A Comprehensive Framework for Detecting Sybils and Spammers on Social Networks

Lixin Fu
Department of Computer Science, University of North Carolina at Greensboro, Greensboro, NC 27412, U.S.A.

Keywords: Sybil Detection, Spam Detection, Social Networks.

Abstract: Social media becomes a common platform for millions of people to communicate with one another online. However, some accounts and computer generated robots can greatly disrupt the normal communications. For example, the fake accounts can simultaneously "like" or "dislike" a tweet, therefore, distort the true nature of the attitudes of real human beings. They collectively respond with similar or the same automate messages to influence sentiment towards certain subject or a tweet. They may also generate large amounts of unwanted spam messages including the irrelevant advertisements of products and services. Even worse, some messages contain harmful phishing links that steal people's sensitive information. We propose a new system that can detect these disruptive behaviours on OSNs. *Our methods* is to integrate several sybil detection models into one prediction model based on the account profiles, social graph characteristics, comment content, and user feedback reports. Specifically, we give two new detection algorithms that have better prediction accuracy than that of the state--of-the-art systems and real time performance. In addition, a prototype system including the software modules and real and synthetic data sets on which comprehensive experiments may confirm our hypothesis. Currently most sybil detection algorithms are based on the structural connections such as few connections of densely connected Sybil communities to normal nodes. Their detection accuracy is mixed and not well. Some algorithms are based on machine learning. The different approaches are separated. We expect our new model will more accurately detect the disruptive behaviour of fake identities with high positive rates and low false negative rates.

1 INTRODUCTION

Social media has gained tremendous growth in the past years. When the Pew Research Center started tracking in 2005, only 5% of American adults had used at least one platform but currently about 69% use some type of social media. Among the 1.65 billion active monthly Facebook users, about 75% visit daily. In 2016, Instagram added 100 million new users in six months on top of existing 500 million. Facebook raked in $9.16 billion advertising revenue in the second quarter, up 84% from the same quarter last year.

Nowadays OSNs (Online Social Networks) has become an essential part in our daily communications. Most popular OSNs include Facebook, WhatsApp, Wechat, Instagram, Tumblr, Twitter, Pinterest, LinkedIn, etc. By simply providing a valid email address, one can join these social networks. A user profile can contain many fields (e.g. Facebook has username, name, schools attended, work places, birth and residence locations, birthday, relationships, events, photos, videos, etc.) or few fields (e.g. Twitter has only name, username, Bio, location, website, photo, birthday). The date of account creation is recorded.

After joining the platform, one can then build associations. A social graph G(V, E), where V is the set of the accounts or users as nodes, and E is the set of associations as the edges. For some platforms the graph is undirected. For example, in Facebook, a user can send "friend" requests to other users who can accept or decline. Once a request is accepted then an edge between the two nodes is added to the graph. The association can be revoked by either side at any time. Fig. 1. shows an example of a portion of the graph.

The social graphs for some platforms are directed. For example, a Twitter user can follow any other user without the need to request and accept. Afterwards, the user can "unfollow" or block any of her followers. Fig. 2 shows an example digraph of Twitter.

Fu, L.
A Comprehensive Framework for Detecting Sybils and Spammers on Social Networks.
DOI: 10.5220/0006800602290236
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018)*, pages 229-236
ISBN: 978-989-758-298-1

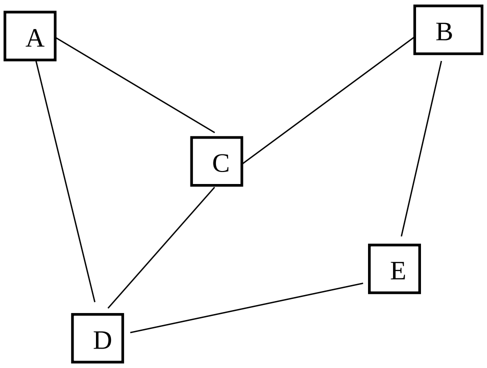

Figure 1: A sample social graph of Facebook.

