On the Local Nature of Sybil Defense in Online Social Networks

ABSTRACT
This paper explores the fundamental limits of using only the structure of social networks to defend against sybil attacks. We derive the number of attack edges needed to foil sybil defenses based on each of the known statistical properties of social graphs: our results suggest that it may be impossible to use this properties to identify with high probability both sybil and honest nodes. We then explore a more modest goal for social defenses: to use local graph properties, such as community structure, to provide an honest node with a way to white-list a set of trustworthy nodes. While recent literature claims that any community detection algorithm may prove equally useful, we find that the choice of community detection protocol can dramatically affect whether sybil nodes are successfully kept from being white-listed. We demonstrate both analytically and experimentally that Local Page Rank can very effectively exploit a community’s local properties to keep sybil nodes at bay.

1. INTRODUCTION
This paper challenges the increasingly popular notion [5, 20–23], that analyzing the global topological structure of real-life social networks can be an effective tool in defending decentralized systems against realistic sybil attacks [6]. There are excellent reasons for this popularity. First, the possibility that malicious users may forge an unbounded number of sybil identities (also called sybil nodes) that are indistinguishable from honest ones is a fundamental threat to any distributed system that relies on voting to determine the outcome of a collaborative task. This threat is particularly acute in a decentralized system, where it may be impractical or even impossible to employ techniques [14] that depend on a single central authority to certify which identities are legitimate. Sybil defenses based on social networks (which we will also refer to as social defenses from now on) are attractive because they promise protection without requiring any centralized infrastructure: honest nodes can verify with high probability whether another node is honest or sybil by taking local steps that leverage global properties of social networks. A second reason for the popularity of social defenses is that they offer their guarantees within a rigorous and elegant theoretical framework, which allows them to provide precise bounds on the severity of attacks (measured in the number of attack edges that sybil nodes can fool honest nodes into accepting) that they are able to withstand before they stop working. Finally, these techniques come very close to meeting Douceur’s gold standard of sybil defense—“Ideally, the system should accept all legitimate identities but no counterfeit entities” [6]: with high probability, an honest node will accept almost all honest nodes while rejecting all but a bounded number of sybil nodes. Need, rigor, and elegance make for a compelling combination, and higher-level sybil-proof applications relying on this style of sybil defense are beginning to surface [12, 13, 17].

As we will see in greater detail in Section 2, current social defenses rely on two fundamental assumptions: first, that the cut between honest and sybil nodes in the social graph is sparse, implying that it must be hard for sybil nodes to create attack edges; second, that the portion of the graph that involves only honest nodes is fast mixing: informally, this implies that any subset of honest nodes is well connected to the rest of the honest nodes and, formally, it is equivalent to requiring that the conductance (defined in Section 2) of the subgraph of honest nodes be asymptotically a constant.

The plausibility of these assumptions has recently come under scrutiny. Recent evidence [1, 4] suggests that sybil nodes may be more successful at creating attack edges than social defenses are willing to give them credit for; if so, these defenses may be less effective in practice than one may hope for. Even more crucially, several papers [3, 10, 11] have questioned whether the conductance of social networks is in fact a constant: if not, the entire elegant theoretical framework on which current social defenses rest would collapse (see Section 2), and honest nodes would be left with no rigorous guarantees. Indeed, as Vishwanath et al. point out in their recent analysis of social network-based sybil defenses [21] “it
is not known ... if there are fundamental limits to using only the structure of social networks to defend against Sybils”.

In this paper, we make three main contributions to the ongoing debate on the merits of social-based sybil defense.

First, we offer a precise answer to the question raised by Viswanath and his colleagues. We consider four well-known global properties of the structure of a social network (popularity, small-world structure, clustering coefficient and conductance) (Section 3), and for each property \( P \) we derive bounds on the number of attack edges needed by an adversary to connect a network \( S \) of sybil nodes to a network \( H \) of honest nodes in such a way that the new joint graph becomes indistinguishable with high probability from \( H \) with respect to \( P \). Further, we confirm empirically the fundamental fragility of current social defenses, which assume constant conductance, when applied to social graphs that have not been pre-processed to eliminate nodes with low degree [22]. We find that the such preprocessing, while it produces a new graph with the conductance required by current social defenses, can sometimes remove much more than half of the nodes in the graph, altering its structural characteristics much more than we assume an adversary can!

Second, in light of the these fundamental limits, we advocate a reappraisal of the goals of social-based sybil defense: since Douceur’s gold standard is out of the reach of any social defense mechanism that relies on any of the known global properties of social networks, we submit that the new goal of social-based sybil defenses should be to determine whether by using local graph properties, such as its community structure, it is at least feasible to provide an honest node with a way to white-list a set of trustworthy nodes.

We are not the first to observe the potential of community detection as a tool for sybil defense: indeed, after observing that community detection is at the core of all existing social defenses, Vishwanath et al. conclude that “it is possible to use off-the-shelf community detection algorithms to find Sybils” [21]. We find otherwise. Our experiments show that the choice of community detection protocol can dramatically affect whether sybil nodes are successfully kept from being white-listed: in particular, we find that the sybil-defense capability of the community detection algorithm proposed in [21] can be crippled by a simple attack involving only two attack edges. In retrospect, this sensitivity should perhaps not be surprising, since there exist multiple ways to formalize the qualitative notion of community, and different protocols look for communities using different criteria.

Third, we demonstrate that Local Page Rank (henceforth LPR) [18] can be effective tool for local sybil defense. As all other social defense techniques, LPR defends against sybil nodes by leveraging conductance, the property of social networks that is harder for an adversary to circumvent: indeed, our bounds show that the higher the conductance, the higher the number of attack edges the adversary needs to foil the defense. Unlike previous techniques, however, LPR does not rely on the conductance of the entire graph, which is the minimum conductance of the cut induced by any subset that includes at most half of the nodes in the graph; instead, LPR can offer greater assurance to sets of nodes with high conductance. Furthermore, LPR’s effectiveness is totally independent of the controversial assumption that conductance is constant. We will discuss the theoretical properties that make LPR a good candidate for a principled approach to sybil defense and will demonstrate its effectiveness experimentally.

