Uncovering Social Network Sybils in the Wild

Zhi Yang
Computer Science Dept.
Peking University
Beijing, China P. R.
yangzhi@net.pku.edu.cn

Christo Wilson
Computer Science Dept.
UC Santa Barbara
Santa Barbara, CA 93106
bowlin@cs.ucsb.edu

Xiao Wang
Computer Science Dept.
Peking University
Beijing, China P. R.
wangxiao@net.pku.edu.cn

Tingting Gao
Renren Inc.
Beijing, China P. R.
tingting.gao@renren-inc.com

Ben Y. Zhao
Computer Science Dept.
UC Santa Barbara
Santa Barbara, CA 93106
ravenben@cs.ucsb.edu

Yafei Dai
Computer Science Dept.
Peking University
Beijing, China P. R.
dyf@net.pku.edu.cn

ABSTRACT

Sybil accounts are fake identities created to unfairly increase the power or resources of a single user. Researchers have long known about the existence of Sybil accounts in online communities such as file-sharing systems, but have not been able to perform large scale measurements to detect them or measure their activities. In this paper, we describe our efforts to detect, characterize and understand Sybil account activity in the Renren online social network (OSN). We use ground truth provided by Renren Inc. to build measurement based Sybil account detectors, and deploy them on Renren to detect over 100,000 Sybil accounts. We study these Sybil accounts, as well as an additional 560,000 Sybil accounts caught by Renren, and analyze their link creation behavior. Most interestingly, we find that contrary to prior conjecture, Sybil accounts in OSNs do not form tight-knit communities. Instead, they integrate into the social graph just like normal users. Using link creation timestamps, we verify that the large majority of links between Sybil accounts are created accidentally, unbeknownst to the attacker. Overall, only a very small portion of Sybil accounts are connected to other Sybils with social links. Our study shows that existing Sybil defenses are unlikely to succeed in today's OSNs, and we must design new techniques to effectively detect and defend against Sybil attacks.

Categories and Subject Descriptors

C.2 [**General**]: Security and protection (e.g., firewalls); J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Measurement, Security

Keywords

Sybil Accounts, Online Social Networks

IMC'11, November 2–4, 2011, Berlin, Germany.
Copyright 2011 ACM 978-1-4503-1013-0/11/11 ...$10.00.

1. INTRODUCTION

Sybil attacks [4] are one of the most prevalent and practical attacks against distributed systems. In this attack, a user creates multiple fake identities, known as Sybils, to unfairly increase their power and influence within a target community. Distributed systems are ill-equipped to defend against this attack, since determining a tight mapping between real users and online identities is an open problem. To date, researchers have demonstrated the efficacy of Sybil attacks against P2P systems [10], anonymous communication networks [1], and sensor networks [14].

Recently, online social networks (OSNs) have also come under attack from Sybils. Researchers have observed Sybils forwarding spam and malware on Facebook [5] and Twitter [6], as well as infiltrating social games [13]. Looking forward, Sybil attacks on OSNs are poised to become increasingly widespread and dangerous as more people come to rely on OSNs for basic online communication [9, 12] and as replacements for news outlets [8].

To address the problem of Sybils on OSNs, researchers have developed algorithms such as SybilGuard [24], SybilLimit [23], SybilInfer [3], and SumUp [17] to perform decentralized detection of Sybils on social graphs. These systems detect Sybils by identifying tightly connected communities of Sybil nodes [18].

Recent work showed that one of the key assumptions of community-based Sybil detectors, fast mixing time, does not hold on social graphs where edges correspond to strong real-world trust (*e.g.,* DBLP, Physics co-authorship, Epinions, *etc.*) [11]. Thus, community-based Sybil detectors do not perform well on "trusted" social graphs. To date, however, no large scale studies have been performed to validate the assumptions of community-based Sybil detectors on *untrusted* social networks such as Facebook and Twitter.

In this paper, we describe our efforts to detect, characterize and understand Sybil account activity in Renren, the largest OSN in China. In Section 2, we use ground truth data on Sybils provided by Renren Inc. to characterize Sybil behavior. We identify several behavioral attributes that are unique to Sybils, and leverage them to build a measurement based, real-time Sybil detector. Our detector is currently deployed on Renren's production systems, and between August 2010 and February 2011 it led to the identification and banning of over 100,000 Sybil accounts.

In Section 3 we analyze the graph structural properties of Sybils on Renren, based on the 100,000 Sybils identified by our detector, as well as 560,000 more identified by Renren using prior techniques. Most interestingly, we find that contrary to prior conjecture, Sybil accounts in Renren do not form tight-knit communities:

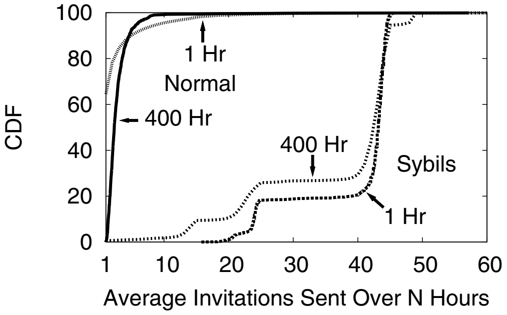

Figure 1: Average friend invitation frequency for Sybil and normal users, over two time scales.

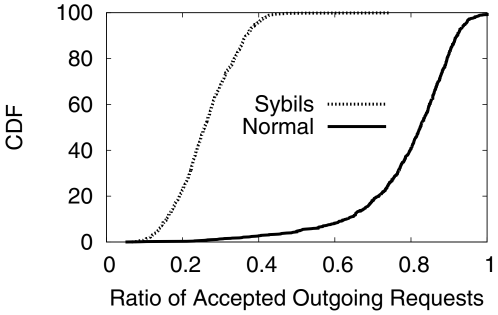

Figure 2: Ratio of accepted *outgoing* friend requests.

>70% of Sybils do not have *any* edges to other Sybils at all. Instead, attackers use snowball sampling to identify and send friend requests to popular users, since these users are more likely to accept requests from strangers. This strategy allows Sybil accounts to integrate seamlessly into the social graph.

We analyze the remaining 30% of Sybils that are friends with other Sybils, and discover that 69% (65,000 accounts) form a single connected component. By analyzing the creation timestamps of these edges, we determine that this component formed accidentally, and not due to coordinated efforts by attackers. We briefly survey several popular Sybil management tools, and show that large Sybil components can form naturally due to bias in the snowball sampling techniques these tools use to locate targets for friending.

