SybilFuse: Combining Local Attributes with Global Structure to Perform Robust Sybil Detection

Extended Abstract

Peng Gao
Princeton University
pgao@princeton.edu

Binghui Wang
Iowa State University
binghuiw@iastate.edu

Neil Zhenqiang Gong
Iowa State University
neilgong@iastate.edu

Sanjeev R. Kulkarni
Princeton University
kulkarni@princeton.edu

Prateek Mittal
Princeton University
pmittal@princeton.edu

1 INTRODUCTION

Our systems today are vulnerable to Sybil attacks, in which an attacker injects multiple fake accounts into the system [6]. Recently, the increasing popularity of online social networks has made them attractive targets for Sybil attacks. The attacker can leverage Sybil accounts to disrupt democratic election and influence financial market via spreading fake news [9, 10], as well as compromise system security and privacy via propagating social malware, carrying out phishing attacks, and learning users' private data [1, 11].

An important thread of research proposes to mitigate Sybil attacks using social network-based trust relationships. The key insight is in a network where edges represent strong trust relationships, it is hard for the attacker to connect to benign users. As a result, the number of edges between benign users and Sybils (i.e., *attack edges*) is limited. Schemes such as SybilGuard [16], SybilLimit [15], SybilInfer [5], SybilRank [4], CIA [14], SybilBelief [8], and SybilSCAR [13] rely on such *strong trust assumption* and separate the benign and Sybil regions by identifying communities [12]. Íntegro [2] extends SybilRank by incorporating victim prediction (i.e., benign accounts that connect to Sybils) in random walks.

While these schemes have pioneered the use of social network structure for Sybil defenses, the actual deployment of these ideas in real-world networks remains controversial. First, real-world social networks do not necessarily have strong trusts. Previous works [1, 3] showed that Sybils could befriend benign users on Facebook at a large scale. Ghosh et al. [7] showed that on Twitter, a link farming phenomenon is widespread, in which certain benign accounts blindly follow back to accounts who follow them. On such weak trust social networks, the number of attack edges can be larger than what is typically considered in previous works, making it challenging to separate the benign region from the Sybil region. Second, Íntegro relies on the strong assumption that the number of victims is small and that the victims can be accurately predicted, which may not hold on certain real-world topologies like Twitter.

2 THE SYBILFUSE FRAMEWORK

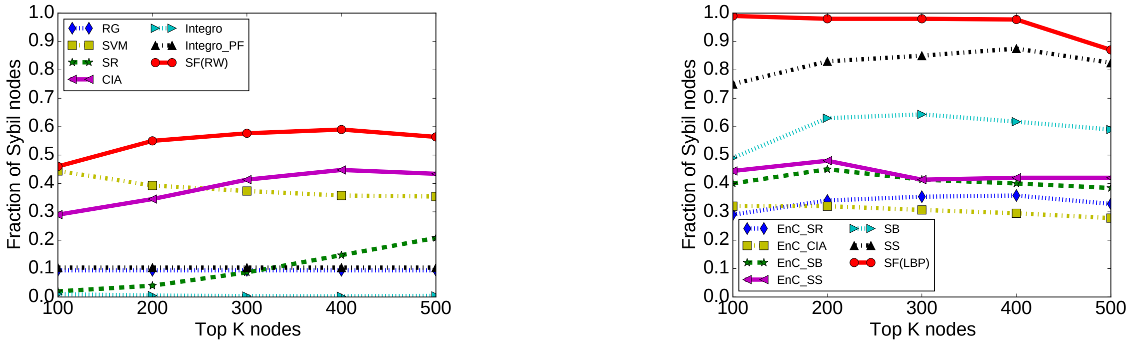
In this work, we propose *SybilFuse*, a defense-in-depth framework that leverages heterogeneous sources of information to perform robust Sybil detection. Different from existing approaches that assume strong trust networks [4, 5, 8, 13–16] or assume strong victim prediction [2], SybilFuse overcomes these limitations by adopting a collective classification scheme. Given social network data as input, SybilFuse first leverages *local attributes* to train local classifiers. Local node classifier predicts a trust score for each node, which indicates the probability of that node to be benign. Local edge classifier predicts a trust score for each edge, which indicates the probability of that edge to be a non-attack edge. These local trust scores are then combined with the *global structure* through weighted trust score propagation. Existing approaches do not leverage rich local information and treat edges equally, thus do not work well when the number of attack edges exceeds their assumption. In contrast, SybilFuse captures local account information in node trust scores, and propagates these scores through the global structure. During the score propagation, SybilFuse utilizes edge trust scores to enforce unequal weights, so that attack edges will have reduced impact on the propagation. After propagation completes, final trust scores of accounts are used for Sybil classification and ranking.

