On The Effectiveness of Sybil Defenses Based on Online Social Networks

David Koll
University of Goettingen
Goettingen, Germany
koll@cs.uni-goettingen.de

Jun Li
University of Oregon
Eugene, USA
lijun@cs.uoregon.edu

Joshua Stein
University of Oregon
Eugene, USA
jgs@cs.uoregon.edu

Xiaoming Fu
University of Goettingen
Goettingen, Germany
fu@cs.uni-goettingen.de

***Abstract*—A Sybil attack can inject many forged identities (called Sybils) to subvert a target system. Among various defense approaches, of particular attention are those that explore the online social networks (OSNs) of users in a target system to detect or tolerate Sybil nodes. Albeit different in their working principle, all these approaches assume it is difficult for an attacker to create attack edges to connect Sybils with honest users. However, researchers have found that an attacker can employ simple strategies to obtain many attack edges. In this work we revisit the state-of-the-art, OSN-based Sybil defenses, and point out their strengths and weaknesses due to the impact of the new properties. We find these defense approaches are vulnerable to attackers under the new scenario, and in many cases a Sybil node only needs to obtain a handful of attack edges to disguise itself as a benign node.**

I. Introduction

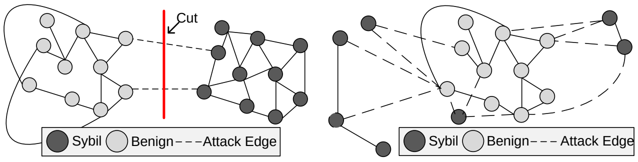
In the past decade, the research community has extensively studied a new kind of malicious behaviour. Introduced as the *Sybil attack*, the attacker tries to subvert a system by forging multiple identities. When orchestrating many forged identities, the attacker can manipulate the system in several dimensions. A good example are reputation or voting systems, in which the attacker-controlled identities can outvote the benign users.

One major concept which has attracted many researchers [2], [3], [6]–[8], [11] is to defend against such attacks by using information provided by the graphs of Online Social Networks (OSNs), i.e., the structure of social relations among participants of the network. The main idea is that identities controlled by the attacker will have difficulties establishing social relations with benign users. They reason that alongside these relations some sort of trust between both ends of the relation must exist, which is rarely found between Sybil and honest nodes. Thus, although there may be many such relations among the Sybil identities themselves, there should be only few from Sybils to the community of benign users. We call these edges *attack edges*. This intuition can be formalized in the following assumption:

Assumption 1: *Although an attacker can create an arbitrary number of Sybil identities in social network, she cannot establish an arbitrary number of attack edges to the densely connected non-Sybil identities.*

As a result of this assumption, the social graph is supposed to offer a *small cut* between both regions.

However, researchers have observed a variety of behaviours of both attackers and benign users which lead to a possible invalidation of Assumption 1. Attackers can easily create links to benign users by simply sending out link-establishing requests in OSNs. The success rates can reach 90% for specifically forged profiles or engineered bots [1], which enables attackers to create millions of attack edges [10]. Additionally, benign users are easily tricked to sending out requests to the forged identity with simple attacks [5]. Moreover, almost 75% of links originating at Sybils are connected to benign users, and not to Sybil nodes, which leaves the Sybil community structure not as densely connected as thought before [10]. To summarize these findings we propose a different assumption:



(a) Assumption 1 *(b)* Assumption 2

Fig. 1: Juxtaposition of Assumptions

Assumption 2: *Rather than connecting with other Sybils, an attacker is able to establish an increasing amount of social relations to benign users, and becomes more and more integrated within the community of benign users.*

Figure 1a shows how Sybil defense approaches picture the OSN graph. This graph provides an easily identifyable minimal-cut, which is defined by a few attack edges, between a benign region and a densely connected Sybil region. As a juxtaposition, Figure 1b shows a represenation of what recent research suggests is more likely.

II. OSN Based Sybil Defenses Under Pressure

Our work revisits seven major Sybil defense approaches that use social relations to prevent Sybil attacks. We analyze these approaches both qualitatively and quantitatively with regards to their efficiency under Assumption 2. Our key finding is that these approaches are indeed vulnerable to the changed scenario. The reason for this is that all schemes exploit the same structural properties: Most *Sybil detection* approaches employ some variation of a random walk on the graph to *identify* Sybil nodes. The main idea is that random walks are unlikely to traverse one of the few attack edges or—in other words— that Sybils are not well reachable from honest nodes. However, these approaches now face better connected Sybils than thought before, which leads to difficulties in distinguishing between honest and Sybil nodes.

