Hemant Ojha et al. 2019, 7:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

A Survey on Social Bot Detection various Features and Techniques

M. Tech. Scholar Hemant Ojha, Dr. Jitendra Agrawal
Dept. of Computer Science & Engineering
SOIT Rgpv MP, Bhopal, India.

Abstract

This paper presents the study of various methods for detection of fake profiles. In this paper a study of various papers is done, and in the reviewed paper we explain the algorithm and methods for detecting fake profiles for security purpose. The main part of this paper covers the security assessment of security on social networking sites. This paper gives a brief survey of social bot detection challenges. Here features of fake profiles are collect. Hence paper reveals the potential hazards of malicious social bots, reviews the detection techniques within a methodological categorization and proposes avenues for future research.

Keywords: Online Social Networks (OSNs), Twitter, Spammers, Legitimate users.

I. INTRODUCTION

During the past two decades, we have progressively turned to the Internet and social media to find news, share opinions, and entertain conversations (Morris and Ogan, 1996; Smith and Brenner, 2012)[1]. What we create and consume on the Internet impacts all aspects of our daily lives, including our political, health, financial, and entertainment decisions. This increased influence of social media has been accompanied by an increase in attempts to alter the organic nature of our online discussions and exchanges of ideas. In particular, over the past 10 years we have witnessed an explosion of social bots (Lee et al., 2011; Boshmaf et al., 2013)[2], a presence that doesn't show signs of decline. Social bots are social media accounts controlled completely or in part by computer algorithms. They can generate content automatically and interact with human users, often posing as, or imitating, humans (Ferrara et al., 2016a)[3]. Automated accounts can be harmless and even helpful in scenarios where they save manual labor without polluting human conversations. In this paper, however, we focus on those actors in online social networks that surreptitiously aim to Manipulate public discourse and influence human opinions and behavior in an opaque fashion. While more traditional nefarious entities, like malware, attack vulnerabilities of hardware and software, social bots exploit human vulnerabilities, such as our tendencies to pay attention to what appears to be popular and to trust social contacts (Jun et al., 2017)[4]. Unlike other social engineering attacks, such as spear phishing (Jagatic et al., 2007)[5], bots can achieve scalability through automation. For example, multiple accounts controlled by a single entity can quickly generate posts and make specific content trend or amplify misinformation. They can trick humans and engagement-based ranking algorithms alike, creating the appearance that some person or opinion is popular.

Therefore, defending from social bots raises serious research challenges (Boshmaf et al., 2012)[5]. Manipulation of public opinion is not new; it has been a common practice since the dawn of humanity. The technological tools of all eras — printed media, radio, television, and the Internet — have been abused to disseminate misinformation and propaganda. The deceptive strategies employed on all these types of channels share striking similarities [6](Varol and Uluturk, 2018). Nevertheless, social media are particularly vulnerable because they facilitate automatic interactions via software. As a result, social media platforms have to combat a deluge of attacks. Face book recently announced that 1.5 billion fake accounts were removed over six months in 2018.1 even a very low miss rate could leave millions of accounts available to be used as bots. In this light, it is not surprising that as many as 9–15% of active Twitter accounts were estimated to be bots in 2017 (Varol et al., 2017a)[7], and that social bots are responsible for generating two thirds of links to popular websites (Wojcik et al., 2018)[8]. Public interest in social bots has also dramatically increased during the past few years.

II. RELATED WORK

McCord et.al. [9] used user based features like number of friends, number of followers and content based features like number of URLs, replies/mentions, retweets, hashtags of collected database. Classifiers namely Random Forest, Support Vector Machine (SVM), Naive Bayesian and K-Nearest Neighbour have been used to identify spam profiles in Twitter. Method has been validated on 1000 users with 95.7% precision and 95.7% accuracy using the Random Forest classifier and this classifier gives the best results followed by the SMO, Naive Bayesian and K-NN classifiers. Limitation of this approach is that for considered dataset reputation feature has been showing wrong results i.e. it is not able to differentiate spammers and non-spammers, unbalanced dataset has been used so Random Forest is giving best results as this classifier is generally used in case of unbalanced dataset, and finally the approach has been validated on less dataset.

