Sybil Detection Using Latent Network Structure

GRANT SCHONEBECK, University of Michigan
AARON SNOOK, University of Michigan
FANG-YI YU, University of Michigan

Sybil attacks, in which an adversary creates a large number of identities, present a formidable problem for the robustness of recommendation systems. One promising method of sybil detection is to use data from social network ties to implicitly infer trust.

Previous work along this dimension typically a) assumes that it is difficult/costly for an adversary to create edges to honest nodes in the network; and b) limits the amount of damage done per such edge, using conductance-based methods. However, these methods fail to detect a simple class of sybil attacks which have been identified in online systems. Indeed, conductance-based methods seem inherently unable to do so, as they are based on the assumption that creating many edges to honest nodes is difficult, which seems to fail in real-world settings.

We create a sybil defense system that accounts for the adversary's ability to launch such attacks yet provably withstands them by:

(1) Not assuming any restriction on the number of edges an adversary can form, but instead making a much weaker assumption that creating edges from sybils to most honest nodes is difficult, yet allowing that the remaining nodes can be freely connected to.
(2) Relaxing the goal from classifying all nodes as honest or sybil to the goal of classifying the "core" nodes of the network as honest; and classifying no sybil nodes as honest.
(3) Exploiting a new, for sybil detection, social network property, namely, that nodes can be embedded in low-dimensional spaces.

Additional Key Words and Phrases: Sybil attack; latent social network; complex contagion

1. INTRODUCTION

The creation of multiple false identities, so-called sybil attacks [Douceur 2002], can enable actors undo influence in recommendation systems or other algorithms that harness user-generated data [Mobasher et al. 2006]. Controlling even a small portion of the alleged user-base can enable nefarious actors to hide their ill-gotten influence over recommendation systems [Yu et al. 2009]. Such recommendation systems might be used to classify spam, recommend products, or filter user-generated content (e.g. on an online-social networking site). Due to society's increasing reliance on the results of harnessing user-generated content/feedback (e.g. "big data"), guarding the veracity of the results will become increasingly important. Manipulation can have economically important (such as product recommendation) and politically important (as a public

The first author was support by NSF Career Award #1452915, the second by a Google Faculty Award, and the third by NSF Award #1535912 and a Facebook Faculty Award.

Authors addresses: G. Schoenebeck, A. Snook, F. Yu, Computer Science & Engineering Department, Univ. of Michigan; email: {schoeneb, asnook, fayu}@umich.edu.

EC'16, July 24–28, 2016, Maastricht, The Netherlands. ACM 978-1-4503-3936-0/16/07 ...$15.00.

http://dx.doi.org/10.1145/2940716.2940747

show of support) outcomes which provides rational actors incentives to manipulate outcomes to match their desires.

This has been recognized as a problem and addressed in the literature via a variety of methods (see Section 1.2). This paper focuses on a particularly promising method of using network ties to (implicitly) infer trust.

The models of prior work tend to restrict the adversary by making an ***edge-limiting*** assumption: the number of ties that the adversary can forge between sybils and honest nodes is restricted [Danezis and Mittal 2009; Tran et al. 2011; Wei et al. 2012; Yu et al. 2008, 2006].

Armed with the edge-limiting assumption and additionally assuming that the honest nodes of a network are "well-connected," these works show that one of two outcomes occurs: A) The adversary does not create many sybils; B) The adversary creates many sybils, but there is a detectable "sparse cut" in the graph. This sparse cut is caused by the assumption that there are few edges between the many sybil nodes and the honest nodes. Moreover, it is unique due to the assumption that the honest nodes are well-connected.

Thus, even if a powerful adversary can create many sybils, and moreover, endow them with high degree by connecting them with each other, the adversary cannot well integrate the sybils back into the rest of the network due to the limited number of ties that the adversary can forge between sybils and honest nodes.

While this defence does indeed (provable) protect against certain types of sybil attacks, the edge-limiting assumption seems to be too strong in practice [Alvisi et al. 2013]. Indeed Yang et al. [2014] recently showed evidence that in the RenRen social network, sybil attacks did not look like those that the prior work was anticipating, but instead were characterized by isolated sybils connected by many edges to honest nodes. We call these ***periphery attacks*** for reasons that will be made clear shortly. In periphery attacks, the number of sybils is only a fraction of the number of edges, yet Yang et al. [2014] found many sybil nodes in such an attack pattern. As such, these attacks violate the edge-limiting assumption; so the guarantees of the conductance-based sybil defences appear not to apply. Indeed Alvisi et al. [2013] showed via simulation on a real network, that the conductance-based defences do a poor job defending against such attacks.

Such attacks seem difficult to attenuate, in particular because often time the majority of nodes in a social network have a similar appearance. For example Leskovec et al. [2009] showed that networks have a "core/periphery" structure, with many nodes on the periphery poorly connected to the core of the network, which was difficult to partition. Additionally, Yardi et al. [2009] showed that in Twitter, the majority of nodes in Twitter only had a few friends, and that the spammers looked like-wise. Alvisi et al. [2013] looked into a collection of network topology properties and showed that the only one that was useful to sybil detection is conductance, which failed in thwarting periphery attacks.

1.1. Our Contribution

We create a framework that accounts for the adversary's ability to launch periphery attacks. Additionally, we create a network topology based sybil defense system that both accounts for and provably withstands periphery attacks. Our work builds upon and advances prior work in three main ways:

(1) We replace the edge-limiting assumption with a new assumption: A random fraction of the honest nodes are ***compromisable*** and can easily be tricked into connecting with sybil nodes; but the remainder of the honest nodes are ***trustworthy*** and will refuse connections from sybils. With such an assumption, periphery at-

tacks are easy for an adversary to launch. The adversary can test which nodes are gullible, and then connect to them at will with his sybil network.

(2) We relax the goal from classifying all nodes as honest or sybil to the goal of classifying the "core" nodes of the network as honest; and classifying no sybil nodes as honest. Our model acknowledges the difficulty of differentiating between the "periphery" nodes of the honest network and nodes that are part of a coordinated periphery sybil attack. Indeed this seems impossible to do with only information about network topology.

(3) We identify a new network property namely, that they can be embedded in low-dimensional spaces as useful to detecting sybils. For a sybil to "blend in" with the core of the topology structure of a network it is not enough that he has many ties; rather the sybil needs a large number of ties amongst other nodes that are "close" in the network. A sybil that connects to random nodes, will not have a "location" in the network the way an honest node might.

The zero false positives is important because even a few sybils can distort recommendations [Yu et al. 2009].

For many applications, like learning algorithms, or implicit community voting algorithms, having white-listed nodes is enough [Alvisi et al. 2013]. The system needs a representative sample of nodes. If the nodes on the periphery are not counted, then, as long as the nodes in "core" are sufficiently numerous, the system can succeed. For other applications (e.g. spam detection), such a classification might not be enough. In such settings, other tools must be used (e.g. user feed-back on spam; setting participation limitations for new nodes, etc).

While new to sybil detection literature, our model is well grounded in the social network literature. We will give additional theoretical justification in Section 3.2 after we define the specific network model. Finally, in Section 6 we run experiments on a variety of data sets, verifying that our assumptions about network structure hold in practice.

1.2. Related Work

Well-mixed networks. A growing number of works look to using network topology to aid in sybil detection.

Yu et al create SybilGuard [Yu et al. 2006] and SybilLimit [Yu et al. 2008], which use a random walk technique to bound the number of sybils that an adversary can produce for each edge that they can produce to honest nodes. This bound is $O(\sqrt{n}\log(n))$ for SybilGuard and was improved to $O(\log(n))$ in SybilLimit. However, in our setting where we do not restrict the number of edges that sybils can make to honest but gullible nodes, these guarantees are empty.