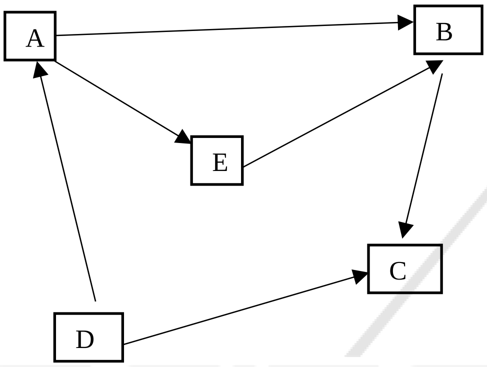

Figure 2: A sample social graph of Twitter.

The times of the associations (follow, request, accept) can be recorded into the vendor's databases so that one can research on the time patterns of these events. Once the associations are established, the users can post messages which will appear in the timelines of the followers or friends. They can also appear at the timelines of the mentioned users. Each post or twit can be reposted, liked, commented/replied. Other users can also report any abusive behaviour such as spamming, harassment, discrimination, and violence.

Albeit great convenience and popularity, the social networks are attacked by all kinds of disruptive behaviours. DoS (Denial of Service) (Gao, 2010) attack attempts to consume the victim's machine resources so that the intended user is unable to perform normal work. The attacks can be originated from multiple sources composed of botnets. The phishing attacks integrate masqueraded links into posts that invite users to input log-in credentials and intercept them. Once username and password are obtained, they will be used to hijack or compromise user accounts. From the compromised accounts, the posts containing phishing or advertisement links are further spread. In online voting, surveys, and recommendation systems, the designed behaviour of sybils may greatly corrupt the results. On Youtube network, users can comment and reply to comments. If a user repeatedly sends long comments, or a large group of sybils send many similar comments, the other comments would be drowned and be hard to find.

Social spam is unwanted content showing on social networks. First, the commercial spams contain large amounts of advertisements for selling products or services that are irrelevant to current discussion of topics. Second, fake accounts called spambots, botnets, or sybils are created to increase the number of followers or friends for a better credibility. It is estimated that as many as 40% of social network accounts are fake (Kharif, 2016). The fake accounts often follow the celebrities and public figures. If they follow back, the spammer will gain credibility and influence. The fake accounts or real but compromised accounts may launch coordinated attacks. For example, they may twist the number of "likes" or "dislikes." They may submit repeated comments with the same or similar texts. They may write reviews to products or services that are not sincere or true, thus misleading normal customers. Third, the comments may contain insulting aggressive languages referred as "cyberbullying", hate speech, threats, and exposure of personally identifiable information.

This paper is to develop new efficient, more accurate sybil detection algorithms that can combine the advantages of existing detection algorithms targeting only a portion of social network data. Various sybil detection algorithms can be categorized into four types: profile-based (P), graph-based (G), content-based (C), and user feedback report-based (R). They explore the sybil patterns according to the profile features, social graph structures, comment content, and users' reports. Currently, these algorithms are separated and reflect sybil characteristics only from certain aspect. Our *rationale* is that the new detection system may integrate these otherwise unrelated technologies into one predictive model. Our hypothesis is that by combining the best technologies on different parts of the social network data the new synopsis system will have a better prediction accuracy in terms of higher positive rate and lower negative rate. Specifically, we propose the following:

- **Investigate the state-of-the-art detection algorithms or develop new detection algorithms that focus only on user profiles, social graphs, contents, or user feedback.** From each category we choose or modify a top algorithm or develop a new one if necessary. The results of this aim will be integrated in the next phase.

- **Develop a new weighted linear model that combines the prediction results from the previous phase**. The prediction results of the component algorithms are normalized into a fraction number in [0,1], representing the likelihood of being a Sybil.
- **Develop a new voting mechanism to derive a comprehensive prediction**.

Our work creates a new family of algorithms that can better predict disruptive behaviour in OSNs, thus greatly enhancing the security of widely used social media.