Our analysis of Sybil behavior and characteristics demonstrates that existing Sybil defenses are unlikely to succeed on today's untrusted OSNs. This opens the door for the development of new techniques to effectively detect and defend against Sybil attacks.

2. DETECTING SYBILS

In this section, we set the backdrop for our data analysis. First, we briefly introduce the Renren OSN and describe the role of Sybil accounts in Renren. Second, we describe experiments characterizing Sybil accounts on a verified ground-truth dataset provided by Renren. Finally, we describe and build a real-time Sybil account detector deployed on Renren, and show how it led to the large Sybil dataset we analyze in the remainder of the paper.

2.1 The Renren Network and Sybil Accounts

With 120 million users, Renren[1] is the largest and oldest OSN in China, and provides functionality and features similar to Facebook. Like Facebook, Renren started in 2005 as a social network for college students in China, then saw its user population grow exponentially once it opened its doors to non-students. Renren users maintain personal profiles, upload photos, write diary entries (blogs), and establish bidirectional social links with friends. The most popular type of user activity is sharing blog entries, which can be forwarded across social hops like "retweets" on Twitter.

As its user population has grown, Renren has become an attractive venue for companies to disseminate information about their products. This has created opportunities for Sybil accounts to spam advertisements for companies, a growing trend observed by the analytics team at Renren. The increased prevalence of spam on Renren mirrors similar findings from Facebook [5] and Twitter [6].

To effectively attract friends and disseminate advertisements, most Sybil accounts on Renren blend in extremely well with normal users. They tend to have completely filled user profiles with realistic background information, coupled with attractive profile photos of young women or men, making their detection quite challenging.

Prior to this project, Renren had already deployed a suite of orthogonal techniques to detect Sybil accounts, including: using thresholds to detect spamming, scanning content for suspect keywords and blacklisted URLs, and providing Renren users with the ability to flag accounts and content as abusive. However, these techniques are generally ad-hoc, require significant human effort, and are effective only after spam content has been posted. To improve security for their users, Renren began a collaboration with our research team in December 2010 to augment their detection systems with a systematic, real-time solution. To support the project, Renren provided full access to user data and operational logs on their servers, as well as allowing us to test and deploy research prototypes of Sybil detectors on their operational network.

Defining Sybils. In this study, as in prior work [3, 17, 23, 24], we are interested in detecting and deterring the use of mass Sybil identities by malicious users. We broadly define Sybils as fake accounts created for the purpose of performing spam or privacy attacks against normal users. We observe that the main goal of Sybils is to increase the power of the attacker by amassing friend links to normal users, thus integrating themselves into the social graph. Attackers create many Sybils to increase their coverage of the graph, as well as to combat attrition from Sybils getting banned. Although penetrating the graph is simply a precursor for other malicious activity, our work is agnostic to these secondary goals, as well as the specific methods and tools used to create and manage the Sybils.

Our definition of Sybils *does not* include fake accounts generated by users for benign purposes, such as preserving privacy and anonymity, acting on behalf of young children, separating work and personal identities, *etc*. These "benign Sybils" act just like normal accounts, and therefor do not fall under our definition of malicious Sybils. As discussed in Section 2.3, benign Sybils are unlikely to be flagged by the techniques proposed in this work.

2.2 Characterizing Sybil Accounts

Our approach to building a real-time Sybil detector begins by first identifying features that distinguish Sybil accounts from normal users. To help, Renren provided us with two sets of user accounts, containing 1000 Sybil accounts and 1000 non-Sybil accounts, respectively. The Sybil accounts were previously identified using existing mechanisms. A volunteer team carefully scrutinized all accounts in both sets to confirm they were correctly classified by looking over detailed profile data, including uploaded photos, messages sent and received, email addresses, and shared content (blogs and web links).

[1] `http://www.renren.com`

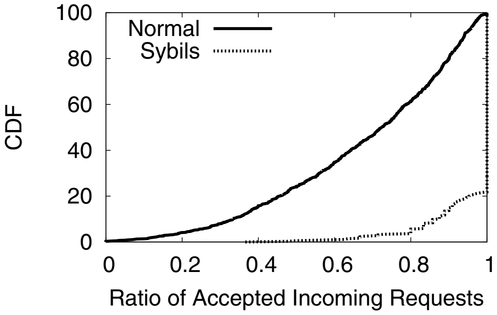

Figure 3: Ratio of accepted *incoming* friend requests.

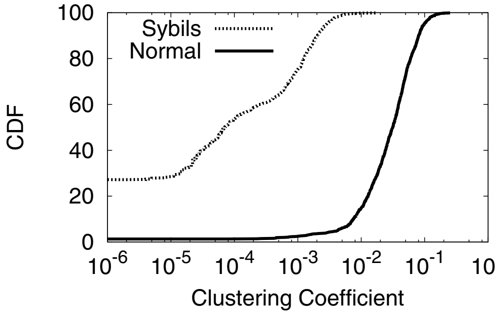

Figure 4: Clustering coefficient for users' 50 first friends.

		SVM Predicted		Threshold Predicted	
		Sybil	**Non-Sybil**	**Sybil**	**Non-Sybil**
True	**Sybil**	98.99%	1.01%	98.68%	1.32%
	Non-Sybil	0.66%	99.34%	0.5%	99.5%

Table 1: Performance of SVM and threshold classifiers.

Using this dataset as our ground truth, we searched for behavioral attributes that may serve to identify Sybil accounts. After examining a wide range of attributes, we found four potential identifiers. We describe them each in turn, and illustrate how they characterize accounts in our ground truth dataset.

Invitation Frequency. Invitation frequency is the number of friend requests a user has sent within a fixed time period (*e.g.,* an hour). Figure 1 shows the friend invitation frequency of our dataset, averaged over long term (400 hour) and short term (1 hour) time scales. Since adding friends is a goal for all Sybil accounts, they are much more aggressive in sending requests than normal users. There is a clear separation: accounts sending more than 20 invites per time interval are Sybils. This result holds true at both long and short time scales, meaning that invitation frequency can be used to detect Sybils without significant delays. For example, a threshold of 40 requests/hour can identify $\approx$70% of Sybils with no false positives. Prior to our work, Renren deployed a threshold based detector that forces users to solve a captcha if they send $\geq$50 requests in a day, which explains the apparent upper limit on friend requests.