2. THE SYBIL-* APPROACH TO SYBIL DEFENSE

Yu et al. jump-started the area of socially-based sybil defense with their seminal SybilGuard protocol [23]. The techniques used by SybilGuard to realize their elegant vision of decentralized sybil defense depend on a rigorous theoretical framework (henceforth referred to as the sybil-* approach), which has since been adopted by several other protocols either to improve SybilGuard’s guarantees [5, 20, 22], or to build sybil-proof applications [12, 13, 17].

Since our paper is motivated by the belief that the assumptions on which that framework rests are dangerously strong, we begin by reviewing these techniques and by discussing the crucial role that the corresponding assumptions play in guaranteeing their soundness.

2.1 Picking whom to trust

The verification process that an honest node \( u \) uses in the sybil-* approach to determine whether it can trust another node \( v \) is based, at its core, on a simple 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.

Each of protocol in the sybil-* family apply this sampling strategy in different ways and to different parts of the graph. SybilLimit [23] samples edges; SybilGuard [22] and Gatekeeper [20] sample nodes in the graph; SybilInfer [5] uses the random walk to build a Bayesian model for the likelihood that a trace \( T \) was initiated by an honest node. In the following text we provide an overview of how SybilLimit [23] applies the random sampling of edges to identify honest users. While the details of the discussion are specific to SybilLimit, the intuition for how the structural properties of the graph allow the random sampling to be effective is common to the entire sybil-* family.
Suppose $u$ and $v$ belong to the same graph, one with $n$ nodes and $m$ edges. Both nodes select an edge at random: $u$ accepts $v$ if they pick the same edge.

Of course, the probability that this happens is very low, $\frac{1}{m}$. To boost it we can do as follows. 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. This probability is,

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

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. We now have a simple protocol that 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 they trust each other. This basic idea can be refined further to come up with a test that succeeds with overwhelming probability with small sized edge sets.

This technique can be immediately used to test membership in graphs. We have two disjoint graphs and two nodes, but we do not know if they belong to the same graph. If vertices are bound to pick the edge set in their own graph, the mechanism above provides the test we are looking for: if the two vertices live in different graphs the chance that they trust each other is zero, otherwise it is given by Equation (1).

But how can we implement the test in a distributed social network? It is here that conductance enters into the picture. Intuitively, the conductance of a set $S$ of vertices, denoted as $\varphi(S)$, in a given network $G = (V, E)$ is the ratio between the number of edges going out from $S$ and the number of edges inside $S$. More precisely,

$$\varphi(S) := \frac{|\{(u, v) : u \in S, v \in V - S\}|}{\sum_{u \in S} \text{deg}(u)}.$$

The quantity at the numerator is the size of the cut induced by $S$, while the denominator is the so-called volume of $S$, the sum of the degrees of vertices contained in $S$.\footnote{There is one more technicality in the definition that we avoid for sake of clarity.} Intuitively, the conductance of a set $S$ gives the probability that a random walk that has entered $S$ will leave it at the next step. The conductance of a graph is the minimum conductance of a set (where the minimum is taken over all sets of size at most $|V(G)|/2$) and is denoted as $\varphi(G)$. It is known that graphs whose conductance is constant have a very interesting property: their mixing time \cite{16}, which quantifies how fast the ending point of a random walk approaches the stationary distribution in which the walk’s ending point is independent of its beginning, is $O(\log n)$.

If the conductance of the graph is constant, then we can specify a new practical procedure that $u$ and $v$ can follow to select an edge at random: start a random walk of length $\Theta(\log n)$ from each vertex and pick the last edge on this walk. Repeat the process $\sqrt{m}$ times, and if $u$ and $v$ belong to the same graph, the probability that they will end up trusting each other is again given by Equation (1), and it is zero otherwise.

Suppose now that we insert just one edge at random between the two graphs. With this change, if the two graphs are large, there is a tiny chance that vertices in two copies will select the same edge— one random walk could enter the other graph and collide—but intuitively it is a very small chance. The membership test is still robust.

In the context of sybil defense, the two initially disjoint graphs comprise, respectively, honest nodes and sybil nodes. Sybil-* protocols use the membership test to determine whether an honest node $u$ should accept another node $v$ as honest. How many random bridges (the attack edges mentioned in the Introduction) can one insert between the graphs and still retain the robustness of the test?

Yu et al. show \cite{23} that if a sybil graph attaches itself to the honest one via $o(\frac{n}{\log n})$ edges the test is still reliable.

Hence, there are two fundamental assumptions that underpin the sybil-* approach to sybil-defense. The first is that the cut between the sybil and the honest region—the set of attack edges—is random and suitably sparse. The second, crucial one, is that the conductance is constant (and hence very high\footnote{The conductance of a class of graphs, say, social networks, can be a function of graph size. For slow functions like $\frac{1}{\log n}, k > 1$ it can be considered already quite high and if it is constant it is very high. A small conductance would be a value of, say, $\frac{1}{\sqrt{m}}$.}): it is this property that ensures that random walks of $\Theta(\log n)$ steps will end in a random edge. If this assumption is not satisfied, then the entire theoretical framework on which sybil-* protocols depend breaks down.

### 2.2 Gathering clouds

Motivated by the popularity of the sybil-* several recent papers have tested the robustness of the assumptions on which it depends. The results have not been encouraging.

**Random and sparse sets of attack edges** As we pointed out in the introduction, users of many social networks appear to be much more causal in accepting random edges than sybil-* protocols would like them to. In a 2007 study \cite{1}, Sophos showed that over 40% of the
Facebook users contacted by Freddi Staur, a small green plastic frog, were willing to accept his request for friendship. Less naive users are at risk as well. In a recent paper [4], Bilge et al. describe a system that automatically crawls popular social networking sites, collects information on their users, automatically create profiles, and send the cloned users’ friends contact requests and personal message; their experiments demonstrate that the system can successfully induce those friends to accept contact requests from whomever controls the cloned identities. These result suggests that automatic attack based on social engineering may fool even users who are careful only to accept contact requests from their (as it happens, pseudo) friends.

