A Survey on the Spam Issue in Twitter

Soufiane Maguerra*
Azedine Boulmakoul*
Lamia Karim**
Hassan Badir***

*LIM/IOS, FSTM, Hassan II University of Casablanca, Mohammedia, Morocco
{maguerra.soufiane,azedine.boulmakoul}@gmail.com,
**Higher School of Technology EST Berrechid, Hassan 1st University, Morocco
lkarim.lkarim@gmail.com
***National School of Applied Sciences Tangier, Abdelmalek Essaâdi University, Morocco
hbadir@gmail.com

Abstract. Social networks are being leveraged by cyber-criminals to cover a wider range of victims. In Twitter, spammers create several bots and behave in different patterns according to their desired aims. Particularly, spammers can spread malicious links leading to malware or phishing sites. Achievable by engaging in social bonds or responding to trending topics (hashtags). Spammers either spam in an individual manner, otherwise in coordinated communities with a clear insight. Decidedly, reinforcing cyber-security in Twitter is an indispensable fact. Several researchers have been studying the different aspects of spamming in Twitter. This paper includes a background over the information handled in Twitter, and a detailed survey over the papers dealing with the spam issue. The discussed papers have been published from 2010 to 2018. In contrast to other surveys, this paper is not limited to the detection of spammers but it also discusses the approaches to the detection of spam communities, compromized accounts, collective attention spam, and the extraction of cybercrime knowledge. Hence, this study can be considered as an essential step for the design of a unified spam detection framework.

1 Introduction

The misuse of social media lead to an exponential rise of cybercrime victims. Cyber-criminals leverage the privileges of social media to range over a wider field of victims. Consequently, malicious links are invading social walls. Crackers gather information about accounts to gain their trust; then, oblige them to pay ransoms for safety. Information concerning system vulnerabilities, hacking tools and techniques are been leaked to expand the cyber-criminals community. Thus, the primal necessity to secure social networks.

As a primal study, we focus on Twitter characterized by a total number of 1.3 billion accounts with 330 million active ones per month, 500 million tweets sent per day with 6000 sent every second, 100 million active users per day, and an estimate of 23 million bots (Sikandar G,

2018). In the year 2017, Twitter estimated 3.2 million bots per week, and 450,000 suspicious log-ins per day. Bots are mainly exploited for spreading fraud, and Twitter suspended from June to September over 117,000 malicious bots approximately responsible for 1.5 billion low-quality tweets (Feldman, 2017). In particular, Twitter has its own native spam detection system; however, spammers still succeed to deceive the detection process. A fact explained by their legitimate behavior, dynamic state, and their effort to continuously study the native approaches for vulnerabilities and gaps. More specifically, spammers escape the follow limit policy either by purchasing followers (Yang et al., 2011) or hiding their presence by creating bonds and manipulating other spamming bots (Bindu et al., 2018). Otherwise, accounts are being compromised and leveraged for sending spams, e.g., in 2016 millions of Twitter accounts were offered by a hacker for sale (Whittaker, 2016).

Furthermore, legitimate accounts can become malicious with time by learning from other cyber-criminals. Knowledge about software vulnerabilities about software and exploits is being described in tweets either by spammers or victims. Tweets tend also to have an informal nature without following neither specific syntactic rules nor grammatical ones. In this paper, a background over Twitter is yielded with a survey describing our current knowledge about the works that have been interested in solving the spam issue in twitter. In detail, we discuss Twitter functionalities and native spam rules, Twitter regular and Streaming APIs. The survey reviews some of the papers that have been published from 2010 to 2018. In contrast to other surveys such as Kaur et al. (2016), Wu et al. (2017), Katpatal and Junnarkar (2018) and Katpatal and Junnarkar (2018); this survey is not limited to a single aspect of the spam detection in twitter. It reviews studies of the detection of spam communities, compromized accounts, collective attention spam, and the extraction of cybercrime knowledge.