These works are typically called "conductance-based" and require an assumption that the network of honest nodes is well-mixing (and thus has high conductance). The intuition is that if there are many sybil nodes, but not many edges between the sybils and the honest nodes, then these algorithms will find a sparse cut. The well-mixing assumption is required to ensure that this sparse cut is unique.

Since these original works, several other works have made improvements along certain dimensions. Danezis and Mittal [2009] create SybilInfer which, instead of classifying nodes as safe or unsafe, using Bayesian reasoning, outputs confidence. Unlike aforementioned conductance-based work SybilInfer is a centralized algorithm. They point out that the run times of the prior, distributed works are very slow because they detect one sybil at a time and show that SybilInfer scales better. Likewise, Wei et al. [2012] propose SybilDefender which uses random walks, but is centralized and has improved scaling properties. They also suggest looking at tie strength as a method for improving results. Tran et al. [2011] propose Gatekeeper which achieves the same worst-case bound as SybilLimit, but improves upon it when the number of honest-sybil edges is very small.

Clustered Honest Networks. All of these works must assume that the network among honest nodes is well-mixing. The SybilLimit [Yu et al. 2008] paper provides some empirical evidence for this, but the claim is generally disputed. For example, Viswanath et al. [2011] analyze the state of current network-based Sybil defenses, showing that they rely on local community structure, and have trouble when their are cuts in the honest networks because they have difficulty distinguishing between the natural partitions in the network of the honest nodes, and the sparse cuts between the sybils and honest nodes. They propose borrowing techniques from the community detection literature.

Alvisi et al. [2013] also believe that the network will be too fragmented to employ the previous techniques, and show rigorous theoretical bounds to substantiate this claim. Without the "well-mixing" assumption, they fear the problem may be intractable as distinguishing between honest and sybil communities seems impossible. For example, consider the extreme case where all communities, both sybil and honest, are small and disjoint. Instead of sybil detection, they suggested "personalized white-lists". They point out that there is no need to distinguish between sybil and honest communities as long as you use the recommendation of each community for the nodes in it. A drawback of this is that if some communities are small, there may not be enough data to provide optimal recommendations. Like Alvisi et al, this work provides a white-listing strategy. However, we provide a global (not local) white list, and the honest nodes our model cannot classify are nodes on the periphery that belong to no community.

Cai and Jermaine [2011] also address the problem of potential community structure within the honest nodes. Their algorithm first partitions the network into disjoint communities, and then tries to ferret out the honest communities from the sybil communities by embedding them into a low dimensional space. They argue that the sybil communities will be on the periphery of this latent community graph. To get this result, their model assumes 1) the network of honest and sybil nodes partitions into well-structured and detectable communities, 2) that honest nodes connect to nodes in other communities according to a latent network of communities, and 3) that some communities are easy for sybils to attach to, while other communities are difficult for sybils to attach to. Our work differs in several ways. Most fundamentally, their algorithm does not guard against periphery attacks. In fact, their model does not allow periphery attacks because they make a necessary (in their setting) edge-limiting assumption. Moreover, they use machine learning techniques and thus do not obtain rigorous security results. Finally, our network models differ: our model of latent structure applies to the nodes and not communities; and in our model which nodes are vulnerable is decided at the node level rather than the community level.

Other Strategies. An increasing sequence of works looks at information beyond the social graph such as users' click-stream data [Wang et al. 2013]; entry and exit times [Noh et al. 2014], number of rejected friend requests [Alvisi et al. 2013], etc [Yang et al. 2014]. It is clear that they currently provide large practical benefits [Yang et al. 2014]. Moreover, they can be usefully combined with network topology based techniques [Alvisi et al. 2013]. Thus it seems like this is a useful orthogonal direction to pursue in ensuring the validity of recommendations. However, a key disadvantage of many of these techniques is that they rely on an uninformed adversary, that does not understand the behavior of honest nodes well enough to mimic them. Thus, their usefulness may wane as they are increasingly deployed and understood.

Viswanath et al. [2015] suggests detection can be conducted by using individual's temporal behaviour statistics which would be encoded in the time stamp of individual's reputation score. The goal here is to severely delay the potential effect of sybils rather than to eliminate it outright.

Another approach is to integrate sybil detection together with opinion aggregation (e.g. SumUp in Tran et al. [2009]). A key advantage here is that the sybil nodes do not have to be completely eliminated; but instead can be "down-weighted". However, a disadvantage of such approaches is that if they depend too sharply on the specific aggregation method, they loose some generality.

Another, somewhat disjoint, line of inquiry is for the setting where a central authority can restrict the entry of sybils through some verification or payment (e.g.Von Ahn et al. [2003] or Netflix). And defense in sensor networks [Lv et al. 2008; Yin and Madria 2007] where the solution concepts offered are light-weight cryptography (so that it can be efficiently executed).

2. PRELIMINARIES

A ***metric space*** is an ordered pair $M = (V, d)$ where V is a set and d is a metric on V mapping $V \times V$ to $\mathbb{R}^+$ such that for any $u, v, w \in V$, the following holds: $d(u, v) \geq 0$; $d(u, u) = 0$; $d(u, v) = d(v, u)$; and $d(u, v) \leq d(u, w) + d(w, v)$. We say that $M' = (V', d')$ is a ***metric subspace*** of $M = (V, d)$ if $V' \subseteq V$ and $d' = d|_{V' \times V'}$. We only consider finite metric spaces, i.e. $|V| \in \mathbb{N}$.

A ***metric graph*** $G = (V, E, d)$ is an undirected graph with distances defined between all pairs in V such that (V, d) is a metric space.

We define $B_M(u, r) = \{x \in V : d(u, x) < r\}$ as a ball with radius r centered at u in metric space M. We will often drop the subscript when it is clear from context, and denote $B(u, 1)$ by $B(u)$.

To capture the idea of low dimension in such a metric space, we use the notion of doubling dimension defined as follows: the ***doubling dimension*** $dim(M)$ of a metric space $M = (V, d)$ is the minimum k such that every ball of radius r is covered by 2^k balls of radius $R/2$; i.e. $\forall c \in V, r > 0, B(c, r) \subseteq V$, there exists $c_1, c_2, ..., c_m$ where $m \leq 2^k$ such that $B(c, r) \subseteq \bigcup_i B(c_i, r/2)$.

The doubling dimension is a very general definition of dimension. When it is applied to Euclidean vector spaces, it recovers the usual definition of dimension, but it also can apply to arbitrary metric spaces. Additionally, note that all finite metric spaces have finite doubling dimensions.

We define the ***neighbors*** of u in metric graph (V, E, d) to be $N(u) = \{v : (u, v) \in E\}$, and the ***core neighbors*** of u to be $CN(u) = B(u) \cap N(u)$, i.e the neighbors of u at distance at most 1.

3. SYBIL DETECTION FRAMEWORK

3.1. Metric Space Properties

We first define some properties of a metric space $M = (V, d)$ which we will make use of throughout.

Definition 3.1. The ***density*** of a metric space is $den(M) = \min_{u \in V} |B_u|$ which is the minimum cardinality of a unit ball.

Definition 3.2. We say that U is an r***-code*** of a metric space $M = (V, d)$ if $U \subseteq V$ and $\forall u, v \in U, d(u, v) > r$ and $V \subseteq \bigcup_{u \in U} B(u, r)$. That is U is a maximal set of points of distance strictly more than r from each other.