Even if the number of attack edges can be contained within the bounds required by sybil-* defense, Viswanath et. al [21] point out that the adversary may not establish those edges by linking to random nodes in the network, but instead target nodes that are closer to the honest node that is trying to determine whom to trust. Their experiments show that these targeted Sybil attacks can rather handily fool sybil-* defenses.

**Constant conductance**
The crucial assumption of constant conductance has been put into question by several recent papers. Studying snapshots of several social networks (including Livejournal, Flickr, the Messenger chat graph, and LinkedIn) Leskovec et al observe [10, 11] that their conductance is at most, respectively, $\frac{1}{1000}$, $\frac{1}{800}$, $\frac{1}{500}$ and $\frac{1}{100}$, indicating that conductance behaves more as an inverse poly-logarithmic factor of $n$ than a constant. Mohaisen et. al have cast similar doubts, having found the mixing time to be higher (and hence the conductance lower) than previously thought [3].

3. **FUNDAMENTAL LIMITS OF SOCIAL NETWORK BASED SYBIL DEFENSE**

To help us achieve some perspective on the limitations of the sybil-* protocols, we take on the fundamental question asked by Viswanath [21] Are there fundamental limits to using solely the global structure of social networks to defend against Sybil attacks? To answer it, we consider the four known structural properties of social networks—popularity, the small world property, clustering coefficient, and conductance, and, having specified an attack model, we determine for each property how hard it is for an adversary to attach a sybil graph $S$ to an honest graph $H$ in such a way that the structural properties of the resulting graphs are indistinguishable from the ones of $H$.

3.1 **Attack model**

We assume a given network $H$ of honest nodes and a network $S$ of sybil nodes whose topology is under total control of the adversary (while $H$ is fixed). The adversary tries to set up $m := |E(H)|$ potential attack edges that connect $H$ with $S$ as follows. The endpoint in $H$ is chosen by preferential attachment, i.e. $u \in V(H)$ is chosen with probability

$$\frac{\text{deg}_H(u)}{2m}$$

where $\text{deg}_H(u)$ denotes $u$’s degree. The other endpoint is chosen in $S$ by preferential attachment. The potential edge is inserted with probability $p$ and discarded with probability $1 − p$. This is the usual model for sybil attacks parametrized by $p$. We account for the outcome of the social engineering experiments of [4] by allowing $p$ to be constant.

The network that results from joining $H$ and $S$ we denote with $G$. Note that the expected number of attack edges is $pm$. It follows easily from large deviation theory that the number of attack edges is also concentrated around this value.

In this paper, we will evaluate the impact of this attack on the seven graphs described in Table 1. For all graphs we consider the largest strongly connected component. Table 2 describes the statistical properties of these graphs before and after they are attacked.

<table>
<thead>
<tr>
<th>Graph Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AstroPh [9]</td>
<td>Collaboration network of Arxiv Astro Physics</td>
</tr>
<tr>
<td>HE Physics [8]</td>
<td>Citation network of Arxiv High Energy Physics</td>
</tr>
<tr>
<td>EuAll [9]</td>
<td>Email network from an EU research institution</td>
</tr>
<tr>
<td>Enron [8]</td>
<td>Email communication network from Enron</td>
</tr>
</tbody>
</table>

**Table 1: Evaluated social graph.**

3.2 **Popularity**

The structural property we consider, popularity, captures the well-known fact that the distribution of degrees of nodes in a social network is heavy-tailed, as in a power law or a lognormal distribution. We find that it is not hard for the adversary to make sure that $G$’s popularity distribution is statistically indistinguishable from $H$:

**Theorem 1.** If the degree distribution of the graph of honest nodes $H$ follows a long tailed distribution so does $G$, the graph under sybil attack.

We leave the proof to the Appendix of [2] The intuition is that, except for very few nodes, with high probability,
Table 2: Statistical properties of the largest strongly 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 3.1.

<table>
<thead>
<tr>
<th>Graph</th>
<th>Nodes</th>
<th>Edges</th>
<th>AttackEdges</th>
<th>90% diameter</th>
<th>Clustering Coeff</th>
<th>Est. Conductance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AstroPh</td>
<td>35806</td>
<td>395925</td>
<td>0</td>
<td>4.60</td>
<td>0.66</td>
<td>0.005111</td>
</tr>
<tr>
<td>... p = 0.1</td>
<td>35806</td>
<td>413799</td>
<td>19855</td>
<td>4.92</td>
<td>0.61</td>
<td>0.043478</td>
</tr>
<tr>
<td>Epinions</td>
<td>26588</td>
<td>100120</td>
<td>0</td>
<td>5.84</td>
<td>0.23</td>
<td>0.020408</td>
</tr>
<tr>
<td>... p = 0.1</td>
<td>53176</td>
<td>201270</td>
<td>1030</td>
<td>6.73</td>
<td>0.23</td>
<td>0.005180</td>
</tr>
<tr>
<td>HE Physics</td>
<td>30356</td>
<td>346631</td>
<td>34689</td>
<td>5.55</td>
<td>0.27</td>
<td>0.037037</td>
</tr>
<tr>
<td>... p = 0.1</td>
<td>60712</td>
<td>727915</td>
<td>0</td>
<td>4.88</td>
<td>0.27</td>
<td>0.037037</td>
</tr>
<tr>
<td>Enron</td>
<td>33696</td>
<td>180811</td>
<td>0</td>
<td>5.01</td>
<td>0.71</td>
<td>0.004525</td>
</tr>
<tr>
<td>... p = 0.1</td>
<td>67392</td>
<td>363425</td>
<td>1803</td>
<td>5.09</td>
<td>0.70</td>
<td>0.005190</td>
</tr>
<tr>
<td>EuAll</td>
<td>32430</td>
<td>54397</td>
<td>0</td>
<td>6.61</td>
<td>0.64</td>
<td>0.017857</td>
</tr>
<tr>
<td>... p = 0.1</td>
<td>64860</td>
<td>109331</td>
<td>537</td>
<td>5.02</td>
<td>0.51</td>
<td>0.004884</td>
</tr>
<tr>
<td>Mashdot</td>
<td>70999</td>
<td>365572</td>
<td>0</td>
<td>4.84</td>
<td>0.10</td>
<td>0.004761</td>
</tr>
<tr>
<td>... p = 0.1</td>
<td>141998</td>
<td>734844</td>
<td>3700</td>
<td>5.48</td>
<td>0.10</td>
<td>0.005069</td>
</tr>
<tr>
<td>Wiki-Talk</td>
<td>92117</td>
<td>360767</td>
<td>36347</td>
<td>4.78</td>
<td>0.12</td>
<td>0.004761</td>
</tr>
</tbody>
</table>