Lee et. al.[10] deployed social honey pots consisting of genuine profiles that detected suspicious users and its bot collected evidence of the spam by crawling the profile of the user sending the unwanted friend requests and hyperlinks in MySpace and Twitter. Features of profiles like their posting behavior, content and friend information to develop a machine learning classifier have been used for identifying spammers. After analysis profiles of users who sent unsolicited friend requests to these social honey pots in MySpace and Twitter have been collected. LIBSVM classifier has been used for identification of spammers. One good point in the approach is that it has been validated on two different combinations of dataset – once with 10% spammers+90% non-spammers and again with 10% non-spammers+90% spammers. Limitation of the approach is that less dataset has been used for validation.

Viswanath et al. [11] discover that dependency on community detection makes more vulnerable to Sybil attacks where honest identities conform strong communities. Because Sybils can infiltrate honest communities by carefully targeting honest accounts. That is, Sybils can be hidden as just another community on OSN by setting up a small number of the targeted links. The targeted links are the links given to the community which contains the trusted node. They make an experiment by allowing Sybils to place their links closer to the trusted node instead of random nodes, where closeness is defined by ranking used by the community detection algorithm they employ. Hence, Sybil nodes are high ranked in the defence scheme. Naturally, it leads to Sybils being less likely to be detected for that attack model because Sybils are appeared as part of the local community of the trusted node.

Boshmaf et al. [12] point out that structure-based Sybil detection algorithms should be designed to find local community structures around known honest (non-Sybil) identities, while incrementally tracking changes in the network by adding or deleting some nodes and edges dynamically in some period for better detection performance.

Chu et al. [13] make a study on profiling human, bot, and cyborgs2. They observe the difference among them in terms of tweet content, tweeting behaviour, and account properties like external URL ratio.

Benevenuto et. al. [14] detected spammers on the basis of tweet content and user based features. Tweet content attributes used are - number of hash tags per number of words in each tweet, number of URLs per word, number of words of each tweet, number of characters of each tweet, number of URLs in each tweet, number of hash tags in each tweet, number of numeric characters that appear in the text, number of users mentioned in each tweet, number of times the tweet has been retweeted. Fraction of tweets containing URLs, fraction of tweets that contains spam words, and average number of words that are hashtags on the tweets are the characteristics that differentiate spammers from non spammers.

Gee et. al. [15] utilized this feature and detected spam profiles using classification technique. Normal user profiles have been collected using Twitter API and spam profiles have been collected from "@spam" in Twitter. Collected data was represented in JSON then it was presented in matrix form using CSV format. Matrix has users as rows and features as columns. Then CSV files were trained using Naive Bayes algorithm with 27% error rate then SVM algorithm has been used with error rate of 10%. Spam profiles detection accuracy is 89.3%. Limitation of this approach is that not very technical features have been used for detection and precision is also less i.e. 89.3% so it has been suggested that aggressive deployment of any system should be done only if precision is more than 99%.

Yang et al. [16] collect Sybil accounts from Renren as ground-truth data set. Then, they analyse it by using network-based and structured-based features such as network clustering coefficient, incoming and outgoing request rate.

III. TECHNIQUES OF BOT DETECTION

1. Traffic-based Detection

The P2P bots communicate with many other peer bots to push/pull commands, send harvested information and receive updates; thus continuously generating large traffic [5]. Various traffic-based detection techniques have been proposed, which examine the network traffic and focus to observe the traffic patterns.

2. Behavior-based Detection

A comprehensive analysis of botnet measurements by Rajab et al. [17] reveals the structural and behavioral properties of botnets. Bots may also possess many inherent features, maintain the persistent connections to communicate with other peer bots and receive the commands from botmaster via C&C server(s). It is observed that the network behavior characteristics of P2P botnets are closely tied to the underlying architecture and operation mechanisms

3. DNS-based Detection

The bots possess a group activity as a key feature and frequently use DNS to rally C&C servers, launch attacks and update their codes. Bots of same botnet contact the same domain periodically leading to similar DNS traffic which is distinct from legitimate users [7].