The rest of the paper is organized as follows. Section 2 surveys related work in the field. Section 3 provides detailed description of our disruption detection algorithms and system, respectively. Experiment design is presented in Section 4. In Section 5, we conclude the work.

2 RELATED WORK

The sybils tend to "follow" each other to create the feeling of that they have a few followers. Since we are uncertain if a user is a sybil, we can assign a probability of being a sybil to any user. They network will form an embedded probability network as well.

The graph-based algorithms mostly use random-walk and mixing time to detect sybils. The sybils are false identities created by computers or software to disrupt (Douceur, 2002). A common assumption of these algorithms is that the normal nodes are fast mixing and tightly connected, and the sybils tend to be mutually connected while only few edges called "attack edges" connect the normal region and sybil region. Among these algorithms are SybilGuard (Yu, 2008), SybilLimit (Yu, 2010), SybilInfer (Danezis, 2009), GateKeeper (Tran, 2011), SybilShield (Shi, 2013), SybilDefender (Wei, 2013), SybilFence (Cao, 2013), SumUp (Tran, 2009).

Through random routes in social networks, SybilGuard (Yu, 2008) exploits the small set of attack edges to bound the number of sybil identities allowed to be created, and the size of each sybil group so that the malicious influence is limited. However, SybilGuard suffers from the high false negative rates. In comparison with SybilGuard, SybilLimit (Yu, 2010) reduced the number of sybil nodes accepted by a factor of Θ (n $^{1/2}$) and achieved a near optimal guarantee. One weakness of SybilLimit is that it relies on the length parameter *w* of random walk, whose correct estimation is critical to security bounds.

SybilInfer (Danezis, 2009) uses Bayesian inference to compute the probability of a node being a sybil node. Different from SybilGuard and SybilLimit which are decentralized systems, SybilInfer is a centralized approach. The time complexity of SybilInfer is O(V^2logV), where *V* is the number of nodes in the social graph. This scheme is not feasible for many real OSNs with large sizes. SybilShield (Shi, 2013) greatly improved the false positive rates against the sybil attacks in OSNs with multiple communities by using agent nodes.

As a centralized mechanism, SybilDefender (Wei, 2013) limits the number of random walks thus achieving better time performance and scalability. The sybil node detection and sybil community detection algorithms limit the number of attack edges by relationship rating.

Based on landing probability of random walks, SybilRank (Cao, 2012) is a sybil-likelihood ranking scheme for an OSN user. It handles multi-communities by applying multiple seeds. It terminates after O (log n) steps. The time complexity of SybilRank is O(nlogn). SybilFence (Cao, 2013) limits the social edges on the users that have negative feedback.

Z. Yang et al (Yang, 2014) find that on RenRen, a Chinese social network, the previous assumption of the sybil topology is not correct. The sybils actually do not mutually connect to a tightly-knitted community so much. The edges among the sybils are accidental. Instead, they develop a threshold-based simple detector that is very effective with low false positive rates.

Clustering and latent variables are used to detect spammers (Yanbin, 2010, Mukherjee, 2013). Gao et. al (Gao, 2010) give a clustering algorithm for similar posts in Facebook and incorporate multiple validation technologies such as URL de-obfuscation, re-direction, blacklisting, wall post keyword search, and URL-grouping, to detect and characterize OSN spam campaigns. Their experiments on Facebook include a data set that contain a sizable 187 million wall posts from 3.5 million distinctive users. Over 70% of the 200,000 malicious attacks are phising while majority of the 57,000 malicious accounts are compromised ones instead of sybils. The attacks are mostly bursty.

With mixed labelled and unlabelled data, semi-supervised learning can be used to detect sybils. SybilBelief (Gong, 2014) relies on belief propagation and random Markov fields.

Wang et. al (Wang, 2013) gives a click stream model at the server side to detect sybils that may have different clicking behaviour from that of normal users. Measured by the distances between two different click streams, similarity weights are calculated, and the graph is partitioned into clusters. The experiments on data sets from RenRen and LinkedIn validate the high detection accuracy of the new click stream method. They use Support Vector Machine (SVM) to learn important features that differentiate from the click streams of Sybil and real users.