Definition 3.3. We define the ***volume*** of a metric space $M = (V, d)$ to be $vol(M) = \max\{|U| : U \text{ is a 2-code of } M\}$.

We show a natural relation between the density, the volume, and the cardinality of a metric space.

LEMMA 3.4. *Let $M = (V, d)$ be a metric space with density $den(M)$ and volume $vol(M)$. Then*

$$den(M) \cdot vol(M) \leq |V|.$$

PROOF. Let Y be a 2-code of M such that $|Y| = vol(M)$. On the one hand we have that

$$den(M) \cdot vol(M) \leq \sum_{y \in Y} |B(y)|$$

because for any $v \in V$, $den(M) \leq |B(y)|$ (by Definition 3.1) and $vol(M) = |Y|$ (by Definition 3.3).

On the other hand, we have that

$$\sum_{y \in Y} |B(y)| = | \bigcup_{y \in Y} B(y)| \leq |V|$$

because the $B(y)$ are disjoint—recall that for all $x, y \in Y$ we have $d(x, y) > 2$—and $\bigcup_{y \in Y} B(y) \subseteq V$. □

Here we provide an efficient algorithm to compute an approximation of the largest 2-code.

LEMMA 3.5. *Let $M = (V, d)$ be a metric space and $dim(M) = k$, then there exists a polynomial algorithm f, such that $f(M)$ is a 2-code and $\frac{vol(M)}{4^k} \leq |f(M)| \leq vol(M)$.*

PROOF. Let Y be the maximum 2-code of M, then by definition $|Y| = vol(M)$. The algorithm f iteratively inserts a node x into X, and removes all the nodes in $B(x, 2)$. Therefore each pair in X has distance more than 2, and $|X| \leq |Y| = vol(M)$ by definition.

On the other hand, consider a 1-code Z of metric space M, because $\forall u, v \in Y, d(u, v) > 2$, every unit ball of Z contains at most one $y \in Y$. Thus

$$|Y| \leq |Z|.$$

Moreover because $dim(M) = k$, $B(x, 2)$ can be covered by 4^k $\frac{1}{2}$ balls, and each $\frac{1}{2}$ ball can contains at most 1 element in Z. Thus

$$|Z| \leq 4^k |X|.$$

Putting this together $|X| \leq |Y| = vol(M) \leq Z \leq 4^k |X|$ which yields the lemma. □

Definition 3.6. Given a metric space $M = (V, d)$, we define a graph $H_r(M) = (V, E)$ where $(u, v) \in E$ if $d(u, v) \leq r$.

Definition 3.7. If $H_1(M)$ is connected, we say a metric space M is ***hyper-connected***.

This characterizes the metric space as “well connected” so that for all pairs of nodes there exists a sequence of points such that the distance between each pair of consecutive nodes is less than 1.

Definition 3.8. We say that $\hat{M} = (\hat{V}, \hat{d})$ is a ***core space*** with density Δ of a metric space $M = (V, d)$ if $\hat{M}$ is a submetric of M; density $\Delta = \min_{v \in \hat{V}} |B_M(v)|$; and $H_1(\hat{M})$ is connected.

This idea of a core space is important, because we only hope to classify nodes in the "core" of the network, not those in the periphery. This is a somewhat connected region with density above some threshold.

3.2. Network of Honest Nodes

In this section we both highlight exactly what we require of honest networks and provide motivation for this model.

We will consider metric graphs that are generated on top of a metric $M = (V, d)$ on n points. We would like that these points a) have doubling dimension bounded by some parameter k; and b) have a "large" core space $\hat{M}$ with density Δ where Δ is again a parameter.

The edges of the graph are generated by including each possible edge (u, v) where $d(u, v) \leq 3$, with probability ρ. Any additional edges may then be added to the graph after the outcomes of these random edges are realized.

Recapping, the important parameters are n, the number of nodes; k the doubling dimension; Δ the density of the core; and ρ, the minimum probability that edges appear between nodes close in the metric.

We think that this is a rather general model that is well-justified. First, the assumption that nodes are embedded in a low-dimensional space where nearby nodes are connected is implicit in many well-regarded network models. For example, in the Watts-Strogatz model [Watts and Strogatz 1998] nodes are arranged on ring (which is just a one-dimensional lattice) and any two nodes within some distance d on the ring, are connected via an edge with some probability that is a parameter of the model. Similarly, Kleinberg's Small World Model [Kleinberg 2000] has the nodes embedded into a low dimensional lattice structure where nodes are connected to neighbors. Additionally, Kumar et al. [2006] allows an arbitrary metric space with low doubling dimension and requires an additional property which is similar to our core space requirement. Though the latent space Abraham et al. [2013] considers is not necessary a metric space, our method can be easily applied to their model, because once having the distance function of all categories, we can removed individuals which fail to have enough common neighbors in all categories.

A host of other works from the mathematical, computer science, sociology, and statistics communities have also mathematically modeled social networks as coming from a low-dimension latent space and use the guiding principal that nodes which are "closer" in the latent space are more likely to be attached [Abraham et al. 2013; Clauset et al. 2008; Fraigniaud et al. 2010; Handcock et al. 2007; Hoff et al. 2002; Kermarrec et al. 2011; Krivitsky et al. 2009; Raftery et al. 2012; Sarkar et al. 2011; Sarkar and Moore 2005].

The intuition behind these models is that the location of a node in a metric space encodes some key properties of the individual, e.g. geographic location, income, political beliefs on a spectrum, education level, etc; and that these attributes are sufficient so that when individuals are "close" in this space, they are likely (with probability ρ) to be friends. Notice that in most of the aforementioned models, nodes are *always* neighbors with the nearby nodes in the metric; where as we only require that nearby nodes are neighbors with some constant, non-zero probability.

Furthermore, there is evidence of the accuracy of such models [Adamic and Adar 2005; Backstrom, Sun, and Marlow Backstrom et al.; Butts 2003; Liben-Nowell et al. 2005; McFarland and Brown 1973; Mok et al. 2007]. In Section 6, we provide our own experimental result which confirms that, for the networks we look at, they can be fruitfully embedded in a low-dimensional latent space. An additional feature of our model is that additional edges may be added to the graph in any, even *adversarial*, manner.

Second, our model additionally requires that the nodes be sufficiently dense in the metric. Notice that most of the aforementioned models have the nodes spread out uniformly, so their are no sparse regions of the network. We additionally relax this assumption and only require that there is a "large" dense region. To a first approximation, this dense region is the area we will be able to white-list; while nodes in sparse regions may not be included in the white list. The necessity of dealing with sparse regions is empirically motivated by aforementioned findings of Leskovec et al. [2009], Alvisi et al. [2013], and Yardi et al. [2009] which all identify nodes on the periphery with low-degree and/or that can be disconnected from the network by only removing a few edges.

3.3. Detection Game

In this section we propose a formal model for sybil detection as a game with two agents: the *adversary* and the *distinguisher*.

The adversary will be given a metric graph G. We say that the nodes of G are the ***honest*** nodes. This set of honest nodes is partitioned into a set of ***compromisable*** nodes C that the adversary can attach to and a set of ***trustworthy*** nodes T that the adversary cannot attach to. The adversary must output a new metric graph G' which is the same as G except that the adversary can add up to Σ ***sybil*** nodes and any edges that it likes except those between trustworthy nodes and sybil nodes.

The distinguisher will then be given the adversary's output graph (as well as some parameters), and must create a white-list of as many nodes as possible without including any sybil nodes.

