq m$

Table 2. The number g of attack edges needed to circumvent the four properties.

2.4. Discussion

None of the structural properties of social graphs that we have considered provides an impregnable defense against sybil attacks in general, or even against the specific attack we have assumed. However, as Table 2 shows, when a graph under attack is observed through the lens of conductance, the adversary has to work much harder to look inconspicuous. These results both motivate and justify the insight of Yu and his collaborators to rely on conductance in the work that jump-started sybil defense via social networks [Yu et al. 06]. We review their approach, its successes, and what we believe to be ultimately its fundamental limitations, in the next section.

3. Leveraging Conductance Toward Universal Sybil Defense

The vision behind the seminal work of Yu and his collaborators was to develop a decentralized approach to *universal sybil defense*, with the goal of allowing honest users to correctly assess with high probability the honesty of every other user in the system. False positive and false negatives would still be possible, but they would be few and, further, their number would be bound within a rigorous theoretical framework. This compelling vision, first articulated in the SybilGuard protocol [Yu et al. 06], was further refined in their later work on the SybilLimit protocol [Yu et al. 08] and has inspired several other efforts in sybil defense [Danezis and Mittal 09, Tran et al. 11, Wei et al. 12, Cao et al. 12].

We begin this section by discussing the main intuition underlying these techniques and the guarantees that they provide; we then proceed to discuss the crucial role that a set of key assumptions play in ensuring those guarantees, and present evidence suggesting that the assumptions do not appear to hold in actual social graphs.

3.1. Picking Whom to Trust

The verification process that an honest node u uses in the above protocols to determine whether it can trust another node v is based, at its core, on the following idea: use a random walk to sample some portion of the graph uniformly at random and identify which nodes to trust on the basis of that sample. Different protocols apply this sampling strategy in different ways and to different parts of the graph. SybilLimit [Yu et al. 08] samples edges; SybilGuard [Yu et al. 06] and Gatekeeper [Tran et al. 11] sample nodes in the graph; SybilInfer [Danezis and Mittal 09] uses the random walks to build a Bayesian model for the likelihood that a trace T was initiated by an honest node. In the remainder, we provide an overview of how SybilLimit [Yu et al. 08] applies the random sampling of edges to identify honest users. Although the details of the discussion are specific to SybilLimit, the intuition for how the structural properties of the graph make random sampling effective is common to this entire family of protocols.

Let us consider a particularly simple version of the sybil detection problem. We are given two disjoint graphs H and S—the graph of honest and, respectively, sybil nodes; an honest vertex u—the seed; and a vertex v. Our task is to determine whether v belongs to H or to S. Suppose that both nodes select an edge at random, subject to the constraint that they must pick an edge from the graph they belong to: u accepts v if they pick the same edge.

If the vertices belong to different graphs, the test is perfect: the probability that u accepts v is 0. Otherwise, the probability of collision is very low, $\frac{1}{m}$, but it can be boosted thanks to the classic birthday paradox. Vertex u picks a set S_u of, say, $\sqrt{m}$ distinct edges, while v picks a set S_v of $\sqrt{m}$ edges independently at random: now u accepts v if there is a collision (i.e., $S_u \cap S_v \neq \varnothing$). This probability is

$$1 - \Pr(\text{no collision}) = 1 - \left(1 - \frac{1}{\sqrt{m}}\right)^{\sqrt{m}} \sim 1 - \frac{1}{e}, \tag{3.1}$$

a good probability of success. Note now that the set S_u can itself be picked at random. Since $|S_u| = \sqrt{m} \ll m$, almost all edges will be distinct. This simple protocol succeeds with good probability: each vertex picks a set of $\sqrt{m}$ edges independently and uniformly at random. If the two sets intersect, then u accepts v, otherwise it does not. The protocol is symmetric and can be used by both u and v to determine whether to trust one another. This basic idea can be further refined to obtain a test that succeeds with overwhelming probability with small-sized edge sets.

With this protocol, the probability that an honest seed accepts a sybil node remains 0, while the probability of accepting another honest node can be pushed to 1 at an acceptable computational cost. But how can we implement the test

in a distributed fashion? It is here that *mixing time*, and hence conductance, enter the picture. A simple approach is to take a random walk in the graph—which, in the interest of efficiency, should be very short—and pick the last edge on the walk. This is a correct implementation of the previous protocol provided that the graph is fast mixing. Indeed, as we saw in Section 2.2, if a graph is fast mixing, the probability that a random walk of length $O(\log(n))$ ends in u is approximately $\frac{\deg(u)}{2m}$. If we pick a random edge $e = (u, v)$ incident to the final vertex of the walk, the edge is picked with probability approximately equal to

$$\frac{\deg(u)}{2m}\frac{1}{\deg(u)} + \frac{\deg(v)}{2m}\frac{1}{\deg(v)} = \frac{1}{m},$$

which means that each edge is picked uniformly at random.

In reality, however, H and S are connected through the attack edges that nodes in S have convinced nodes in H to accept: it is then possible that a random walk starting from $v \in S$ will traverse an attack edge, enter H, and pick one of the edges selected by $u \in H$. The intuition is that, as long as the cut between H and S is sparse, in such situations it is sufficiently unlikely that the mechanism continues to function with good probability. Indeed, as we already mentioned, recent work has proved that as long as the number of attack edges is bound by $o(\frac{n}{\log n})$, then this approach can reliably distinguish between honest and sybil nodes [Yu et al. 06].

3.2. Limitation of the Model

There are, then, two fundamental assumptions that underlie this elegant approach toward decentralized universal sybil defense. The first is that the cut between the sybil and honest region—the set of attack edges—is suitably sparse. The second is that the mixing time of the honest region is $O(\log(n))$. The combination of these two assumptions ensures the high probability that random walks of $\Theta(\log n)$ steps will end in a random edge in the honest region.

Recent literature has cast doubts on whether these assumptions hold in practice. Social graphs do not seem to be fast mixing after all [Mohaisen et al. 10], and fake identities are accepted as friends with much higher probability than anticipated [Bilge et al. 09, Yang et al. 11], implying that the set of attack edges is not as sparse as assumed. We then ask: to what degree are SybilLimit-like protocols sensitive to their assumptions about sparse cuts and mixing time?

To answer this question, using SybilLimit [Yu et al. 08] as representative (we find that the behavior of other SybilLimit-like protocols is similar), we produce,

as in the recent work of Viswanath et al., a ranking of nodes with respect to a given *verifier node* u, in decreasing order of trust: the first node in the ranking is the node that u trusts the most [Viswanath et al. 10]. We then measure the defensive efficacy of SybilLimit by using three metrics, well known in the field of information retrieval, that appear very natural in this context: *precision*, *recall*, and *ROC*. In particular, we define the precision at position k as the fraction of honest nodes among the k nodes that the protocol ranks the highest. Similarly, we define the recall at position k as the ratio between the number of honest nodes among the top k positions in the ranking and the total number of honest nodes in the network.

Another well-known accuracy measure, employed in our analysis, is the ROC index, which measures the probability that a randomly chosen honest node be considered more trustworthy than a randomly chosen sybil one. A probability of 1 corresponds to the ideal case in which every honest node is ranked higher than any sybil one; a probability of 0 indicates the reverse case; a random ranking corresponds to 0.5 probability.

3.2.1. Sensitivity to Mixing Time. SybilLimit-like protocols do not operate on raw social networks: they are to be used only on networks that have been preprocessed by iteratively removing all nodes with degree lower than five [Yu et al. 06]. Table 3 shows the statistical properties of the graphs we use in our experiments.

Mohaisen et al. were the first to observe that this step, while boosting the mixing time of social graphs to the level required by SybilLimit to be effective, can also reduce the size of the graph [Mohaisen et al. 10]. Table 3 confirms this observation: in the case of Wiki-Talk, the preprocessing step removes over 85% of the nodes. Removed nodes are effectively considered sybils by the protocol, and although those nodes might still be able in some circumstances to enlist other nodes in the network as proxies [Yu et al. 08], it is unclear in general how removed nodes can safely take advantage of honest nodes' resources and vice versa [Mohaisen et al. 10].

Figure 2 shows the impact of the preprocessing step on the performance of SybilLimit. Preprocessing increases the performances of SybilLimit in most networks, with the notable exception of the Enron network, where preprocessing *decreases* SybilLimit's performance: in this small and incomplete network (e-mail between contacts outside of the company is not available) eliminating low-degree nodes ends up disrupting severely the connectivity of the honest region.

3.2.2. Sensitivity to Sparse Cuts. Figure 3 plots SybilLimit's precision versus recall for the preprocessed Facebook dataset—a similar behavior is observed with all

Graph	Nodes	Edges	Diameter	90% Diameter	Clustering Coeff	Est. Conductance
AstroPh	17903	196972	14	4.99	0.67	0.010
... preprocessed	12118	162232	10	4.46	0.58	0.017
DBLP	718115	2786906	20	7.43	0.73	0.016
... preprocessed	191172	1438509	15	5.97	0.60	0.020
Enron	33696	180811	12	4.83	0.70	0.005
... preprocessed	9357	86656	10	4.90	0.47	0.005
Epinions	26588	100120	16	5.98	0.23	0.020
... preprocessed	5624	57341	7	3.89	0.18	0.040
EuAll	32430	54397	9	4.57	0.52	0.031
... preprocessed	1106	8569	5	3.49	0.18	0.222
Facebook	63392	816886	12	5.15	0.25	0.020
... preprocessed	40757	632597	7	4.43	0.23	0.023
Renren	33294	705248	11	4.29	0.23	0.032
... preprocessed	22032	473443	7	3.77	0.21	0.031
Slashdot	70999	365572	11	4.84	0.10	0.023
... preprocessed	17993	183406	8	3.82	0.03	0.027
Wiki-Talk	92117	360767	9	4.63	0.14	0.047
... preprocessed	13069	133343	5	3.78	0.06	0.333

Table 3. Statistical properties of the graphs before and after preprocessing. Preprocessing drastically reduces the graphs' sizes and significantly alters their structural properties.

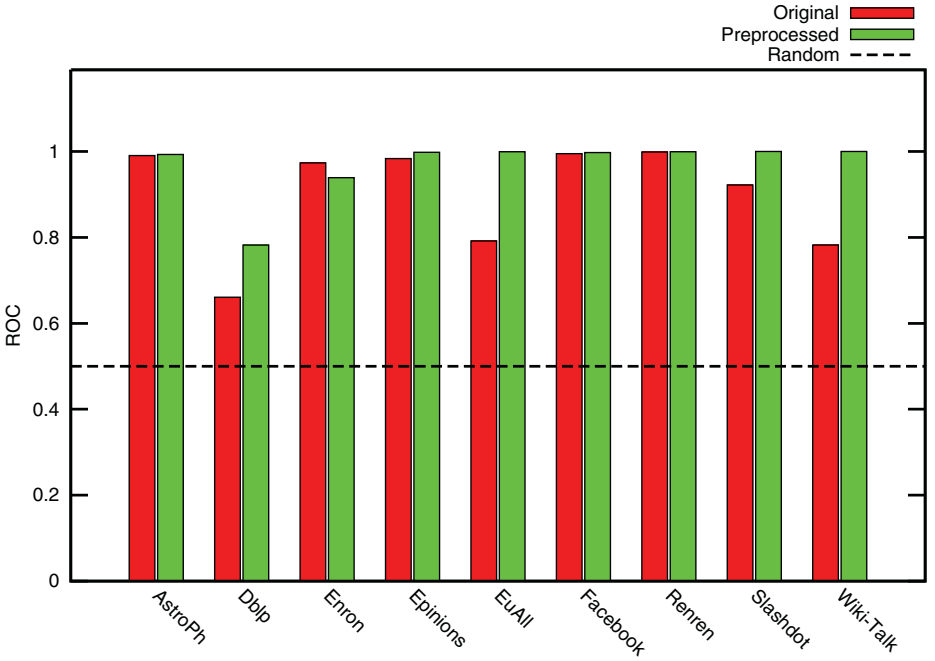


Figure 2. The ROC of the SybilLimit on each of the social networks we consider, when the graphs are attacked with attack strength $p = 0.01$. Other SybilLimit-like protocols show qualitatively similar results.

other networks in our dataset. SybilLimit proves very effective when the number of attack edges is within the theoretical bound (which corresponds to $p = 0.01$). Once the bound is exceeded, however, the performance of SybilLimit falls rather quickly: the algorithm can no longer ensure that at most $\log(n)$ sybil nodes per attack edge are admitted, leading to a sudden drop in the precision observed in our experiments.

3.3. Discussion

The goal of universal decentralized sybil defense with strong theoretical guarantees, which has driven early research on sybil defense via social networks, rests on assumptions (short mixing time and cut sparseness) whose validity is at best dubious. What to do? In a recent survey [Yu 11], Yu suggests a couple of ways forward: one could offer sybil defense only to the nodes in the core of the social graph, in effect institutionalizing the removal of nodes that are not well connected, or one could simply renounce the elegant theoretical worst-case claims

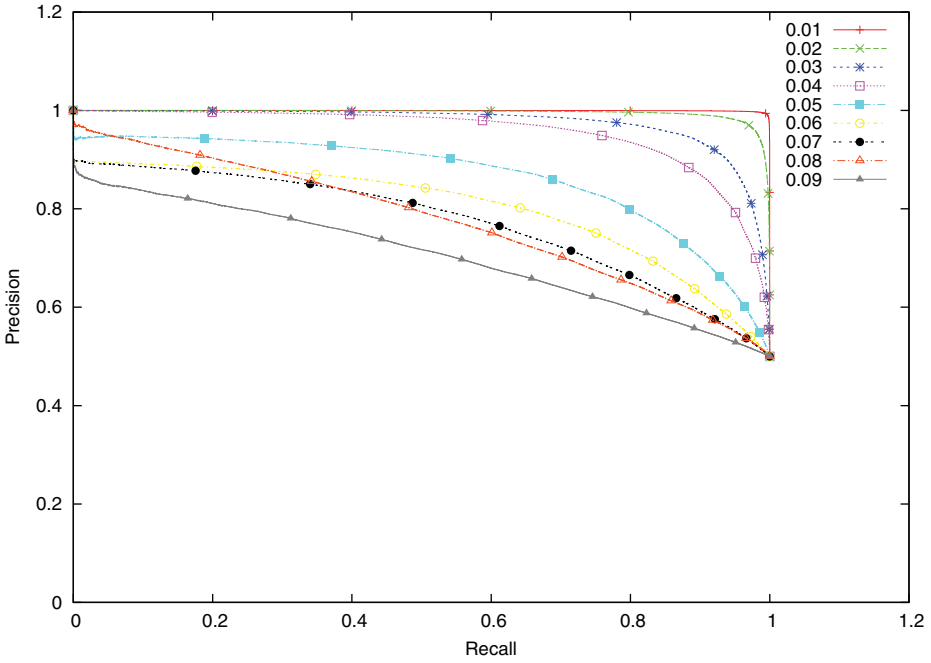


Figure 3. Precision vs. recall of SybilLimit for different values of p (ranging from 0.01 to 0.10). The number of attack edges is pm. The theoretical guarantee of SybilLimit-like holds only for $p = 0.01$. The results are shown for the Facebook network. As the number of attack edges goes beyond the prescribed limit, there is a significant drop in performance. The same test made with other graphs shows similar results.

of the current framework and rely instead on "weaker but less clean assumptions" [Yu 11]. In the next section, we explore a third alternative that offers every honest node a useful degree of sybil protection without compromising on elegance and rigor.