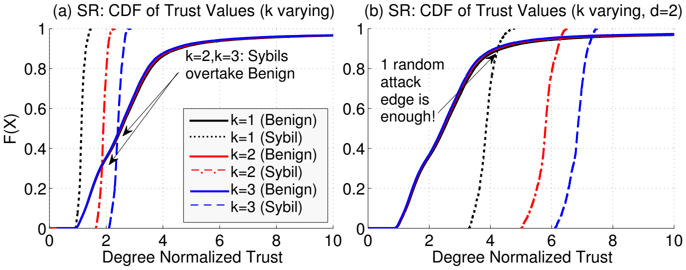


Fig. 2: SybilRank under Pressure

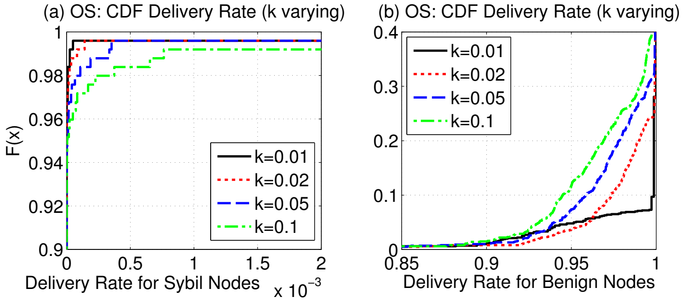


Fig. 3: Ostra under Pressure

Figure 2 shows these difficulties exemplary for SybilRank [2], based on experiments on a real-world Facebook dataset [9] as a CDF of the trust rankings. For SybilRank to work efficiently, the trust obtained by the honest nodes should clearly be greater than the trust obtained by Sybils (i.e., benign users should obtain better rankings than Sybils). In our study, we let each Sybil establish an increasing (small) amount of attack edges k towards benign nodes. While SybilRank can distinguish between both classes of nodes quite well if k does not exceed one randomly placed edge, it is not able to do so at $k = 2$ already (Figure 2a)). In this case, the Sybil CDF 'overtakes' the benign one, indicating mixed-up scores of the two classes. The situation gets worse when the attacker can place attack edges closer to the trust seeds of SybilRank. If it is able to place an edge in $d = 2$ hops away from such a seed, this one edge is already sufficient for the Sybil to disguise itself as an honest node. We observe similar problems in all other Sybil detection approaches under investigation.

Sybil tolerance approaches on the other hand are designed to *limit the impact* of possible Sybils in a system. The main idea of these schemes is to assign a certain capacity to each edge in the network and to subsequently penalize suspicious edges. For instance, the spam-prevention approach Ostra [6] assigns credits to links between users, where each link has two dependent credit values, one for each direction. Only if Ostra can find a path with available credit from a sender to the receiver, the message can be sent. Credit will be deducted from each traversed link on the path in the direction of message transmission, while the same amount of credit is added in the opposite direction. Ostra's feedback mechanisms ensure that only unwanted messages will have an effect on the credit balances. The main idea to limit the influence of Sybils is that the feedback on messages will quickly deplete the capacity on attack edges, leaving Sybils unable to distribute any spam afterwards. In our experiments, we measure the performance of Sybil tolerance schemes when facing a *relative* number of attack edges k in the system (e.g., $k = 0.01$: for 100 regular edges, one attack edge is created). Figure 3a shows that Ostra performs quite well with regards to spam prevention, as only a small fraction of spam eventually reaches its destiniation, regardless of the number of attack edges k. However, as other Sybil tolerance schemes, Ostra faces more specific issues. For instance, penalizing *whole* paths of spam messages punishes regular edges as well. Depending on how many attack edges the attacker can establish, this can lead to benign communication being blocked. Figure 3b shows that even a small $k = 0.01$ can lead to 5% of honest users not being able to send 5% of their messages. This effect multiplies the larger k grows.

III. Future Work

Our results show that current defense schemes may not provide the desired functionality in a new emerging threat scenario. In our future work, we are going to see whether simple modifications to the schemes are sufficient to re-enable their resiliency. Otherwise, we will research the creation of a Sybil resistant social graph based on different criteria. For instance, each link could be enriched with metadata describing the strength of the link. Attack edges may experience less communication between both ends and could further be of shorter life times [4]. A scheme, which goes beyond pure structural properties, should be more resilient—even if attackers can create vast amounts of attack edges.

References

[1] Bilge, L., Strufe, T., Balzarotti, D., and Kirda, E. All Your Contacts are Belong to us: Automated Identity Theft Attacks on Social Networks. In *WWW'09*.

[2] Cao, Q., Sirivianos, M., Yang, X., and Pregueiro, T. Aiding the Detection of Fake Accounts in Large Scale Social Online Services. In *NSDI'12* (2012).

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

[4] Gilbert, E., and Karahalios, K. Predicting Tie Strength with Social Media. In *Proceedings of the 27th international conference on Human factors in computing systems* (New York, NY, USA, 2009), CHI '09, ACM, pp. 211–220.

[5] Irani, D., Balduzzi, M., Balzarotti, D., Kirda, E., and Pu, C. Reverse Social Engineering Attacks in Online Social Networks. In *DIMVA'11* (2011).

[6] Mislove, A., Post, A., Druschel, P., and Gummadi, K. P. Ostra: Leveraging Trust to Thwart Unwanted Communication. In *NSDI'08* (2008).

[7] Tran, N., Li, J., Subramanian, L., and Chow, S. Optimal Sybil-resilient Node Admission Control. In *INFOCOM'11* (2011).

[8] Tran, N., Min, B., Li, J., and Subramanian, L. Sybil-resilient Online Content Voting. In *NSDI'09* (2009).

[9] Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. On the Evolution of User Interaction in Facebook. In *WOSN'09*.

[10] Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., and Dai, Y. Uncovering Social Network Sybils in the Wild. In *IMC'11* (2011).

[11] Yu, H., Gibbons, P. B., Kaminsky, M., and Xiao, F. SybilLimit: a near-optimal Social Network Defense against Sybil Attacks. *IEEE/ACM Trans. Netw. 18* (2010), 885–898.