Definition 3.9. Let $A : (G, C, p, \rho, \Sigma) \to G'$ be a (possibly random) function where $G = (V, E, d)$ and $G' = (V', E', d')$ are metric graphs, $C \subseteq V$ is a set of "compromisable" nodes, p, ρ are real values between 0 and 1, and $\Sigma > 0$. We say that A is an ***adversary*** if for every input G, C, p, ρ, Σ:

(1) $|S| < \Sigma$ where $S = V' \setminus V$.
(2) The distance function d' is a metric that extends d to $V \cup S$.
(3) $E \subseteq E'$ but E' contains no edges from $V \backslash C$ to S. However E' may contain additional edges between V and itself, between S and itself, and between C and S.

Our definition limits the adversary in two keys ways: first, he can only introduce so many sybil nodes. Such a condition is necessary because otherwise the adversary could just create a completely new graph on a disjoint set of vertices which is identical to the original graph; no detection algorithm could distinguish the ordinal graph from the identical facsimile. Second, the adversary can only connect sybils to the original network via compromisable nodes. The intuition is that some set of nodes can by tricked or bribed into connecting with the sybils. The remaining vertices are more trustworthy, concerned, aware, and/or vigilant and are thus immune from the adversaries attempts to connect. This aligns with the observations of Yang et al. [2014] that software toolkits which facilitate the creating of sybil nodes for the Renren cite were available and would attempt to identify network nodes that would likely accept a sybil's tie request (e.g. nodes with extremely large degree).

Note especially that the adversary can also add ties between honest nodes. This is not meant to model that the adversary could or would actually compel honest nodes to add a tie (though it does capture this as well). Rather it is meant to model that, apart from the ties in the network that we assume to exist from the low-dimensional embedding (that are included in G and cannot be removed), the rest of the graph is adversarial bad. In actuality, we think that the graph on the honest nodes would come from nature. However, we do not wish to prescribe anything more about the honest graph other than that nodes which are "close" in the low-dimensional latent space are often connected; and may be connected in a way that is not helpful to the "distinguisher."

Definition 3.10. A ***distinguisher*** D is a (possibly random) function $D : (G', p, den(M), vol(M)) \rightarrow W$ where $G' = (V', E', d')$ is a metric graph, $p, den(M), vol(M)$ are real valued parameters, and $W \subseteq V'$.

Now we formally define a *detection game* on a metric space $M = (V, d)$.

Definition 3.11. We define a ***detection game*** Γ with input $(M, p, \rho, \Sigma, A, D)$ where M is a metric space, p, ρ are real values between 0 and 1, $\Sigma > 0$, A is an adversary, and D is a defender as follows,

(1) Based on $M = (V, d)$, a metric graph $G = (V, E, d)$ is instantiated where E is created by independently including each edges (u, v) with probability p if $d(u, v) < 3$, and otherwise with probability 0. [Note that in Step 3, the adversary can add *any* additional ties it likes between honest nodes in an attempt to thwart the distinguisher. At that point the adversary knows which nodes are trustworthy and compromisable, so the additional edges can depend on those labels.]
(2) We randomly partition V into two sets T (for trusthworthy) and C (for compromisable). Each agent $v \in V$ will, independently, be included in set C with probability ρ and in set T otherwise.
(3) The adversary A creates a new metric graph $G' = A(G, C, p, \rho, \Sigma)$.
(4) The distinguisher D outputs a list of nodes W with input $(G', p, den(M), vol(M))$
(5) If $W \subseteq V$ we say that the distinguisher ***succeeds with score*** $|W|$; otherwise, if $W \cap S \neq \emptyset$ we way that the distinguisher *fails*.

We note that we give the distinguisher help via the parameters $p, den(M), vol(M)$. In general, we do not feel this assumption is overly restrictive, as distinguisher could likely learn these over time.

We also node that the detection game maps onto our definition of honest networks in Section 3.2. In particular, this gives the adversely (perhaps unrealistic) power to manipulate the graph of honest nodes by adding additional edges between any pair of vertices even after the random edges have been realized and the compromisable nodes have been determined. However, this only makes our results stronger.

4. SYBIL DETECTION ALGORITHM

In this section present our main result by exhibiting a detection algorithm that provably works when the adversary is restricted to only using a fixed number of sybil nodes.

THEOREM 4.1. *Fix* $0 < \epsilon < \frac{1}{\sqrt{2}}$ *and let* $\Gamma(M, p, \rho, \Sigma, A, D)$ *be a detection game where* p, ρ *are probabilities such that* $\frac{1+\epsilon}{1-\epsilon}\rho < p$, $0 \leq \Sigma$, *and* M *is a metric space that has* n *nodes and doubling dimension* k *with core-space* $\hat{M} = (\hat{V}, \hat{d})$ *with density* Δ *with* $m = |\hat{V}|$. *Then if*

$$\Sigma < (1 - \epsilon)\frac{p}{2 \cdot 128^k} den(\hat{M}) \cdot vol(\hat{M}) - (1 + \epsilon)\rho n$$

there exists a detection algorithm D *such that for any adversary* A *the detection algorithm* D *will succeed with score at least* m *with probability*

$$1 - n^2 \exp(-\frac{\epsilon^2}{2} p\Delta) - n \exp(-\frac{\epsilon^2}{3}\rho\Delta) - \exp(-\frac{\epsilon^2}{3}\rho n).$$

Note that the size of the white-list is at least as large as the dense core of M. The parameters of the theorem can cover a variety of settings. For example, if $\Delta = \omega(\frac{\log(n)}{p\epsilon^2})$ and $\epsilon^2\rho = o(n/\log n)$, then the probability of error is negligible (less than the inverse of any polynomial).

To the end of proving Theorem 4.1, we propose the *detection algorithm* which is specified in Algorithm 1.

ALGORITHM 1: Detection algorithm

```
   Input: $G' = (V', E', d')$, $p$, and $\Delta, vol(\hat{M})$
   Output: $W$, denoting the white-listed nodes.
1  Find a 2-code $Y$ of $H_2(V', d')$ by the algorithm in Lemma 3.5.
2  Obtain $(V'', E'', d'')$ from $G'$ by iteratively finding nodes $u \in V'$ where
   $|N_{G'}(u) \cap B_{G'}(u, 2)| < (1-\epsilon)p \cdot \Delta$ and removing these nodes and all incident edges.
3  for $y \in Y$ do
4      $G_y(V_y, E_y, d_y) \leftarrow (V'', E'', d'')$
5      $W_y \leftarrow \emptyset$
6      $U_y \leftarrow \emptyset$
7      while $U_y = \emptyset$ and $|B_{G_y}(y)| \geq \Delta$ or $\exists v \in U_y$ such that $\exists u \in B_{G_y}(v) \setminus U_y$ where $|B_{G_y}(u)| \geq \Delta$
        do
8          if $U_y = \emptyset$ then
9              $u \leftarrow y$
10         else
11             Set $u$ to be some $u$ from Step 7
12         end
13         $U_y \leftarrow U_y \cup \{u\}$
14         for $v \in B_{G_y}(u, 2)$, and $v \notin W_y$ do
15             if $|N_{G_y}(v) \cap B_{G_y}(u)| > (1-\epsilon)p|B_{G_y}(u)|$ then
16                 $W_y = W_y \cup \{v\}$
17             else
18                 Remove $v$ and all its edges from $G_y$.
19             end
20         end
21     end
22 end
```

Before we dig into the proof we sketch the intuition behind the detection algorithm. Verification goes as follow: the algorithm pretends that there is no sybil node in the starting region $B(y)$ for some y from Step 3 and attempts to certify nodes $v \in B(y, 2)$ by checking whether they have many neighbors in $B(y)$. Then the algorithm moves to a different ***center*** u in Step 13 and verifies the region $B(u, 2)$. Doing this, it will iteratively remove the sybils on the boundary; allowing it to grow a white-listed region in the graph to cover the entire core.