2 Background

Twitter accounts have each a specific profile including a screen name, creation date, description, profile and header photo, URL, and a timeline of tweets. Each account can view public tweets posted by others without restriction. In contrast, protected tweets can be visualized only by accepted accounts. Unless a user protects its tweets by becoming private, he can be followed by any other user. Hence, obtaining friends and receive their tweets in the proper home timeline. A tweet can be retweeted by another user with the possibility of adding a comment; in this case, it is denoted as quote. The poster of the original tweet does not receive any notification about the replies unless he is mentioned in it. A user can mention any user even private accounts in a tweet by adding the symbol @ plus its screen name, e.g., @account. However, only the followers of both accounts can see the tweet in their home timelines. Still, if the user retweets the tweet it will become visible to the followers. An account is characterized by its number of friends and followers who have followed its account. Accounts can be manipulated manually or by using the Twitter API. The accounts manipulated automatically are denoted as bots. Bots can be used by companies for profit by sharing their products or their news. However, spammers leverage bots for spreading spam. In 2017, Twitter estimated 3.2 million bots per week, and 450,000 suspicious log-ins per day. Bots are mainly exploited for spreading fraud, and Twitter suspended from June to September over 117,000 malicious bots approximately responsible for 1.5 billion low-quality tweets (Feldman, 2017). In particular, Twitter has its own native spam detection system.

2.1 Twitter Rules and Countermeasures

Twitter defines spam as any abnormal behavior with the purpose of misleading or deceiving people. The spammer is a user who violates the Twitter rules by possessing several accounts, posting malicious links and mainly tweets involving links, repeatedly tweeting duplicates or in response to trending topics, aggressively adding users in lists, using other profiles information, aggressively mass following and mass unfollowing other accounts, harmfully replying to tweets, abusively mentioning users to get their attention and the one of their followers. People can report spammers as a countermeasure leading to suspension when a large number of blocks and complaints has been affirmed.

Twitter has taken severe countermeasures in response to spam. The Twitter Trust and Safety Council takes part in them by ensuring safe expressivity. The council involves more than 40 organizations and experts from 13 regions with an aim to ban cyberbullying, inaccurate information, and terrorism. In 2014, Twitter announced Botmaker a distributed system that contributed to a 40% reduction of spam with approximately zero false positives rate. Botmaker takes consideration of a set of rules denoted as bots. The bots respect a Botmaker rule language, and are composed of a condition triggerring a specific action. The system came to surpass the knowledge of malicious users over their past anti-spam actions. Avoiding high latency and preserving the real-time tweeting facets while handling billions of events per day. Botmaker aims to prevent the creation of spam, to reduce its visibility as much as possible, and quickly react to novel spam attacks (Raghav, 2014).

2.2 Twitter API

Users can leverage the Twitter API for manipulating several accounts without direct contact with the platform by the aid of a software. The API offers all possible native functionalities including posting, searching, modifying, replying, etc. Researchers exploit often this API to get a hold of the data. Another option is the Twitter Streaming API yielding real-time streams of tweets. Tweets are restricted to 280 characters. Hence, URL shorteners are used when a URL is included in the tweet. Users are free to choose either the native t.co service or others such as bit.ly, goo.gl, etc. Despite the efforts of the services to verify the links before shortening them, spammers still leverage this option to feed the service with hidden malicious links.

3 State of the Art

3.1 Works Conducted on Social Honeypots in Twitter

Lee et al. (2010a) introduced the Social Honeypot Project with an aim to identify spammers. Social honeypots capture spammers in social networks. In twitter, they harvested 500 spammers on a period ranging from August 2009 to September 2009. They trained classifiers using user demographics, contributed content, activities, and connections. The trained classifiers achieved an accuracy of 88.98%.

In another work (Lee et al., 2010), they manually studied the behavior of the 500 spammers. Concluding that the spammers target specific users, repeatedly posting duplicates, and disseminate spam or phishing links. Moreover, they lured 131 spammers in 2 weeks from the 24th November 2009 to the 8th December 2009. After studying the behavior of the spammers,

they observed that spammers react to trending topics, post legitimate tweets, and periodically change their number of friends and followers. More importantly, they discovered that only 8% of the spammers have been suspended after an average time of 49 hours.