4. Communities

The theoretical guarantees offered by the protocols discussed so far hold only as long as honest nodes are closely connected to one another everywhere in the social graph and the cut between honest and sybil nodes is sparse. Empirical evidence suggests a different reality: social graphs consist of communities, each a tightly knit subnetwork. Indeed, it is quite conceivable that the cut between two tightly knit communities of honest nodes A and B be as sparse as the cut

between A and the sybil region: to an honest node in A using a protocol in the style of SybilLimit, a sybil node would then be indistinguishable from an honest node in B [Viswanath et al. 10, 12a].

Although these considerations argue against the universal sybil defense, they suggest an alternative goal: to provide each honest node u with the ability to white-list a trustworthy set of nodes—vis. those in the community to which u belongs. This new goal can be more precisely stated as follows:

Problem 4.1. Let u be an honest user and C be a subset of honest vertices in the social graph such that: (a) $u \in C$, (b) the graph induced by C has mixing time τ and (c) there are at most $o(|C|\tau^{-1})$ edges between C and the rest of the social graph. We want an algorithm (ideally, amenable to an efficient distributed implementation) that, given u and the social graph, can distinguish almost perfectly between the nodes in C and the nodes outside of C.

We make two observations. First, the problem of universal sybil defense is a special case of Problem 4.1 in which $\tau = O(\log n)$, and C is the entire honest region. Second, sybil defense appears, informally, to reduce to the task of detecting the "community" C of the honest seed u.

The fundamental affinity between community detection and sybil defense has been first observed by [Viswanath et al. 10]. After pointing out that, from the perspective of an honest node, SybilLimit-like protocols separate the social graph in two communities—honest nodes and sybils—they go on to ask a natural follow-up question: can off-the-shelf community detection algorithms be used to detect sybils? Their answer is mixed: on the one hand, they show that a generic community detection algorithm due to [Mislove et al. 10] (also a coauthor in [Viswanath et al. 10]) achieves results comparable to those of SybilLimit-like protocols on both a synthetic topology and a real-life Facebook social graph; on the other, they observe that attackers wise to the community substructure of the honest portion of the social graph can manage, as we discussed above, to make the sybil region appear indistinguishable from a subnetwork of honest nodes.

We believe that a first step toward a more conclusive answer is to recognize that casting the problem simply in terms of generic community detection leaves it underspecified. Although intuitively compelling, the notion of community is ambiguous, as are the many community detection algorithms found in the literature, each aiming for a subtly different notion of community, clearly indicated by [Fortunato 09]. But what should be the basis of a notion of community that can be used effectively for the sybil defense?

4.1. The Minimum Conductance Cut

A somewhat obvious candidate to serve in this role is conductance. Conductance is difficult to tamper with (see Section 2), and it is intimately related to mixing time, a critical property to leverage against sybil attacks (see Section 3).

It is tempting to define the problem of sybil defense in terms of the *minimum conductance cut problem* found in the community detection literature:

Problem 4.2. Let $G = (V, E)$ be an undirected graph. Find a set $C \subset V$ whose conductance $\phi(C)$ is as close as possible to $\phi(G)$, the minimum conductance of the graph.

If we believe that the honest region is fast mixing and that it is connected to the sybil region via a sparse cut, then the set C should be very close to capturing precisely the entire honest region. This view is of course too simplistic and can lead to community detection algorithms that can be circumvented by an adversary using far fewer attack edges than needed to dupe SybilLimit-like protocols. Mislove's algorithm [Mislove et al. 10], a community detection algorithm that has been used in the context of sybil defense [Viswanath et al. 10], provides an interesting example.

Mislove's algorithm is a heuristic algorithm that finds small conductance cuts—which is, in essence, analogous to finding an approximate solution to Problem 4.2. Note that finding an approximate cut is the best one can hope for, unless $P = NP$. The set C is built greedily. Starting from a vertex u, the algorithm grows C by incorporating the vertex v connected to C that results in a set $C \cup \{v\}$ with minimal conductance. If no neighboring vertex decreases the conductance, then the algorithm adds the vertex that increases it the least.[7]

Although this simple heuristic appears to capture the intuition behind Problem 4.2, it fails against the following simple attack. Let v be an honest node, that has no neighbor of degree smaller than 3. We create the sybil region with nodes $s_0, s_1, \ldots, s_n$ as follows:

- s_0 and s_1 are connected to v.
- for every $i \leq n - 2$, s_i is connected with the next two sybil nodes in the sequence, s_{i+1} and s_{i+2}, and also with the previous two, s_{i-1} and s_{i-2}.

[7] The original proposal for Mislove's algorithm [Mislove et al. 10] relies on a normalized conductance metric, but in the context of sybil defense the protocol is evaluated using just conductance [Viswanath et al. 10]. For consistency, we follow the approach of the second work.

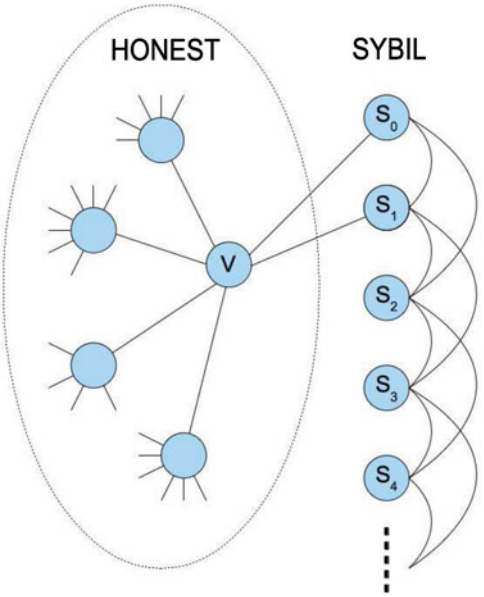


Figure 4. Two edge attack.

Figure 4 illustrates the attack, involving only the two attack edges connecting v to s_0 and s_1, that results in Mislove's algorithm deterministically admitting every node of the sybil region.[8]

At the beginning, the best choice for the algorithm is to add the lowest-degree node s_0: the value of the ratio that defines the conductance of the new set decreases, as its numerator is increased by only two edges (those from s_0 to s_1 and s_2), while its denominator is incremented by three. After this fatal mistake, the best node to add becomes s_1, which raises the numerator again by two and the denominator by four. Proceeding in a similar way leads to admitting the entire sequence $s_2, \ldots s_n$.

[8] Furthermore this attack can be modified to withstand also the preprocessing defined in Section 3.2. For instance, to avoid a preprocessing of nodes with degree < 5, the attacker can add in the sybil region a series $s_0, s_1, \ldots, s_n$ of sybil nodes as before. Each sybil node s_i is connected to the previous four sybil nodes $s_{i-1}, \ldots, s_{i-4}$ (if they exist) and the four consecutive sybil nodes $s_{i+1}, \ldots, s_{i+4}$ (if they exist). Furthermore s_0, s_1, and s_n are connected to v. In this setting it is possible to see that if initially v picks s_0, it will then pick all the nodes in the sybil region in sequence. If node v has no honest neighbor of degree 5 (after preprocessing), then the entire sequence of sybil nodes is admitted before any of his honest neighbors.

4.2. Discussion

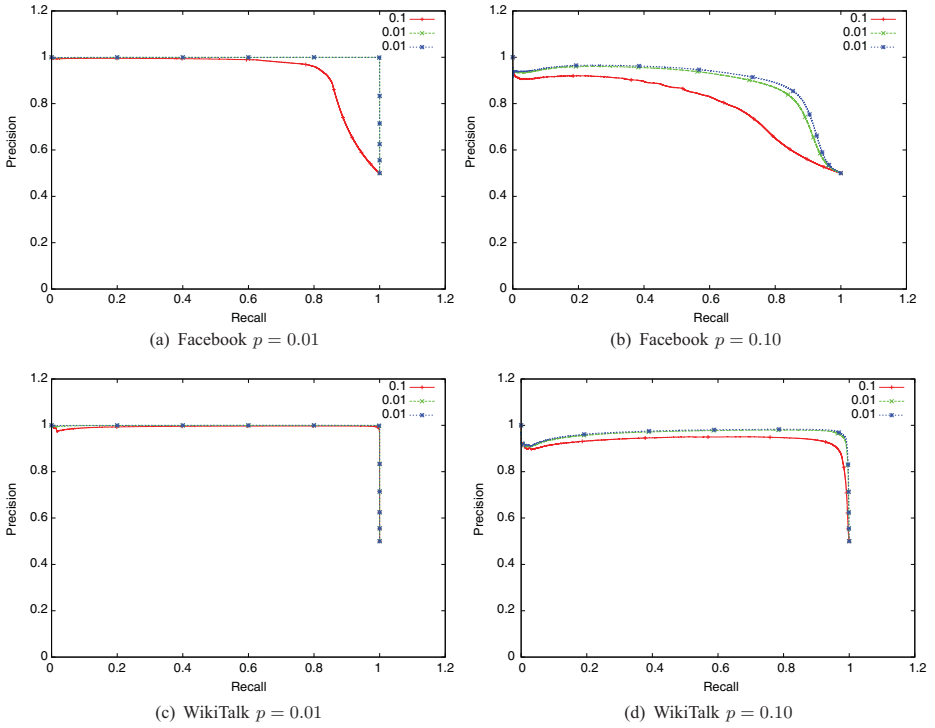
Reframing sybil defense to leverage the community substructure that exists in social graphs requires a deep understanding of the relationship between sybil defense and conductance—in essence, understanding when a solution to Problem 4.2 is also a solution to Problem 4.1. The key to the approach we explore in subsequent sections relies, at a local scale, on a technique central to the efforts towards universal sybil defense discussed in Section 3: random walks.

5. Fast-Mixing Communities

Because of its tight connection with the theory of random walks, the minimum conductance cut problem, which we have used to formalize the intuitive relationship between sybil defense and community detection, has been studied in depth. Indeed, as we will see, a recently proposed sybil-defense algorithm [Cao et al. 12] is based on a well-known random walk algorithm previously developed to answer certain foundational issues in the theory of algorithms [Spielman and Teng 04].

Problem 4.2, as we have called it, is NP-hard [Garey and Johnson 79] and from the point of view of approximation, a series of results have established various nontrivial approximation guarantees [Sinclair and Jerrum 89, Leighton and Rao 99, Arora et al. 09]. In our context, however, these sophisticated algorithms do not appear to be directly applicable. They are not obviously parallelizable, an essential scalability requirement given the huge size of real-life social networks. A second, more subtle, drawback is that their running time is polynomial in the size of the entire graph. In contrast, there exist methods whose time complexity depends only on the size of the set of trustworthy nodes that we are trying to determine, which we expect to be significantly smaller than the size of the entire network.

Spielman and Teng developed the first such "local" algorithm [Spielman and Teng 04]. Very roughly, their idea is to associate a weight with each node and to identify as part of the community all nodes whose weight exceeds a certain threshold. To determine the weight of a node, they effectively run many truncated random walks of the same length $t \in \tilde{O}(\frac{1}{\phi})$, all originating from the same node (the *seed*): a node's weight is given by the frequency with which it is visited divided by its degree. The potential of this algorithm for sybil detection becomes evident once one interprets the weight of a node v as a measure of the trust that the seed node puts in v. Indeed, the recent sybil detection protocol SybilRank [Cao et al. 12] is essentially an implementation of the algorithm of Spielman and Teng, run using multiple seed nodes.



(a) Facebook $p = 0.01$ (b) Facebook $p = 0.10$

(c) WikiTalk $p = 0.01$ (d) WikiTalk $p = 0.10$

Figure 5. Impact of varying α. Precision vs. recall graph with Facebook-New Orleans dataset under (a) a weak attack (edge density $p = 0.01$) and (b) a strong attack (edge density $p = 0.1$). Figures (c) and (d) refer to a weak and a strong attack, respectively, in the WikiTalk graph.

Since the work of Spielman and Teng, however, the use of truncated random walks for computing low conductance cuts has been further refined. In particular, [Andersen et al. 07] originate many random walks from the honest seed, as in the previous algorithm [Spielman and Teng 04], but the length of their random walks, instead of being fixed, is determined by means of a (geometrically distributed) random variable. This algorithm has a property that is extremely useful in our context: it identifies a region around the honest seed whose conductance is smaller than what is computable with the approach used in SybilRank.

[Andersen and Peres 09] and, very recently, [Gharan and Trevisan 12] have proposed further improvements. It is not immediately obvious, to us at least, if these algorithms can be used by an honest seed to rank other nodes according to its trust in them. For this reason, we will focus henceforth on the method

proposed by Andersen, Chung, and Lang, which naturally computes such a ranking [Andersen et al. 07].

5.1. Discussion

Formalizing community detection in terms of Problem 4.2 allows us to draw from the rich literature on algorithms based on randomwalks. Among them, the algorithm of Andersen, Chung, and Lang stands out for the combination of its features: it supports node ranking; the cut it computes has smaller conductance than most of its peers; its running time depends on the size of the community, not that of the graph; and it is easy to parallelize. In the next section we will see that this algorithm solves Problems 4.1 and 4.2 simultaneously, i.e., it is able to identify a community of honest nodes containing the honest seed, without being lured into the sybil region. Further, we will prove the first theoretical guarantees concerning the performance of a community detection algorithm in the context of sybil defense and show experimentally that the algorithm is quite competitive with the state of the art.

6. Personalized PageRank and Local Defense

In this section we analyze the "variable length" random walk method of [Andersen et al. 07], ACL henceforth, and show that it provides both formal and experimental guarantees for our localized vision of social sybil defense: white listing of the community to which our honest node belongs.

ACL is based on the Personalized PageRank (PPR) random walk, whose definition we now review. Starting from an initial seed vertex v, at each step in the walk a pebble returns to node v with probability α and moves to a uniformly random neighbor of its current location with probability $1 - \alpha$. This random walk has a unique stationary distribution [Andersen et al. 07] that we denote as $\mathrm{ppr}_{\alpha,v} := (\mathrm{ppr}_{\alpha,v}(v_1), \ldots, \mathrm{ppr}_{\alpha,v}(v_n))$. Clearly, this distribution depends on the starting node v and the *jumpback* parameter α. We will drop these subscripts when they are clear from the context.