G. Wang et. al (Wang, 2012) employ human assessments to develop a crowdsourcing based detection method. They collect datasets from both RenRen and Facebook. Their experiments show that filters opinioned by experts as ground truth can winnow off turkers with low accuracy.

SybilExposer (Misra, 2016) includes a community extraction algorithm and community ranking algorithm to detect potentially multiple sybil communities from honest communities where the sybil communities have much lower ratio of inter-community degree to intra-community degree. The time complexity of SybilExposer is O (n + c log c), where *n* is the number of nodes in the social graph and *c* is the total number of extracted communities. Their experiments on selected data sets show that SybilExposer has a higher positive rate and a lower negative rate than SybilDefender, SybilShield, and SybilRank have.

UNIK (Tan, 2013) is an unsupervised spam detection algorithm that can deal with increased levels of spam attacks. It first uses a SD2, a community-based sybil detection algorithm to detect spammers on the social graph, to form a whitelist of URLs shared by non-spammers. It then trims the matching whitelist edges in the user link graph and treats the remaining high-degree nodes as spammers. Based on experiments on a data set with 176-thousand users collected within 10-month period, UNIK has a lower false positive rate and lower false negative rate than those of AutoRE (Xie, 2008) and FBCluster (Gao, 2010). UNIK is suitable for detecting heavy spam attacks of groups, forums or blogs but not for attacks from compromised accounts.

Zeng (Zheng, 2015) gives a supervised algorithm SVM (Support Vector Machine) to detect spammers. Experiments on data sets with 30 thousand users and 16 million messages on Sina Weibo, a Chinese Social Network, show that it has better accuracy than other classifiers such as decision trees and Bayesian network. For user features, spammers tend to have fewer followers, more messages per day, more URLs per messages on average, and shorter life. From content perspective, spam messages have fewer reposts, likes, mentions, hashtags, and comments. Rather than detecting from only certain aspect or source of data, we propose a comprehensive detection algorithm that can combine the results of multiple detectors, thus improving the prediction accuracy.

3 DISRUPTION DETECTION ALGORITHMS AND SYSTEM

First, we need to develop the detection algorithms that are based only on user profiles, social graphs, content, and user feedback reports separately.

3.1 Profile-based Approach

The research problems of detection algorithms based on profiles are:

- What features are more relevant to abnormal detection?
- Which classification algorithm to use?
- How to get or estimate the real normal or abnormal labels also known as "ground truth?"

According to our related work research, we plan to choose the following features in the profiles:

- Name/username of the account. The spammers tend to use names that are similar to existing celebrity names who have large number of friends or followers. This is understandable. A familiar name or public figure has already earned public trust that helps to attract new friends or followers for spamming or other fraudulent purposes. From the social network data set, we can either choose top *k* account names with the highest node degrees in the social graph, or choose the names of the accounts with a degree at least *d*. We can measure the name similarity between the account name and any chosen celebrity account name either by editing distance or if a name is a substring of another.
- Profile photo. We may use the third-party image matching system to compare the profile photo if any to those of the celebrity accounts.
- Acceptance request ratio. Spam accounts intend to make a lot of requests but gain few acceptances because people hesitate to accept strangers. A low ratio raises a red flag. For OSNs with directed social graph

e.g. Twitter we can use the followers to following ratio.

- Number of friends or followers. Sybil accounts usually have a low degree of associations.
- Account age. Spammers usually have shorter lives because they may be detected and blocked.

As for the classification algorithms, we plan to compare SVM with RainForest in terms of performance and prediction accuracy and then choose a better one. To label the accounts as spam or not spam, we plan to extract a random sample and let experts manually assign classes. Alternative plan is to treat the blocked accounts as spam accounts and others as normal ones.

3.2 Graph-based Approach

The research problem here is how to detect sybil communities from regular communities composed of real users from the social graph. As we have pointed in the second section, this type of algorithms assumes that the sybil communities tend to have links with one another to form an impression of having friends or followers; regular communities also tend to have natural intra-community connections, but there are few attack edges between sybil nodes and normal nodes. Fig. 3 shows an example of a social graph that has multiple sybil and normal communities.