In (Lee et al., 2010b) they studied 500 users in a period ranging from August 2009 to September 2009. After analysing the dataset, they concluded different behaviors of spammers and classified them into duplicates, promoters, phishers, adult content spreaders and, lastly, friend infiltrators behaving first in a legitimate manner to gain followers; then, start spreading spams. For the classification process, they obtained an accuracy of 88.98% with Decorate.

The same authors conducted a long term study in (Lee et al., 2011). They deployed 60 social honeypots in a period of 7 months ranging from December 2009 to August 2010. The dataset contained 36,043 users, they filtered the users who followed more than one honeypot to end up with 23,869 users. Then, they removed the accounts that were quickly suspended by Twitter to assemble 22.223 spammers. After clustering, they observed high numbers of followers friends, imitation of legitimate behavior, and balancing between the number of friends and followers. In the classification process, they attended high accuracy by using the Random Forest with boosting and bagging standards.

Yang et al. (2014) conducted the first study over understanding the taste of spammers by using 96 decoys with 24 different patterns. Their work aims to understand how to construct effective social honeypots. In five months, they collected 1077 and 440 accounts that respectively have followed and mentioned a decoy. After a deep analysis, they concluded that spammers are attracted to users who post to trending or specific topics, and have an interest on famous accounts.

3.2 Evasive Spammers Tactics and More Robust Features

Yang et al. (2011, 2013) were the first to analyse how spammers evade detection process. They constructed a dataset containing 500,000 accounts. The dataset was assembled by using the Twitter Streaming API; then, use the Twitter API to get information about the users, their followers and friends. To get the spammers, they verified the posted links by using Google Safe Browsing. If a link is not detected, then they use the Capture-HPC client-side honeypot. To filter the spammers, they employed a spam ratio with a threshold of 10%. The filter yielded to over 2,933 accounts; then after manual verification, they ended up with 2,060 spammers. After analysing the spammers, they observed different evasion tactics discussed infra.

— **profile-based** : gaining more followers either by purchase, exchange or by controlling other bots; posting more tweets.
— **content-based** : mixing normal tweets with malicious ones and posting heterogeneous tweets.

Afterwards, they evaluated existing classification features and proposed 10 new features more robust to the evasion strategies (see table 1). After evaluation, they found out that automation-based feature have a medium robustness; betweenness centrality, clustering coefficient, and followings to median neighbor's followers followings have high robustness.

3.3 Collective Attention Spam

Lee et al. (2012) conducted the first research over the detection of collective attention spam tweets in their early stages. Collective attention spam differs from bulk email and inten-

tioned social spam because the victims harm them selfs without direct contact from attackers by taking interest in trending topics. Their dataset involves 5.3 million tweets belonging to 1.5 million users. The data was gathered in 11 days starting from September to October 2011 by checking every 5 minutes for messages involving trending topics. They discovered that Twitter suspended over 17,411 accounts which posted over 136,255 tweets. To prove that the suspended accounts' tweets belong to spammers, two judges manually classified a sample of over 400 tweets, half of them belonged to suspended accounts. They obtained an accuracy of 96,5%; hence, they assumed that the suspended accounts belong to spammers. After studying the suspended accounts, they observed that more than 50% of the accounts are more interested in posting spams than creating bonds. The features used for the classification include urls, hashtags, mentions, is retweet, length of tweet, and length of payload (after removing mentions, urls, and hashtags). Additionally, they used bag of words and concluded that they can improve the classification outcome; however, at the cost of time complexity. The classification is achieved via Random Forest and they observed that the best training time for identifying following spams is 3 hours after the topic becomes trending.