Intuitively, it is as if, starting from the honest seed, we performed many random walks whose length is determined by means of a geometric random variable: a random walk has length k with probability $\alpha(1-\alpha)^{k-1}$. The expected length of each walk is α^{-1}, meaning that long walks are rare and short walks in the neighborhood of the seed are common. In this fashion, the nodes in the "community" to which the seed belongs should be visited most frequently. Nodes are assigned a score proportional to the number of times they are visited.

ACL introduces an additional step to the PPR computation: the score assigned to the vertices is given by

$$\text{score}_{\alpha,v}(u) := \frac{\text{ppr}_{\alpha,v}(u)}{\deg(u)}, \tag{6.1}$$

for all vertices u. This step, also used in the algorithm of [Spielman and Teng 04], ensures that the score acquired by u is not inflated by its unusually high degree.

In the following, we will prove theoretical guarantees for the ACL score. It is interesting to note that they hold only for the ACL score and not for the PPR score (for which several nodes of high volume may be included in the first positions of the ranking). The ACL algorithm proceeds by sorting the nodes in V in descending order of $\text{score}_{\alpha,v}$. Although ACL is originally motivated by finding a low conductance cut, the properties enjoyed by such ranking can be exploited in the context of sybil defense as well, as the rest of the section shows.

Intuitively, the ranking computed using a honest node v as the seed defines, from the point of view of v, an ordering of the nodes in V, from the most trustworthy to the least.

This ranking is significantly more robust than that obtained by methods based on PageRank (see for example EigenTrust [Kamvar and Schlosser 03] and TrustRank [Gyöngyi et al. 04]): because a random walk can be reset only to the seed node, this ranking is immune to all attacks to PageRank based on exploiting random walks that jump back to a spam node [Cheng and Friedman 06]. Notably, in the context of sybil defense, ACL solves Problem 4.1: it computes a low-conductance cut containing the honest seed and almost no sybil nodes. The next subsection is devoted to proving the following theorem, which formalizes this result.

Theorem 6.1. *Let C be a set of vertices such that the graph it induces is connected and has mixing time τ and with $|\text{cut}(C)| \in o(\text{vol}(C)\tau^{-1})$. Let $1/2 > \epsilon > 0$ be a constant and let $\alpha := (10\tau)^{-1}$. Given a seed v, define*

$$S_v := \{u : \text{score}_{\alpha,v}(u) > (2\text{vol}(C)e^{1/10})^{-1}\}$$

(this is the set of nodes that obtain high enough ACL score). Then, there exists a subset $C' \subset C$ such that $\text{vol}(C') \geq (1-\epsilon)\text{vol}(C)$ and such that, if $v \in C'$, then $\text{vol}(S_v \cap C) \geq (1 - o(1))\text{vol}(C)$ and $\text{vol}(S_v \setminus C) = o(\text{vol}(C))$.

Notice that here, and in the rest of the section, when referring to the mixing time of the graph induced by the set C, we write τ in place of $\tau(\epsilon)$ where $\epsilon \in O(\frac{1}{n})$ (see Definition 2.1).

Some comments are in order. Theorem 6.1 provides mathematical guarantees on the security of the ACL ranking in the context of sybil defense. If we let the set C of the statement be any connected subset of the honest region, and denote with τ its mixing time, the theorem says that the ACL score computed using most of the nodes in C as seeds recovers C almost perfectly in the first positions of the ranking, essentially achieving the goal envisioned by Problem 4.1.

Notice that the guarantees of Theorem 6.1 are expressed in terms of volume and not, as has been the custom in prior articles on sybil defense (see for instance [Yu et al. 06, 08, Tran et al. 11]) in terms of number of nodes. However, if we assume that $\mathrm{vol}(C) \in O(|C|)$, then the guarantees given in terms of volume translate into the familiar ones expressed in terms of number of nodes. Because social graphs consist mostly of low-degree nodes, we expect this condition to be roughly satisfied in practice, as our experiments on the performances of ACL confirm. More formally, it can also be shown, for instance, that if the graph follows a power-law distribution [Albert and Barabási 02] with exponent greater than two, then this condition holds.

6.1. Security Guarantees of ACL

In this section we prove Theorem 6.1. Our results are heavily based on previous work [Andersen et al. 07, Zhu et al. 13]: for completeness, we present here the full proof of all the statements and discuss the security implications in detail.

Referring to the statement of Theorem 6.1, we use the following notation: C is a subset of nodes that induces a connected component, denoted as $G[C]$, with mixing time τ and cut $\mathrm{cut}(C) \in o(\frac{\mathrm{vol}(C)}{\tau})$. Intuitively, C is the community of the honest seed, connected to the rest of the social graph by means of a somewhat sparse cut. The rankings PPR and ACL will be computed with respect to $\alpha := (10\tau)^{-1}$.

To prove the theorem, we first lower bound the ACL score for all nodes inside S_v and then we upper bound the aggregate PPR score of the nodes outside C. More specifically, we first prove the following lemma, which shows that the total score that can be absorbed by the complement of the community C containing the honest seed is negligible.

Lemma 6.1. (Boundedness Lemma). *Let C be a set of vertices such that the graph it induces is connected, has mixing time τ, and its cut is such that $|\mathrm{cut}(C)| \in o(\mathrm{vol}(C)\tau^{-1})$; and let $\alpha := (10\tau)^{-1}$. Then, for any positive constant $0 < c_1 < \frac{1}{2}$, there exists a*

subset C' of C, such that $\mathrm{vol}(C') \geq (1 - c_1)\mathrm{vol}(C)$ *and such that*

$$\sum_{u \in V \setminus C} \mathrm{ppr}_{\alpha,v}(u) = o(1),$$

where v is any (seed) node in C' (the $o(1)$ term goes to zero as C grows).

Proof. Let $b(i,t)$ be the random variable describing the following event: a random walk of length t, starting on node i, crosses an edge in $\mathrm{cut}(C)$ during the walk. To upper bound $P[b(i,t)]$, we will use a technique inspired by [Yu et al. 06].

Suppose that the walk starts from the stationary distribution restricted to the subgraph C (considering also the edges that leave C): then, the probability of crossing any edge in the cut in a given step is equal to $\frac{\mathrm{cut}(C)}{\mathrm{vol}(C)}$. So, by the union bound, the probability of crossing the cut in one of the t steps is smaller than or equal to $\frac{t \cdot \mathrm{cut}(C)}{\mathrm{vol}(C)}$.

Let p_i be the probability that we visit vertex i in the stationary distribution. Since $p_i = \deg(i)/\mathrm{vol}(C)$, we have:

$$\sum_{i \in C} P[b(i,t)] \frac{\deg(i)}{\mathrm{vol}(C)} = \sum_{i \in C} P[b(i,t)]\, p_i \leq t \frac{|\mathrm{cut}(C)|}{\mathrm{vol}(C)}.$$

So,

$$\sum_{i \in C} P[b(i,t)] \deg(i) \leq t |\mathrm{cut}(C)|.$$

Now, this inequality implies that there is a set $C' \subseteq C$ of nodes of volume at least $(1 - c_1)\mathrm{vol}(C)$, with constant $0 < c_1 < \frac{1}{2}$, such that for any $i \in C'$ we have $P[b(i,t)] \leq t \frac{\mathrm{cut}(C)}{(1-c_1)\mathrm{vol}(C)}$. Otherwise, we would get a contradiction because $\sum_{i \in C'} P[b(i,t)] \deg(i) > t |\mathrm{cut}(C)|$.

For $1 \leq i \leq n$, let x_i be the indicator vector for node v_i (i.e., a vector whose components are all set to 0 except for the ith one, which is set to 1). With abuse of notation we write x_v for the indicator vector of the node v. We can now describe the PPR calculation in matrix form as in [Andersen et al. 07]:

$$\mathrm{ppr}_{\alpha,v} = \alpha \sum_{t=0}^{\infty} (1 - \alpha)^t x_v W^t, \tag{6.1}$$

where $\text{ppr}_{\alpha,v}$ defines the PPR vector with jump-back probability α and seed node v. W is the standard random walk transfer matrix.[9] W_{ij} is the probability of reaching node j, in a single step of the walk, starting from node i.

Let $B = \sum_{v_i \in V \setminus C} x_i$, we have that:

$$\text{ppr}_{\alpha,v}(B) = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t x_v W^t(B). \tag{6.2}$$

Now suppose that $v \in C'$. Note that the probability of landing in a node in $V \setminus C$ at step t starting from v is upper-bounded by the probability of crossing the cut during a walk of length t. Hence,

$$\begin{aligned}
\text{ppr}_{\alpha,v}(B) &\leq \alpha \sum_{t=0}^{\infty} (1-\alpha)^t t \frac{\text{cut}(C)}{(1-c_1)\text{vol}(C)} \\
&\leq \alpha \frac{\text{cut}(C)}{(1-c_1)\text{vol}(C)} \sum_{t=0}^{\infty} (1-\alpha)^t t \\
&\leq \frac{\alpha}{\log^2(1-\alpha)} \frac{\text{cut}(C)}{(1-c_1)\text{vol}(C)} \\
&\leq \frac{1}{\alpha} \frac{\text{cut}(C)}{(1-c_1)\text{vol}(C)}.
\end{aligned}$$

By choosing $\alpha = \frac{1}{10\tau}$,

$$\begin{aligned}
&\leq 10\tau \frac{1}{c_1 \text{vol}(C)} \, o\left(\frac{\text{vol}(C)}{\tau}\right) \\
&= o(1)\,.
\end{aligned}$$

□

Note that since the score of each node is obtained by dividing the ppr probability by the degree (whose value, by the completeness hypothesis, is at least equal to 1), the previous lemma provides also a bound on the total score of nodes in $V \setminus C$.

We have showed that the overall score assigned to nodes in $V \setminus C$ is proportional to the size of the cut and strictly bounded by $o(1)$. We now prove that most of the nodes in C receive a constant fraction of the overall score.

[9] The ACL algorithm [Andersen et al. 07] is actually defined in terms of a lazy version of the walk, in which at every step there is a probability of 1/2 of remaining in the same node. For the purpose of this study the two definitions are equivalent up to a simple change in α, so for simplicity here we use the standard random walk.

As in the statement of Theorem 6.1, let S_v denote the set of nodes that receive a high ACL score with respect to a seed v, that is, $S_v := \{u : \text{score}_{\alpha,v}(u) > (2\text{vol}(C)e^{1/10})^{-1}\}$.

Lemma 6.2. (Coverage Lemma). *Let C be a set of vertices such that the graph it induces is connected, has mixing time τ, and its cut is such that $|\text{cut}(C)| \in o(\text{vol}(C)\tau^{-1})$; and let $\alpha := (10\tau)^{-1}$. Then, for any positive constant $0 < c_1 < \frac{1}{2}$, there is a set $C' \subseteq C$ such that $\text{vol}(C') \geq (1 - c_1)\text{vol}(C)$ and such that $\text{vol}(S_v \cap C) \geq (1 - o(1))\text{vol}(C)$, for $v \in C'$.*

Proof. Observe that, by setting $\alpha = \frac{1}{10\tau}$, a sizable fraction of the PPR random walks will be longer than τ, the mixing time of $G[C]$. More precisely, let $l(t)$ be the probability of a PPR random walk that is t steps long. Since the lengths of the random walks follow a geometric distribution, we have that $l(t) = \alpha(1 - \alpha)^t$ and consequently,

$$\sum_{t=\tau}^{\infty} l(t) = (1 - \alpha)^{\tau}.$$

Consider the set $C' \subseteq C$ for which we showed in the previous lemma that, for any $v \in C'$, the probability of crossing the cut for a t-step long walk starting in a node in C' is bounded by $t\frac{\text{cut}(C)}{c_1|\text{vol}(C)|}$.

Fix a node $v \in C'$ and let $v_i \neq v$ be any other node in C. We want to determine a lower bound on the score assigned to node v_i by PPR if we compute it using v as seed.

As already mentioned, we have

$$\text{ppr}_{\alpha,v}(v_i) = \alpha \sum_{t=0}^{\infty} (1 - \alpha)^t x_v W^t x_i,$$

where W is the standard random walk transfer matrix. If we restrict our attention to random walks longer than the mixing time, we obtain the lower bound

$$\text{ppr}_{\alpha,v}(v_i) \geq \alpha \sum_{t=\tau}^{\infty} (1 - \alpha)^t x_v W^t x_i.$$

So, in order to find a good lower bound to $\text{ppr}_{\alpha,v}(v_i)$, we would like to know the probability that a random walk of length t, for $t > \tau$, ends in v_i. Note that this would be easy in the graph induced by C, because we know that its mixing time is τ, while it is not immediately obvious when we consider the edges going out of C (whether attack or nonattack edges). But from the previous lemma we

know that the total PPR score leaking from C is $o(1)$. This implies that very few random walks "leak" probability outside of C. Let us suppose that no random walk leaves C and denote by ppr′ the ppr score in this setting. Then, since the mixing time of $G[C]$ is τ, we can compute ppr′:

$$\mathrm{ppr}'_{\alpha,v}(v_i) \geq \alpha \sum_{t=\tau}^{\infty} (1-\alpha)^t \left(\frac{\deg(v_i)}{\mathrm{vol}(C)} - \frac{1}{\epsilon} \right) \geq \alpha \sum_{t=\tau}^{\infty} (1-\alpha)^t \left(\frac{\deg(v_i)}{(1+\delta)\mathrm{vol}(C)} \right).$$

For any positive $\delta > 0$ it follows that,

$$\begin{aligned} \mathrm{ppr}'_{\alpha,v}(v_i) &\geq \frac{\deg(v_i)}{(1+\delta)\mathrm{vol}(C)} \left(\alpha \sum_{t=\tau}^{\infty} (1-\alpha)^t \right) \\ &\geq \frac{\deg(v_i)}{(1+\delta)\mathrm{vol}(C)} \left(1 - \frac{1}{10\tau} \right)^{\tau} \\ &\geq \frac{\deg(v_i)}{(1+\delta)\mathrm{vol}(C)} \, e^{-1/10}. \end{aligned}$$

We know from Lemma 6.1 that the total score distributed by walks that cross the edges in the boundary of C is at most $o(1)$. From the previous chain of inequalities, each node in C has $\mathrm{ppr}' \in \Omega(\frac{1}{\mathrm{vol}(C)})$. So, even if we remove the score distributed by walks that cross the cut, there exists a set $C'' \subseteq C$ with $\mathrm{vol}(C'') \geq (1 - o(1))\mathrm{vol}(C)$ for which each node v_i in C'' has PPR score greater than $\frac{\deg(v_i)}{2\mathrm{vol}(C)} \left(e^{-1/10}\right)$ and ACL score larger than $\frac{1}{2\mathrm{vol}(C)} \left(e^{-1/10}\right)$ □

Note that these lemmata imply the existence of a gap between the score of the nodes inside and those outside of C. We leverage this gap to prove Theorem 6.1.