The remaining difficulty is to find a good starting point y. In Step 3, we say $y \in V$ is a ***good starting point*** if $B(y) \cap S = \emptyset$ and $|B(y)| \geq \Delta$, and say $y \in V$ is a ***bad starting point*** if $B(y) \cap S \neq \emptyset$ and $|B(y)| \geq \Delta$. The main idea is that the adversary cannot corrupt every region of the graph with many nodes. Thus after Step 2 there will be many regions of the graph with no sybils. In Step 1, we get a maximal independent set corresponding to a 2-code of (V', d') which ensures that we are exploring many diverse regions of the network.

The proof can be separated into two parts:

(1) (completeness/soundness) If y from Step 3 is a good starting point, then with high probability, this algorithm will white-list every honest node in the core space and no sybil nodes will be white-listed;
(2) (majority) There are many $y \in Y$ that are good starting points, and not too many bad starting points.

We first prove three lemmas about structural properties of the network that occur with high probability. The first of these lemmas shows that if node v is near a node u with many nodes within unit distance, then node v has large degree. The second says that if node v has many nodes within unit distance, then v does not (fractionally) have too many compromisable nodes within unit distance. The third lemma bounds the total number of compromisable nodes.

We will then show that if these properties hold, then our detection algorithm succeeds. The proofs of the following 3 lemmas follow from a simply application of a Chernoff bound, and are deferred to the full version.

LEMMA 4.2. *Let $\Gamma(M, p, \rho, \Sigma, A, D)$ be a detection game, let $n = |M|$, and let $\Delta \in \mathbf{R}_{\geq 0}$. Then with probability $1 - n^2 \exp(-\frac{\epsilon^2}{2} p\Delta)$ for every $u, v \in M$ with $d(u, v) \leq 2$ and $|B_M(u)| \geq \Delta$, it is the case that $|N_G(v) \cap B_M(u)| \geq (1 - \epsilon)p|B_M(u)|$.*

LEMMA 4.3. *Let $\Gamma(M, p, \rho, \Sigma, A, D)$ be a detection game, let $n = |M|$, and let $\Delta \in \mathbf{R}_{\geq 0}$. Then with probability $1 - n \exp(-\frac{\epsilon^2}{3} \rho\Delta)$ for every $u \in M$ with $|B_M(u)| \geq \Delta$, it is the case that $|B_M(u) \cap C| \leq (1 + \epsilon)\rho|B_M(u)|$.*

LEMMA 4.4. *Let $\Gamma(M, p, \rho, \Sigma, A, D)$ be a detection game and let $n = |M|$, then with probability $1 - \exp(-\frac{\epsilon^2}{3} \rho n)$, $|C| < (1 + \epsilon)\rho n$.*

Now notice that by a union bound, the statements of Lemmas 4.2, 4.3, and 4.4 holds with probability $1 - n^2 \exp(-\frac{\epsilon^2}{2} p\Delta) - n \exp(-\frac{\epsilon^2}{3} \rho\Delta) - \exp(-\frac{\epsilon^2}{3} \rho n)$.

We now assume that all these statements hold, and show that when this is the case, our detection algorithm works. The next lemma shows that no honest node within unit distance of a node with high density is removed in Step 2.

LEMMA 4.5. *Let $\Gamma(M, p, \rho, \Sigma, A, D)$ be a detection game where D is our detection algorithm with inputs G', p, Δ, vol. Let $v \in V$ with $|B_G(v)| \geq \Delta$ then, assuming statement of Lemma 4.2 holds, after Step 2, $B_{G''}(u) \cap V = B_G(u) \cap V$.*

The proof is deferred to the full version, but follows from the idea that no node can be the first removed.

LEMMA 4.6. *Let M be a metric space and let $\hat{M}$ be a core space with density Δ. Let $\Gamma(M, p, \rho, \Sigma, A, D)$ be a detection game where D is our detection algorithm with inputs $G', p, \Delta, vol(\hat{M})$. Assume that the conditions in Lemmas 4.2 and Lemma 4.3 are true, and let y be a good starting point. Then the Detection algorithm will output $W_y \subseteq V$. Moreover, if $y \in \hat{M}$ then $\hat{V} \subseteq W_y \subseteq V$*

PROOF. We assume the statements of Lemma 4.2 and Lemma 4.3 and that y is a good starting point and then we will show that the following always hold:

(1) $V_y \cap V = V'' \cap V$,
(2) $W_y \cap S = \emptyset$,
(3) For all $u \in U_y$ and $u' \in B_{G_y}(u)$ where either $|B_{G_y}(u')| \geq \Delta$ or $|B_G(u')| \geq \Delta$, we have $B_{G_y}(u') = B_G(u') \subseteq W_y$.

If we prove this, then, by the second statement, we know that $W_y \subseteq V$. We must also show that if $y \in \hat{M}$ then $\hat{M} \subseteq W_y$. We show something stronger: each node in $\hat{M}$ is eventually included in U_y. This is a stronger statement because, by Statement 3, if $u \in U_y$, then $B_{G_y}(u) = B_G(u) \subseteq W_y$. Say that some node $w \in \hat{M}$ is never added to U_y. By the hyper-connection property of $\hat{M}$ we can create a spanning tree on the nodes of $H_1(\hat{M})$ rooted at y, and let w be a "closest" node to y (in the tree) that is not included and let v be its parent.

However, from the third statement above, we know $B_G(w) = B_{G_y}(w)$ because $d(w, v) \leq 1$, $v \in U_y$ and $|B_G(w)| \geq \Delta$. Thus w will also be processed as a center, and this is a contradiction.

We now show that the three properties always hold via induction on $|U_y|$. For $|U_y| = 0$, the first statement holds because at that point $V_y = V''$; while the second statement holds because $W_y = \emptyset$ and the third statement holds because $U_y = \emptyset$.

We now show the inductive step, that if the three statements hold when $|U_y| = k$, they will also hold when $|U_y| = k + 1$.

Lets say that u is the $k + 1$st node chosen for a center in Step 13. We know that $B_{G_y}(u) = B_G(u)$ either because $u = y$ and then it follows from the fact that y is a good starting point and Lemma 4.5, or because there must exist $w \in U_y$ such that $d(w, u) \leq 1$, and then it follows from the third assumption (note that $|B_{G_y}(u)| \geq \Delta$ because u was chosen to be a center).

Before processing center u, a node $v \in V_y(u) \cap V$ has two cases:

1) $v \notin B_{G_y}(u, 2)$ then v will certainly be in V_y after the process;

2) if $v \in B_{G_y}(u, 2)$, since Lemma 4.2 holds, we have $|N_G(v) \cap B_G(u)| \geq (1-\epsilon)p|B_G(u)|$. Because $B_{G_y}(u) = B_G(u)$, we have also have $|N_{G_y}(v) \cap B_{G_y}(u)| \geq (1-\epsilon)p|B_{G_y}(u)|$. Thus $v \in V_y$ holds after the process, and that proves $V_y \cap V = V'' \cap V$, and $B_{G_y}(u, 2) \subseteq W_y$.

