

Identification of Sybil Communities Generating Context-Aware Spam on Online Social Networks

Faraz Ahmed and Muhammad Abulaish

Center of Excellence in Information Assurance
King Saud University, Riyadh, Saudi Arabia
{fahmed.c, mabulaish}@ksu.edu.sa

Abstract. This paper presents a hybrid approach to identify coordinated spam or malware attacks carried out using sybil accounts on online social networks. It also presents an online social network data collection methodology, with a special focus on Facebook social network. The pages crawled from Facebook network are grouped according to users' interests and analyzed to retrieve users' profiles from each of them. As a result, based on the users' page-likes behavior, a total number of six groups has been identified. Each group is treated separately and modeled using a graph structure, termed as *profile graph*, in which a node represents a profile and a weighted edge connecting a pair of profiles represents the degree of their behavior similarity. Behavior similarity is calculated as a function of *common shared links*, *common page-likes*, and *cosine similarity of the posts*, and used to determine weights of the edges of the profile graph. Louvain's community detection algorithm is applied on the profile graphs to identify various communities. Finally, a set of statistical features identified in one of our previous works is used to classify the obtained communities either as malicious or benign. The experimental results on a real dataset show that profiles belonging to a malicious community have high closeness-centrality representing high behavioral similarity, whereas those of a benign community have low closeness-centrality.

Keywords: Social network analysis, social network security, sybil community detection.

1 Introduction

Online social networking sites have attracted a large number of internet users. Among many existing Online Social Networks (OSNs), Facebook and Twitter are the most popular social networking sites with over 800 million and 100 million active users, respectively. However, due to this popularity and existence of a rich set of potential users, malicious third parties have also diverted their attention towards exploiting various features of these social networking platforms. Though, the exploitation methodologies vary according to the features provided by the social networking platforms, malware infections, spam, and phishing are the most common security concerns for all of these platforms. In addition, a number of social botnets have emerged that utilize social networking features to

spreading infections as command and control channels [1], [2], [3]. The root cause of all these security concerns is the social network sybils or fake accounts created by malicious users to increase the efficacy of their attacks that are commonly known as *sybil attacks* [4]. Generally, an attacker uses multiple fake identities to unfairly increase its ranking or influence in a target community. Moreover, several underground communities exist, which trade sybil accounts with users and organizations looking for online publicity [5], [6]. Recent studies have shown that with the increase in the popularity of social media, sybil attacks are becoming more widespread [7]. Several sybil communities have been reported so far that forward spam and malwares on Facebook [8] network. In online social networks, third-party nodes are most vulnerable to sybil attacks, where the third-party nodes are communities and groups on OSN platforms which bring together users from different real-world communities on the basis of their interests. In case of Facebook, a third-party node can be defined as a Facebook community page which is used to connect two users from entirely different regions. Sybil accounts hired for carrying out spam campaigns target such vulnerable nodes. Recently, the rapid increase in the number of spam on popular online social networking sites has attracted the attention of researchers from security and related fields.

Though a significant amount of research works has been reported for the detection and characterization of spam on Facebook and Twitter networks [9], [10], [11], [12], [13], the existing techniques do not focus on the detection of coordinated spam campaigns carried out by the communities of sybil accounts. Similarly, several techniques have been presented for the identification of sybil communities [4], [14], [15],[16], [17], but all of them focus on the decentralized detection of sybil accounts. Moreover, the existing techniques are based on two common assumptions about the behavior of sybil nodes. Firstly, sybil nodes can form edges between them in a social graph and secondly, the number of edges connecting sybil and normal nodes is less as compared to the number of edges connecting either only normal nodes or only sybil nodes. These assumptions were based on the intuition that normal users do not readily accept friendship requests from seemingly unknown users. Although empirical studies from [17] showed existence of such sybil communities in the Tuenti social network, another study of Renren social network [7] showed that sybil nodes rarely created edges between themselves. This implies that the community behavior of sybil nodes in a social graph is mercurial and the assumption that sybil nodes form communities cannot be generalized [18].