Figure 1: Degree distribution of the Epinions graph before and after attack. Note that the attack shift the distribution up and to the right, but does not change the shape of the lines.

after the attack the degree of every node will not change by much. Indeed, since endpoints of attack edges are chosen according to (2), the expected degree increase of a vertex in $H$ is

$$\frac{\deg_H(u)}{2m}pm = 2p\deg_H(u)$$

because the vertex is chosen with probability $\frac{\deg_H(u)}{2m}$ and there are $pm$ attack edges. The adversary can shape the sybil region $S$ in any way it wants to make it fit the desired target distribution. Figure 1 shows a plot of the degree distribution in the Epinions social network before and after two attacks in which attack edges are inserted respectively with probability $p = .01$ and $p = 0.1$. Note that the curves produced by the two attacks are barely distinguishable from each other and that their shape is virtually identical to that of the curve produced by the original Epinions. The curves under attack are shifted up, since the attack doubles the size of the graph, and a little bit to the right, since by and large all vertices with a given degree degree increase it by the same amount after the attack (the equation above shows this is true in expectation but it also holds with high probability).

3.3 Small world property

The celebrated small world property is one of the most interesting and most studied properties of social networks. It refers to the fact that their diameter—the longest distance between any two nodes in the network—is small. It is clear that if, say, the adversary picks $S = H$ and it succeeds in inserting just one attack edge the diameter can at most double.

**FACT 2.** A sybil attack can at most double the diameter.

The adversary, having control of the sybil region $S$ can effect this change in a piecemeal fashion. Since a slow doubling of the diameter can be entirely physiological, the small world property cannot be relied upon to develop effective sybil defense schemes. The 90% diameter column of Table 2 confirms, for several real-life social networks, how the diameter is barely affected by the attacks.

3.4 Clustering coefficient
The third property is related to the so-called clustering coefficient, a measure of how closely-knit social networks are. The clustering coefficient of a vertex $u$, denoted here as $c_u$, is computed as follows. Let $k$ be the maximum number of edges between neighbors of $u$. Then, $k = \binom{\deg(u)}{2}$. And let $f_u$ be the actual number of edges between neighbors of $u$, i.e. $f_u := |\{xy : x \in N_u, y \in N_u\}|$. Then,

$$c_u := \frac{f_u}{k}.$$ 

When we associate the vertex of a social network with the user that it represents, the clustering coefficient is measure of how many of that user’s friends are friends among them; more precisely, it is the ratio between the actual number of friendships between the friends of a user and the number of all possible friendships between them. In a social network the clustering coefficient is high with respect to what one would find in a purely random network because social forces tend to create friendships between the friends of a person (a friend of a friend is my friend). The clustering coefficient of a network is the minimum clustering coefficient of any vertex, i.e. $c(G) := \min_{u \in V(G)} c_u$. At the outset, the clustering coefficient appears very promising as a bulwark against sybil attacks because attack edges reduce its value. But, it turns out, not enough to be detected. Indeed, while the clustering coefficient of social networks is typically high, it varies significantly from network to network. The authors of [10] show that for LiveJournal it is about 0.35, that of the social network of Messenger chat exchanges is about 0.09, while that of the actor collaboration network of IMDB is 0.79. Thus, if an attack modifies the clustering coefficient by a small multiplicative factor the change is hard to detect, especially if made very gradually. This however is exactly what the adversary can accomplish. This intuition is captured precisely in the following result.

**Theorem 3.** Let $H$ be the graph of honest nodes and let $G$ be the network under sybil attack. Also, let $\alpha := \frac{3}{2}(1+p)^2$ and $\beta := 3(1+2p)^2$, where $p$ is the probability that an attack edges is accepted. Then,

$$E[c(G)] \geq \alpha c(H)$$

where $E[c(G)]$ denotes the expected value of $c(G)$, and

$$c(G) \geq \beta c(H)$$

with high probability.

Although we leave the proof to the Appendix of [2] the theorem’s implications are disappointingly clear: the clustering coefficient is not a good basis for sybil defense, since even after the attack its value cannot drop by too much. The Clustering Coeff column of Table 2 confirms the theorem’s predictions.

### 3.5 Conductance

At last we analyze conductance, the property that is leveraged by sybil-* protocols to achieve random sampling. We ask the question: how many attack edges are needed for the adversary to maintain the same conductance? The following theorem provides the answer.

**Theorem 4.** Let $H$ denote a network of $n$ honest nodes and $m$ edges such that $\varphi(H)m = \Theta(\log n)$, and let $S$ denote a network of $n'$ sybil nodes with $m'$ edges such that $\varphi(S) \geq \varphi(H)$ and $m' \leq m$. Suppose further that the adversary is able to set up at least 

$$\varphi(G)m \log \varphi(G)^{-1}$$

attack edges. Then, with high probability, $\varphi(G) = \Theta(\varphi(H))$.

The theorem (whose rather technically-involved proof can be found in the Appendix of [2]) generalizes to graphs of arbitrary conductance the $o\left(\frac{n}{\log n}\right)$ bound proved by Yu et al. for the special case of graphs with constant conductance. From the theorem we learn that if there are at least $\varphi(H)m \log \varphi(G)^{-1}$ attack edges the conductance of the graph will remain very nearly the same, with high probability, and so it will be impossible to even detect that the network is under sybil attack. It is possible to show that this number of attack edges is also necessary for the adversary to foil detection with high probability.

Table 2 confirms the theorem’s message that, once an adversary succeeds in generating sufficient attack edges, he can circumvent any technique that attempts to detect sybil nodes by looking for significant changes in global conductance. In particular, we see that, as expected, the conductance drops significantly under a weak attack ($p = 0.01$), providing leverage for sybil detection; under a strong ($p = 0.1$), attack however, the conductance may actually increases because, by adding random attack edges, the adversary enlarges every cut with some probability, including the cut with minimum conductance which defines the conductance of the entire network.