Proof of Theorem 6.1. From Lemma 6.1, we have that the nodes in $V \setminus C$ have aggregate PPR score in $o(1)$; furthermore, all nodes in C'' have score at least $\frac{1}{2\mathrm{vol}(C)}(e^{-1/10})$. The PPR score of a node of degree d in $V \setminus C$, computed using as seed node $v \in C'$, must be larger than $\frac{d}{2\mathrm{vol}(C)}(e^{-1/10})$ to be in the set S_v: thus, the total volume of nodes in $S_v \setminus C$ is $o(\mathrm{vol}(C))$. Hence the claim follows.[10] □

6.1.1. Comparison with the State of the Art. In the theoretical framework that underpins SybilLimit and its ilk, the honest region $H \subset V$ is assumed to be fast mixing, i.e., $\tau = O(\log(|H|))$. Let g be the number of attack edges connecting honest and sybil nodes.

By setting $\alpha = \frac{1}{10\log(|H|)}$ and choosing $C = H$, we have $\mathrm{cut}(C) = g$. Suppose to have $g = o(\frac{|H|}{\log(n)})$, as in the assumption of SybilLimit [Yu et al. 08]. As $|H| =$

[10] Note that the theorem would not hold if we used the PPR score directly.

$O(\mathrm{vol}(H))$, in a connected graph ACL is able to accept a slightly larger number of attack edges than SybilLimit: $O(\frac{\mathrm{vol}(H)}{\log(|H|)})$ vs $O(\frac{|H|}{\log(|H|)})$. Note, however, that ACL guarantees are expressed in terms of the volume of H rather than the number of its nodes.

Moreover, with the additional assumption that $\mathrm{vol}(C) = O(|C|)$ discussed in the previous section, Theorem 6.1 guarantees that for any positive constant $0 < c_1 < \frac{1}{2}$, the ranking given by $\mathrm{score}_{\alpha,v}(u)$, for a fraction $1 - o(1)$ of nodes $u \in V$, contains in the first $|V|$ positions all but a $1 - o(1)$ fraction of good nodes, essentially matching both the number of attack edges and the guarantees of SybilLimit.

The consequences of our theoretical results can be summarized as follows.

- Under the hypotheses of SybilLimit-like protocols, the performance of ACL is comparable with the state of the art.
- In the more general setting where only a subset of the honest region is assumed to be wellconnected, ACL can guarantee that a subset of honest nodes is trusted more than sybil nodes.
- In harder settings, there is an explicit trade-off between the mixing time of the honest region and the number of attack edges that the network can handle.

6.2. Computing the Ranking

Algorithm 1. ApproxACL(v, α, ϵ)

$\mathrm{ppr}(u) = 0\ \forall u \in V$
$r(v) = 1$
$r(u) = 0\ \forall u \in V \setminus \{v\}$
$Q = \{v\}$
while $|Q| \neq 0$ **do**
  Extract u from Q.
  while $r(u) \geq \epsilon \mathrm{deg}(u)$ **do**
    $\mathrm{ppr}, r = Push_u(ppr, r)$
    Insert in Q all the nodes w in the neighborhood of u such that $r(w) \geq \epsilon\, \mathrm{deg}(w)$.
  end while
end while
$\mathrm{score}_{\alpha,v}(u) = \frac{\mathrm{ppr}(u)}{\mathrm{deg}(u)}\ \forall u \in V$
return $\mathrm{score}_{\alpha,v}$

The PPR distribution can be expressed as the solution of a system of linear equations, and it can be computed or approximated very efficiently in parallel (see, for instance, [Fogaras et al. 05] and [Bahmani et al. 11]). Here we present the push-flow algorithm of Andersen et al., which computes an approximation of the ACL score and possesses many desirable properties [Andersen et al. 07]. The algorithm, which we name ApproxACL, for Approximated ACL score, has three input parameters: the starting honest vertex v, the jump back probability α, and the error parameter ϵ. ApproxACL computes a vector $q^{\epsilon}_{\alpha,v} := (q_1, \ldots, q_n)$ that is an approximation of the ACL score vector $\text{score}_{\alpha,v}$. ApproxACL first computes an approximation of the ppr stationary distribution as follows. The algorithm starts with an amount of "residual PPR score" equal to 1 from the starting node v. This residual score flows from the source node to the rest of the network with a series of "trickle" operations. Each push-flow operation simulates one step of the PPR random walk by transferring a small amount of residual score from a vertex u to its neighbor w in proportion to the probability that the random walk moves from u to w in one step. For each node v, ApproxACL keeps track of two quantities: a $\text{ppr}(v)$ value and a residual value $r(v)$. The former is the current approximation of the PPR of the node v, and the latter is the amount of total residual amount of "score" that the node is allowed to distribute to itself and to its neighbors. Once the approximated PPR distribution is computed, the algorithm divides the stationary distribution probability of each node by the degree to compute the approximated ACL score.

The algorithm is described as Algorithm 1 (for a full discussion see [Andersen et al. 07]).

Algorithm 2. $Push_v(\text{ppr}, r)$

Ensure: The new updated vectors ppr_{new} and r_{new} are such that $\text{ppr}_{\text{new}} = \text{ppr}$ and $r = r'$ with the following exceptions:
$\text{ppr}_{\text{new}}(v) = \text{ppr}(v) + \alpha r(v)$
$r_{\text{new}}(v) = \frac{1-\alpha}{2} r(v)$
for all $u \in V$: $(u, v) \in E$ **do**
$r_{\text{new}}(u) = r(u) + \frac{1-\alpha}{2\deg(v)} r(v)$
end for
return ppr_{new} e r_{new}

How does the behavior of ApproxACL change as a function of the parameters α and ϵ? Theorem 6.1 tells us how we should set the value of α. The dependence on ϵ is also reasonably straightforward. The parameter ϵ measures how far we are from the actual ACL score. Clearly, smaller values of ϵ imply longer running times. The good news is that this dependence on precision is linear: it is possible

	δ			
ϵ	$= 10^{-4}$	$= 10^{-5}$	$= 10^{-6}$	$= 10^{-7}$
$= 10^{-3}$	0.84	0.83	0.82	0.82
$= 10^{-4}$		0.81	0.79	0.79
$= 10^{-5}$			0.73	0.73
$= 10^{-6}$				0.99

Table 4. Kendall-Tau distance correlation between an ϵ-ranking and a δ-ranking for the Facebook snapshot.

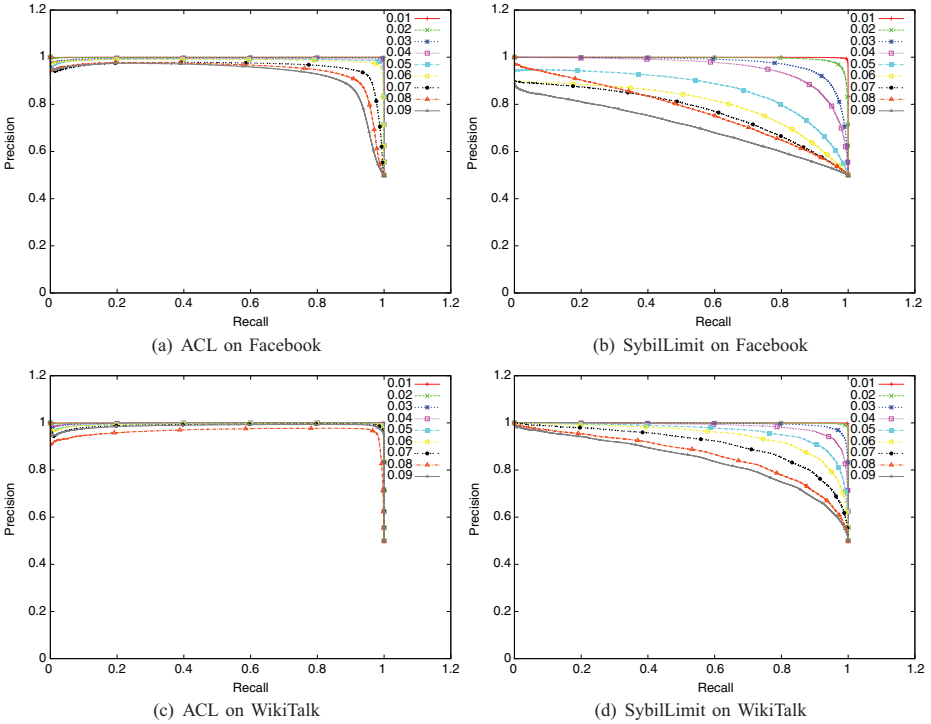
The index is a real number between $+1$ (perfect concordance) and -1 (reverse order). A value of 0 indicates that one ranking is a random permutation of the other. Similar high correlation was observed for different snapshots of social networks.

to show that the running time of the algorithm is $O(\frac{1}{\alpha\epsilon})$ and therefore, for a given α, the running time is $O(\frac{1}{\epsilon})$.

A second consequence of the choice of ϵ comes from the way the push-flow algorithm works. It can be shown that all vertices w whose probability $\mathrm{ppr}(w)$ in the stationary distribution is smaller than ϵ receive a score of 0 from ApproxACL. When ApproxACL stops, nodes with a nonzero ppr value define a connected component around the source, whereas the score of all outside vertices is 0. It is interesting to see what happens when ApproxACL is run with two values $\epsilon < \delta$. If we produce the ACL-ranking in the two cases, then the nonzero portion of the ϵ-ranking is longer than the corresponding prefix of the δ-ranking. The surprising finding is that these rankings are very stable, in the following sense. Let $u_1^\epsilon, \ldots, u_n^\epsilon$ and $u_1^\delta, \ldots, u_n^\delta$ be the two rankings. Then these two rankings are almost the same. This can be measured for instance with the Kendall–Tau distance, as reported in Table 4. This is a very useful property in the context of the sybil defense. It says that if we want to identify quickly a set of trusted nodes, we can do so simply by using a larger value of ϵ. Because the running time of the protocol is dependent on the values of α and ϵ and not on the size of the graph, this allows ApproxACL to effectively scale in situations where partial node rankings suffice.

6.3. Comparative Evaluation

Our key question in evaluating ACL is to determine whether it succeeds in expanding the guarantees offered by today's social defense systems in two



(a) ACL on Facebook (b) SybilLimit on Facebook

(c) ACL on WikiTalk (d) SybilLimit on WikiTalk

Figure 6. The impact of varying the attack strength, respectively, on Facebook (a,b) and WikiTalk graph (c,d). Results for SybilLimit are reported on preprocessed graphs whereas for ACL we use raw graphs.

directions: (i) withstanding denser attacks, and (ii) providing high-quality sybil defense without relying on the assumption that the entire graph is fast mixing.

6.3.1. Method and Environment. Viswanath et al. observe that, despite their peculiarities, sybil defense schemes are based on the same fundamental principle—community detection—and produce highly correlated results [Viswanath et al. 10]. Hence, for the sake of clarity, the experiments we report compare ACL only against SybilLimit, which we use as the SybilLimit-like champion. Although SybilLimit performed better than its peers, our experiments with SybilGuard, SybilInfer, and Gatekeeper returned qualitatively similar results.

The graphs we use to compare their performances are generated by subjecting social networks that we assume to include only honest nodes to the attack described in Section 2.3. We then run ACL and SybilLimit on the resulting graphs, rank the nodes using the same methodology discussed in Section 3, and

measure precision (the percentage of nodes in the prefix of the ranking that are honest) and recall (the percentage of honest nodes that are in the ranking's prefix) from the perspective of 10 randomly chosen seeds. We report the average of the values we obtain.

We configure SybilLimit to have $1.5\sqrt{m}$ random walks of length $1.5\log(n)$, where m is the number of edges in the final graph. ACL is configured with $\alpha = 10^{-3}$ and ϵ sufficiently small to label every node in the attacked graph with nonzero weight. (Figure 5 shows the results for other configuration of alpha. Notice that the results are qualitatively similar across a wide range of parameter settings.) For DBLP $\epsilon = 10^{-7}$; for all other graphs $\epsilon = 10^{-6}$ suffices. In Section 6.3.4, where we report the results of the other algorithms as well, we set the length of the random walks in SybilGuard as $1.5\log(n)$ and the number of ticket sources in Gatekeeper as 400.

6.3.2. ACL Tolerates Denser Attacks. Figure 6 shows the degree to which ACL and SybilLimit succeed in defending the Facebook and WikiTalk graphs when the attack strength, measured as the percentage p of attack edges in the graph, varies from $p = 0.01$ to $p = 0.1$. Note that, to respect the "operating range" of each protocol, the results we report for ACL are obtained on the *original* Facebook graph, while the results from SybilLimit apply to the *preprocessed* Facebook graph.

We observe that the ability of ACL to correctly classify nodes degrades gracefully as the attack increases in strength, remaining relatively high even when $p = 0.1$. Indeed, for the Facebook graph, the selectivity of ACL under an attack of strength $p = 0.05$ is comparable to SybilLimit's with an attack of $p = 0.01$. The performance of SybilLimit, on the other hand, decreases rather rapidly as the attack strength increases.

6.3.3. ACL Does Not Need Preprocessing. Figure 7 shows the protection offered by ACL and SybilLimit to the DBLP, Epinions, Facebook, Slahdot, RenRen, and WikiTalk graphs for an attack where $p = 0.01$. For ACL, we report only results from the raw graph. For SybilLimit we report results from both the raw and preprocessed graphs.

Without preprocessing, ACL achieves high precision at high recall. SybilLimit's performance, however, is mixed. For most graphs, SybilLimit provides excellent protection as long as the graphs are preprocessed. When the graphs are not preprocessed, the offered coverage degrades to varying extents. The degradation in coverage for Facebook and RenRen is negligible; for Epinions the degradation is minor but noticeable.