4. Graph-based Detection

The graphical structure is an inherent feature of the botnets and is useful to understand how botnets communicate internally. The graphical analysis of the botnet communication network can be used to find the characteristic patterns of the botnets. The P2P C&C communications graph exhibit the topological features useful for traffic classification and botnet detection.

5. Data Mining-based Detection

The data mining techniques can be used to detect an anomaly i.e., the unusual or fraudulent behavior. Data mining techniques are used for malicious code detection and intrusion detection. Many authors has used classification and clustering techniques to efficiently detect botnet C&C traffic.

6. Generic Frameworks

A number of general botnet detection frameworks have been proposed based on behavior monitoring and traffic correlation analysis. BotMiner is a general framework for botnet detection [14]. The system detect botnets based on network packets and flow analysis. It relies on behavior monitoring and traffic correlation analysis that is mostly applicable at a small scale and does not scale well, because it requires analysis of vast amounts of fine-grained information.

IV. FEATURE EXTRACTION

Data collected using the Twitter API are distilled in 1,150 features in six different classes.

1. User-based features.

Features extracted from user metadata have been used to classify users and patterns before (Mislove et al. 2011; Ferrara et al. 2016a). We extract user-based features from meta-data available through the Twitter API. Such features include the number of friends and followers, the number of tweets produced by the users, profile description and settings.

2. Friends features.

Twitter actively fosters interconnectivity. Users are linked by follower-friend (followee) relations. Content travels from person to person via retweets. Also, tweets can be addressed to specific users via mentions. We consider four types of links: retweeting, mentioning, being retweeted, and being mentioned. For each group separately, we extract features about language use, local time, popularity, etc. Note that, due to Twitter's API limits, we do not use follower/followed information beyond these aggregate statistics.

3. Network features.

The network structure carries crucial information for the characterization of different types of communication. In fact, the usage of network features significantly helps in tasks like political Astroturf detection (Ratkiewicz et al. 2011). Our system reconstructs three types of networks: retweet, mention, and hash tag co-occurrence networks. Retweet and mention networks have users as nodes, with a directed link between a pair of users that follows the direction of information spreading: toward the user retweeting or being mentioned. Hashtag co-occurrence networks have undirected links between hash tag nodes when two hashtags occur together in a tweet. All networks are weighted according to the frequency of interactions or co occurrences. For each network, we compute a set of features, including in- and out-strength (weighted degree) distributions, density, and clustering. Note that out-degree and out-strength are measures of popularity.

4. Temporal features

Prior research suggests that the temporal signature of content production and consumption may reveal important information about online campaigns and their evolution. To extract this signal we measure several temporal features related to user activity, including average rates of tweet production over various time periods and distributions of time intervals between events.

5. Content and language features

Many recent papers have demonstrated the importance of content and language features in revealing the nature of social media conversations. For example, deceiving messages generally exhibit informal language and short sentences (Briscoe, Appling, and Hayes 2014). Our system does not employ features capturing the quality of tweets, but collects statistics about length and entropy of tweet text. Additionally, we extract language features by applying the Part-of-Speech (POS) tagging technique, which identifies different types of natural language components, or POS tags. Tweets are therefore analyzed to study how POS tags are distributed.

6. Sentiment features

Sentiment analysis is a powerful tool to describe the emotions conveyed by a piece of text, and more broadly the attitude or mood of an entire conversation. Sentiment extracted from social media conversations has been used to forecast offline events including financial market fluctuations (Bollen, Mao, and Zeng 2011), and is known to affect information spreading.