Note that computing the conductance is NP-Hard. The conductance values that we report are approximate and were obtained using the the approximation method proposed by Leskovec et al. [10, 11].

### 3.6 Discussion

The results of this section present a fairly depressing outlook for any hope of using global properties of social networks to meet Doucet’s gold-standard of sybil defense—to identify, with high probability, whether a node is honest or sybil—as none of the properties provides anything approaching a silver bullet. We make three corollary observations. First, conductance is by far the most useful social network property for sybil defense, as it is the one that, to be foiled, requires the great-
Table 3: Statistical properties of the graphs before and after pre-processing. Pre-processing drastically reduces the graphs’ size and significantly alters their structural properties.

<table>
<thead>
<tr>
<th>Graph</th>
<th>Nodes</th>
<th>Edges</th>
<th>Diameter</th>
<th>90% Diameter</th>
<th>Clustering Coeff</th>
<th>Est. Conductance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AstroPh</td>
<td>17903</td>
<td>190972</td>
<td>11</td>
<td>5.29</td>
<td>0.61</td>
<td>0.011010</td>
</tr>
<tr>
<td></td>
<td>11091</td>
<td>172103</td>
<td>8</td>
<td>4.14</td>
<td>0.60</td>
<td>0.014925</td>
</tr>
<tr>
<td>Epinions</td>
<td>26588</td>
<td>100120</td>
<td>14</td>
<td>5.84</td>
<td>0.23</td>
<td>0.020408</td>
</tr>
<tr>
<td></td>
<td>4989</td>
<td>57326</td>
<td>7</td>
<td>3.76</td>
<td>0.18</td>
<td>0.044084</td>
</tr>
<tr>
<td>HE Physics</td>
<td>30356</td>
<td>346613</td>
<td>12</td>
<td>5.15</td>
<td>0.29</td>
<td>0.037037</td>
</tr>
<tr>
<td></td>
<td>23190</td>
<td>301940</td>
<td>8</td>
<td>4.81</td>
<td>0.26</td>
<td>0.065819</td>
</tr>
<tr>
<td>EuAll</td>
<td>32430</td>
<td>54397</td>
<td>8</td>
<td>4.60</td>
<td>0.52</td>
<td>0.030928</td>
</tr>
<tr>
<td></td>
<td>1180</td>
<td>7906</td>
<td>6</td>
<td>3.63</td>
<td>0.19</td>
<td>0.181028</td>
</tr>
<tr>
<td>Enron</td>
<td>33696</td>
<td>180811</td>
<td>11</td>
<td>4.68</td>
<td>0.71</td>
<td>0.004525</td>
</tr>
<tr>
<td></td>
<td>9149</td>
<td>89851</td>
<td>10</td>
<td>4.74</td>
<td>0.45</td>
<td>0.009524</td>
</tr>
<tr>
<td>Slashdot</td>
<td>70999</td>
<td>365572</td>
<td>11</td>
<td>4.84</td>
<td>0.10</td>
<td>0.023256</td>
</tr>
<tr>
<td></td>
<td>18076</td>
<td>208536</td>
<td>6</td>
<td>3.74</td>
<td>0.03</td>
<td>0.076923</td>
</tr>
<tr>
<td>Wiki-Talk</td>
<td>92117</td>
<td>360767</td>
<td>9</td>
<td>4.63</td>
<td>0.14</td>
<td>0.047619</td>
</tr>
<tr>
<td></td>
<td>13523</td>
<td>147397</td>
<td>6</td>
<td>3.74</td>
<td>0.06</td>
<td>0.372960</td>
</tr>
</tbody>
</table>

Figure 2: The precision of SybilLimit at when recall is 95% on each of the social networks we consider when the graphs are attacked with attack strength \( p = 0.01 \). Other sybil-* protocols show qualitatively similar results.

All global social-defense protocols are subject to the bound on the number of attack edges given in Theorem 4. As we have seen in Section 2, the theoretical framework that sustains sybil-* protocols crucially depend on one additional assumption: that the network is fast mixing, or, equivalently, that conductance is constant. This is far from a weak assumption. Conductance in general can go to zero as a function of \( n \), the number of nodes in the network: assuming it constant means that it is very high, and, as we have already observed, social networks do not seem to demonstrate the required high values. The authors of the sybil-* suite of protocols appear to give an implicit nod to this fact when, before applying their protocol to a social network, they pre-process it to eliminate all vertices of small degree [5, 20, 22, 23]: the effect of this step is to boost the conductance of the graph that is left to the level necessary for their algorithm to work properly.

In this section, we show that there exists a basic tension between the need for this pre-processing step and the stated goal of sybil-* protocols to identify, with high probability, almost all honest nodes (and hence, almost all sybil ones). We accomplish this in two steps. We first show the impact of pre-processing on the structural properties of social networks; then, we measure the impact of (the lack of) pre-processing on the robustness of sybil-* protocols.

**How pre-processing changes the network** Table 3 shows the results of the pre-processing step for the social networks listed in Table 1. We make two observations. First, the size of the graphs is dramatically reduced. For the case of Wiki-talk, our largest network, the pre-processing step removes from the graph, and from the opportunity of taking advantage of sybil protection, over 85% of the nodes in the original network. Second, comparing Table 3 with Table 2 shows that the pre-processing step significantly alters the structural properties of the net-
Figure 3: Precision plotted against node ranking for an attack on Wiki-Talk with $p = 0.01$. This results in a graph with a total attack edge count equal to 1% of the edges in the initial graph.

**Figure 2** shows the precision achieved SybilLimit [22] for each of the considered social networks both without (black bar) and with (white bar) pre-processing. We select $k$ so that the recall of SybilLimit (defined as the fraction between the number of honest nodes correctly identified and the total number of honest nodes) is at 95%, meaning that at most 5% of honest nodes are not recognized as such. We set $p = 0.01$, which, for the networks that we consider, results in a number of attack edges that is within the $o\left(\frac{n}{\log n}\right)$ bound for which SybilLimit is designed. We see that while SybilLimit achieves near perfect precision after pre-processing, its precision can drop dramatically without it.