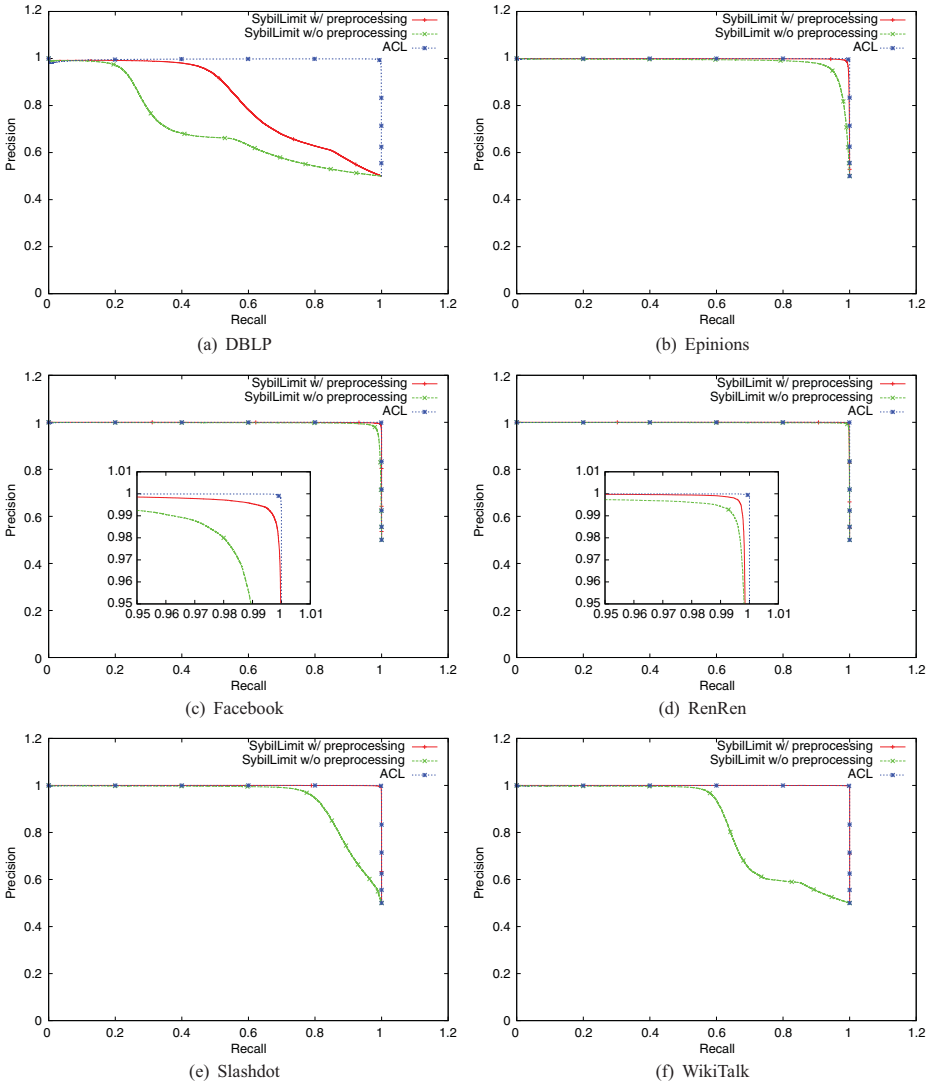


Figure 7. The precision-recall trade-offs for ACL and SybilLimit on DBLP, Epinions, Facebook, Slahdot, RenRen, and WikiTalk, with $p = 0.01$. Results for ACL are reported for the raw graphs. Results for SybilLimit are reported for both raw and preprocessed graphs.

SybilLimit performs poorly on DBLP with or without preprocessing, though preprocessing the graph does provide a significant boost. We speculate that this poor performance is the side effect of the relatively high mixing time observed by recent work [Mohaisen et al. 10].

6.3.4. A Second Attack Model. In this section we compare the algorithms using an attack model widely used in the literature [Danezis and Mittal 09, Wei et al. 12]. The number of attack edges g is fixed, and random honest nodes are declared to be sybil until we achieve g attack edges. Then, more sybil nodes are created from scratch until a total of γ sybils is reached. These γ sybils are then connected to one another via a scale-free topology: similar to other recent sybil defense literature [Wei et al. 12], our attack uses the scale-free topology of Barabási-Albert. We run each experiment ten times, and report the average values of precision and recall.

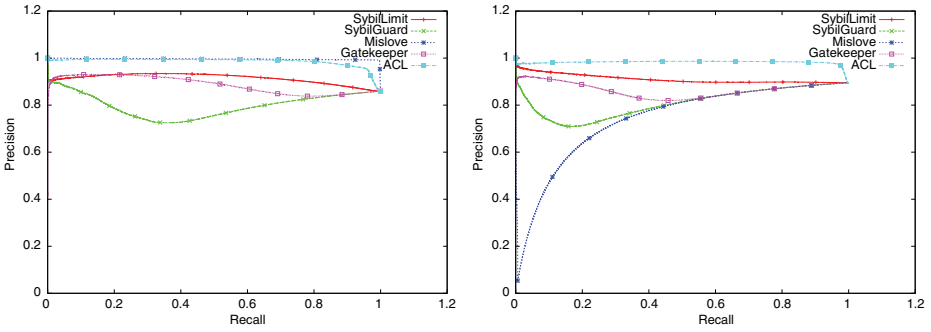
Figure 8 shows the results for our Facebook and WikiTalk graph and $g = 50{,}000$ and $\gamma = 10{,}000$. In the Facebook graph, ACL and Mislove are essentially perfect, outperforming all other algorithms (Gatekeeper, SybiLimit, and SybilGuard). In the WikiTalk graph, Mislove is outperformed by the other algorithms. The large performance difference between the two graphs confirms the sensitivity of Mislove 's algorithm to the graphs' topology (see Section 4) and supports similar observations made in the recent literature [Cao et al. 12]. We also ran experiments with other graphs and obtained similar results.

6.4. Local vs. Global Detection

We have shown that ACL is very effective in practice to address Problem 4.1. Building a universal sybil defense system for community-structured networks, however, remains an open problem.

In a recently published paper, Cao et al. suggested to expand defensive coverage by relying on multiple trusted seed nodes instead of a single one [Cao et al. 12]. More precisely, suppose there are several trusted seeds evenly distributed among communities of honest nodes; it is then possible to merge the local ranking of the nodes to get a unified global ranking of the nodes in the network.

Although effective in practice, the use of multiple seeds does not immediately lead to strong theoretical guarantees, even assuming that all seeds are honest nodes. For example, suppose we can prove, as it is typical for ACL, that a $1 - o(1)$ fraction of the honest seeds will assign a negligible fraction of the overall score to sybil nodes and distribute the rest evenly across the honest region. There is always, however, a fraction of unlucky honest seeds for which such guarantees are impossible—e.g., seeds at the boundary between the honest



(a) Facebook (b) WikiTalk

Figure 8. The precision of ACL and the other algorithms on the Facebook and WikiTalk graphs with the attack model described in Section 6.3.4 with $g = 50000$ and $\gamma = 10000$. Results are shown for the average of 10 random trials.

and sybil regions. Unfortunately, because of the arbitrary nature of the sybil region, walks originating from these nodes might produce an unconstrained (and adversarial) probability distribution among the sybil nodes.

This is true not only for the ACL algorithm, but virtually for any sybil defense algorithm that relies on random walks and mixing time (see for instance [Yu et al. 06, 08, Cao et al. 12]).

Unfortunately, it is not clear how such unlucky choice of seeds will affect the overall ranking. In fact, notice that while the *correct* seeds will distribute evenly the score among honest nodes, the *wrong* ones might concentrate the score to a smaller, but still significant, subregion of the sybil graph, thus letting such nodes overtake the first positions of the ranking.

Nevertheless, we think that the use of multiple seeds is a promising research direction, and recent literature [Cao et al. 12] has empirically verified the effectiveness of this approach in real-world scenarios.

6.5. Discussion

We have shown experimentally that ACL can identify quite accurately the community of a given honest seed and that it provides formal guarantees for the rankings it produces. Although it is effective at solving Problem 4.1, as we are about to see in the next section, ACL is still vulnerable to some simple, primitive sybil attacks that are encountered in deployed social networks—a stark reminder of the gap that, while narrowing, still exists between the theoretical assumptions that underpin the theory behind the current state of the art in sybil defense, and the reality of sybil attacks encountered in the wild. The existence of this gap does

not, in our view, belittle the importance of the current theoretical tools, as it is only by understanding their strengths and limitations that one can obtain a firmer grasp of the problem of social sybil defense. It does, however, point to a concrete challenge, and the next section outlines an approach that we believe can prove fruitful in addressing it.

7. Social Sybil Defense Against Real-World Attacks

Our appraisal in Section 2 of the resilience of different structural properties of social graphs indicated that leveraging the complementary notions of mixing time and conductance is the most promising line of defense against sybil attacks; furthermore, techniques based on this approach can provide impressive end-to-end guarantees. Yet one key question remains: how effective are these techniques against actual sybil attacks?

Although data on sybil attacks in deployed social networks is not readily available, two recent papers have included experience reports that shed light on the types of attacks that occur in the wild.

Cao et al. report to have successfully used SybilRank to identify sybil users in the Tuenti social network [Cao et al. 12]. They observe large clusters of sybil users in regular topologies (star, mesh, tree, etc.) that are connected to the honest communities through a limited number of attack edges. They also report that an unspecified fraction of the remaining accounts are sybil, but to preserve confidentiality they are unable to report on the number or characteristics of those accounts.

Yang et al.'s experience in analyzing the RenRen social network is significantly different [Yang et al. 11]: they did not observe any large clusters of well-connected sybil nodes connected in turn to the honest subgraph through a small set of attack edges, as would be expected by the sybil defense techniques we have surveyed; instead, they find isolated sybil nodes each connected to the honest subgraph through a large number of attack edges.

The simple attack observed in the RenRen social network is problematic for conductance-based protocols. We simulated the attack on our Facebook graph by introducing s isolated sybil nodes and by allowing the attacker to attempt to establish m potential attack edges by selecting both the honest and the sybil endpoint uniformly at random (m and n denote the number of edges and of vertices, respectively, in the Facebook graph). As usual, each potential attack edge is accepted with probability p. In the experiment we set $s = pn$ so that the order of magnitude of the average degree of sybil and honest nodes is the same. To assess the results, we used the well-known ROC index, defined in Section 3.2. The results show that, even for a very small number of attack edges ($p = 0.01$), every

protocol performs poorly: the ROC is 0.45 for SybilLimit, 0.44 for SybilGuard, 0.34 for Mislove, 0.49 for Gatekeeper, and 0.37 for ACL. Notice that a 0.50 ROC is consistent with a random ranking.

These results are not coincidental, as the vulnerability of conductance-based techniques to an attack where each sybil node can create more than one attack edge is fundamental: indeed, Yu et al. proved a lower bound of $\Theta(1)$ to the number of attack edges per sybil node that any mixing-time-based algorithm might tolerate [Yu et al. 08].

These experiences indicate that although today's socially based sybil defenses are designed to provide the theoretically best defense posture, they may be also easily circumvented.

7.1. Defense in Depth

To overcome this impasse, we believe that effective sybil-defense mechanisms should embrace a strategy inspired by the notion of defense in depth [Stytz 04]: rather than relying solely on techniques based on conductance, they should include a portfolio of complementary detection techniques. For example, Yang et al. observe that it is possible to spot sybil nodes by tracking their clustering coefficient (see Section 2) and the rate at which their requests of friendship are accepted: both of these measures in the RenRen graph are significantly higher for honest nodes than for sybils (in the case of the clustering coefficient, this is because a single sybil node that randomly issues friendship requests is unlikely to have many friends who are themselves friends with each other) [Yang et al. 11]. As a rule of thumb, Yang et al. suggested to report as sybil those users whose friendship-request acceptance rate is less than 50% and whose clustering coefficient is below 1/100. They report that this is sufficient to correctly identify more than 98% of the sybil nodes, with a false positive rate of less than 0.5%. Note that, while these results sound impressive, they are not cause for unconditional celebration, as it is quite easy for a slightly more sophisticated adversary to circumvent both checks by issuing friendship requests to other sybil nodes under his control. But, at the very least, checks like these make the life of the attacker more difficult and prevent more sophisticated defenses from being trivially sidestepped. Indeed, they can even nudge the attacker, who might like it or not, toward the kind of attacks where conductance-based method can start to be effective. For instance, simply adding a defense layer that monitors the rate of friendship acceptance introduces a bound (albeit loose) on the conductance of the cut between honest users and sybil nodes.

In particular, assume that honest users accept sybil request with probability p and that the threshold of accepted requests below which a node is flagged as sybil is T. Then, in our attack model, the following simple result holds:

Proposition 7.1. *Suppose that honest nodes accept friendship from a sybil node with probability p. Then, to have average acceptance ratio larger than T and avoid detection, a sybil node must create $\frac{T-p}{1-T}$ edges to fellow sybil nodes for every tentative attack edge aimed at an honest node.*

Proof. For a given sybil node, let δ be the ratio between the number E_s of edges connecting to other sybil nodes and E_h, the number of attack edges that a sybil attempts to create with honest node, i.e., $\delta = \frac{E_s}{E_h}$.

Note that in expectation the total number of edges that the sybil node will successfully create is $E_s + pE_h$, so its average acceptance rate is $\frac{E_s+pE_h}{E_s+E_h} = \frac{\delta E_h+pE_h}{\delta E_h+E_h} = \frac{\delta+p}{\delta+1}$. So to have the average acceptance ratio larger than T, we have that $\frac{\delta+p}{\delta+1} > T$ and hence $\delta > \frac{T-p}{1-T}$. □

For example, if honest users accept friendship requests with probability $p = 10\%$ and $T = 50\%$ [Yang et al. 11], then each sybil node should have eight links to sybil nodes for every attack edge to avoid detection.

Proposition 7.1 bounds the conductance of the cut between honest and sybil nodes, in the sense that whenever the sybil region has fewer edges than the honest region, the conductance of the cut is at most $2p\frac{1-T}{T-p}$.

Although this bound on conductance is loose, it is encouraging that such limitation to the attacker can be obtained based on a fairly primitive measure such as the rate of friendship acceptance. We speculate that in the near future new defense layers based on advanced machine-learning and profiling techniques [Stein et al. 11] will force a sybil attacker who wants to escape detection to generate sybil regions that ever more accurately resemble honest regions, connected through a sparse cut of attack edges to the honest portion of the graph: in other words, exactly the scenario suitable for conductance-based sybil defense.

8. Conclusions

This work has traced the evolution of social sybil defenses from the seminal work of [Yu et al. 06] to the developments of the last several years [Yu et al. 08, Danezis and Mittal 09, Tran et al. 11, Cao et al. 12] to recent reports [Yang et al. 11, Cao et al. 12] that detail their usage in practice.

We have identified two main trends in the literature. The first is based on random walk methods whose goal is to identify fast-mixing (sub)regions that contain the honest seed. The implicit assumption is that social networks under sybil attacks must exhibit a simple structure—a fast-mixing region of honest nodes connected via a sparse cut to the sybil region. We have seen how this initial

simplified picture of the world has progressively become more nuanced, leading to methods based on random walks that are able to cope with a more complex world consisting of a constellation of tightlyknit, fast-mixing communities loosely connected among themselves and to the sybil region.

The other trend that we have discussed considers sybil defense as an instance of community detection. Although we have revealed the limitation of this approach, we have been able to enucleate its core validity.

As we have shown with our discussion on Personalized PageRank, the two approaches can go hand in hand to yield more robust sybil defense protocols that are competitive with the state of the art. The discussion has highlighted the importance of the body of literature that studies foundational issues on random walks. As we have shown, both algorithms and useful conceptual tools can be distilled from it and successfully deployed in the context of sybil defense.