In this work, the authors utilize the rich corpus of prior research works on spam detection and sybil community identification as a basis and present a hybrid approach to identify coordinated spam or malware attacks carried out using sybil accounts. The proposed approach is independent of the assumptions discussed above by the previous researchers. Although the proposed approach is generic in nature, this paper focuses on the sybil accounts present on Facebook social network for experiment and evaluation purposes. The contributions of this paper can be summarized as follows:

- An online social network data collection methodology is introduced which is based on the intuition that sybil accounts under the control of a single user tend to attack different nodes of the same community; they may not be connected to each other, but may have a common target.
- A new social graph generation mechanism is presented, in which a node represents a profile and an edge represents an association between a pair of connecting profiles. The weight of an edge is determined as a function of the features extracted from the content of linked profiles. In this way, the weight of an edges is independent of the actual friendship link between the profiles, and consequently profiles with similar behavior are interlinked together to form a single group.
- Each group of related profiles is modeled as a social graph and analyzed independently using a community detection algorithm.
- A statistical approach is applied on the obtained communities from each profile group to identify sybil communities.

The rest of the paper is organized as follows. After a brief review of the existing state-of-the-art techniques for spam identification in online social networks in Section 2, Section 3 presents a data collection methodology from Facebook social network. Section 4 presents the profile grouping methodology to generate various groups of similar profiles in the original social network dataset. This Section also presents the experimental results obtained from a real dataset and their analyses. Finally, Section 5 presents conclusions.

2 Related Work

A significant number of research works has been reported in last few years for spam detection on online social networks. In [19], the authors proposed a real time URL-spam detection scheme for Twitter. They proposed a browser monitoring approach, which takes into account a number of details including HTTP redirects, web domains contacted while constructing a page, HTML content being loaded, HTTP headers, and invocation of JavaScript plug-ins. In [11], the authors created honey-profiles representing different age, nationality, and so on. Their study is based on a dataset collected from the profiles of several regions, including USA, Middle-East, and Europe. They logged all types of requests, wall posts, status updates, and private messages on Facebook. Based on the users' activities over social networking sites, they distinguished spam and normal profiles. The authors in [12] utilized the concept of *social honeypot* to lure content polluters on Twitter. The harvested users are analyzed to identify a set of features for classification purpose. The technique is evaluated on a dataset of Twitter spammers collected using the *@spam* mention to flag spammers. In [8], the authors analyzed a large dataset of wall posts on Facebook user profiles to detect spam accounts. They built wall posts similarity graph for the detection of malicious wall posts. Similarly, in [13] the authors presented a thorough analysis of profile-based and content-based evasion tactics employed by Twitter spammers. The authors proposed a set of 24 features consisting of graph-, neighbor-,

automation-, and timing-based features that are evaluated using different machine learning techniques. In [20] and [10], the authors proposed combination of content-based and user-based features for the detection of spam profiles on Twitter. In order to evaluate the importance of these features, the collected dataset is fed into traditional classifiers.

Similar to spam detection on online social networks that has received a lot of attention from researchers, a significant effort have been diverted towards the detection of sybil accounts. Initial studies [4], [14], [15], [16], [7] focus on detecting sybil users. However, individual users do not pose a great threat to normal users of OSNs. The situation becomes alarming when a large number of sybil accounts generate a coordinated attack. Several techniques have been presented to detect groups of accounts coordinating with each other [4], [14], [15],[16], [17]. All these techniques focus on the decentralized detection of sybil accounts. Moreover, they are based on two common assumptions: i) sybil nodes can form edges between them in a social graph, and ii) the number of edges connecting sybil and normal nodes is less as compared to the number of edges connecting either only normal nodes or only sybil nodes. However, later studies have shown that these assumptions cannot be applied in general [7]. Despite the presence of rich amount of works for spam and sybil detection, there has been little attention towards the identification of sybil accounts that are particularly responsible for spam proliferation. Therefore, this paper focuses on the detection of coordinated spam campaigns that are carried out by sybil accounts under the control of a single user.