On the other hand, let $s \in B_{G_y}(u, 2)$ be a sybil node. Then s can only connect to the compromised nodes in $B_{G_y}(u)$ because, by assumption, $B_{G_y}(u) = B_G(u)$, which contains no sybil nodes. Formally, we see:

$$\begin{aligned} |N_{G_y}(s) \cap B_{G_y}(u)| &\leq |C \cap B_{G_y}(u)| = |C \cap B_G(u)| && (1)\\ &< (1-\epsilon)\rho|B_G(u)| = (1-\epsilon)\rho|B_{G_y}(u)| && (2)\\ &\leq (1+\epsilon)p|B_{G_y}(u)|. && (3) \end{aligned}$$

The first equality is from the assumption that $B_{G_y}(u) = B_G(u)$, the second inequality is from the assumption of Lemma 4.3, and the final inequality is because $\frac{1+\epsilon}{1-\epsilon}\rho < p$. This proves $W_y \cap S = \emptyset$.

It remains to show that part 3) holds. We break the analysis into two cases by partitioning U_y into $U_y \setminus \{u\}$ and $\{u\}$.

First, let $v \in U_y \setminus \{u\}$ and let $u' \in B_{G_y}(v)$ where $|B_{G_y}(u')| \geq \Delta$ or $|B_G(u')| \geq \Delta$. Then, by the inductive hypothesis, after the time the v was processed, we had that $B_{G_y}(u') = B_G(u') \subseteq W_y$. No node in W_y is ever removed, so this still must be the case.

Second, let $u' \in B_{G_y}(u)$ with $|B_{G_y}(u')| \geq \Delta$ or $|B_G(u')| \geq \Delta$ and so that u' was not considered above. Then, we must show $B_{G_y}(u') = B_G(u') \subseteq W_y$.

Note that combining the facts that $B_{G_y}(u') \subseteq B_{G_y}(u, 2)$ and $B_{G_y}(u, 2) \subseteq W_y$ (argued above) we see, that $B_{G_y}(u') \subseteq W_y$. Using that $B_{G_y}(u') \subseteq W_y$ and $W_y \cap S = \emptyset$ we see that $B_{G_y}(u') \cap S = \emptyset$, which means that $B_{G_y}(u')$ has no sybils and so $B_{G_y}(u') \subseteq B_G(u')$. This additionally implies that $|B_G(u')| \geq \Delta$.

It remains to show that $B_G(u') \subseteq B_{G_y}(u')$. Intuitively, the one problem we could encounter is that some nodes of $B_G(u')$ might have been removed in Step 2. However, this does not happen. Rather $B_G(u') = B_{G''}(u') \cap V$ because $|B_G(u')| \geq \Delta$ and so by Lemma 4.5 $B_{G''}(u') \cap V = B_G(u')$. We use this to get:

$$B_G(u') = B_{G''}(u') \cap V = B_{G''}(u') \cap V'' \cap V \tag{4}$$
$$= B_{G''}(u') \cap V_y \cap V = B_{G_y}(u') \cap V \subseteq B_{G_y}(u') \tag{5}$$

The third equality is because $V'' \cap V = V_y \cap V$, as proved above.

Putting everything together we have $B_G(u') = B_{G_y}(u') \subseteq W_y$ and this concludes the proof of the lemma. □

LEMMA 4.7. *(Majority) Let $\Gamma(M, p, \rho, \Sigma, A, D)$ be a detection game and assume that the condition in Lemma 4.4 is true, and let Y be the 2-code D gets after step 1, then at most $\frac{|Y|}{2\cdot 4^k} \leq \frac{vol(\hat{M})}{2\cdot 4^k}$ points in Y are bad starting points.*

PROOF. Suppose the lemma is false. Then we consider the subset $Y' \subseteq Y$ such that every $y' \in Y'$ is a bad starting point and

$$|Y'| \geq \frac{|Y|}{2 \cdot 4^k}.$$

We consider some $X \subseteq Y'$ such that X is a 8-code for Y'. For each $x \in X$, $|B(x, 8) \cap Y'| \leq 8^k$ because $B(x, 8)$ can be covered by less than 8^k unit balls by definition of doubling dimension and each unit ball contains at most 1 element of Y'. The cardinality of 2-code Y is greater than $\frac{vol(\hat{M})}{4^k}$ by Lemma 3.5. Thus

$$|X| \geq \frac{|Y'|}{8^k} \geq \frac{|Y|}{2 \cdot 32^k} \geq \frac{vol(\hat{M})}{2 \cdot 128^k} \tag{6}$$

By the assumption of Y' every $x \in X \subseteq Y$ is a bad starting point which means $\forall x \in X, \exists s_x \in B(x)$ which is a sybil node, and since s_x survives after step 2, $|B(s_x, 2)| > (1-\epsilon)p \cdot \Delta$. Moreover, for all $x, z \in X$ and $x \neq z$, $d(s_x, s_z) \geq d(x, z) - d(x, s_x) - d(z, s_z) > 8 - 4 = 4$, $B(s_x, 2), B(s_z, 2)$ are disjoint. On one hand,

$$|C \cup S| \geq |\bigcup_x B(s_x, 2)| \geq |X|(1-\epsilon)p \cdot \Delta$$

Using (6) and the condition on S, we get

$$|C| \geq -|S| + (1-\epsilon)p\frac{vol(\hat{M})\Delta}{2 \cdot 128^k} > (1+\epsilon)\rho n$$

On the other hand, by Lemma 4.4, $|C| \leq (1+\epsilon)\rho n$, and so we get a contradiction. □

Now we can prove the Theorem 4.1

PROOF. First, we note that the statements of Lemmas 4.2, 4.3, and 4.4 hold with probability

$$1 - n^2 \exp(-\frac{\epsilon^2}{2}p\Delta) - n\exp(-\frac{\epsilon^2}{3}\rho\Delta) - \exp(-\frac{\epsilon^2}{3}\rho n) \tag{7}$$

In the case that y is a good starting point, we never add a sybil nodes to W_y by Lemma 4.6. By Lemma 4.7, there are only $\frac{|Y|}{2\cdot 4^k}$ bad starting points in Y. Thus no sybil meets the threshold in Step 7 to be included in W.

However, by Lemma 3.5, for any node v in the $\hat{M}$-core, there are $\frac{vol(\hat{M})}{4^k}$ start nodes in Y. Moreover, less than $\frac{vol(\hat{M})}{2\cdot 4^k}$ of them can be bad. Thus at least $\frac{vol(\hat{M})}{2\cdot 4^k}$ of them are good. By Lemma 4.6 for these y, $\hat{M} \subseteq W_y$ and thus, $\hat{M}$ will be included in W. □

5. SYBIL DETECTION WITH A TRUSTWORTHY SEED

We also consider the additional assumption that the distinguisher is given one trustworthy node as advice. In the full version, we show that we can obtain similar results as Theorem 4.1 but with no limit on the number of sybil nodes (nor the doubling dimension). However, to take advantage of this advice, we will lose a fair bit in the trade-off between the parameters of ρ and p—the fraction of nodes that are compromisable and the fraction of edges present between honest nodes which are close in the underlying latent space, respectively. Before we required that $\frac{1+\epsilon}{1-\epsilon}\rho < p$, but in this case, we will require $\frac{1+\epsilon}{1-\epsilon}\rho < p^3$.

6. EXPERIMENTS

We previously noted that our assumptions hold in many generative models, and general versions of these assumptions are typically assumed to be true. We conducted several experiments to evaluate and further study our specific assumptions in different online communities and social networks.

In our main theorems, we assume the social network G can be embedded into a low doubling dimension space $M(G, d)$; that a large fraction of nodes forms a *core space* with density Δ, that every node in the core space has at least Δ nodes whose distance is smaller than 3, and the edges between the node and nearby nodes form independently with probability p. Collectively, we refer to this set of assumptions our *low-dimension assumption*.