We plan to use a clustering algorithm SybilExposer (Misra, 2016) to extract and rank the communities.

Initially each node itself is a community. The algorithm iteratively merges nearby nodes into a larger community so that the partition modularity is increased. The merging process continues until the cardinality cannot further be reduced. In the second phase, the communities are sorted according to their probabilities of being sybil. The ratio of inter-community to intra-community is used as ranking index. The sybil communities have lower ratios.

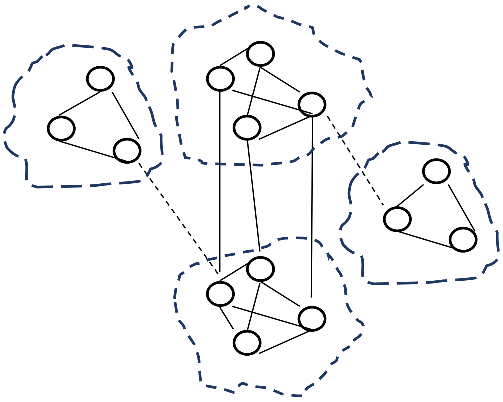

Figure 3: Multiple sybil communities and multiple normal communities.

3.3 Content-based Approach

OSNs allow users to post comments and reply comments. This is the main place where the spammers put their messages and embedded links to achieve their disruptive purposes. The research problem is how to detect spam messages and underlying spam accounts according to the content of the comments. This is related to text analysis. Spam account often posts the same messages repeatedly or the sybil accounts post the same or similar messages in coordination to bombast the communities. They post more frequently than normal users and the messages often contain redirection links that lead to other pages for product or service promotion, or pages to phish account identities. We plan to focus on the following features:

- Number of message per day
- Post times. The spam messages are often bursty. If many messages are posted in sleep hours, they may be spammers.
- Number of URLs per message on average
- Message similarity. The challenge is that the number of messages is very large and pair-wise comparison will take long time to find the clusters of messages that share the same URL link or have similar content in the messages. We can form a content graph where the nodes are messages and if two messages are similar, an edge between them is added. Once one message is found to be spam, then we may find a community of spam accounts.
- Filtering words. Spam or harmful messages may contain aggressive words, abusive words, porn words, sexual assault words, hate speech, discriminative words of race, religion, sex orientation, etc. Such words frequently shown in spam messages are collected automatically or manually into a list of filtering words.
- Blacklists of URLs or URL patterns.

Other content features related to each message have user feedback statistics:

- Number of likes (or dislikes)
- Number of reposts (or re-twits, shares)

- Number of comments
- Number of mentions

These statistics also give a hint of whether a message is spam or underlying account is sybil since a software generated message may not attract interest from people thus have fewer responses.

3.4 User Feedback Reports

Many OSNs such as Facebook and twitter allow users to report problems in the following categories:

- Spam content
- Not interested
- Sensitive image
- Harmful or abusive
- Discrimination against race, gender, orientation, religion
- Threat, assault, violence, suicide
- Offending, disrespectful, pornography
- Private information

For each message, if any, the number of times reported in each of the problem categories given above is recorded. For each account, the reported sample messages are also stored and counted. The research problem is to build a model to compute the likelihood of being a spam message or a spam account based on the reports.

3.5 Weighted Linear Model

From sections 3.1 to 3.4, we choose or develop the state-of-the-art detection algorithms called Alg_P, Alg_G, Alg_C, Alg_R based on the user profiles, social graph, message content, and user reports respectively. Suppose Pr_P, Pr_G, Pr_C, and Pr_R are the corresponding Sybil probabilities from these algorithms. We can construct a linear model to predict the final probability *Pr* of being a spam or sybil account as follows:

$$Pr = W_P\, Pr\_P + W_G\, Pr\_G + W_C\, Pr\_C + W_R\, Pr\_R \;,$$

where W_P, W_G, $W_{C,}$ W_R are the weights of corresponding probabilities. The weights reflect the preferences to each algorithm. Fig. 4 shows the architecture of this new detection model.