Outgoing Requests Accepted. A second distinguishing feature is the fraction of outgoing friend requests confirmed by the recipient. The CDF shown in Figure 2 shows a distinct difference between Sybils and normal users. In general, non-Sybil users generally have high accepted ratios with an average of 79%. On average, however, only 26% of all friend requests sent by Sybil accounts are accepted. This is unsurprising, since normal users typically send invites to people with whom they have prior relationships, whereas Sybils target strangers.

Despite prior studies that show users accept requests indiscriminately [15, 16], our results show that most users can still effectively identify and decline invitations from Sybils. The fact that some users still accept requests from Sybils is explained by two factors. First, most Sybils target members of the opposite sex by using photos of attractive young men and women in their profiles. While women make up 46.5% of the overall Renren user population, they make up 77.3% of the Sybils in our dataset. Second, Sybils typically target popular, high degree users who are more likely to be careless about accepting friend requests from strangers. We further explore this point in Section 3.4.

Incoming Requests Accepted. Figure 3 plots a CDF of users by the fraction of incoming friend requests they accept. The incoming requests accepted by non-Sybil users are spread across the board. In contrast, Sybil accounts are nearly uniform in that they accept all incoming friend requests, *e.g.,* 80% of Sybils accepted all friend requests. In fact, many of the Sybils with $<$100% accept rate fall into this category because Renren banned them before they could respond to all outstanding requests. However, since Sybil accounts receive few friend requests, this mechanism can incur a significant delay before detecting Sybils.

Clustering Coefficient. Clustering coefficient (cc) is a graph metric that measures the mutual connectivity of a user's friends. Since normal users tend to have a small number of well-connected social cliques, we expect them to have much higher cc values than Sybil accounts, which are likely to befriend users with no mutual friendships. Figure 4 plots the CDF of cc values for each user's first 50 friends (sorted by time). As expected, non-Sybil users have cc values orders of magnitude larger than Sybil users (average cc values of 0.0386 and 0.0006 respectively). Since cc can be computed based on invitations only (*i.e.,* user responses are not required) it can potentially perform well as a real-time Sybil detection metric.

2.3 Building and Running a Sybil Detector

Our analysis results seem to indicate that a threshold based scheme can effectively detect most Sybil accounts. Our next step is to verify this assertion by comparing the efficacy of a simple threshold detection approach against a more complex learning algorithm.

We apply a support vector machine (SVM) classifier to our ground truth dataset of 1000 normal users and 1000 Sybils. We randomly partition the original sample into 5 sub-samples, 4 of which are used for training the classifier, and the last used to test the classifier. The results in Table 1 show that the classifier is very accurate, correctly identifying 99% of both Sybil and non-Sybil accounts. We compare these results to those of a threshold-based detector: *outgoing requests accepted ratio* $< 0.5 \wedge$ *frequency* $> 20 \wedge$ *cc* < 0.01. Our results show that a properly tuned threshold-based detector can achieve performance similar to the computationally expensive SVM.

Real-time Sybil Detection. Our analytical results using the ground-truth dataset led to the design of an adaptive, threshold-based Sybil detector that identifies Sybil accounts in near real-time. The detector monitors all accounts using a combination of friend-request frequency, outgoing request acceptance rates, and clustering coefficient. It uses an adaptive feedback scheme to dynamically tune threshold parameters on the fly[2]. Tuning the thresholds minimizes the likelihood of false positive classifications of normal accounts as Sybils. Because our system works by detecting abnor-

[2] We omit details of the adaptive scheme for Renren's security and confidentiality.

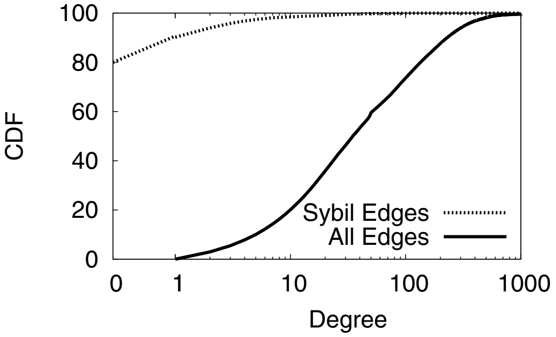

Figure 5: The degree of Sybil accounts.

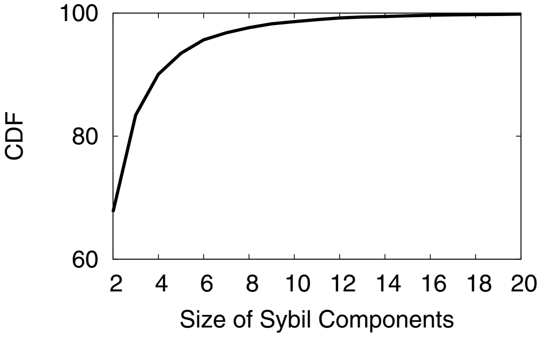

Figure 6: The size of connected Sybil components.

mal behavior in friending or content dissemination, it is unlikely to detect benign Sybils that behave like normal users.

After offline testing, Renren deployed our Sybil detection mechanism in late August 2010, and it has been in continuous operation ever since. From August 2010 to February 2011, Renren administrators used our mechanism to detect and subsequently ban ∼100,000 Sybil accounts in Renren. In addition to these accounts, Renren provided us with data on ∼560,000 accounts that were detected and banned using prior techniques from 2008 to February 2011. For the remainder of this paper, we will use all of these Sybil accounts (660,000 in all) to study the behavior of Sybil accounts.

Sybil Account Behavior. We confirmed that the Sybil accounts identified by our detector are actually malicious by analyzing the content generated by these accounts. Offline analysis confirmed that 67% of content generated by Sybils trips Renren's spam detectors (*e.g.,* suspicious keyword filter, blacklisted URLs, *etc.*). Of the remaining accounts, the vast majority were banned before they had a chance to generate any content. Analysis of spam keywords and campaign clusters produces results that are consistent with prior work on OSN spam [5, 6], and we omit the results for brevity.