3 Dataset

Based on the analyses reported in [21], it is found that a significant amount of spam posts on Facebook are directed towards those Facebook pages that are publicly accessible and any user in the network can post on them. Spammers generally utilize such openly accessible public pages to spread spam in the network. This type of spam spreading mechanism not only relieves the spammers from their dependence on friendship requests, but also increases the number of target users. Once a spam post is visible on a page’s wall, it can be visible to every user who *likes* that page. In addition, users’ page-like information can help spammers to spread context-aware spam through Facebook pages in normal user communities. Recently, there has been an increasing number of evidence about the existence of underground communities trading groups of accounts that carry spam campaigns [18]. Therefore, a group of accounts bought by a party could be used for a single purpose, resulting in a high correlation in their behavior.

This work exploits the intuition that a spam targeting a community is most likely the spam generated by a community. A dataset [21] containing Facebook spam profiles is analyzed to identify Facebook pages that have been mostly targeted by spammers. As a result, a total number of 14 Facebook pages is found that are heavily spammed by the spam profiles identified in [21]. All these pages are accessed to identify active users and to group profiles based on the

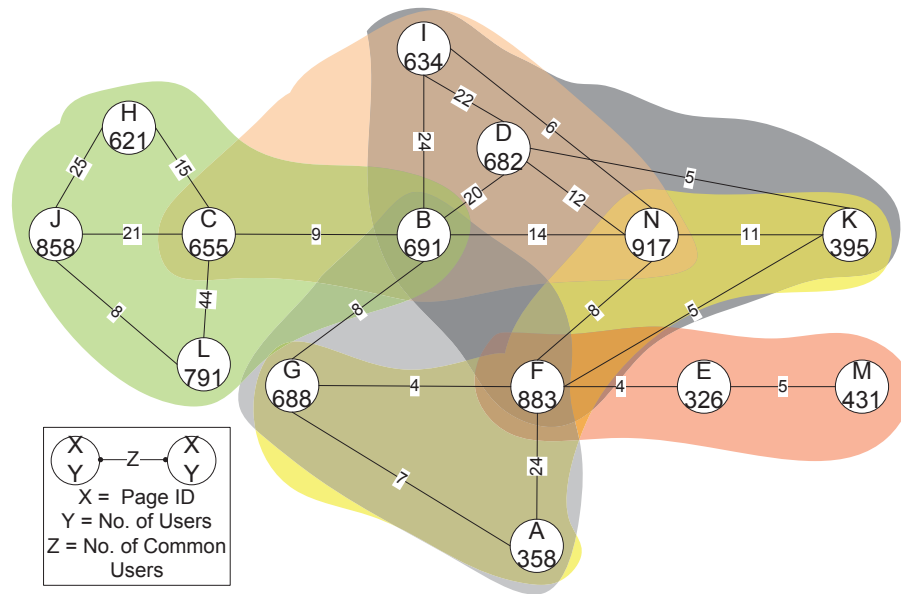


Fig. 1. Graphical illustration of Facebook pages and users

number of their common pages-likes. Figure 1 shows a graph in which each node represents a Facebook page and the weight associated to an edge represents the number of users commonly shared by the connected pair of nodes. The weights of the edges can be used to divide the users into various groups based on their interests in the network. In Figure 1, there are six groups of pages that are close to each other in terms of their common interests, and each group is treated separately for detection of profiles that are under the control of a single spammer and generate context-aware spam towards a community of normal users. Table 1 shows the various groups and the number of users belonging to each of them. The names of the groups in Table 1 has been derived from the node levels used in Figure 1. The next Section explains the methodology used to identify the groups of sybil accounts.