Given a social graph $G = (V, E)$, we denote S_v as the trust score of node $v \in V$, which quantifies the probability that v is benign. We denote $S_{u,v}$ as the trust score of edge $(u, v) \in E$, which quantifies the probability that node u and node v take the same label (i.e., homophily strength). The computation of S_v can be done through training local node classifiers (e.g., SVM, Logistic Regression) using various attributes (e.g., degree, clustering coefficient). The computation of $S_{u,v}$ can be done through training local edge classifiers or measuring the similarity between linked nodes using various similarity metrics (e.g., Cosine, Jaccard, Adamic-Adar).

Score propagation is done through either weighted random walk or weighted loopy belief propagation. In weighted random walk, the update equation is: $S^{(i)}(v) = \sum_{(u,v)\in E} S^{(i-1)}(u) \frac{S_{u,v}}{\sum_{(u,w)\in E} S_{u,w}}$, where initial scores $S^{(0)}(v)$ are set to be local node scores S_v. After $d = O(\log n)$ steps of power iteration (n is the number of nodes), we obtain the final node score: $S_v^F = S^{(d)}(v)$. In weighted loopy belief propagation, a pairwise Markov Random Field model is constructed by initializing node potentials ($\psi_v(X_v)$) and edge potentials ($\psi_{u,v}(X_u, X_v)$) with local trust scores. The message update equation is then adopted for $d = 5 \sim 10$ iterations to update belief

MIS2, 2018, Marina Del Rey, CA, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4



(a) Random walk-based approaches **(b) LBP-based approaches and ensemble methods**

Fig. 1: Fraction of Sybils among top K nodes of all evaluated methods.

scores $bel_v(X_v = x_v)$. The final node scores are then obtained by normalizing $bel_v(X_v = x_v)$: $S_v^F = \frac{bel_v(X_v=1)}{bel_v(X_v=1)+bel_v(X_v=-1)}$.

3 LABELED TWITTER EVALUATION

We evaluate SybilFuse against existing approaches in a real-world, labeled Twitter network obtained from [14]. Since it is easy for the attacker to manipulate one-way directed edges, we preprocess the original directed network to an undirected one by retaining an undirected edge if both directions exist. After preprocessing, the network consists of 8,167 nodes and 54,146 edges, with verified 7,358 benign nodes and 809 Sybil nodes. We observe that: (1) the number of attack edges is large (40,001), with around 49 attack edges on average per Sybil. (2) the number of victims is large (5,546), which takes more than 75% of benign nodes. Thus, the assumptions that previous approaches require are not satisfied.

We compute three discriminative node features using the original directed network: (1) incoming requests accepted ratio: $Req_{in}(v) = \frac{|In(v) \cap Out(v)|}{|In(v)|}$, where $In(v)$ ($Out(v)$) represents the set of all incoming (outgoing) edges of v; (2) outgoing requests accepted ratio: $Req_{out}(v) = \frac{|In(v) \cap Out(v)|}{|Out(v)|}$; (3) local cluster coefficient: $CC(v) = \frac{|\{(i,j):i,j \in Nei(v),(i,j) \in E\}|}{|Nei(v)|(|Nei(v)|-1)}$), where $Nei(v)$ represents the set of neighbors of v. We map these features to the corresponding nodes in the undirected network. We randomly sample 50 benign and Sybil nodes and train a SVM classifier with RBF kernel using *LIBSVM*. We then output probabilistic estimates as local node scores. Since extracting discriminative edge features from this dataset is difficult, we set local edge scores to be 0.9 by default to model homophily.