TAB. 1: Novel Features

graph	Local Clustering Coefficient[1]	$\frac{2\|e^v\|}{K_v(K_v-1)}$	spammers tend to randomly choose their victims; hence, the resulting graph would be farther than constituting a clique
	Betweenness Centrality[2]	$\frac{1}{(n-1)(n-2)}\sum_{s\neq v\neq t\in V}\frac{\delta_{st}(v)}{\delta_{st}}$	spammers tend to be in the middle of several shortest paths in the graph by following several unrelated accounts
	Bi-directional Links Ratio	$\frac{N_{bilink}}{N_{friends}}$	a large number of users do not follow spammers back
neighbors	Average Neighbors' Followers	$\frac{1}{\|N_{fing}(v)\|}\sum_{u\in N_{N_{fing}(v)}} N_{fer}(u)$	legitimate account follow users with higher follower rate
	Average Neighbors' Tweets	$\frac{1}{\|N_{fing}(v)\|}\sum_{u\in N_{N_{fing}(v)}} \|tweets_{fer}(u)\|$	legitimate account follow users with higher tweet rate
	Followings to Median Neighbors' Followers[3]	$\frac{N_{fing}}{M_{nfer}}$	spammers do not guarantee the quality of their friends
timing	Following Rate	number of tweets per a specific time	spammers tend to follow a large number of accounts in a short period
automaion	API Ratio	$\frac{\|API\ tweets\|}{\|tweets\|}$	spammers tend to be manipulating bots by means of the API
	API URL Ratio	$\frac{\|API\ weets\ with\ URLs\|}{\|API\ tweets\|}$	malicious tweets usually contain links
	API Tweet Similarity	average similarity of tweets posted by the API	when spammers exploit the API their tweets tend to have common payload

Note: The features colored in red tend to have a highe value for spammers. In contrast, the blue color means smaller values.
[1] K_v is the degree of the vertex v and e^v is the number of links between its neighbors.
[2] n is the number of vertices in the graph, the number of shortest paths from s to t is δ_{st}, and the shortest paths passing from v are $\delta_{st}(v)$.
[3] M_{nfer} is the median of number of followers characterizing the friends.

In an advanced work (Lee et al., 2013), they conducted the first comprehensive study over collective attention spam trying to study the following subjects :

— the vulnerability of Twitter to collective attention spam,
— the effectiveness of spams,
— the access of victims to their interest and their exposition to spam,
— the countermeasures that can be deployed by a system and their effectiveness.

Differently than other works to find answers, they simulated a Twitter social system because they believed that snapshots of twitter data are not sufficient to accurately understand collective attention spam. For seeding the simulation parameters, they assembled a dataset of 17,275,961

tweets belonging to 3,989,563 users. The tweets involve 354 topics and the collection period started from September 2011 to November 2011. They considered the suspension of accounts to accurately set the simulation parameters related to the posting spam probabilities. The system simulated legitimate users who search the tweets by recency or relevance. After evaluation, they observed that spammers can orchestrate their behavior to be more effective and that their damage can be reduced while applying early stage countermeasures.

3.4 Detection of Spam Communities

Yang et al. (2012) conducted the first study of analysing the inter and outer interactions of social spammers groups in Twitter. They observed the strong interactions between social spammers by measuring the graph density. Additionally, hidden social hubs supporting and controlling other spammers. For their research, they collected data using the Twitter Streaming API in a period ranging from April 2010 to july 2010. They identified in the social relationships criminal supporters by using a proposed Maclious Relevance Score Propagation Algorithm (Mr.SPA) that outputs criminal supporters while considering a specific threshold. The algorithm measures how closely an account supports spammers. The obtained supporters were identified to be members of these categories :

— **butterflies** with a higher number of followers and friends, the butterflies ussually do not give much attention to requests and usually follow back;
— **promoters** usually having high followers friends ratio and use the service to promote their self or products;
— **dummies** posting interesting legitimate tweets in few numbers and characterized by a large number of followers.

Bindu et al. (2018) conducted the first study to detect spam communities in Twitter. They employed the past assembled social honeypot dataset of Lee et al. (2011). An algorithm constituting of different parts has been proposed to achieve the detection process. The algorithm handels a multilayer social network graph composed of two layers. The users follower/friend relationships are modeled in the first layer, and in the second contains users and their tweets. Base spammers are identified by checking their unique URLs ratio compared to a specific threshold. In parallel, they employ the LA and IS^2 algorithm to identify hidden spammer communities. The communities are considered as hypergraphs. Then, for each community they check for base spammers. Each base spammer's maximum clique is identified, the clique contains victims, spammers and legitimate users. To extract the spammers, they measure the Local Clustering Coefficient because spammers tend to have a lower value. The extracted actors form the suspect spammers set. For each suspect a spam score is measured and compared to a specific threshold. The score is measured by considering the Jaccard Similarity Index between the base spammer and suspect URL, the average neighbors' followers, URL tweet ratio, and longevity of the account.