Table 1. Various profile groups along with the number of users

Groups	FEM	BGFA	BNDIC	AFGNK	JHLCB	DBINKF
n	1	2	3	4	5	6
No. of users	1631	2575	3465	3166	3482	4059

4 Methodology

To detect communities of sybil accounts generating context-aware spam, the rich amount of textual information contained in each profile is used to generate an undirected-weighted social graph, in which a node represents a profile and an edge connecting a pair of nodes represents a link between them. The connections initiated through a friendship request are independent of the links created in the actual social graph. A total number of three important features has been used to determine the weight of an edge in the social graph. For each group of profiles identified in Section 3, a social graph is generated as $G = (V_n, E, W)$, where n represent the group id, V_n is the set of profiles in group n , $E \subseteq V \times V$ is the set of edges, and $W \subseteq \mathfrak{R}$ is the set of weights assigned to edges. For each node $v \in V_n$, a 3-dimensional feature vector comprising *profile similarity*, *page likes*, and *URLs shared* is generated, which is then used to calculate the weight of an edge $e_{ij} = (v_i, v_j)$. Further details about the features and weight calculation process are presented in the following Subsections.

4.1 Social Graph Generation

To generate social graph, a set of features has been identified to determine the weight of an edge highlighting the degree of similarity of the connected profiles. The following paragraphs present a detailed discussion on the identified features and edge’s weight calculation mechanism.

Profile Similarity: The profile similarity of a pair of connected users represents the degree of match in their posts. This is calculated as a similarity index, I_s , for each edge $e_{ij} = (v_i, v_j)$ that connects a pair of nodes. The similarity index uses vector-space model to represent users’ posts and applies *cosine* function to measure their degree of similarity. The first criteria for two profiles to be similar is the number of times a profile has posted on its own wall. For example, a profile v_i with a large number of posts as compared to a profile v_j with a small number of posts on their own walls is clearly dissimilar. In this elimination process, the posts from other profiles on the subject profile’s wall are not considered. For two profiles v_i and v_j containing x and y number of posts, respectively, a *squareness measure*, as shown in Equation 1, is used to determine the eligibility of the two profiles to be considered for further comparison. In Equation 1, S_{ij} is the squareness measure of nodes v_i and v_j , which must be greater than or equal to 4 before considering them for similarity index calculation.

$$S_{ij} = x/y | x > y \quad (1)$$

For the nodes v_i and v_j that fulfil the squareness measure criterion, the similarity index is calculated as follows. Considering x and y as the number of posts of v_i and v_j on their own walls, respectively, a cosine similarity matrix C of dimensions $x \times y$ is created in such a way that each post of v_i is compared with all the posts of v_j . For cosine similarity, each post is converted into a *tf-idf* feature vector, where *tf-idf* of a term t is calculated using Equations 2 and 3. In

these equations, d is the post under consideration, D is the set of posts present in nodes v_i and v_j , and $tf(t, d)$ is calculated as the number of times t appears in d .

$$tf-idf(t, d, D) = tf(t, d) \times idf(t, D) \quad (2)$$

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (3)$$

For any two posts a and b with their corresponding $tf-idf$ feature vectors A and B , the value of an element c_{ij} of the matrix C is calculated as a cosine similarity using Equation 4, where l is the length of feature vectors.

$$c_{ij} = \frac{\sum_{k=1}^l A_k \times B_k}{\sqrt{\sum_{k=1}^l (A_k)^2} \times \sqrt{\sum_{k=1}^l (B_k)^2}} \quad (4)$$

Finally, after smoothing the values of the matrix C using Equation 5, the similarity index for the edge $e_{ij} = (v_i, v_j)$ is calculated as the normalized cardinality of the set of non-zero elements in C , as shown in Equation 6.

$$c_{ij} = \begin{cases} 1 & \text{if } c_{ij} > 0.1 \\ 0 & \text{if } c_{ij} < 0.1 \end{cases} \quad (5)$$

$$I_s = \frac{|\{c_{ij} \in C | c_{ij} = 1\}|}{x \times y} \quad (6)$$

Page-Likes: This feature is similar to the feature considered in [21]. However, in this work, the value of this feature is normalized along the lines of the similarity index normalization process. This feature captures the *page-likes* behavior of the users in a social network. For an edge $e_{ij} = (v_i, v_j)$, the common *page-likes* of v_i and v_j , P_{ij} , is calculated as a fraction of the *page-likes* commonly shared by them, as given in Equation 7. In this equation, P_i and P_j represent the set of page-likes by nodes v_i and v_j , respectively.