False Positives. To assess false positives, we examine feedback to Renren's customer support department. Renren operates a telephone number and e-mail address where customers whose accounts have been banned for abuse can attempt to get the account reinstated. Complaints are evaluated by a human operator, who determines if the account was banned erroneously.

We use the complaint rate, measured as the number of complaints per-day divided by the number of accounts banned per-day, as an upper-bound on false positives. During the two week period between Dec. 13-26, 2010, Renren received ∼50 complaints per-day, with the complaint rate being ∼0.015, which is extremely low. Of these complaints, manual inspection confirms that 48% of the accounts are Sybils, meaning that attackers attempted to recover Sybils by abusing the account recovery process. The majority of the remaining complaints can be attributed to compromised accounts. Thus, the true false positive rate is even less than the daily complaint rate.

3. SYBIL TOPOLOGY

In this section we analyze the graph topological characteristics of Sybils on Renren. In particular, we are interested in analyzing whether Sybils in the wild are vulnerable to identification using the community-based Sybil detectors that have been proposed by researchers. Our results show that Sybils on Renren do not conform to the assumptions of existing work. Analysis of the degree distribution of Sybil accounts demonstrates that, contrary to expectations, the vast majority of Sybils do not form social links with other Sybils. Furthermore, temporal analysis of social links between Sybils indicates that these connections are often formed randomly by accident, rather than intentionally by attacker.

3.1 Sybil Community Detectors

SybilGuard [24], SybilLimit [23], SybilInfer [3], and SumUp [17] are all algorithms for performing decentralized detection of Sybil nodes on social graphs. At their core, all of these algorithms are based on two assumptions of Sybil and normal user behavior:

1. Attackers can create unlimited Sybils and form edges between them. Edges between Sybils are beneficial since they make Sybils appear more legitimate to normal users.
2. The number of edges between Sybils and normal users will be limited, since normal users are unlikely to accept friend requests from unknown strangers.

Under these assumptions, Sybils tend to form tight knit clusters, since the number of edges between Sybils is greater than the number of edges connecting to normal users. We refer to edges between Sybils as *Sybil edges*, while edges connecting Sybils and normal users are called *attack edges*.

Sybil detection algorithms identify Sybil clusters by locating the small number of edge cuts that separate the Sybil region from the social graph. SybilGuard, SybilLimit, and SybilInfer all leverage specially engineered random walks for this purpose, while SumUp uses a max-flow approach. Although all of these algorithms are implemented differently, it has been shown that they all generalize to the problem of detecting communities of Sybil nodes [18].

Although these four algorithms have been shown to work on synthetic graphs (*i.e.,* real social graphs with Sybil communities artificially injected), to date no studies have demonstrated their efficacy at detecting Sybils in the wild. In the following sections, we examine the characteristics of Sybils on Renren in order to ascertain whether they are amenable to identification by community-based Sybil detectors.

3.2 Sybil Edges

We begin our analysis of Sybil topology by examining the degree distribution of Sybil accounts on Renren. Our goal is to test the most basic assumption of community-based Sybil detectors: do Sybils in the wild form tight-knit communities? In order for Sybils to cluster, they must have at least one edge connecting to another Sybil, otherwise they will be disconnected.

Figure 5 shows the degree distribution of all 667,723 Sybil accounts. When all edges are considered, the degree distribution is unremarkable: it follows the same general trend that has been observed on numerous other OSNs [21].

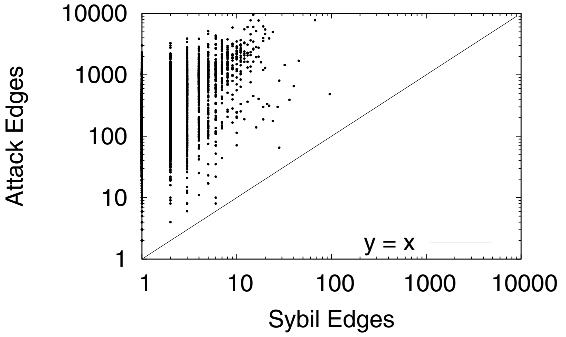

Figure 7: Scatter plot of Sybil edges vs. attack edges for Sybil components on Renren.

Sybils	Sybil Edges	Attack Edges	Audience
63,541	134,941	9,848,881	6,497,179
631	1153	104,074	21,014
68	67	7,761	7,702
51	50	15,349	15,179
37	40	14,431	13,886

Table 2: Statistics for the five largest Sybil components.

However, when we restrict the distribution to only edges between Sybils, we discover an unexpected result: only 20% of Sybils are friends with one or more other Sybils. This indicates that the vast majority of Sybils do not demonstrate any sort of clustering behavior with other Sybils. Rather, most Sybils only form attack edges, and thus totally integrate into the normal social graph.

3.3 Sybil Communities

We now shift our focus to the minority of Sybils that do connect to other Sybils. Although we can conclude from Figure 5 that most Sybils in the wild do not obey the key assumption of community-based Sybil detectors, it is still possible that the connected minority are vulnerable to community detection. Thus, we now seek to answer the following questions: what are the characteristics of Sybil communities on Renren, and would community-based Sybil detectors be able to identify them?

To bootstrap our analysis, we construct a graph consisting solely of Sybils with at least one edge to another Sybil. The resulting graph is highly fragmented: it consists of 7,094 separate connected components. Figure 6 shows the size distribution of these Sybil components. The distribution is heavy tailed: although 98% of Sybil components have less than 10 members, the vast majority of Sybil accounts belong to a single, large connected component. Table 2 lists the details for the five largest Sybil components.

In order for Sybil communities to be identifiable by existing algorithms, they must form tight knit communities. Put another way, the number of Sybil edges inside the community must be greater than the number of attack edges that connect to the normal population. However, as shown in Table 2, this assumption does not hold for the largest Sybil components on Renren.

Figure 7 shows a scatter plot comparing the number of Sybil edges and attack edges in each Sybil component on Renren. All components are above the 45° line, meaning that they have more attack edges than Sybil edges. Thus, no components meet the requirements for detection using existing community-based Sybil identification algorithms.