3.6 Voting Mechanism

An alternative of weighted linear model for sybil detection is to establish a voting mechanism if the output of the above four algorithms is a binary decision (true or false) instead of probabilities. The first strategy is strict: whenever any one of the four detectors predicts true, the final verdict is true. The advantage of this scheme is that the positive rate is high i.e. the true sybils will be recognized. However, this may lead to higher false alarms i.e. the normal accounts are mistakenly labelled as sybils. The second strategy is to use majority vote: true if at least three detectors vote true. The architecture of this approach is similar to Fig. 4 except that the last module is a voting model. If the outputs of the above four algorithms are fraction probabilities, a cut-off (above 0.5) is regarded as true otherwise false. One of the two strategies is then used to compute the final prediction.

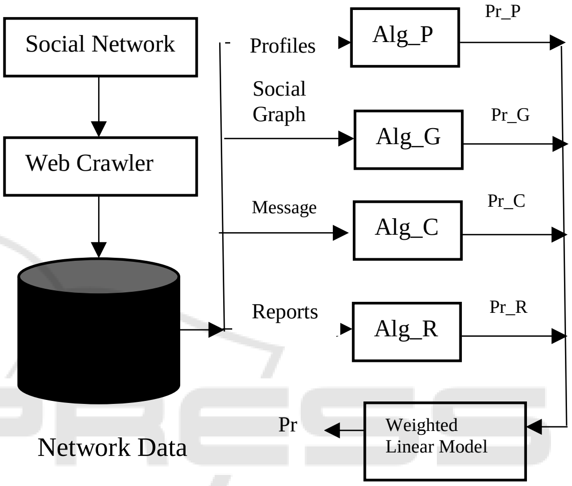

Figure 4: Architecture of weighted linear detection.

4 EXPERIMENT DESIGN

4.1 Data Set and Feature Collection

Most social media platforms have a public development API that allows users to download the messages and various kinds of statistics. We need to develop a web crawler to browse the social network graph systematically to collect user profiles. We plan to use BFS (breadth first search) algorithm to perform graph traversal from different seeds scattered from different geographical areas. Once the crawler has collected the data, we organize into tables which include total number of nodes, number of edges, number of messages, number of clusters, etc. We also experiment on the real data sets available in research community as well as synthetic data sets that have controllable parameters.

4.2 Evaluation Criteria

The main criteria to measure the effectiveness include precision, recall, F-measure. Suppose *a* is the number of spammers correctly labelled and *b* is the number of spammers mis-labelled as normal. Value *c* is number of non-spammers classified as spammers (false alarm) and *d* is number of non-spammers correctly classified. Then precision p is a/(a+c) and recall r is a/(a+b). The F-measure is 2pr/(p+r). True positive rate measures and false negative rate are other criteria. We also compare the runtimes of the algorithms.

4.3 Automatic Generation of Sybils and Simulation of Group Attacks

Of course, we can manually create some valid emails and use them to join the social network platforms to create some accounts we can control. The problem is that the number of sybil accounts is limited and time consuming. We need automate this procedure to generate large number of controllable accounts. The vendors may have restrictions or are able to detect these events. More research is required.

5 CONCLUSION

In this work we have proposed and investigated a comprehensive disruption detection technique that can integrate the otherwise separate detection strategies that use only part of available information. We propose two such schemes: weighted linear model and voting model. A system architecture that supports the two new algorithms is also proposed. Though the final detailed implementation and experiments are not completed, we expect our new exploratory algorithms and prototype will greatly advance the disruption detection technology.

REFERENCES

Douceur, J. R., Druschel, P., 2002. "The Sybil attack" in Peer-to-Peer Systems, Heidelberg, Germany:Springer, pp. 251-260, 2002.

Yu, H., Kaminsky, M., Gibbons, P. B., Flaxman, A. D., 2008. "SybilGuard: Defending against Sybil attacks via social networks", *IEEE/ACM Trans. Netw.*, vol. 16, no. 3, pp. 576-589, Jun. 2008.