We also compare our solutions with real-world attacks. We believe that the defense-in-depth approach that we have advocated as a response to this challenge can be facilitated by moving from the original vision of offering individual honest users decentralized and universal sybil defense [Yu et al. 06, 08] toward defense techniques that assume that the defender has complete knowledge of the social graph topology [Cao et al. 12, Yang et al. 11] and can deploy the kind of parallelizable implementations suitable for handling the large graphs of online social networks. In particular, social network operators are in a position to use machine learning techniques, user profiling, and monitoring of user activity to gain additional knowledge that can help them filter sybil attacks not wellsuited for detection using techniques based on random walks, community detection, and their combination. Still, as attackers increase in sophistication, claims of a silver bullet should be met with healthy skepticism. As the arms race between attackers and defenders continues, it will be increasingly important that new defense mechanisms clearly state the kind of attack they aim to withstand, a landscape that too often is blurred.

Acknowledgments. We thank Bimal Viswanath and Alan Mislove for the code of Mislove's algorithm, Nguyen Tran for the Gatekeeper code, and Krishna Gummadi for his comments on an early draft.

Funding. Lorenzo Alvisi is supported by the National Science Foundation under Grant No. 0905625. Alessandro Epasto is supported by the Google European Doctoral Fellowship in Algorithms, 2011. Alessandro Panconesi is partially supported by a Google Faculty Research Award and by the EU FET project MULTIPLEX 317532.

Appendix

A.1 Proof of Theorem 2.5

We prove Theorem 2.5, whose statement we now recall.

Theorem 2.5. *Let $H = (V, E)$ be a connected simple graph such that $\phi(H)\text{vol}(V) \in \Omega(\log n)$, $\phi(H) \leq \frac{1}{e}$ and let $S = (V', E')$ be another connected simple graph with $\phi(S) \geq \phi(H)$. Suppose further that $\phi(H)\text{vol}(V) \leq \text{vol}(V') \leq \text{vol}(V)$.*
Let $G_F = (V_F, E_F)$ be the union of S, H and let g be the number of random attack edges between H and S, whose endpoints are selected with probability proportional to the degrees of the nodes. Then if $\log(\frac{1}{\phi(V)}) \cdot \phi(V) \cdot \text{vol}(V) \leq g \leq \text{vol}(V')$ we have that, with high probability,

$$\phi(G_F) \in \Omega(\phi(V)).$$

To prove the theorem we need to show that the probability that all sets of volume smaller than $\frac{1}{2}\text{vol}(V_F)$ have conductance in $\Omega(\phi(H))$ is $1 - o(1)$.
We start by defining some useful notation.

Definition A.1. For any *disjoint* subsets A, B of V_F, let $C_F(A, B)$ be the number of edges with one endpoint in A and one in B in the final graph G_F. More formally:

$$C_F(A, B) := |(x, y) \in E_F : x \in A, y \in B|.$$

Similarly let $C_H(A, B)$ and $C_S(A, B)$ be the analogous for graph H and S, respectively.

Definition A.2. For any $K \subseteq V_F$, let $C_H(K)$ be the number of edges with one endpoint in $K \cap V$ and the other in $V \setminus K$. More formally:

$$C_H(K) := C_H(K \cap V, V \setminus K).$$

Similarly for the graph S, let $C_S(K)$ bet $C_S(K) = C_S(K \cap V', V' \setminus K)$.

In the rest of the section, unless otherwise specified, $\text{vol}(K)$, $\phi(K)$, $\text{cut}(K)$ without subscript refer to the volume, conductance, and cut of $K \subseteq V_F$, respectively, in the graph G_F. On the other hand, $\text{vol}_H(K)$, $\phi_H(K)$, $\text{cut}_H(K)$ refer to the volume, conductance, and cut of the subset $K \cap V$, respectively, in the graph H, and a similar convention is adopted for the graph S.

Definition A.3. Let $\boldsymbol{K}$ be the family of subsets of V_F such that for any $K_F \in \boldsymbol{K}$, $\text{vol}(K_F) \leq \frac{1}{2}\text{vol}(V_F)$ and $G[K_F]$ is connected.

We use the following proof strategy. First, we show that we can restrict our attention to only the subsets of V_F in $\boldsymbol{K}$. Then, we partition these subsets and derive a probabilistic bound on their cut in G_F. Using these bounds, we show that the probability

that any subset K_F of V_F with $\mathrm{vol}(K_F) \leq \frac{1}{2}\mathrm{vol}(V_F)$ has conductance $\phi(K_F) \in \Omega(\phi(H))$ is $1 - o(1)$.

We begin by showing that we can restrict our attention to the family of sets $\boldsymbol{K}$ such that for any $K_F \in \boldsymbol{K}$ the induced graph $G_F[K_F]$ is connected, thus motivating the definition of $\boldsymbol{K}$.

Lemma A.1. *Let $G_F[K_F]$ be the subgraph induced by K_F and let $K_{F_1}, K_{F_2}, \cdots, K_{F_R}$, with $R > 1$, be the connected components of $G_F[K_F]$. If $\phi(K_{F_i}) \geq \alpha$ for all i with $1 \leq i \leq R$ then $\phi(K_F) \geq \alpha$.*

Proof.

$$\phi(K_F) = \frac{|\mathrm{cut}(K_F)|}{\mathrm{vol}(K_F)} = \frac{\sum_i |\mathrm{cut}(K_{F_i})|}{\sum_i \mathrm{vol}(K_{F_i})} = \frac{\sum_i \phi(K_{F_i})\mathrm{vol}(K_{F_i})}{\sum_i \mathrm{vol}(K_{F_i})} \geq \frac{\sum_i \alpha \mathrm{vol}(K_{F_i})}{\sum_i \mathrm{vol}(K_{F_i})} = \alpha.$$

□

We proceed by defining a useful partitioning of $\boldsymbol{K}$.

Definition A.4. Let $\boldsymbol{K_{k,k'}} \subseteq \boldsymbol{K}$ be the family of subsets of V_F such that $K_F \in \boldsymbol{K_{k,k'}}$ if and only if $C_H(K_F) = k$ and $C_S(K_F) = k'$.

Definition A.5. For any subset $K_F \subseteq V_F$ let Y_{K_F} be the number of attack edges with two endpoints in K_F, let $X_{V \cap K_F}$ be the number of attack edges with one endpoint in $V \cap K_F$ and the other one in $V_F \setminus K_F$, and, finally, let $X_{V' \cap K_F}$ be defined in a specular way.

Given the previous definitions, we can now proceed with stating a few lemmas, whose proof we postpone to the last subsection of this appendix.

Lemma A.2. *Let C be a constant larger than 12000, $vol_S(V') \leq \frac{vol_H(V)}{12}$ and let $\phi(H) < \frac{1}{e}$. For any $0 \leq k, k' \leq \frac{1}{C}\lceil \phi(H) \cdot \mathrm{vol}(V') \rceil$ we have*

$$|\boldsymbol{K_{k,k'}}| \leq \exp\left(\frac{1}{384}\log\left(\frac{1}{\phi(H)}\right) \cdot \phi(H) \cdot \mathrm{vol}_S(V')\right).$$

Lemma A.3. *Under the assumptions of Theorem 2.5, let $K_F \in \boldsymbol{K}$.*

Y_{K_F} has expected value

$$E[Y_{K_F}] = g\frac{\mathrm{vol}_H(K_F)}{\mathrm{vol}_H(V)}\frac{\mathrm{vol}_S(K_F)}{\mathrm{vol}_S(V')}.$$

Moreover

$$P(Y_{K_F} > 3\mathrm{vol}(K_F) + E[Y_{K_F}]) \leq \exp\left(-3\mathrm{vol}_H(V)\right).$$

Based on these lemmas, we can now prove the main result.

Proof of Theorem 2.5. We want to prove that, for all $K_F \subseteq V_F$ such $\text{vol}(K_F) \leq \frac{1}{2}\text{vol}(V_F)$ we have $\phi(V_F) \in \Omega(\phi(H))$ with probability $1 - o(1)$. We begin by splitting $\boldsymbol{K}$ in 4 families of sets. Then we consider each family separately and we prove that the probability that for any set $K_F \in \boldsymbol{K}$ the probability that $\phi(K_F) \notin \Omega(\phi(H))$ is asymptotically smaller than the size of the family to which it belongs. This, in turn, will imply the result.

From Lemma A.1 we know that we can restrict our attention to sets $K_F \subseteq V_F$ whose induced subgraph $G_F[K_F]$ consists of single connected components. The lemma shows in fact that if we can prove a lower bound on the conductance of all such connected components of $G[K_F]$, then the bound applies also to $\phi(K_F)$.

We now proceed to prove that for any k, k' all subsets $K_F \subseteq V_F$ in $\boldsymbol{K_{k,k'}}$ have conductance $\Omega(\phi(H))$ with probability $1 - o(1)$.

We start with an easy example to warmup; notice that in this simple case the result holds with probability 1:

a) $\boldsymbol{K_{0,0}}$ Let us consider the elements $K_F \in \boldsymbol{K_{0,0}}$. If $C_H(K_F)$ is 0, since H is connected, we have that $K_F \cap V$ is either equal to V or to the empty set. Similarly, we can see that $K_F \cap V'$ is either equal to V' or to the empty set. Recall that $\boldsymbol{K_{0,0}} \subseteq \boldsymbol{K}$ contains only sets K_F such that $\text{vol}(K_F) \leq \frac{1}{2}\text{vol}(V_F)$ and that, by assumption, $\text{vol}_H(V) \geq \text{vol}_S(V')$. Then, the only two possible elements of $\boldsymbol{K_{0,0}}$ are V and V'.

If $V \in \boldsymbol{K_{0,0}}$ we have $\text{vol}_H(V) = \text{vol}_S(V')$. Otherwise $\text{vol}(V) = \text{vol}_H(V) + g > \text{vol}_S(V') + g = \text{vol}(V')$, but since $\text{vol}(V_F) = \text{vol}(V) + \text{vol}(V') < 2\text{vol}(V)$ then $\text{vol}(V) > \frac{1}{2}\text{vol}(V_F)$, which contradicts the definition of $\boldsymbol{K}$.

As $\text{vol}_H(V) = \text{vol}_S(V')$ we have $\phi(V) = \frac{g}{\text{vol}(V)} = \frac{g}{\text{vol}(V')} \geq \log(\frac{1}{\phi(H)})\phi(H)$.

Similarly, if $V' \in \boldsymbol{K_{0,0}}$ we have $\phi(V') = \frac{g}{\text{vol}(V')} \geq \log(\frac{1}{\phi(H)})\phi(H)$.

b) General Case We now consider the general case in which $k \geq 0$ and $k' \geq 0$, recall $k = C_H(K_F)$ and $k' = C_S(K_F)$. Recall the definitions of Y_{K_F},$X_{V \cap K_F}$ and $X_{V' \cap K_F}$, given in Definition A.5.

We have

$$\phi(K_F) = \frac{C_H(K_F) + C_S(K_F) + X_{V \cap K_F} + X_{V' \cap K_F}}{\text{vol}_H(K_F) + \text{vol}_S(K_F) + X_{V \cap K_F} + X_{V' \cap K_F} + 2Y_{K_F}} \tag{A.1}$$

$$= \frac{C_H(K_F) + C_S(K_F) + X_{V \cap K_F} + X_{V' \cap K_F}}{\text{vol}_H(K_F) + \text{vol}_S(K_F) + g}. \tag{A.2}$$

To find the right bound for the conductance, it is necessary to consider separately the following four cases that cover the relative size of the volumes of $V \cap K_F$ and $V' \cap K_F$ in relation to V and V', respectively.

1) *Large* $V \cap K_F$ *and small* $V' \cap K_F$: when $\text{vol}_H(K_F) \geq \frac{3\text{vol}_H(V)}{4}$ and $\text{vol}_S(K_F) < \frac{3\text{vol}_S(V')}{4}$.

2) *Smal* $V \cap K_F$ *and large* $V' \cap K_F$: when $\text{vol}_H(K_F) < \frac{3\text{vol}_H(V)}{4}$ and $\text{vol}_S(K_F) \geq \frac{3\text{vol}_S(V')}{4}$.

3) *Small* $V \cap K_F$ *and small* $V' \cap K_F$: when $\text{vol}_H(K_F) < \frac{3\text{vol}_H(V)}{4}$ and $\text{vol}_S(K_F) < \frac{3\text{vol}_S(V')}{4}$.

4) *Large* $V \cap K_F$ *and large* $V' \cap K_F$: when $\mathrm{vol}_H(K_F) \geq \frac{3\mathrm{vol}_H(V)}{4}$ and $\mathrm{vol}_S(K_F) \geq \frac{3\mathrm{vol}_S(V')}{4}$.

Case 1. *Large* $V \cap K_F$ *and small* $V' \cap K_F$. In this case we have $\mathrm{vol}_H(K_F) \geq \frac{3\mathrm{vol}_H(V)}{4}$ and $\mathrm{vol}_S(K_F) < \frac{3\mathrm{vol}_S(V')}{4}$. Recall that $K_F \in \boldsymbol{K}$ has $\mathrm{vol}(K_F) \leq \frac{1}{2}\mathrm{vol}(V_F)$, where $V_F = V \cup V'$. So we have,

$$\mathrm{vol}_H(K_F) \leq \frac{1}{2}\mathrm{vol}(V_F) \leq \frac{1}{2}\left(\mathrm{vol}_H(V) + \mathrm{vol}_S(V') + 2g\right),$$

by the assumption $\mathrm{vol}_S(V') \geq g$. Now,

$$\mathrm{vol}_H(K_F) \leq \frac{1}{2}\left(\mathrm{vol}_H(V) + 3\mathrm{vol}_S(V')\right).$$

Hence, from the assumption on the size of $K_F \cap V$, we have

$$\frac{3}{4}\mathrm{vol}_H(V) \leq \frac{1}{2}\left(\mathrm{vol}_H(V) + 3\mathrm{vol}_S(V')\right),$$

and

$$\mathrm{vol}_S(V') \geq \frac{1}{12}\mathrm{vol}_H(V).$$

Now, suppose that k or $k' \geq \frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil$, for a large constant C, by equation A.1 we get:

$$\begin{aligned}\phi(K_F) &\geq \frac{\frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil}{\mathrm{vol}_H(K_F) + \mathrm{vol}_S(K_F) + 2g} \geq \frac{\frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil}{12\mathrm{vol}_S(V') + \mathrm{vol}_S(V') + 2\mathrm{vol}_S(V')} \\ &\geq \frac{\frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil}{15\mathrm{vol}_S(V')} \geq \frac{1}{15C}\phi(H),\end{aligned}$$

where the last inequality comes from $\mathrm{vol}(V') \geq \mathrm{vol}_S(V')$. Hence we have $\phi(K_F) \in \Omega(\phi(H))$.