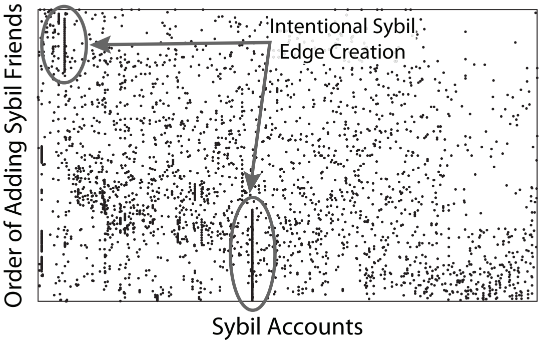

Figure 8: The order of adding Sybil friends for 1,000 Sybils. Each column represents an individual Sybil.

3.4 Sybil Edge Formation

We now examine the processes driving the formation of Sybil edges on Renren. In particular, we seek to determine if edges between Sybil nodes are intentionally created by attackers. If they are, then this means that community detection may still be a viable approach to detecting Sybils on OSNs. However, if Sybil edges are not created intentionally, then this raises a new question: what process drives the accidental creation of Sybil edges?

Temporal Characteristics. One simple litmus test for identifying intentional Sybil edge creation is examining the order in which edges were established. If Sybil edges are formed intentionally by attackers, then we would expect to see them created sequentially, before friend requests are sent out to normal users. This behavior maximizes the utility of Sybil edges by giving Sybils the appearance of "normal" friend relations, thus (potentially) deceiving normal users into accepting friend requests from Sybils.

Figure 8 shows the order in which edges were created for 1,000 random Sybils drawn from the largest Sybil component on Renren (containing 63,541 Sybils). For each Sybil i with n edges, we construct the sequence $\langle f_1, f_2, \ldots, f_n \rangle$, where f_i is an edge, and the sequence is sorted chronologically by creation time. Each column of the figure shows the sequence of edge creations for a particular Sybil, with black dots representing Sybil edges.

As shown in Figure 8, the order of Sybil edge creation is almost uniformly random. This indicates that the vast majority of Sybil edges in the large component were formed accidentally: attackers had no intention to link Sybils together and form a connected component. Intentionally created connections between Sybils appear as solid vertical lines in the figure. We highlight two examples in Figure 8 by circling them. It is unclear why a tiny minority of Sybils exhibit correlated behavior; we are currently studying this behavior as part of our ongoing work.

Sybil Degree. In order to reinforce the idea that the vast majority of Sybil edges in the large component are not intentionally created, we plot the degree distribution of the large component in Figure 9. 34.5% of Sybils only connect to 1 other Sybil, and 93.7% connect to ≤ 10. It is unlikely that an attacker would expend the effort to link Sybils in such a loose way, since these edge counts are not high enough to make Sybils appear legitimate to normal users.

Snowball Sampling. At this point we have established that attackers do not create the vast majority of Sybil edges intentionally; instead, they appear to occur randomly by accident. To understand how this happens, we conducted a brief survey of three software tools used to manage Sybil accounts on Renren. The details for

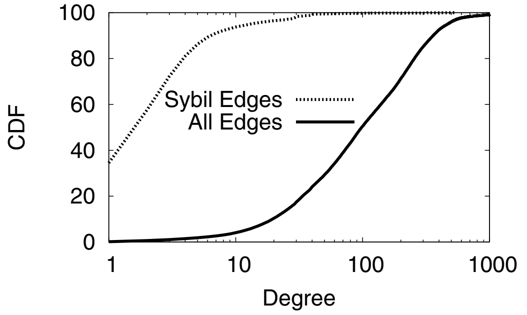

Figure 9: Degree distribution of the largest Sybil component.

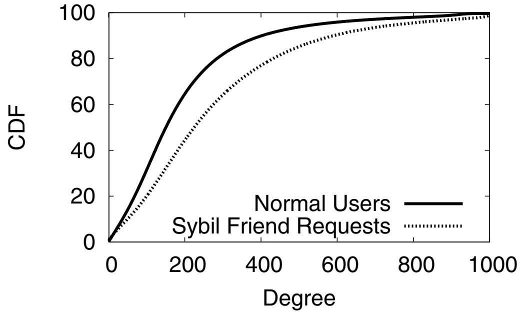

Figure 10: Degree distribution of normal users (from [7]) and targets of friend requests from Sybils.

Tool Name & URL	Platform	Cost
Renren Marketing Assistant V1.0 http://www.duote.com/soft/30348.html	Windows	$37
Renren Super Node Collector V1.0 http://www.snstools.com/snstool/86.html	Windows	Contact Author
Renren Almighty Assistant V5.8 http://www.sns78.com/	Windows	Contact Author

Table 3: Popular Sybil creation and management tools.

each tool are given in Table 3. The purpose of these tools is to automate the process of creating Renren accounts, forming edges between the Sybils and other users, and posting content en-mass.

The documentation for the tools in Table 3 state that they select targets for friending by performing snowball sampling on the social graph to locate popular users. Although we cannot be certain whether the Sybils in our dataset were created using the tools in Table 3, we can show that Sybils on Renren do bias friend requests towards high degree nodes. Figure 10 shows the degree distribution for all users that received friend requests from Sybils, and illustrates that it is skewed towards high degrees when compared to the actual degree distribution of the Renren population [7].

Based on the advertised functionality of these tools, and the results in Figure 10, we can surmise that Sybil edges are created accidentally due to two factors. First, the goal of Sybils is to accrue many friends by sending out numerous friend requests. If a Sybil is successful, it becomes popular by virtue of its large social degree. Second, the snowball sampling performed by Sybil management tools is intentionally biased towards locating popular users. Thus, it is likely that these tools will, unbeknownst to the attacker, occasionally select Sybil nodes to send friend requests to. As shown in Figure 3, Sybils almost always accept incoming friend requests, hence when this situation arises a Sybil edge is likely to be created.

4. RELATED WORK

OSN Spam. Recent studies have characterized the growing OSN spam problem on Facebook [5] and Twitter [6]. These studies rely on offline heuristics to identify spam content in status updates/tweets, as well as aberrant behavior that is indicative of spamming. The authors locate millions of spam messages on each OSN, and use them to analyze the large scale, coordinated spam campaigns. In contrast, our study is focused on the graph topological characteristics of malicious accounts, rather than spam content.

OSN Spam Detection. Various techniques borrowed from e-mail spam detection have been applied to OSN spam. Webb et al. use honeypot accounts on MySpace to trap spammers who attempt to friend them [20]. Our results indicate that unless social honeypots are engineered to appear popular, they are unlikely to be targeted by spammers.