Figures 3(a) and 3(b) show a similar drop in precision once pre-processing is not performed. In this case, we focus just on the Wiki-Talk network and show the performance of different sybil-* protocols with and without pre-processing. The black vertical line in the figures corresponds to a value of $k$ equal to the total number of honest nodes: in other words, a protocol that achieved perfect precision up to the vertical line would be able to recognize every honest node (perfect recall). One should not draw conclusions from this figure about the relative performance of sybil-* protocols: although Gatekeeper outperforms SybilGuard and SybilLimit for Wiki-Talk, we found that for other networks, such as Slashdot, the protocols trade places. The only intended conclusion from this figure is that the absence of preprocessing can severely impact the sybil-* protocols.

5. LOCAL COMMUNITY DETECTION

We draw two lessons from our results so far. First, there exist fundamental limitations to what can be achieved in terms of sybil defense by leveraging only the global properties of social networks. Second, while conductance is the property most suitable for sybil defense, there is a high price to be paid for basing sybil defense on the strong assumption of constant conductance: trying to keep true to Douceur’s gold standard (almost perfect precision and recall) requires sybil-* protocols to operate under conditions where the random sampling techniques that these protocol rely upon may no longer be sound.

The way forward, we believe, is one that makes the best possible use of the limited, but nonetheless real, ability to leverage conductance for sybil detection—and can do so without requiring that conductance be constant.

In Section 3 we have already observed that while the conductance of an entire graph may be small—because it is the smallest conductance of any cut in the graph, the conductance of a set within the graph may be much higher. This leads us to propose a new approach to socially-based sybil defense: rather than aiming for indentifying with high probability almost all honest and sybil nodes as such, we settle for the more limited goal of guaranteeing that honest nodes, with high probability, will be able to white list a set of trustworthy nodes—and, of course, will not need constant conductance to do so.
Local Sybil defense requires community detection algorithms that identify communities with two important properties:

1. The community must be easy to split away from the rest of the graph (i.e. be separated from the rest of the graph by a small cut)

2. The community must be difficult to split itself (i.e. have no small internal cut).

In the next section we discuss an algorithm, LPR, which has these properties.

Refining the problem. Viswanath et al. propose community detection as a solution to the classic statement of Sybil defense which requires separating all (most) honest users from all (most) Sybil users. We believe that the path forward is to target a weaker notion of local Sybil defense in which we attempt to accept as many honest nodes as possible while rejecting all (most) Sybil users. Note that the difference between the two goals is subtle. The traditional Sybil defense goal is only as resilient as the global conductance—i.e. the smallest conductance of any cut in the graph—while local Sybil defense can provide stronger protection to communities with higher local conductance.

The potential of community detection as a defensive mechanism against sybil attacks has been first observed by Viswanath et al. [21]. They note that at the core of all sybil-* protocols lies a community detection problem: the goal of each honest user is to create exactly two communities, one that includes almost all honest nodes, the other that is overwhelmingly comprised of sybil nodes. If sybil-defense reduces to community detection, then it becomes possible to apply to sybil defense the large body of research on community detection. Indeed, they conclude that off-the-shelf community detection algorithms, such as the one by Mislove [15] that they study extensively, are likely to lead to effective defenses. We find otherwise.

Not all community detection is created equal. Mislove’s algorithm builds communities greedily. Starting from a vertex $v$, the algorithm computes a set by adding to the current set the vertex (connected to the current set) that increases the conductance the most. If no neighboring vertex increases the conductance, then the algorithm stops. $^3$

$^3$We note that the original proposal for Mislove’s algo-

In light of these results, we believe that it is appropriate to reappraise the goals of socially-based sybil defense. The conclusions we draw from the above lessons is that any socially-base

Refining the problem. Viswanath et al. propose community detection as a solution to the classic statement of Sybil defense which requires separating all (most) honest users from all (most) Sybil users. We believe that the path forward is to target a weaker notion of local Sybil defense in which we attempt to accept as many honest nodes as possible while rejecting all (most) Sybil users. Note that the difference between the two goals is subtle. The traditional Sybil defense goal is only as resilient as the global conductance—i.e. the smallest conductance of any cut in the graph—while local Sybil defense can provide stronger protection to communities with higher local conductance.

The potential of community detection as a defensive mechanism against sybil attacks has been first observed by Viswanath et al. [21]. They note that at the core of all sybil-* protocols lies a community detection problem: the goal of each honest user is to create exactly two communities, one that includes almost all honest nodes, the other that is overwhelmingly comprised of sybil nodes. If sybil-defense reduces to community detection, then it becomes possible to apply to sybil defense the large body of research on community detection. Indeed, they conclude that off-the-shelf community detection algorithms, such as the one by Mislove [15] that they study extensively, are likely to lead to effective defenses. We find otherwise.

Not all community detection is created equal. Mislove’s algorithm builds communities greedily. Starting from a vertex $v$, the algorithm computes a set by adding to the current set the vertex (connected to the current set) that increases the conductance the most. If no neighboring vertex increases the conductance, then the algorithm stops. $^3$

$^3$We note that the original proposal for Mislove’s algo-

The potential of community detection as a defensive mechanism against sybil attacks has been first observed by Viswanath et al. [21]. They note that at the core of all sybil-* protocols lies a community detection problem: the goal of each honest user is to create exactly two communities, one that includes almost all honest nodes, the other that is overwhelmingly comprised of sybil nodes. If sybil-defense reduces to community detection, then it becomes possible to apply to sybil defense the large body of research on community detection. Indeed, they conclude that off-the-shelf community detection algorithms, such as the one by Mislove [15] that they study extensively, are likely to lead to effective defenses. We find otherwise.