Note that because our low dimensional assumptions are stochastic in nature, we cannot exactly test them empirically. Instead we will ensure that a node connects to some p fraction of neighbors within distance 3 (rather than a random set of neighbors).

6.1. Dataset Description

Our experiments use all 4 social network data sets on Stanford Network Analysis Project with between $4,000$ and $100,000$ nodes: this includes networks collected from Facebook [McAuley and Leskovec 2012] and Twitter [McAuley and Leskovec 2012] as well as the Wikipedia voting network [Leskovec et al. 2010] and the Epinion network [Richardson et al. 2003].

The data sets are more completely described in the full version, but we summarize the basic statistics of the network data sets we used in Table I.

Social network	Facebook	Wiki-vote	Twitter	Epinion
Nodes	4,039	7,115	81,306	75,879
Edges	88,234	103,689	1,768,149	508,837
Average degree	21.85	14.57	21.75	6.70
Nodes in 6-core	3,478	3,343	58,787	13,911
Edges in 6-core	86,492	94,179	1279,919	303,324

Table I: Data set statistics

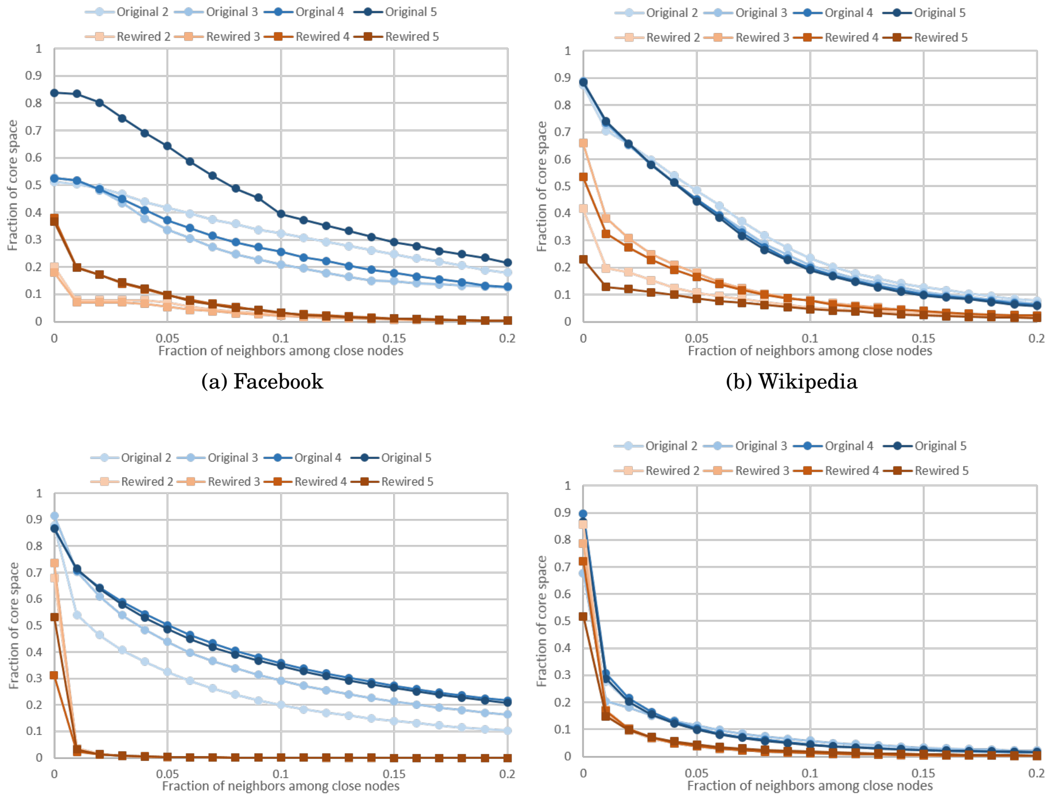
6.2. Implementation Details

To test the low-degree assumption on each network we used spectral embedding techniques to embed the 6-core of the graph into $\mathbb{R}^d$, and then measured the ***core-fraction*** of the resulting metric graph as follows: for given parameters r, p, and Δ, we first removed all nodes in the 6-core that either a) did not have Δ nodes within distance r, or b) were not neighbors with a p fraction of the nodes within distance r. We then created a graph of the remaining nodes by connecting those within distance r. We output the size of the largest component divided by the size of the 6-core, the core-fraction.

We additionally, randomly “rewired” the 6-core of each graph and again embedded that into $\mathbb{R}^d$, and then measured properties of the resulting metric graphs.

In the full version, we give additional details and explanations for the above procedures.

6.3. Experimental Results



(a) Facebook

(b) Wikipedia

(c) Twitter

(d) Epinion

Fig. 1: The relation between fraction of core space to graph under required fraction of neighbors among close nodes and in different dimensions.

The results with $\Delta = 10$, $0 < p < 0.2$, and $d = 2, 3, 4$, and 5 (recall d is the dimension of the embedding) are shown in Figure 1. We generally found that there is a large fraction of nodes in core space, with the Twitter and Facebook networks embedding more effectively than E-pinions or Wikipedia. This is promising because these data sets are the closest to traditional social networks.

Note that the charts only show the fraction of the 6-core in the core. Table I additionally shows the fraction of the nodes in the 6-core, which averages around 50% but varies greatly between datasets. For example, even when we required that a core node be connected to 20% of the close nodes in the Facebook data set, about 22% of the 6-core nodes remained in the core. Because in this dataset over 86% of the nodes are in the 6-core, this means that about 19% of the nodes are in the core. If we only require that a core node be connected to 10% of the close nodes, then the overall fraction of core nodes jumps to 34%.

The exception was the E-pinions network. In this network, even when we only required a core node be connected to 5% of the close nodes only 12% of the 6-core remained in the core. Also, unique to this network is that the 6-core only represented about 18% of the nodes. So at this point only 2% of the nodes are in the core. While we cannot definitively say, we postulate that one reason for this failure is the low average degree of the E-pinion network, which is less than half of any other network. Additionally, we note that even though the numbers are small, the faction of nodes in the core of the E-pinions network is still a factor of 10 greater than in the rewired E-pinions network.

The dimension for which we embedded a network seemed not to make a systematic difference, though it seemed like slightly larger dimensions were more effective in Twitter and Facebook.

In the rewiring setting, the experiments show that the embeddings of the rewired networks do not do as well placing neighbors close by. This indicates that the link structure in original networks contains features that the rewired networks do not. In particular, the core-fraction of the rewired networks when $p = 0.2$ was about 0.46% in Facebook, 2.2% in Wikipedia, 0.01% in Twitter, and 0.39% in E-pinion.

7. FUTURE WORK

Our work has several limitations that we hope can be addressed in future work. First, there exist cases where Lemma 4.3 fails for only a single node, yet the adversary can add arbitrarily many sybils. Ideally the performance would degrade more gracefully. Second, this fact implies that our techniques require $p\Delta$, the number of local ties, to grow with the natural log of the number of vertices, which, in practice, requires a fairly dense network. If $p\Delta$ is a small constant, then for large enough networks, Theorem 4.1 is vacuously true. Improvements by probabilistic analysis (e.g. chaining instead of union bound) seem limited. Rather, a new detection algorithm is required for this setting.

Additionally, we assume that the metric is known, but future work could relax this assumption. While several results (e.g. [Abraham et al. 2013; Backstrom, Sun, and Marlow Backstrom et al.; Clauset et al. 2008; Handcock et al. 2007; Hoff et al. 2002; Kermarrec et al. 2011; Krivitsky et al. 2009; Raftery et al. 2012; Sarkar et al. 2011; Sarkar and Moore 2005]) reconstruct latent spaces given a network, it is not clear if they can do this reliably in the presence of sybil nodes. However, the nodes may have attributes that can be leveraged to this end.

