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.