Not all community detection is created equal. Mislove’s algorithm builds communities greedily. Starting from a vertex $v$, the algorithm computes a set by adding to the current set the vertex (connected to the current set) that increases the conductance the most. If no neighboring vertex increases the conductance, then the algorithm stops. $^3$

$^3$We note that the original proposal for Mislove’s algo-

In the following, we show how Mislove’s’ greedy algorithm can be tricked into labeling an arbitrary number of Sybil identities as honest using only two attack edges. Let $v$ be an honest node, the victim, such that it has no neighbor of degree 2 or 3. Assume also without loss of generality and for clarity of exposition that $v$ has no neighbor of degree 1. We create the sybil region with nodes $s_0, s_1, \ldots, s_n$ as follows:

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

Figure 4 illustrates the attack. At the beginning, the best choice for the algorithm is to pick $s_0$. This is because it is the lowest degree node and hence gives the minimal conductance. The numerator of the conductance is increased by only 2 cutting edges while the denominator is incremented of 3. After this fatal error the best node to add becomes $s_1$, which raises the numerator again by 2 and the denominator by 4. This continues admitting the whole sequence $s_2, \ldots, s_n$. Thus Mislove’s algorithm will deterministically admit every node in the sybil region, regardless of its size, with only two attack edges.

Notice that this is exactly the situation that sybil-star is good at: the sybil region is separated from the honest one by a very sparse cut. Although the attack is guaranteed to succeed and cripple Mislove’s algorithm entirely, Figure 5 graphically shows the outcome of the attack precisely because of its greedy nature. We note that the sybil-* approaches identify a minimal conductance when they begin accepting the Sybil region indicating the discovery of the small cut separating the Sybil and
Figure 5: Performance of Mislove’s algorithm, SybilGuard, SybilLimit, and Gatekeeper when the WikiTalk graph is attacked with the two-edge attack. In all figures the x axis corresponds to the node ranking produced by the protocol. (a) The y axis corresponds to the precision of the first x nodes in the ranking. (b) The y axis corresponds to the conductance of the cut between the first x nodes in the ranking and the rest of the graph.

honest regions. Mislove’s algorithm steps into the Sybil region immediately and fully explores the Sybil region before transitioning to accept the rest of the graph.

These results suggest that, although community detection can be a useful notion when dealing with sybil attacks, we cannot just blindly use any off-the-shelf algorithm—an important aspect of Sybil defense that is not inherent to community detection is bounding the number of accepted Sybil users.

A blueprint for local sybil defense Although community detection is a well-developed field, there is still no agreed upon formal mathematical characterization of what actually specifies a community. From this perspective, it should perhaps in retrospect not be surprising that different algorithms may perform differently in terms of sybil defense, depending on the criteria they use to define a community.

We believe that Local Sybil defense requires community detection algorithms that identify communities with two important properties:

1. The community must be easy to split away from the rest of the graph (i.e. be separated from the rest of the graph by a small cut)
2. The community must be difficult to split itself (i.e. have no small internal cut).

In the next section we discuss an algorithm, LPR, which has these properties.

6. FROM GLOBAL TO LOCAL DEFENSE

We have seen some of the shortcomings of state-of-the-art and discussed some fundamental limitation of an approach based on global properties. Here we propose LPR (Local Page Rank [18]) as a possible step forward. In view of the fundamental limitations of social defense that we discussed with respect to the possibility of attaining Doucer’s “gold standard”, we revise our goal as follows. Rather than trying to classify each and every node as either honest or sybil, let us try to white-list a set of trusted nodes from the point of view of a given node.

After all a sybil region and a distant community might look the same for very good reasons, but should appear as very different from one own’s community. The problem here is that “community” is a compelling but vague notion, without a precise mathematical definition. Pragmatically, given a node u, as also suggested in [21], we will rank the nodes in terms of “trustworthiness”, leaving to the user to decide the cut-off point. The ranking however will have very strong properties that will assist us in making the right choice.

LPR has been used with good success to combat spam [24]. Spam-detection is a problem similar to sybil defense and it is in somewhat surprising that this connection has not been observed before. Similar to our case, several proposed solutions are based on the fact or rather, the hope, that spam-farms constitute a topological anomaly in the structure of the web. Our motivation however really comes from the fact that LPR possesses certain mathematical properties that are very useful in our context. Before discussing them let us briefly review the basic definition of LPR. LPR is the stationary distribution of the following random walk: starting from a vertex u, a pebble walks about the graph as follows. From the current position the pebble is moved back to u with probability \( \alpha \), otherwise, with probability \( 1 - \alpha \) it moves to one neighbor of the current position chosen uniformly at random. More formally, if \( x \) is the current node and \( N_x \)
is its set of neighbours, then the next position is $u$ with probability $\alpha$ and it is $y \in N_x$ with probability $\frac{1-\alpha}{\deg(x)}$. It is well-known that this random walk converges to a unique stationary distribution. Such a distribution assigns to every vertex $x$ in the graph a probability $p_x$. Note that this distribution depends on two parameters, $\alpha$ and the starting point $u$. This distribution enjoys several mathematical properties, that can roughly be described as follows (see for instance [18] for a formal and precise discussion): Given a node $x$ in a network $G = (V, E)$,

1. LPR computes a set $S_x$ whose (local) conductance is a good approximation of the conductance of the graph, which is a global minimum. Thus, LPR tends to compute a set of low conductance.

2. The sub-network induced by the set of vertices $S_x$ has high conductance.

3. The stationary distribution that LPR computes can be seen as a way to assign “trust” to the vertices of the network. The total trust of the set $V - S_x$ (the complement of $S_x$ where we hope to have confined the sybil nodes) is certified to be bounded from above by the conductance of $S_x$ (which, by the 1st item, we know to be low).

Previous sybil defense algorithms were able to guarantee the first property, but the combination of all three is more effective. Property 1 says that the set $S_x$ and its complement are separated by a sparse cut, which intuitively should capture the edges between sybil and honest nodes. As we have seen, this is also one of the motivations of sybil-star. Our new Property 2 says that $S_x$ is a very tightly knit community within, and hence hard to penetrate by sybil intruders. Property 3 is very interesting. The probability distribution computed by LPR assigns to each node $u$ in the network a “trust” score $t_u$ (the total trust so distributed is normalized and sums up to 1). We use LPR to rank the nodes according to their trustworthiness, and declare the first positions of the ranking to be the set $S_x$ of trusted nodes. Intuitively, the aim of the adversary is to place as many sybil nodes as possible in the first positions of the ranking. However, in combination with the first two, Property 3 ensures that the adversary has limited total trust: if it decides to distribute it among many sybil nodes none of them will be ranked high. The other option would be to concentrate its limited trust budget on few nodes, which might make it into the first positions but will be very few. In other words, the adversary is at most able to insert a few Byzantine nodes amidst a large community of honest ones, a very useful property for a distributed system.