One direction for improvement in practice, is to attempt to detect which nodes are compromisable. This seems plausible since the adversary must already do this (or risk their sybils having many rejected tie requests—a tell-tail sign [Yang et al. 2014]).

Another direction for future work is to extend a core idea of this paper—that while sybil nodes can connect to many neighbors, they cannot necessary connect to the correct combination of neighbors—to other network models, for example networks with discrete community structures.

REFERENCES

Ittai Abraham, Shiri Chechik, David Kempe, and Aleksandrs Slivkins. 2013. Low-distortion Inference of Latent Similarities from a Multiplex Social Network. In *SODA 2013*. 1853–1883.

Lada Adamic and Eytan Adar. 2005. How to search a social network. *Social networks* 27, 3 (2005), 187–203.

L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi. 2013. SoK: The Evolution of Sybil Defense via Social Networks. In *Security and Privacy (SP), 2013 IEEE Symposium on*. 382–396.

Lars Backstrom, Eric Sun, and Cameron Marlow. Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *WWW 2010*. 61–70.

Carter T Butts. 2003. *Predictability of large-scale spatially embedded networks*.

Zhuhua Cai and Christopher Jermaine. 2011. The latent community model for detecting sybil attacks in social networks. In *VLDB*.

Aaron Clauset, Cristopher Moore, and Mark EJ Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 7191 (2008), 98–101.

George Danezis and Prateek Mittal. 2009. SybilInfer: Detecting Sybil Nodes using Social Networks.. In *NDSS*.

John R Douceur. 2002. The sybil attack. In *Peer-to-peer Systems*. 251–260.

Pierre Fraigniaud, Emmanuelle Lebhar, and Zvi Lotker. 2010. Recovering the Long-range Links in Augmented Graphs. *Theor. Comput. Sci.* 411, 14-15 (2010), 1613–1625.

Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. 2007. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170, 2 (2007), 301–354.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. 2002. Latent space approaches to social network analysis. *Journal of the american Statistical association* 97, 460 (2002), 1090–1098.

Anne-Marie Kermarrec, Vincent Leroy, and Gilles Trédan. 2011. Distributed social graph embedding. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 1209–1214.

Jon Kleinberg. 2000. The Small-world Phenomenon: An Algorithmic Perspective. In *STOC 2000*. 163–170.

Pavel N Krivitsky, Mark S Handcock, Adrian E Raftery, and Peter D Hoff. 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks* 31, 3 (2009), 204–213.

Ravi Kumar, David Liben-Nowell, and Andrew Tomkins. 2006. Navigating Low-Dimensional and Hierarchical Population Networks. In *Algorithms ESA 2006*. Lecture Notes in Computer Science, Vol. 4168. 480–491.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed networks in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1361–1370.

Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. 2009. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* 6, 1 (2009), 29–123.

David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America* 102, 33 (2005), 11623–11628.

Shaohe Lv, Xiaodong Wang, Xin Zhao, and Xingming Zhou. 2008. Detecting the sybil attack cooperatively in wireless sensor networks. In *Computational Intelligence and Security, 2008. CIS'08. International Conference on*, Vol. 1. 442–446.

Julian J McAuley and Jure Leskovec. 2012. Learning to Discover Social Circles in Ego Networks.. In *NIPS*, Vol. 2012. 548–56.

David D McFarland and Daniel J Brown. 1973. Social distance as a metric: a systematic introduction to smallest space analysis. *EO Laumann. Bonds of Pluralism: The Form and Substance of Urban Social Networks: John Wiley* (1973), 213–252.

Bamshad Mobasher, Robin Burke, and Jeff J Sandvig. 2006. Model-based collaborative filtering as a defense against profile injection attacks. In *AAAI*, Vol. 6. 1388.

Diana Mok, Barry Wellman, and Ranu Basu. 2007. Did distance matter before the Internet?: Interpersonal contact and support in the 1970s. *Social networks* 29, 3 (2007), 430–461.

Giseop Noh, Hayoung Oh, Young-myoung Kang, and Chong-kwon Kim. 2014. PSD: Practical Sybil detection schemes using stickiness and persistence in online recommender systems. *Information Sciences* 281 (2014), 66–84.

Adrian E Raftery, Xiaoyue Niu, Peter D Hoff, and Ka Yee Yeung. 2012. Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics* 21, 4 (2012), 901–919.

Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. 2003. Trust management for the semantic web. In *The Semantic Web-ISWC 2003*. 351–368.

Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W Moore. 2011. Theoretical Justification of Popular Link Prediction Heuristics.. In *IJCAI*, Vol. 22. 2722.

Purnamrita Sarkar and Andrew Moore. 2005. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter* 7, 2 (2005), 31–40.

Dinh Nguyen Tran, Bonan Min, Jinyang Li, and Lakshminarayanan Subramanian. 2009. Sybil-Resilient Online Content Voting.. In *NSDI*, Vol. 9. 15–28.

Nguyen Tran, Jinyang Li, Lakshminarayanan Subramanian, and Sherman SM Chow. 2011. Optimal sybil-resilient node admission control. In *INFOCOM*. 3218–3226.

Bimal Viswanath, Muhammad Ahmad Bashir, Muhammad Bilal Zafar, Simon Bouget, Saikat Guha, Krishna P Gummadi, Aniket Kate, and Alan Mislove. 2015. Strength in Numbers: Robust Tamper Detection in Crowd Computations. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*. 113–124.

Bimal Viswanath, Ansley Post, Krishna P Gummadi, and Alan Mislove. 2011. An analysis of social network-based sybil defenses. *ACM SIGCOMM Computer Communication Review* 41, 4 (2011), 363–374.

Luis Von Ahn, Manuel Blum, Nicholas Hopper, and John Langford. 2003. CAPTCHA: Using hard AI problems for security. In *Advances in CryptologyEUROCRYPT 2003*. 294–311.

Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y Zhao. 2013. You Are How You Click: Clickstream Analysis for Sybil Detection.. In *Usenix Security*. 241–256.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of small-worldnetworks. *nature* 393, 6684 (1998), 440–442.

Wei Wei, Fengyuan Xu, Chiu C Tan, and Qun Li. 2012. Sybildefender: Defend against sybil attacks in large social networks. In *INFOCOM*. 1951–1959.

Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y Zhao, and Yafei Dai. 2014. Uncovering social network sybils in the wild. *TKDD* 8, 1 (2014), 2.

Sarita Yardi, Daniel Romero, Grant Schoenebeck, and others. 2009. Detecting spam in a twitter network. *First Monday* 15, 1 (2009).

Jian Yin and Sanjay Kumar Madria. 2007. Sybil attack detection in a hierarchical sensor network. In *Security and Privacy in Communications Networks and the Workshops*. 494–503.

Haifeng Yu, Phillip B Gibbons, Michael Kaminsky, and Feng Xiao. 2008. Sybillimit: A near-optimal social network defense against sybil attacks. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. 3–17.

Haifeng Yu, Michael Kaminsky, Phillip B Gibbons, and Abraham Flaxman. 2006. Sybilguard: defending against sybil attacks via social networks. *ACM SIGCOMM Computer Communication Review* 36, 4 (2006), 267–278.

Haifeng Yu, Chenwei Shi, Michael Kaminsky, Phillip B Gibbons, and Feng Xiao. 2009. Dsybil: Optimal sybil-resistance for recommendation systems. In *IEEE Security and Privacy*. 283–298.