Other studies have leveraged Bayesian filters and SVMs to identify spammers on Twitter [2,19,22] and Facebook [16]. These techniques work well on Twitter, since Sybil friending behavior can be identified using publicly available following and followed information. However, detection on OSNs like Facebook and Renren is less successful, since only existing friendships are publicly viewable, while invitation frequency is hidden. Our Sybil detector overcomes this issue by leveraging friend invitation information that is only accessible from within Renren.

5. CONCLUSION

In this paper we make two contributions to the area of Sybil detection on OSNs. First, we use ground-truth data about the behavior of Sybils in the wild to create a measurement-based, real-time Sybil detector. We show that a computationally efficient, threshold-based classifier is sufficient to catch 99% of Sybils, with low false positive and negative rates. We have deployed our detector on Renren's production systems, and to date it has led to the identification and banning of over 100,000 Sybil accounts.

Our second contribution is a first-of-its-kind characterization of Sybil graph topology on a major online social network. Using edge creation information for over 660,000 Sybil accounts on Renren, we show that Sybils on Renren do not obey behavioral assumptions that underlie previous work on decentralized Sybil detectors. 80% of Sybils do not connect socially to other Sybils, but instead focus on building friendships with normal users. Even in rare cases where Sybils do form connected components, these clusters are loose, rather than tightly knit. Temporal analysis indicates that these Sybil edges are formed accidentally by attackers, rather than intentionally.

Although we cannot be sure that our results generalize to all OSNs, our findings for a traditional, *i.e.,* untrusted, OSN, coupled with results from prior work on trusted OSNs [11], suggest that we should explore new approaches to perform decentralized detection of Sybil accounts on OSNs.

Acknowledgments

We thank the anonymous reviewers for their comments. This material is based in part upon work supported by the National Science Foundation under grants IIS-0916307 and IIS-847925. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work is also supported in part by South Korea National Research Foundation under the World Class University program, and the National Basic Research Program of China (Grant No.2011CB302305).

6. REFERENCES

[1] Bauer, K., McCoy, D., Grunwald, D., Kohno, T., and Sicker, D. Low-resource routing attacks against tor. In *Proc. of Workshop on Privacy in Electronic Society* (Alexandria, VA, 2007).

[2] Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. Detecting spammers on twitter. In *Proc. of CEAS* (Redmond, WA, July 2010).

[3] Danezis, G., and Mittal, P. Sybilinfer: Detecting sybil nodes using social networks. In *Proc of NDSS* (2009).

[4] Douceur, J. R. The Sybil attack. In *Proc. of IPTPS* (March 2002).

[5] Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Y. Detecting and characterizing social spam campaigns. In *Proc. of IMC* (2010).

[6] Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: the underground on 140 characters or less. In *Proc. of CCS* (2010).

[7] Jiang, J., Wilson, C., Wang, X., Huang, P., Sha, W., Dai, Y., and Zhao, B. Y. Understanding latent interactions in online social networks. In *Proc. of IMC* (2010).

[8] Kwak, H., Lee, C., Park, H., and Moon, S. B. What is twitter, a social network or a news media? In *Proc. of World Wide Web Conference* (Raleigh, NC, April 2010).

[9] Lenhart, A., Purcell, K., Smith, A., and Zickuhr, K. Social media and young adults. Pew Research Center, February 2010.

[10] Lian, Q., Zhang, Z., Yang, M., Zhao, B. Y., Dai, Y., and Li, X. An empirical study of collusion behavior in the maze p2p file-sharing system. In *Proc. of ICDCS* (June 2007).

[11] Mohaisen, A., Yun, A., and Kim, Y. Measuring the Mixing Time of Social Graphs. In *Proc. of IMC* (2010).

[12] Murphy, S. Teens ditch e-mail for texting and facebook. MSNBC.com, Aug 2010.

[13] Nazir, A., Raza, S., Chuah, C.-N., and Schipper, B. Ghostbusting facebook: Detecting and characterizing phantom profiles in online social gaming applications. In *Proc. of SIGCOMM WOSN* (June 2010).

[14] Newsome, J., Shi, E., Song, D., and Perrig, A. The sybil attack in sensor networks: Analysis & defenses. In *Proc. of IPSN* (Berkeley, CA, 2004).

[15] Sophos. Sophos Facebook ID probe shows 41% of users happy to reveal all to potential identity thieves. `http://www.sophos.com/pressoffice/news/articles/2007/08/facebook.html`, 2007.

[16] Stringhini, G., Kruegel, C., and Vigna, G. Detecting spammers on social networks. In *Proc. of ACSAC* (Austin, TX, December 2010).

[17] Tran, N., Min, B., Li, J., and Subramanian, L. Sybil-resilient online content voting. In *Proc. of NSDI* (2009).

[18] Viswanath, B., Post, A., Gummadi, K. P., and Mislove, A. An analysis of social network-based sybil defenses. In *Proc. of SIGCOMM* (2010).

[19] Wang, A. H. Don't follow me: Spam detection on twitter. In *Proc. of SECRYPT* (Athens, Greece, July 2010).

[20] Webb, S., Caverlee, J., and Pu, C. Social honeypots: Making friends with a spammer near you. In *Proc of CEAS* (Mountain View, CA, August 2008).

[21] Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., and Zhao, B. Y. User interactions in social networks and their implications. In *Proc. of EuroSys* (April 2009).

[22] Yardi, S., Romero, D., Schoenebeck, G., and Boyd, D. Detecting spam in a twitter network. *First Monday 15*, 1 (2010).

[23] Yu, H., Gibbons, P. B., Kaminsky, M., and Xiao, F. Sybillimit: A near-optimal social network defense against sybil attacks. In *Proc. of IEEE S&P* (2008).

[24] Yu, H., Kaminsky, M., Gibbons, P. B., and Flaxman, A. Sybilguard: defending against sybil attacks via social networks. In *Proc. of SIGCOMM* (2006).

Summary Review Documentation for

“Uncovering Social Network Sybils in the Wild”

Authors: Z. Yang, C. Wilson, X. Wang, T. Gao, B. Zhao, Y. Dai

Reviewer #1

Strengths: The data set is particularly interesting; merely having the ground truth about Sybils is remarkable, and the authors use it well to infer the behavior of Sybil-generators. The results are exceptionally important for social networking research.