$$P_{ij} = \frac{|P_i \cap P_j|}{|P_i \cup P_j|} \quad (7)$$

URL sharing: Like page-likes feature, the value of URL sharing feature of nodes v_i and v_j is calculated as the fraction of the URLs commonly shared by them, as shown in Equation 8. In this equation, U_i and U_j represent the set of URLs shared by nodes v_i and v_j , respectively.

$$U_{ij} = \frac{|U_i \cap U_j|}{|U_i \cup U_j|} \quad (8)$$

Based on the values of the features discussed above, the final weight of edge $e_{ij} = (v_i, v_j)$, $\omega(e_{ij})$, is calculated using Equation 9, where α_1 , α_2 , and α_3 are constants such that each $\alpha_i > 0$ and $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

$$\omega(e_{ij}) = \alpha_1 \times I_s + \alpha_2 \times P_{ij} + \alpha_3 \times U_{ij} \quad (9)$$

4.2 Community Detection

To identify the communities in a social graph, the proposed approach uses the *Louvain* algorithm, which has been implemented as a part of an open source social network analysis tool *Gephi* 0.8.1 [22]. It has been widely used for social network analysis [23]. The algorithm supports community detection in various types of graphs and provides the flexibility to identify communities at different levels of granularity. It implements a greedy approach for optimizing *modularity* of a network divisions. The modularity measures the strength/ability of a network to be divided into groups or communities. Initially, the algorithm optimizes the modularity of smaller individual communities, then nodes from the same communities are added to form a new network in which each node represents a community. This process is repeated until maximum possible modularity is obtained. The result is a hierarchy of communities present in the underlying social graph.

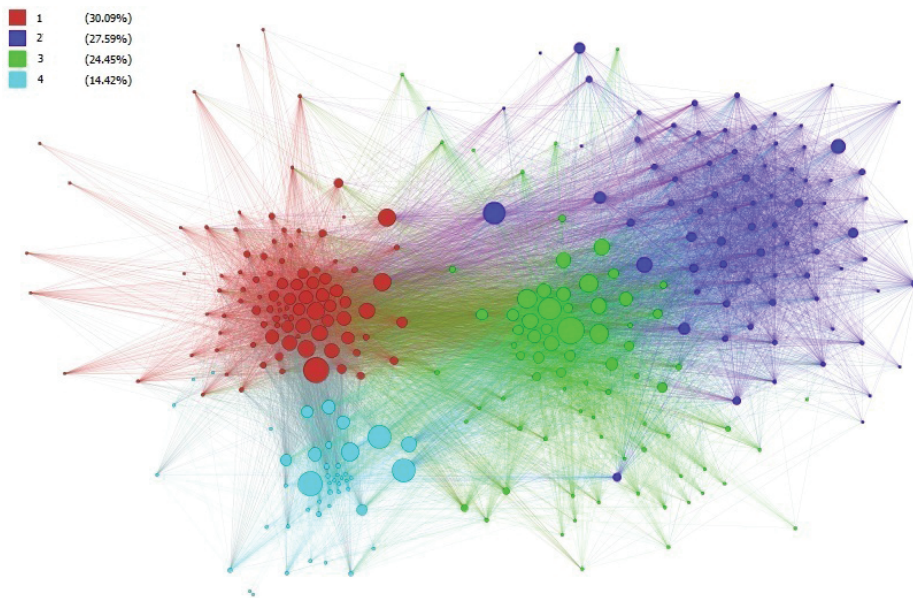


Fig. 2. Community structures in FEM group of profiles

Figure 2 shows a subgraph of the *FEM* group present in the dataset. The graph shows 4 major communities out of total 14 communities obtained through *Louvain* algorithm. In the experiment, the default resolution values of *Louvain*'s implementation in *Gephi* has been used. In Figure 2, the legend describes the percentage of nodes in each community. It can be observed in Figure 2 that nodes with modularity class 2 are dispersed, whereas nodes of classes 1, 3 and 4 are more closely related. In Section 4.3, the analysis has been further extended