Altogether, these properties offer a principled solution to social defense from a local point of view. We now provide a thorough experimental assessment, showing that LPR offers a substantial improvement on the state of the art and is a promising point of departure for further research.

6.1 Sybil defense with Local Page Rank

Our primary goal in this evaluation is to observe the empirical performance of Local Page Rank and compare its performance with existing social defense protocols. As discussed, LPR depends on two parameters, the jump-back probability $\alpha$ and the starting point $u$, and computes a stationary distribution $\{p_x, x \in V(G)\}$. These probabilities are used to rank the vertices with respect to $u$ as follows: nodes are sorted according to

$$\frac{p_x}{\deg(x)}$$

(rather than simply by $p_x$). This normalizes with respect to high degree nodes, that are more likely to be visited
often by virtue of their high degree.

We compare the performance of Local Page Rank Sybil-Guard, SybilLimit, Gatekeeper, and Mislove’s algorithm against attacked graphs from Table 2. We do not preprocess the data sets prior to performing the attack or evaluating the results. We consider attacks with probability of attack edge acceptance \( p = 0.01 \) and \( p = 0.10 \); the former represents a weak attack that is nominally covered by the strong mathematical guarantees of the sybil-* family of protocols while the latter is definitively outside the protective limits of those bounds.

We observe that in both attacked datasets, Local Page Rank performs almost perfectly. The sybil-* protocols, especially SybilLimit and SybilGuard, suffer from the lack of preprocessing, though they remain above 75% precision when \( k = |\text{honest nodes}| \) have been ordered. Mislove’s algorithm performs surprisingly poorly on both graphs. We observed the same performance from Local Page Rank for all data sets when we set \( p = 0.01 \). These results indicate that Local Page Rank is a good option for sybil defense and that it’s performance is not fundamentally reliant on preprocessing to ensure proper graph structure.

**Massive attacks.** We now turn our attention to stronger attacks where \( p = 0.10 \). It is important to note that attacks of this strength employ at least \( \frac{n}{\log n} \) attack edges on our targeted graphs. In the case of AstroPH and HE Physics, there are more attack edges than honest nodes! Figure 7 depicts the results of the five sybil defense algorithms for every data set under this strong attack. Unlike the attacks where \( p = 0.01 \), under these attacks no protocol performs exceptionally well (i.e. near perfect recall and precision).

The sybil-* protocols converge to 50% precision by the time they have accepted half of the graph, indicating that they are accepting nodes randomly from that point on. This indicates that the resulting attack graphs are relatively fast mixing.

Mislove’s algorithm is alternately good or bad—it regularly performs worse than random, primarily accepting nodes from the Sybil community as evidenced by a precision under 50%; it is also the only protocol to have higher precision than LPR on any graph (HE Physics). We attribute the wide variation in the performance of Mislove’s algorithm to its greedy nature. We observe that it very rarely performs randomly, indicating that it does a reasonable job of identifying a community, just not always the right one!

LPR consistently performs better than the sybil-* protocols and provides at least 70% precision at the midway point of the rankings for every graph except HE Physics. Even when performing poorly at that point, LPR has the highest precision early on (> 80% until the 25% point in the rankings) indicating that it is suitable for identifying strong local communities.

**Two-edge attack** We briefly revisit the attack presented in the previous section and observe that Local Page Rank is not vulnerable to it (see Figure 8). This result is not surprising as the two-edge attack was designed to leverage the greedy nature of Mislove’s algorithm. Local Page Rank avoids the trap because the combination of Property 3 and the small cut to the chain of sybil nodes ensures that the sybil portion of the graph receives a very small amount of trust.

**Pyrrhic victories.** The LPR precision lines in Figures 7(g) and 7(f) exhibit a peculiar behavior: towards the beginning of the rankings the precision dips down below 95% before recovering and stabilizing at a higher value. The dip begins around node 100 in every graph—coinciding with the size of the tight communities observed by Leskovec et al. \[11\]. We consistently observe this behavior (though to a lesser extent) on graphs subjected to the \( p = 0.01 \) attack. While this behavior is initially surprising, it is actually indicative of one of the fundamental strengths of LPR.

Recall that the distribution of trust in Local Page Rank is bounded by the size of the cut between the verifier and the community on the far side of the cut. Once Local Page Rank finishes ranking the nodes in the initial tight community, it must step into another community which is nearby. The sybil region may be nearby, but the trust available to that region is bounded by the cut that separates the verifier from the sybil region. The attacker can move some nodes towards the top of the rankings, but doing so dramatically reduces the trust available to other nodes in the region.
Figure 7: Precision versus node ranking for Local Page Rank, SybilGuard, SybilLimit, Gatekeeper, and Mislave’s algorithm when run on the graphs in our test suite with $p = 0.10$. 
7. CONCLUDING DISCUSSION

We believe that this paper contributes to the on-going debate surrounding the possibility of social defense in several ways. First, it highlights the fact that it is still very much a debate. Several interesting insights are already there but a conclusive answer appears to be quite elusive. In this paper we have shown that it is possible to gain some insight into the problem thanks to impossibility or, rather, indistinguishability results. If the adversary can tamper with the network by leaving certain observables well within the accepted range we cannot rely on such observables. The results we presented are very simple but they have the merit of pointing to what we believe is a fruitful line of research that will lead to greater sophistication and interesting insights. We have also shifted the focus from the absolute guarantees of Doucer’s gold standard to a more relativistic and, we believe, realistic point of view. A local approach gives a more realistic hope to fulfill the original program, namely that of developing sybil defense based on social properties. Although our comparative analysis shows the good behaviour of LPR we feel that we are just scratching the surface. The suggestion to consider community detection remains valid, but should be refined and better understood. After all, there is no (nor can there be) a precise definition of “community” and furthermore community detection algorithms are very often heuristic in nature and devoid of a proper understanding. Here a comparative analysis of existing community detection algorithms could be of value.

8. REFERENCES