We can therefore restrict now our attention to the case where $k, k' \leq \frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil$. Consider a single set $K_F \in \boldsymbol{K_{k,k'}}$ for a given pair of k, k'. In expectation over the random choice of attack edges according to the model we have:

$$E[X_{V\cap K_F}] = g\frac{\mathrm{vol}_H(K_F)}{\mathrm{vol}_H(V)}\frac{\mathrm{vol}_S(V'\setminus K_F)}{\mathrm{vol}_S(V')}.$$

This is because, for each of the g attack edges, there is a probability $\frac{\deg(v)}{\mathrm{vol}_H(V)}$ that $v \in V$ will be the endpoint in H and, similarly, there is a probability $\frac{\deg(v')}{\mathrm{vol}_S(V')}$ that the other endpoint will be $v' \in V'$.

Since $\frac{\mathrm{vol}_H(K_F)}{\mathrm{vol}_H(V)} \geq \frac{3}{4}$ and $\frac{\mathrm{vol}_S(V'\setminus K_F)}{\mathrm{vol}_S(V')} = 1 - \frac{\mathrm{vol}_S(K_F)}{\mathrm{vol}_S(V')} \geq \frac{1}{4}$, we have

$$E[X_{V\cap K_F}] \geq \frac{3}{16}g.$$

Further, since attack edges are independent, using the Chernoff bound (Theorem 2.2) we get that

$$P\left[X_{V\cap K_F} \leq \frac{1}{16}g\right] \leq \exp\left(-\frac{1}{192}g\right).$$

Thus, with probability $1-\exp\left(-\frac{1}{192}g\right)$, we have that $\phi(K_F)$ is lower bounded by

$$\begin{aligned}\phi(K_F) &\geq \frac{C_H(K_F)+C_S(K_F)+X_{V\cap K_F}+X_{V'\cap K_F}}{\mathrm{vol}_H(K_F)+\mathrm{vol}_S(K_F)+X_{V\cap K_F}+X_{V'\cap K_F}+2Y_{K_F}}\\ &\geq \frac{\frac{1}{16}g}{\mathrm{vol}_H(K_F)+\mathrm{vol}_S(K_F)+2g} \geq \frac{\frac{1}{16}g}{15\mathrm{vol}_S(V')}\\ &\geq \frac{1}{240}\phi(H),\end{aligned}$$

where the last inequality comes form the bound on $X_{V\cap K_F}$ and the previous considerations on the volume of K_F.

To complete the proof of this case we have to show that the result holds not only for a single set, but for *all* the sets $K_F \in \boldsymbol{K_{k,k'}}$ with $k,k' \leq \frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil$.

By Lemma A.2, we know that

$$|\boldsymbol{K_{k,k'}}| \leq \exp\left(\frac{1}{384}\log\left(\frac{1}{\phi(H)}\right)\cdot\phi(H)\cdot\mathrm{vol}_S(V')\right),$$

for any with $k,k' \leq \frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil$.

Furthermore, there are at most $\mathrm{vol}_H(V)^2 \leq 4n^4$ different pairs k,k' with $k,k' \leq \frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil$ such that $\boldsymbol{K_{k,k'}}$ is not empty. So, using the union bound, we get that:

$$\begin{aligned}&P\left(\exists K_F \subseteq V\cup V' : \phi(K_F) < \frac{1}{160}\phi(H)\right)\\ &\leq 4n^4 P\left(\exists K_F, k, k' : K_F \subseteq \boldsymbol{K_{k,k'}} \wedge \phi(K_F) < \frac{1}{160}\phi(H)\right)\\ &\leq 4n^4 \exp\left(\frac{1}{384}\log\left(\frac{1}{\phi(H)}\right)\cdot\phi(H)\cdot\mathrm{vol}_S(V')\right)\cdot P\left(\phi(K_F) < \frac{1}{160}\phi(H)\right)\\ &\leq 4n^4 \exp\left(\frac{1}{384}\log\left(\frac{1}{\phi(H)}\right)\cdot\phi(H)\cdot\mathrm{vol}_S(V')\right)\cdot\exp\left(-\frac{1}{192}g\right)\\ &\in O\left(\exp\left(-\frac{1}{384}\log\left(\frac{1}{\phi(H)}\right)\cdot\phi(H)\cdot\mathrm{vol}_S(V')\right)\right).\end{aligned}$$

Thus, for all K_F covered by Case 1 we have that with high probability $\phi(K_F) \in \Omega(\phi(H))$.

Case 2. *Large $V'\cap K_F$ and small $V\cap K_F$.* In this case we have $\mathrm{vol}_H(K_F) < \frac{3\mathrm{vol}_H(V)}{4}$ and $\mathrm{vol}_S(K_F) \geq \frac{3\mathrm{vol}_S(V')}{4}$.

If $\mathrm{vol}_H(K_F) \geq \mathrm{vol}_S(K_F)$ we have

$$\begin{aligned}\phi(K_F) &\geq \frac{C_H(K_F)}{\mathrm{vol}_H(K_F)+\mathrm{vol}_S(K_F)+2g}\\ &\geq \frac{C_H(K_F)}{\mathrm{vol}_H(K_F)+3\mathrm{vol}_S(K_F)}\\ &\geq \frac{C_H(K_F)}{4\mathrm{vol}_H(K_F)} \in \Omega(\phi(H)).\end{aligned}$$

So we can restrict our attention to the case when $\mathrm{vol}_H(K_F) < \frac{4}{3}\mathrm{vol}_S(V')$, and the proof of this case mirrors the one for the case above.

Case 3. *Small $S \cap K_F$ and small $H \cap K_F$.* In this case we have $\mathrm{vol}_H(K_F) < \frac{3\mathrm{vol}_H(V)}{4}$ and $\mathrm{vol}_S(K_F) < \frac{3\mathrm{vol}_S(V')}{4}$. For this reason, $\mathrm{vol}_H(K_F) \leq 4\mathrm{vol}_H(V \setminus K_F)$, as $\mathrm{vol}_H(K_F) \leq \frac{1}{2}\mathrm{vol}_H(V)$. Hence,

$$C_H(K_F) \geq \phi(H)\mathrm{vol}_H(K_F).$$

Similarly, $C_S(K_F) \geq \phi(S)\mathrm{vol}_S(K_F)$ and hence the following inequality for $\phi(K_F)$ holds

$$\begin{aligned}\phi(K_F) &= \frac{C_H(K_F)+C_S(K_F)+X_{V\cap K_F}+X_{V'\cap K_F}}{\mathrm{vol}_H(K_F)+\mathrm{vol}_S(K_F)+X_{V\cap K_F}+X_{V'\cap K_F}+2Y_{K_F}}\\ &\geq \frac{\phi(H)\mathrm{vol}_H(K_F)+\phi(S)\mathrm{vol}_S(K_F)+X_{V\cap K_F}+X_{V'\cap K_F}}{\mathrm{vol}_H(K_F)+\mathrm{vol}_S(K_F)+2Y_{K_F}+X_{V\cap K_F}+X_{V'\cap K_F}}\\ &\geq \frac{\phi(H)\mathrm{vol}_H(K_F)+\phi(S)\mathrm{vol}_S(K_F)}{\mathrm{vol}_H(K_F)+\mathrm{vol}_S(K_F)+2Y_{K_F}}.\end{aligned}$$

By Lemma A.3 we know that

$$E[Y_{K_F}] = g\frac{\mathrm{vol}_H(K_F)}{\mathrm{vol}_H(V)}\frac{\mathrm{vol}_S(K_F)}{\mathrm{vol}_S(V')},$$

and that $P(Y_{K_F} > 3\mathrm{vol}(K_F)+E[Y_{K_F}]) \leq \exp(-3\mathrm{vol}_H(V))$.

Thus, $P(Y_{K_F} > 8\cdot max(\mathrm{vol}_H(K_F),\mathrm{vol}_S(K_F))) \leq P(Y_{K_F} > 3\mathrm{vol}(K_F)+E[Y_{K_F}])$ is at most $\exp(-3\mathrm{vol}_H(V))$.

In this case we have a strong probabilistic bound and thus we can use a simpler bound on the size of $\boldsymbol{K}_{\boldsymbol{k},\boldsymbol{k}'}$. In fact it is enough to notice that $|\boldsymbol{K}_{\boldsymbol{k},\boldsymbol{k}'}| \leq 2^{\mathrm{vol}_H(V)}2^{\mathrm{vol}_S(V)} \leq 2^{2\mathrm{vol}_H(V)}$ to get from the union bound that:

$$P(\exists K_F \subseteq V\cup V' : Y_{K_F} > 8\cdot\max(\mathrm{vol}_H(K_F),\mathrm{vol}_S(K_F))) \in O\left(e^{-\mathrm{vol}_H(V)}\right).$$

Thus, with high probability,

$$\phi(K_F) \geq \frac{\phi(H)\mathrm{vol}_H(K_F)+\phi(S)\mathrm{vol}_S(K_F)}{\mathrm{vol}_H(K_F)+\mathrm{vol}_S(K_F)+16\max(\mathrm{vol}_H(K_F),\mathrm{vol}_S(K_F))} \geq \frac{1}{18}\phi(H).$$

Case 4. *Large $S \cap K_F$ and large $H \cap K_F$.* Finally, in this case we have $\mathrm{vol}_H(K_F) \geq \frac{3\mathrm{vol}_H(V)}{4}$ and $\mathrm{vol}_S(K_F) \geq \frac{3\mathrm{vol}_S(V')}{4}$. Note that $g \leq \mathrm{vol}_S(V') \leq \frac{4}{3}\mathrm{vol}_S(K_F)$,

$$\begin{aligned}\phi(K_F) &= \frac{C_H(K_F) + C_S(K_F) + X_{V\cap K_F} + X_{V'\cap K_F}}{\mathrm{vol}_H(K_F) + \mathrm{vol}_S(K_F) + X_{V\cap K_F} + X_{V'\cap K_F} + 2Y_{K_F}} \\ &\geq \frac{C_H(K_F) + C_S(K_F)}{\mathrm{vol}_H(K_F) + \mathrm{vol}_S(K_F) + 2g} \\ &\geq \frac{C_H(K_F) + C_S(K_F)}{\mathrm{vol}_H(K_F) + \mathrm{vol}_S(K_F) + \frac{8}{3}\mathrm{vol}_S(K_F)} \\ &\geq \frac{C_H(K_F) + C_S(K_F)}{\frac{11}{3}(\mathrm{vol}_H(K_F) + \mathrm{vol}_S(K_F))}.\end{aligned}$$

Note that for any four positive positive real numbers we have that if $a/b > c/d$, then $\frac{a+c}{d+b} \geq \frac{c}{d}$, thus we get

$$\phi(K_F) = \min\left(\frac{C_H(K_F)}{\frac{11}{3}\mathrm{vol}_H(K_F)}, \frac{C_S(K_F)}{\frac{11}{3}\mathrm{vol}_S(K_F)}\right) \geq O(\phi(H)).$$

Having covered all four cases, we can then conclude that $\phi(G_F) \in \Omega(\phi(H))$ with high probability, completing the proof. □

A.2 Proofs of Lemmas A.2 and A.3.

Proof Proof of Lemma A.2.

Remember that $K_{k,k'}$ contains the subsets K_F of V_F such that $C_H(K_F) = k$ and $C_H(K_F) = k'$. Notice that once we have selected the k edges between $K_F \cap V$ and $V \setminus K_F$ and the k' ones between $K_F \cap V'$ and $V' \setminus K_F$, we have defined the two cuts in V and V', so we have just four possible sets K_F. Thus, for a given pair k, k', we have at most

$$\begin{aligned}|\boldsymbol{K_{k,k'}}| &\leq 4\left(\binom{|E|}{k}\cdot\binom{|E'|}{k'}\right) \leq 4\binom{|E|}{\frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil}\cdot\binom{|E'|}{\frac{1}{C}\lceil\phi(H)\cdot\mathrm{vol}(V')\rceil} \\ &\leq 4\left(\frac{|E|e}{\frac{1}{C}\phi(H)\cdot\mathrm{vol}(V')}\right)^{\frac{1}{C}\phi(H)\cdot\mathrm{vol}(V')}\left(\frac{|E'|e}{\frac{1}{C}\phi(H)\cdot\mathrm{vol}(V')}\right)^{\frac{1}{C}\phi(H)\cdot\mathrm{vol}(V')} \\ &\leq 4\left(\frac{|E||E'|e^2}{\left(\frac{1}{C}\phi(H)\cdot\mathrm{vol}(V')\right)^2}\right)^{\frac{1}{C}\phi(H)\cdot\mathrm{vol}(V')} \leq \left(\frac{2Ce\delta}{\phi(H)}\frac{2Ce\delta}{\phi(H)}\right)^{\frac{1}{C}\phi(H)\cdot\mathrm{vol}(V')},\end{aligned}$$

where $\delta = \max\left(\frac{|E|}{\mathrm{vol}(V')}, \frac{|E'|}{\mathrm{vol}(V')}\right)$. Thus, by the theorem hypotheses, $\delta = \frac{|E|}{\mathrm{vol}(V')}$. Finally, because of the lower bound on the size of $\mathrm{vol}_S(V')$ we know that,

$$\delta = \frac{|E|}{\mathrm{vol}_S(V')} \leq \frac{12|E|}{\mathrm{vol}_H(V)} \leq 6,$$

and hence,

$$\begin{aligned}|\boldsymbol{K_{k,k'}}| &= \left(\frac{12Ce}{\phi(H)}\right)^{\frac{2}{C}\phi(H)\cdot \mathrm{vol}_S(V')} \\ &\leq \exp\left(\left(\log\left(\frac{12C}{\phi(H)}\right)+1\right)\frac{2}{C}\phi(H)\cdot \mathrm{vol}_S(V')\right) \\ &\leq \exp\Big(\left(\log(12C)+\log\left(\frac{1}{\phi(H)}\right)+1\right)\cdot\frac{2}{C}\phi(H)\cdot \mathrm{vol}_S(V')\Big).\end{aligned}$$

Using the fact that $C > 9000$ we have $|\boldsymbol{K_{k,k'}}| \leq \exp\left(\frac{1}{384}\log\left(\frac{1}{\phi(H)}\right)\cdot\phi(H)\cdot \mathrm{vol}_S(V')\right)$, completing the proof. □

Proof Proof of Lemma A.3.
By definition of Y_{K_F} we have

$$E[Y_{K_F}] = g\frac{\mathrm{vol}_H(K_F)}{\mathrm{vol}_H(V)}\frac{\mathrm{vol}_S(K_F)}{\mathrm{vol}_S(V')}.$$

Let χ be the event $\{Y_{K_F} > 3\mathrm{vol}(K_F) + E[Y_{K_F}]\}$. Using the Chernoff bound (Theorem 2.2) we get that:

$$\begin{aligned}P(\chi) &= P\left(Y_{K_F} - E[Y_{K_F}] > 3\mathrm{vol}(K_F)\right) \\ &= P\left(Y_{K_F} - E[Y_{K_F}] > \left(\frac{3\mathrm{vol}(K_F)}{E[Y_{K_F}]}\right)E[Y_{K_F}]\right) \\ &\leq P\left(|Y_{K_F} - E[Y_{K_F}]| > \left(\frac{3\mathrm{vol}(K_F)}{E[Y_{K_F}]}\right)E[Y_{K_F}]\right) \\ &\leq \exp\left(-\frac{1}{3E[Y_{K_F}]}\left(3\mathrm{vol}(K_F)\right)^2\right) \\ &\leq \exp\left(-\frac{\mathrm{vol}_H(V)\mathrm{vol}_S(V')}{3g\mathrm{vol}_H(K_F)\mathrm{vol}_S(K_F)}\left(3\mathrm{vol}(K_F)\right)^2\right) \\ &\leq \exp\left(-3\mathrm{vol}_H(V)\frac{\mathrm{vol}_S(V')}{g}\frac{\left(\mathrm{vol}(K_F)\right)^2}{\mathrm{vol}_H(K_F)\mathrm{vol}_S(K_F)}\right) \\ &\leq \exp\left(-3\mathrm{vol}_H(V)\right).\end{aligned}$$

The last inequality follows from the fact that $\mathrm{vol}(K_F) \geq \mathrm{vol}_H(K_F), \mathrm{vol}_S(K_F)$ and that $\mathrm{vol}_S(V') \geq g$. □

References

[Albert and Barabási 02] R. Albert, and A. Barabási. "Statistical Mechanics of Complex Networks." *Reviews of Modern Physics* 74:1 (2002), 47–97.