Table 2. Modularity percentages of communities identified for each group of the dataset

Groups	FEM	BGFA	BNDIC	AFGNK	JHLCB	DBINKF
Class-1	30.09	18.25	17	16.33	10.60	20.92
Class-2	27.59	10.72	12.35	11.37	10.43	11.06
Class-3	24.45	9.28	7.22	11.24	7.96	6.48
Class-4	14.42	6.8	4.7	9.79	6, 15	3.03
Class-5	0.18	0.08	3.17	0.25	5.14	2.81

Table 3. Communities with closeness-centrality values

Groups	FEM	BGFA	BNDIC	AFGNK	JHLCB	DBINKF
Class-1	0.651	0.573	0.421	0.546	0.592	0.589
Class-2	0.549	0.600	0.596	0.548	0.609	0.611
Class-3	0.562	0.620	0.582	0.545	0.644	0.590
Class-4	0.628	0.568	0.592	0.548	0.610	0.570
Class-5	0.621	0.448	0.574	0.451	0.601	0.575

to classify the identified communities as sybils or normal. Table 2 shows details about the percentage of nodes in communities along with the highest modularity in each group of the dataset.

4.3 Sybil Community Identification

Once the communities are identified, profiles of each community with the highest closeness-centrality have been processed separately to classify them either as malicious or benign. Table 3 provides the details about the nodes with highest values of closeness-centrality. A set of features and JRip rules identified from a locally crawled dataset have been used to classify the nodes with highest closeness-centrality as malicious or benign. Table 4 shows the final results obtained after identifying communities as normal or malicious on the basis of the nodes' closeness centrality values. After having a close look at the closeness centrality values and the final results, it can be found that, in most of the cases, the nodes of a normal user communities have low closeness centrality values. This mainly happens because the weights assigned to the edges are according to the degree of similarity among the nodes. A higher similarity between a pair of nodes produces a higher weight for the edge connecting them in the social graph. Therefore, in the generated social graph, nodes with high closeness centrality values are similar to the majority of the nodes in the set, and as a result, a higher weight is assigned to all the edges connecting the similar nodes. Moreover, because the sybil accounts are controlled by a single spammer, they have high similarity among them as compared to normal users. Hence, nodes belonging to sybil communities have higher closeness centrality values in comparison to normal users.

Table 4. Communities identified as malicious (M) or benign (B)

Groups	FEM	BGFA	BNDIC	AFGNK	JHLCB	DBINKF
Class-1	M	B	B	B	B	B
Class-2	B	M	M	M	M	M
Class-3	B	M	B	B	M	M
Class-4	M	M	M	M	M	B
Class-5	M	B	M	B	B	B

5 Conclusions

Along the lines of the previous research works, this paper has presented a hybrid approach to detect communities of sybil accounts that are under the control of spammers and generate context-aware spam towards normal user communities. The proposed approach is independent of the assumptions made by the previous efforts and identifies six different profiles groups in the dataset based on the users' interests on Facebook network. The users with most common page-likes have been grouped together for further analysis. Three different types of features have been identified and used to model each group as a social graph in which profiles are represented as nodes and their links as edges. The weight of a link is calculated as a function of the degree of similarity of the nodes. Louvain community detection algorithm is applied on the social graphs to identify communities embedded within them. Thereafter, based on the class (malicious or benign) of the nodes with high closeness-centrality values, the underlying community is marked either as malicious or benign. The obtained results highlight that generally nodes with high closeness-centrality values are malicious and belong to sybil communities, whereas nodes with low closeness-centrality values are benign and constitute normal user communities.

Acknowledgment

The authors would like to thank King Abdulaziz City for Science and Technology (KACST) and King Saud University for their support. This work has been funded by KACST under the NPST project number 11-INF1594-02.