Weaknesses: Sybils in Renren have little utility in forming tightly-knit Sybil communities, unlike other social networks such as recommendation systems. It is unclear that the results on Renren sufficiently invalidate the usefulness of Sybil detection algorithms that rely on the assumption of few attack edges for other social networks.

Sybils are presumably being created for a specific reason, to launch a specific attack; there is no need for Sybil-to-Sybil links to launch such attacks (spamming and/or privacy violation attacks) on Renren, but since there is a use for Sybil-to-Sybil links on something like eBay or Amazon (recommendation poisoning attacks), a similar analysis of the social networks on those sites may have very different results.

Comments to Authors: This is a fantastic paper, and the main complaint is the one about the structure of Sybils being different in different networks. It would be good to see this mentioned in the paper as a limitation of the study, though as presented, this work is incredibly valuable and a significant contribution to the field.

One concern that I have that is perhaps out of the scope of this paper is that the metrics used to characterize Sybil accounts can all be, to some extent, controlled by the Sybil generators through a combination of rate limiting and the addition of more Sybils. A clever attacker who is aware of these detection algorithms could thwart it. Though having such an algorithm may limit the speed with which the attacker can generate Sybil identities, it's not clear that they can't simply create more Sybil identities to compensate. This is not really a criticism: this is a typical result of the ongoing battle against spam, where attackers have the advantage because the defenders generally have to reveal what they're doing to some extent.

Finally, and again this is outside of the scope of the paper, it would be interesting to consider impersonation-style Sybils as well. Perhaps such Sybils are less common or harder to detect, so it may be very difficult to repeat your study on these Sybils. However, that these Sybils are radically different from the more random, optimistic attachment Sybils, both in terms of your detection metrics and in terms of their internal clustering.

Minor notes: It is not clear if existing algorithms missed the 100,000 Sybils that were detected using the techniques in this paper. Also, how were the 560,000 Sybils previously found verified to be Sybils?

Reviewer #2

Strengths: The paper involves a very unique data set. The authors' work is being used in practice.

Weaknesses: A very loose definition of Sybil is used. There is a lack of evidence provided that demonstrates all of the users labeled as Sybils are actually malicious (i.e., the paper fails to demonstrate that Hanlon's razor is not applicable).

Comments to Authors: My primary concerns with the paper start with the definition of what a sybil account is. I agree with the first part: “Sybil accounts are fake identities”; I could even agree with the remainder of the definition (“created to unfairly increase the power or resources of a single malicious user”), if “malicious” were properly defined. However, no definition is provided; it simply lumps all users with multiple online identities together, whether they are hackers intending to steal identities, content or other resources, miscreants attempting to propagate spam in OSNs, or simply teenagers having fun. While the latter activities may be considered malicious for some definitions of the word, it is usually harmless fun. Since the paper does not appear to discriminate between different reasons for having multiple online personalities, the results are not that compelling in my opinion; dealing with hackers and spammers is an important matter, addressing users expressions of their individuality is something completely different.

Stated differently, Hanlon's razor reads “Never attribute to malice that which is adequately explained by stupidity.” The latter is certainly responsible for at least some of the fake identities on OSNs. As a result, the authors need to provide evidence of malice, to distinguish the detrimental cases from the seemingly harmless ones. The paper fails to do this, which is why I cannot rate it higher. I do think that the topic is an important one, so the authors should continue their work on it.

Section 1: (i) Clarify what is meant by “infiltrating social games”? If no financial gain is possible, is creating numerous players for one's self really malicious? (ii) I agree that certain attacks need to be curtailed to prevent the general public from losing confidence in OSNs; however, I don't see that eliminating all instances of “multiple personalities” is going to solve this issue. (iii) What evidence do you have that each of the 100,000 “Sybil users” attempted any sort of malicious activity?

Section 2: (i) If “Renren provided full access to user data and operational logs”, then where is the evidence that the users you labeled as “Sybil” conducted malicious acts? (ii) What evidence is there that the Sybil accounts provided as ground truth were malicious? Did Renren use the same broad definition of malicious? (iii) The last sentence assumes that there are no “stealth” sybils in the “other” population; you have no proof of this.

Section 3: (i) Were any of these Sybils reported for malicious activity, such as sending spam? The paper provides no validation of the claims being made. (ii)- The discussion of Figure 8 assumes that you have positively identified all Sybils, of which there is no proof. (iii) Why would you expect Sybil edges to be created sequentially?

Section 4: While your study may not involve characterizing "spam content", it should involve finding evidence to support your claims that the users labeled as Sybils have actually done at least one malicious activity.

Section 5: The results section did not demonstrate that 99% of Sybils in the data set were identified "with low false positive and negative rates"; this is a key shortcoming of the paper.

Reviewer #3

Strengths: Access to a unique dataset (ground truth for sybil account), reasonable analysis.

Weaknesses: Important details about verifying sybil account (ground truth) and the proposed detection techniques are missing.

Comments to Authors: This clearly written paper benefits from the provided dataset and close collaborations with Renren OSN in China to characterize sybil accounts, uses these characterizations to devise a detection technique. Also using detected sybil accounts to examine their topological properties that appears to debunk the common assumptions about their tight connectivity. Most of the description in the paper seems reasonable. There are a few issues that deserve further clarifications by authors.

1. The key break that enables this study is the availability of confirmed sybil account that are used to bootstrap the identification process. Authors rely on identified sybil account by RenRen based on examination of many attributes (as stated in subsection 2.2). One of the main difficulties is to make sure that an account is indeed a sybil account. This leads to an important question that what tests were conducted on these accounts to ensure that they are indeed sybil? If the identification of sybil accounts are based on the attributes that are presented in subsection 2.3, there is no surprise here. Renren filtered these account based on the four attributes and authors showed that these attributes are different for sybil account. The paper should elaborate on this issue.
2. Authors present only 4 attributes for distinguishing sybil account from many they have examined. It is useful if authors at least briefly mention what other attributes were examined and to what extent they separated sybil from non-sybil accounts, and possibly describe why.
3. Authors suggest that threshold based sybil detection technique works (in subsec 2.3) without providing any information on its success rate. Ironically, the proposed adaptive version of threshold based detection is not described at all!
4. It is not clear to me what the sybil creation and management tools are and how authors got access to them to determine sampling technique used by sybil users. If such tools are available, why are they used to determine many other behaviors of sybil users?
5. One other lingering question is whether all the sybil accounts are related to each other or more likely they belong to different groups. How can this be examined from the connectivity among these accounts?