Precision: Precision value is the ratio of predicted positive user to the total predicted user.

$$\text{Precision} = \left(\frac{\text{True}_{\text{positive}}}{(\text{False}_{\text{positive}} + \text{True}_{\text{positive}})}\right)$$

Recall: The recall is the fraction of relevant users that have been predicted over the total amount of input users. It is also known as Sensitivity or Completeness.

$$\text{Recall} = \left(\frac{\text{True}_{\text{positive}}}{\text{False}_{\text{negative}} + \text{True}_{\text{positive}}}\right)$$

F-Measure: Harmonic mean of precision value and recall value is F-measure.

$$\text{F} - \text{Measure} = \left(\frac{\text{2xPrecisionxRecall}}{(\text{Recall} + \text{Precision})}\right)$$

Accuracy: This act as the percentage of correct prediction from the total set of prediction.

$$\text{Accuracy} = \left(\frac{\text{Correct\_class}}{(\text{Correct\_class} + \text{InCorrect\_class})}\right)$$

V. CONCLUSIONS

With the high demand of image in various fields researchers get attracted for analysis. This paper covers various approaches of bot techniques. As with other online attacks, defending against malicious social bots is an arms race where the objective of the defender is to limit any potential harm or damage, that is, to extend the time at which the system enjoys its safe state. In this paper, we observed that in order to effectively defend against such bots, one has to fix a set of inherent vulnerabilities found in today's OSNs, which collectively represent the enabling factors causing the problem. Here paper has review an feature set of social bot detection. In future a perfect algorithm is with good feature combination is desired which remove haze while image object get can identify easily.

REFERENCES

[1]. Morris, M. and Ogan, C. (1996). The internet as mass medium. Journal of communication, 46(1):39-50.

[2]. Lee, K., Eo_, B. D., and Caverlee, J. (2011). Seven Months with the Devils:A Long-Term Study of Content Polluters on Twitter. In Proc. AAAI Intl. Conf. on Web and Social Media (ICWSM).

[3]. Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016a). The rise of social bots. Communications of the ACM, 59(7):96{104.

[4]. Jun, Y., Meng, R., and Johar, G. V. (2017). Perceived social presence re-duces fact-checking. Proceedings of the National Academy of Sciences,114(23):5976{5981.

[5]. Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. (2007). Social phishing. Communications of the ACM, 50(10):94{100.

[6]. Boshmaf, Y., Muslukhov, I., Beznosov, K., and Ripeanu, M. (2012). Key chal- lenges in defending against malicious socialbots. In Proc. 5th USENIX Conference on Large-Scale Exploits and Emergent Threats (LEET).

[7]. Varol, O. and Uluturk, I. (2018). Deception strategies and threats for online discussions. First Monday, 22(5).

[8]. Wojcik, S., Messing, S., Smith, A., Rainie, L., and Hitlin, P. (2018). Bots in the twittersphere. Pew Research Center, Washington, D.C.

[9]. M. McCord, M. Chuah, Spam Detection on Twitter Using Traditional Classifiers, ATC'11, Banff, Canada, Sept 2-4, 2011, IEEE.

[10]. Kyumin Lee, James Caverlee, Steve Webb, Uncovering Social Spammers: Social Honeypots + Machine Learning, Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, Pages 435–442, ACM, New York (2010).

[11]. B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An analysis of social network-based sybil defenses," ACM SIGCOMM Computer Communication Review, vol. 40, pp. 363-374, 2010.

[12]. Y. Boshmaf, K. Beznosov, and M. Ripeanu, "Graph-based sybil detection in social and information systems," in Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, 2013, pp. 466-473.

[13]. Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?," IEEE Transactions on Dependable and Secure Computing, vol. 9, pp. 811-824, 2012.

[14]. Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, Detecting Spammers on Twitter, CEAS 2010 Seventh annual Collaboration, Electronic messaging, Anti Abuse and Spam Conference, July 2010, Washington, US Grace gee, Hakson Teh, Twitter Spammer Profile Detection, 2010.

[15]. Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network sybils in the wild," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 8, p. 2, 2014.

[16]. M. A. Rajab, J. Zarfoss, F. Monrose and A. Terzis,\ A multifaceted approach to understanding the bot-

[17]. Net phenomenon," in Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement (IMC'06), pp. 41-52, 2006.