Yu, H., Gibbons, P. B., Kaminsky, M., Xiao, F., 2010. "SybilLimit: A near-optimal social network defence against Sybil attacks", *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 885-898, Jun. 2010.

Danezis, G., Mittal, P., 2009. "SybilInfer: Detecting Sybil nodes using social networks", *Proc. NDSS*, pp. 1-15, 2009.

Tran, N., Li, J., Subramanian, L., Chow, S. S., 2011. "Optimal Sybil-resilient node admission control", *Proc. IEEE INFOCOM*, pp. 3218-3226, Apr. 2011.

Shi, L., Yu, S., Lou, W., Hou, Y. T., 2013. "Sybilshield: An agent-aided social network-based Sybil defence among multiple communities", *Proc. IEEE INFOCOM*, pp. 1034-1042, Apr. 2013.

Wei, W., Xu, F., Tan, C. C., Li, Q., 2013. "SybilDefender: A defense mechanism for Sybil attacks in large social networks", *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 12, pp. 2492-2502, Dec. 2013.

Cao, Q., Yang, X., 2013. "Sybilfence: Improving social-graphbased Sybil defenses with user negative feedback", 2013, [online] Available: https://arxiv.org/abs/1304.3819.

Tran, D. N., Min, B., Li, J., Subramanian, L., 2009. "Sybil-resilient online content voting", *Proc. NSDI*, pp. 15-28, 2009.

Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., Dai, Y., 2014. "Uncovering social network Sybils in the wild", *ACM Trans. Knowl. Discovery Data (TKDD)*, vol. 8, no. 1, 2014.

Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B. Y., 2010. "Detecting and characterizing social spam campaigns", *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, pp. 35-47, 2010.

Yanbin, Z., 2010. "Detecting and characterizing spam campaigns in online social networks" in Fall LERSAIS IA Seminar, Pittsburgh, PA, USA, 2010.

Mukherjee, A. et al.,2013. "Spotting opinion spammers using behavioral footprints", *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, vol. 2013, pp. 632-640.

Gong, N. Z., Frank, M., Mittal, P., 2014. "SybilBelief: A semi-supervised learning approach for structure-based Sybil detection", *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 6, pp. 976-987, Jun. 2014.

Cao, Q., Sirivianos, M., Yang, X., Pregueiro, T., 2012. "Aiding the detection of fake accounts in large scale social online services", *Proc. 9th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, pp. 197-210, 2012.

Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., Zhao, B. Y., 2013. "You are how you click: Clickstream analysis for Sybil detection", *Proc. USENIX Secur.*, pp. 1-15, 2013.

Wang, G. et al., 2012. "Social turing tests: Crowdsourcing Sybil detection", 2012, [online] Available: https://arxiv.org/abs/1205.3856.

Misra, S., Tayeen, A. S. M., Xu, W., 2016. "SybilExposer: An effective scheme to detect Sybil communities in online social networks", *Proc. IEEE Int. Conf. Commun. (ICC)*, pp. 1-6, May 2016.

Tan, Enhua, Guo, Lei, Chen, Songqing, Zhang, Xiaodong, and Zhao,Yihong., 2013. UNIK: unsupervised social network spam detection. In *Proceedings of the 22nd*

ACM international conference on Information & Knowledge Management (CIKM '13). ACM, New York, NY, USA, 479-488.

Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G., and Osipkov, I., 2008. Spamming botnet: Signatures and characteristics. In Proc. of SIGCOMM, 2008.

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

Kharif, O., 2016. "'Likejacking': Spammers Hit Social Media". Businessweek. Retrieved 2016-08-05.

Zheng, X., Zeng, Z., Chen, Z., Yu, Y., Rong, C., 2015. Detecting spammers on social networks, Neurocomputing, 2015 - Elsevier

Timm, Carl and Perez, Richard, 2010. Seven Deadliest Social Network Attacks, ISBN: 978-1-59749-545-5, 2010 Elsevier Inc.