Reviewer #4

Strengths: The dataset is unique, and the measurements extracted well presented.

Weaknesses: The paper is imprecise: does not define Sybil, does not describe false positives or negatives, does not claim to detect Sybils other than those generated by commercial tools, oversells its contribution and overstates its conclusions.

Comments to Authors: Great problem domain, great dataset. However, I found this paper frustrating.

First problem (no Sybil definition): The goal of being a Sybil in Renren isn't well described. Why do it? Is it the same reason why one would create Sybils in eBay? Or Twitter? Or for Facebook games? No. At this point, you're dealing with what might be a different type of Sybil than researchers often look at, and in turn, your general statements about sybils aren't well-founded. The only passage present is (too late) on page 6 "the goal of Sybils is to accrue many friends by sending out numerous friend requests." This doesn't seem to be the real goal: popularity for its own sake doesn't make money.

First problem, part 2 (Biased Sybil detection): Are *all* types of Sybils discovered by the tool or just the type that are generated by the tools described on page 6? Are there false users for other reasons present in the data set? Are the Sybils in the known-to-be-Sybil set representative? Or are they just the type of Sybils that Renren administrators find to be the most annoying? How were they found? Was a random set of users chosen to provide truly identifying information to renew their accounts? I find an unbiased Sybil detection scheme unlikely, and no details are provided for how known-Sybils are found.

Second problem (Unfounded generalization): What if the Sybils you deal with are selfish and not that bright while the Sybils researchers have dealt with in the past are clever and evil? Vern Paxson published a lovely paper (How to 0wn the Internet...) that suggested what havoc could be wrought by clever, evil Internet worms, and such a paper states a case for defending against the worst case, rather than the typical case. The expectation in section 3, "contrary to expectations, the vast majority do not form social links with other Sybil accounts" and the conclusion in section 5 is then wildly overstated, "Sybils in the wild do not obey behavioral assumptions that underlie previous work on decentralized Sybil detectors". Is it the case that the Sybils to be detected are those bent on manipulating reputation metrics in other networks while the Sybils here are just spammers? Should the "Sybil" name apply for this? Do there need to be many of them for the attacks in your paper?

Third problem (explanation): WHY do sybils behave in the way discovered? Is it for getting better penetration? Or is it for avoiding detection? Such explanation of what you observe would be easy if the goal of a Sybil in Renren (or the goal of Renren in protecting its network) were better described.

It appears that the main mismatch you've encountered is research literature designed to protect against worst case attacks, while Renren is manipulated for different purposes to by different means.

That said, it's good data, and a reasonable contribution for a short paper. The paper just could have been written to rely on the data and the dataset, as well as to present the measurement methodology well and analyze the measurement errors properly, without the unnecessary generalization to all Sybils in all social networks.

Reviewer #5

Strengths: The real-world data set used in the paper is impressive. The analysis on Sybil identities in section 3 is thorough. And it is interesting to see that the finding from the real dataset contradicts the baseline assumption of many previous works on Sybil attack defense.

Weaknesses: The paper does not suggest any significant methodology for the detection of Sybil attacks in general setup. While there are some evidence that sybil edges are created intentionally, the paper omits discussion on it.

Comments to Authors: This is a well-written paper. The flow of the paper is smooth and the level of detail is adequate (except for Section 2.3). My concerns about the paper are the following:

1. While the paper points out the flawed assumptions in previously suggested Sybil detection schemes, it does not suggest any new methodologies on its own. To me, the methodology outlined in Section 2.3. is not directly applicable outside Renren OSN for the following two reasons: First, for each OSN the tool is applied to, it may require a significant amount of man-hour to hand-pick Sybil and non-Sybil identity data in order to train SVM classifier. Second, its classifier requires prior knowledge on a set of good thresholds on <outgoing requests acceptance ratio, frequency, and cc>.
2. In the introduction and at the beginning of section 3, the paper claims that the links between Sybils are formed by accident. However, in Figure 8, there are evidences that there are edges created intentionally by Sybil accounts. I am curious to see why there are few but obvious attempts from some Sybil accounts to connect among them. Can they be the only real Sybil accounts by any chance?

Response from the Authors

We thank all the anonymous reviewers for their time and comments. The camera-ready version of this paper includes several changes and additions in order to address the reviewer's comments.

First, the language of the paper has been adjusted to make it clear that we are focused on analyzing Sybils on untrusted social networks like Facebook and Renren. We acknowledge that our results may not generalize to trusted OSNs like DBLP or Epinions, where edges correspond to real-world trust between users.

Second, additional text has been added in Section 2.1 to clearly define what we consider to be a Sybil account. Our definition is generic, in agreement with definitions from prior work, and not dependent on particular features of Renren or specific tools used by attackers. We note that benign Sybils created by normal users (e.g. to protect individual privacy) are not included in our Sybil definition, and the new text explains why they unlikely to be caught by the techniques proposed in the paper.

Third, in Section 2.3 we address the question "what are Sybil accounts on Renren used for?" We briefly examine the malicious behaviors of Sybil accounts on Renren, and confirm that the majority are used to disseminate spam. Analysis of the generated spam confirms that it exhibits the same salient properties as prior studies on OSN spam.

Fourth, also in Section 2.3, we have added new paragraphs on the false positives of our Sybil detection methodology. We use the complaint rate to Renren's customer service department as the signal to examine how many accounts are erroneously classified and banned as Sybils. Our results show that the false positive detection rate of our techniques is very low.

Fifth, additional text has been added to Section 3.4 discussing Sybil management tools. We acknowledge that the tools listed in the paper are only a small subset of possible tools, and we have no way of confirming whether the Sybils we identified were created using them. However, we have added a new Figure 10 demonstrating that Sybils do exhibit bias towards high degree nodes when selecting targets to friend, which accords with our understanding of the features of popular Sybil management tools.

Lastly, new details on Renren's Sybil detection methods, prior to the deployment of our tool, have been added to Sections 2.1 and 2.2. These existing measures help to explain artifacts in Figure 1.