References

1. Boshmaf, Y., Muslukhov, I., Beznosov, K., Ripeanu, M.: Key challenges in defending against malicious socialbots. In: Proceedings of the 5th USENIX conference on Large-scale exploits and emergent threats, LEET. Volume 12. (2012)
2. Nagaraja, S., Houmansadr, A., Piyawongwisal, P., Singh, V., Agarwal, P., Borisov, N.: Stegobot: a covert social network botnet. In: Information Hiding, Springer (2011) 299–313
3. Thomas, K., Nicol, D.: The koobface botnet and the rise of social malware. In: Malicious and Unwanted Software (MALWARE), 2010 5th International Conference on, IEEE (2010) 63–70

4. Yu, H., Kaminsky, M., Gibbons, P., Flaxman, A.: Sybilguard: defending against sybil attacks via social networks. In: ACM SIGCOMM Computer Communication Review. Volume 36., ACM (2006) 267–278
5. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Gossip algorithms: Design, analysis and applications. In: INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. Volume 3., IEEE (2005) 1653–1664
6. Danezis, G., Lesniewski-Laas, C., Kaashoek, M., Anderson, R.: Sybil-resistant dht routing. Computer Security–ESORICS 2005 (2005) 305–318
7. Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B., Dai, Y.: Uncovering social network sybils in the wild. Conference on Internet Measurement, 2011 (2011)
8. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.: Detecting and characterizing social spam campaigns. In: Proceedings of the 10th annual conference on Internet measurement, ACM (2010) 35–47
9. Lee, K., Caverlee, J., Cheng, Z., Sui, D.: Content-driven detection of campaigns in social media. (2011)
10. Jin, X., Lin, C., Luo, J., Han, J.: A data mining-based spam detection system for social media networks. Proceedings of the VLDB Endowment 4(12) (2011)
11. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, ACM (2010) 1–9
12. Lee, K., Eoff, B., Caverlee, J.: Seven months with the devils: A long-term study of content polluters on twitter. In: Intl AAAI Conference on Weblogs and Social Media (ICWSM). (2011)
13. Yang, C., Harkreader, R., Gu, G.: Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In: Proceedings of the 14th International Symposium on Recent Advances in Intrusion Detection (RAID11). (2011)
14. Yu, H., Gibbons, P., Kaminsky, M., Xiao, F.: Sybillimit: A near-optimal social network defense against sybil attacks. In: Security and Privacy, 2008. SP 2008. IEEE Symposium on, Ieee (2008) 3–17
15. Danezis, G., Mittal, P.: Sybilinfer: Detecting sybil nodes using social networks, NDSS (2009)
16. Tran, N., Min, B., Li, J., Subramanian, L.: Sybil-resilient online content voting. In: Proceedings of the 6th USENIX symposium on Networked systems design and implementation, USENIX Association (2009) 15–28
17. Cao, Q., Sirivianos, M., Yang, X., Pregueiro, T.: Aiding the detection of fake accounts in large scale social online services. Technical report, Technical Report, http://www.cs.duke.edu/~qiangcao/publications/sybilrank_tr.pdf (2011)
18. Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., Zhao, B.: Social turing tests: Crowdsourcing sybil detection. Arxiv preprint arXiv:1205.3856 (2012)
19. Thomas, K., Grier, C., Ma, J., Paxson, V., Song, D.: Design and evaluation of a real-time url spam filtering service. In: IEEE Symposium on Security and Privacy. (2011)
20. McCord, M., Chuah, M.: Spam detection on twitter using traditional classifiers. Autonomic and Trusted Computing (2011) 175–186
21. Ahmed, F., Abulaish, M.: An mcl-based approach for spam profile detection in online social networks. In: The 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom-2012), IEEE (2012)

22. Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008** (2008) P10008
23. Blondel, V.: The Louvain method for community detection in large networks. <http://perso.uclouvain.be/vincent.blondel/research/louvain.html> (2011) [Online; accessed 11-July-2012].