[Andersen et al. 07] R. Andersen, F. Chung, and K. Lang. "Using PageRank to Locally Partition a Graph." *Internet Mathematics* 4:1 (2007), 1–128.

[Andersen and Peres 09] R. Andersen and Y. Peres. "Finding Sparse Cuts Locally Using Evolving Sets." In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC),* pp. 235–244. New York, NY: ACM, 2009.

[Arora et al. 09] S. Arora, S. Rao, and U. Vazirani. "Expander Flows, Geometric Embeddings and Graph Partitioning." *J. ACM.* 56:2 (2009), Article 5.

[Bahmani et al. 11] B. Bahmani, K. Chakrabarti, and D. Xin. "Fast Personalized PageRank on MapReduce." *SIGMOD '11.* New York, NY: ACM, 2011.

[Barabási and Albert 99] A.-L. Barabási, and R. Albert. "Emergence of Scaling in Random Networks." *Science.* 286:5439 (1999), 509–512.

[Bilge et al. 09] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. "All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks." In *Proceedings of the 18th International Conference on World Wide Web (WWW),* 551–560. New York, NY: ACM, 2009.

[Cao et al. 12] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro. "Aiding the Detection of Fake Accounts in Large Scale Social Online Services." In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI),* p. 15. Berkley, CA: USENIX, 2012.

[Cheng and Friedman 06] A. Cheng and E. Friedman. (2006). "Manipulability of PageRank Under Sybil Strategies." In *First Workshop on the Economics of Networked Systems (NetEcon)*, pp. 75–82.

[Cox and Noble 03] L. Cox and B. Noble. "Samsara: Honor Among Thieves in Peer-to-Peer Storage." In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles SOSP,* pp. 120–132. New York, NY: ACM, 2003.

[Danezis and Mittal 09] G. Danezis and P. Mittal. "SybilInfer: Detecting Sybil Nodes using Social Networks." In *NDSS.* (2009).

[dbl 11] (2011). Dblp. http://dblp.uni-trier.de/xml/.

[Douceur 02] J. Douceur. "The Sybil Attack." In *Revised Papers from the First International Workshop on Peer-to-Peer Systems (IPTPS)*, pp. 251–260. London, UK: Springer-Verlag, 2002.

[Dubhashi and Panconesi 09] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomised Algorithms.* New York, NY: Cambridge University Press, 2009.

[Fogaras et al. 05] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. "Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments." *Internet Mathematics.* 2:3 (2005), 333–358.

[Fortunato 09] S. Fortunato. "Community Detection in Graphs." *CoRR*, abs/0906.0612. (2009).

[Freeman 77] L. C. Freeman. "A Set of Measures of Centrality Based on Betweenness." *Sociometry* 40:1 (1977), 35–41.

[Garey and Johnson 79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NPCompleteness.* New York, NY: W. H. Freeman, 1979.

[Gharan and Trevisan 12] S. O. Gharan and L. Trevisan. "Approximating the Expansion Profile and Almost Optimal Local Graph Clustering." In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science,* pp. 187–196.Washington, D.C.: IEEE, 2012.

[Gyöngyi et al. 04] Z. Gyöngyi, H. Garcia-Molina, and J. O. Pedersen. (2004). "Combating Web Spam with TrustRank." Paper presented at the International Conference on Very Large data Bases (VLDB) 576–587. Toronto, Canada, August 2004.

[Jiang et al. 10] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, and B. Y Zhao. "Understanding Latent Interactions in Online Social Networks." In *IMC '10 Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pp. 369–382. New York: ACM, 2010.

[Klimt and Yang 04] B. Klimt and Y. Yang. "Introducing the Enron Corpus." Paper presented at the First Conference on Email and Anti-Spam (CEAS), Mountain View, CA, 2004.

[Leighton and Rao 99] T. Leighton and S. Rao. "Multicommodity Max-Flow Min-Cut Theorems and their Use in Designing Approximation Algorithms." *J. ACM.* 36:6 (1999), 787–832.

[Leskovec et al. 10] J. Leskovec, D. Huttenlocher, and J. Kleinberg. "Predicting Positive and Negative Links in Online Social Networks." In *Proceedings of the 19th international conference on World Wide Web* (*WWW), 641–650.* New York, NY: ACM, 2010.

[Leskovec et al. 05] J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations." In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDDWS),* pp. 177–187. New York, NY: ACM, 2005.

[Leskovec et al. 07] J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graph Evolution: Densification and Shrinking Diameters." *ACM Transactions on the Web (TWEB)* 1:1 (2007), 5.

[Leskovec et al. 08] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. "Statistical Properties of Community Structure in Large Social and Information Networks." In *Proceedings of the 17th International Conference on World Wide Web*, pp. 695–704. New York, NY: ACM, 2008.

[Leskovec et al. 09] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. "Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters." *Internet Mathematics* 6:1 (2009), 29–123.

[Lesniewski-Laas 10] C. Lesniewski-Laas. A Sybil-Proof One-Hop DHT. In *Proceedings of the 1st Workshop on Social Network Systems (SNS),* pp. 19–24. New York, NY: ACM, 2010.

[Lesniewski-Laas and Kaashoek 10] C. Lesniewski-Laas and M. F. Kaashoek. "Whanau: A Sybil-proof Distributed Hash Table." *In Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation (NSDI)*, p. 8. San Jose, CA: USENIX, 2010.

[Margolin and Levine 05] N. Margolin and B. N. Levine. "Quantifying and discouraging sybil attacks." Technical Report, UMass Amherst, 2005.

[Mislove et al. 10] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. "You Are Who You Know: Inferring User Profiles in Online Social Networks." In *WSDM '10 Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 251–260. New York: ACM, 2010.

[Mitzenmacher and Upfal 05] M. Mitzenmacher, and E. Upfal. *Probability and Computing*. New York, NY: Cambridge University Press, 2005.

[Mohaisen et al. 10] A. Mohaisen, A. Yun, and Y. Kim. "Measuring the Mixing Time of Social Graphs." In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC),* pp. 383–389. New York, NY: ACM, 2010.

[Newman 03] M. E. Newman."Mixing Patterns in Networks." *Physical Review E* 67:2 (2003), 026126.

[Newman and Girvan 04] M. E. Newman and M. Girvan. "Finding and Evaluating Community Structure in Networks." *Physical Review E*. 69:2 (2004), 026113.

[Pouwelse et al. 05] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips. "The Bittorrent P2P File-Sharing System: Measurements and Analysis." In *Peer-to-Peer Systems IV*, 205–216. Berlin, Heidelberg: Springer, 2005.

[Quercia and Hailes 10] D. Quercia and S. Hailes. "Sybil Attacks Against Mobile Users: Friends and Foes to the Rescue." In *Proceedings of INFOCOM,* pp. 1–5. IEEE, 2010.

[Richardson et al. 03] M. Richardson, R. Agrawal, and P. Domingos. "Trust Management for the Semantic Web." In *The Semantic Web–ISWC,* pp. 351–368. Lecture Notes in Computer Science 2870. Berlin, Heidelberg: Springer-Verlag, 2003.

[Sepandar D. Kamvar 03] D. Sepandar, M. T. Kamvar, and H. G.-M. Schlosser. "The Eigentrust Algorithm for Reputation Management in P2P Networks." In *WWW '03 Proceedings of the 12th International Conference on World Wide Web*, pp. 640–651. New York: ACM, 2003.

[Sinclair 92] A. Sinclair. "Improved Bounds for Mixing Rates of Marcov Chains and Multicommodity Flow." *Combinatorics,Probability & Computing*. 1:1 (1992), 351–370.

[Sinclair and Jerrum 89] A. Sinclair and M. Jerrum. (1989). "Approximate Counting, Uniform Generation and Rapidly Mixing Markov Chains." *Inf. Comput.* 82:1 (1989), 93–113.

[Spielman and Teng 04] D. A. Spielman and S.-H. Teng. "Nearly-Linear Time Algorithms for Graph Partitioning, Graph Sparsification, and Solving Linear Systems." In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 81–90. New York, NY: ACM, 2004.

[Stein et al. 11] T. Stein, E. Chen, and K. Mangla. "Facebook Immune System." In *SNS '11 Proceedings of the 4th Workshop on Social Network Systems*, Article 8. New York: ACM, 2011.

[Stytz 04] M. Stytz. "Considering Defense in Depth for Software Applications." *IEEE Security and Privacy Magazine* 2:1 (2004), 72–75.

[Tran et al. 11] N. Tran, N., J. Li, L. Subramanian, and S. Chow. "Optimal Sybil-Resilient Node Admission Control." In *Proceedings of IEEE INFOCOM,* pp. 3218–3226. IEEE, 2011.

[Tran et al. 09] N. Tran, B. Min, J. Li, and L. Subramanian. (2009). "Sybil-Resilient Online Content Voting." In *NSDI* 9:1 (2009), 15–28.

[Viswanath et al. 09] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. "On the Evolution of User Interaction in Facebook." In *Proceedings of the 2nd ACM Workshop on Online Social Networks (WOSN),* pp. 37–42. New York, NY: ACM, 2009.

[Viswanath et al. 12a] B. Viswanath, M. Mondal, A. Clement, P. Druschel, K. Gummadi, A. Mislove, and A. Post. "Exploring the Design Space of Social Network-Based Sybil Defenses." In *4th International Conference on Communication Systems and Networks (COMSNETS),* pp. 1–8, IEEE, 2012.

[Viswanath et al. 12b] B. Viswanath, M. Mondal, K. Gummadi, A. Mislove, and A. Post. "Canal: Scaling Social Network- Based Sybil Tolerance Schemes." In *Proceedings of the 7th ACM European Conference on Computer Systems (EuroSys),* pp. 309–322. New York, NY: ACM, 2012.

[Viswanath et al. 10] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove. "An Analysis of Social Network-Based Sybil Defenses." In *ACM SIGCOMM* 41:4 (2010), 363–374.

[Von Ahn et al. 03] L. Von Ahn, M. Blum, N. Hopper, and J. Langford. "CAPTCHA: Using Hard AI Problems for Security." In *EUROCRYPT' 03 Proceedings of the 22nd International Conference on Theory and Applications of Cryptographic Techniques*, pp. 294–311. Berlin, Heidelberg: Springer, 2003.

[Walpole et al. 93] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye. *Probability and Statistics for Engineers and Scientists*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[Watts and Strogatz 98] D. J. Watts and S. Strogatz. "Collective Dynamics of 'Small-World' Networks." *Nature* 393:6684 (1998), 440–428.

[Wei et al. 12] W. Wei, F. Xu, C. C. Tan, and Q. Li. "SybilDefender: Defend Against Sybil Attacks in Large Social Networks." In *Proceedings of IEEE INFOCOM,* pp. 1951–1959. IEEE, 2012.

[Xu et al. 10] L. Xu, S. Chainan, H. Takizawa, and H. Kobayashi. "Resisting Sybil Attack by Social Network and Network Clustering." In *IEEE Symposium on Applications and the Internet (SAINT),* pp. 15–21. IEEE, 2010.

[Yang et al. 13] C. Yang, R. Harkreader, and G. Gu. "Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers." *IEEE Transactions on Information Forensics and Security* 8:8 (2013), 1280–1293.

[Yang et al. 11] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai. "Uncovering Social Network Sybils in the Wild." In *Proceedings of Internet Measurement Conference* (*IMC),* pp. 259–268. IEEE, 2011.

[Yu 11] H. Yu. "Using Social Networks to Overcome Sybil Attacks." *ACM SIGACT News* 42:3 (2011), 80–101.

[Yu et al. 08] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. "SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks." In *IEEE Symposium on Security and Privacy, (SP 2008)*, pp. 3–17. IEEE, 2008.

[Yu et al. 06] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. "SybilGuard: Defending Against Sybil Attacks via Social Networks." *ACM SIGCOMM Computer Communication Review* 36:4 (2006), 267–278.

[Yuen et al. 11] M.-C. Yuen, I. King, and K.-S. Leung. "A Survey of Crowdsourcing Systems." In *Proceedings of the 3rd IEEE International Conference on Social Computing (IEEE SocialCom)*, Boston, MA, October 9–11, 2011, pp. 766–773.

[Zhu et al. 13] Z. A. Zhu, S. Lattanzi, and V. Mirrokni. "A Local Algorithm for Finding Well-Connected Clusters." In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Atlanta, GA, June 15–21, 2013, pp. 396–404.

Lorenzo Alvisi, Department of Computer Science, The University of Texas at Austin, 2317 Speedway, 2.302, Austin, TX 78712, USA (lorenzo@cs.utexas.edu)

Allen Clement, MPI-SWS, Campus E1 4, D-66123 Saarbruecken, Germany (aclement@mpi-sws.org)

Alessandro Epasto, Department of Computer Science, Sapienza University of Rome, Via Salaria 113, 00198 Rome, Italy (epasto.ale@di.uniroma-1.it)

Silvio Lattanzi, Google NYC, 111 8th Avenue, New York, NY 10011, USA (silviol@google.com)

Alessandro Panconesi, Department of Computer Science, Sapienza University of Rome, Via Salaria 113, 00198 Rome, Italy (ale@di.uniroma1.it)