Evaluated Approaches: For SybilFuse, we propagate local scores through weighted random walks and loopy belief propagation (denoted by *SF(RW)*, *SF(LBP)*). We evaluate the following existing approaches: (1) node classification: *SVM*; (2) structure-based approaches: SybilRank (*SR*), CIA(*CIA*), SybilBelief (*SB*), SybilSCAR (*SS*); (3) extended random walks with victim prediction: *Íntegro*, *Íntegro_PF* (i.e., Íntegro under perfect victim prediction with 100% accuracy); (4) ensemble approaches: *EnC_SR*, *EnC_CIA*, *EnC_SB*, *EnC_SS*. We combine local SVM scores with structure propagation scores in a standard voting scheme; (5) random guess: *RG*.

Evaluation Results: We evaluate the ranking performance of these approaches by ranking all nodes in an ascending order of final scores. Better approaches will rank more Sybil nodes upfront. Fig. 1 shows the fraction of Sybils among top K nodes. We observe that: (1) *SF(RW)* achieves the best performance among all random walk-based approaches; (2) *SF(LBP)* achieves the best performance among all evaluated approaches (> 98% Sybil ranking up to top 400 nodes). By combining local attributes with global structure, SybilFuse significantly outperforms existing approaches, and a better way of combination is through weighted loopy belief propagation.

4 CONCLUSION

In this work, we propose SybilFuse, a framework for robust Sybil detection. Experiments on a real-world Twitter network demonstrate that SybilFuse outperforms existing approaches significantly.

ACKNOWLEDGEMENT

This work was supported in part by DARPA RADICS under contract FA-8750-16-C-0054, and the National Science Foundation under grants CNS-1553437 and CNS-1409415.

REFERENCES

[1] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. 2009. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In *WWW*.

[2] Yazan Boshmaf, Dionysios Logothetis, Georgos Siganos, Jorge Leria, Jose Lorenzo, Matei Ripeanu, and Konstantin Beznosov. 2014. Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs. In *NDSS*.

[3] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. 2011. The socialbot network: when bots socialize for fame and money. In *ACSAC*.

[4] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. 2012. Aiding the Detection of Fake Accounts in Large Scale Social Online Services. In *NSDI*.

[5] G. Danezis and P. Mittal. 2009. SybilInfer: Detecting Sybil Nodes using Social Networks. In *NDSS*.

[6] John R. Douceur. 2002. The Sybil Attack. In *IPTPS*.

[7] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen K. Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna P. Gummadi. 2012. Understanding and Combating Link Farming in the Twitter Social Network. In *WWW*.

[8] Neil Zhenqiang Gong, Mario Frank, and Prateek Mittal. 2014. SybilBelief: A semi-supervised learning approach for structure-based sybil detection. *IEEE TIFS* (June 2014).

[9] Hacking Election. 2016. (May 2016). http://goo.gl/G8o9x0

[10] Hacking Financial Market. 2016. (May 2016). http://goo.gl/4AkWyt

[11] Kurt Thomas, Chris Grier, Vern Paxson, and Dawn Song. 2011. Suspended Accounts in Retrospect: An Analysis of Twitter Spam. In *IMC*.

[12] Bimal Viswanath, Ansley Post, Krishna P. Gummadi, and Alan Mislove. 2010. An Analysis of Social Network-Based Sybil Defenses. In *SIGCOMM*.

[13] Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. 2017. SybilSCAR: Sybil detection in online social networks via local rule based propagation. In *INFOCOM*.

[14] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing Spammer's Social Networks for Fun and Profit. In *WWW*.

[15] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. 2008. SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks. In *IEEE S & P*.

[16] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. 2006. SybilGuard: Defending Against Sybil Attacks via Social Networks. In *SIGCOMM*.