Chen et al. (2017) proposed an unsupervised approach to detect malicious bot communities in real-time Twitter streams. Their approach is based upon four sequential processess :

— **Crawler** for filtering the Twitter Streaming API using specific keywords.
— **Duplication Filter** which hashes the tweets and maps them to groups with similar content while keeping only groups of a size larger than or equal 20.
— **Collector** collecting the 200 most recent tweets of each account using the Twitter API.

— **Bot Detector** responsible for the detection of bots communities in each group by assembling the set C of tweets that are tweeted by a specific number of users; then, to compare the overlap between a users timeline and C.

Vo et al. (2017) are the first to study malicious retweeter groups. They collected data over one year from November 2014 to November 2015 using the Twitter Streaming API. The dataset contained over 1.6 billion tweets belonging to over 21 million users. To extract the retweeter groups they proposed the Attractor+ algorithm. Similarities between users are computed by using 40 cores distributed hadoop system. For the classification, they sampled the dataset and employed three judges to manually label the retweeter groups. Novel group-based are proposed depending on content and temporal aspects to characterize the groups. The classification with XGBoost yielded to a high accuracy.

3.5 Detection of Spammers Using Statistical Relational Learning

The Linconly Laboratory (Campbell Jr et al., 2015) leveraged the linear supervised SVM and Logistic regression techniques to classify english posts extracted from Twitter, Stack Exchange and Reddit. They used the TF-IDF ratios as attributes for the classification process. More importantly, they presented the idea of a Twitter's Users Meta-graph stored in Neo4J which contains information of users and their different relationships. Semi-Supervised classification techniques based on collective inference can be applied over this graph with the aim to classify and infer cyber-criminals.

Rao et al. (2016) were the first to employ Markov Logic Networks to model the social spammer graph while proposing a unified spammer detection framework denoted as SocialKB. A knowledge base is formed containing knowledge of Twitter users; then, a set of first order logic rules are defined to classify the users. Rules based on time, malicious urls, and friendships were constructed. Afterwards using Tuffy, Maximum A Posteriori (MAP) and Marginal Inference queries are applied to the graph for obtaining knowledge about spammers. In June 2016, they collected over 20,000 tweets using the Apache Spark Twitter Streaming utility and to filter the malicious links they used URLBlacklist.com.

3.6 Extracting Cybersecurity Knowledge from Twitter

Mittal et al. (2016) presented a CyberTwitter Framework that mines the twitter's discussions to extract and infer knowledge over vulnerabilities and their solutions. The system is based on a cyber-security ontology including the Unified Cyber-security Ontology, and an intelligence ontology which modelizes the temporal aspects. Tweets are checked for different concepts using the Security Vulnerability Concept Extractor (SVCE), these concepts are then linked to the open knowledge graph. The system can be used to inform users about possible vulnerabilities in their systems. Knowledge about possible ways to protect themselves can be yielded, and possible malicious websites can also be listed to the users by considering the knowledge fed to the system.

3.7 Detection of Compromised Accounts

Egele et al. (2013) realized Compa the first system to detect compromized accounts with a high classification precision. even if they did not post spam content and spread links. The sys-

tem constructs behavioral models for users with a status including more than 10 tweets. These profiles are checked for abnormal behavior and an anomaly score is affected to each profile for a final decision. They proposed methods to compute the anomaly score for each model. The scores are fused to a single value, the weights of the scores are appended while employing Sequential Minimal Optimization. For learning the weights, they manually classified a sample of the dataset by checking links and promoting behaviors. The data used to evaluate the system was assembled from May 2011 to August 2011. The dataset includes over 1.4 billion tweets and they detected 383.613 compromized accounts. The Twitter Streaming API was used to get the tweets and the Twitter API to obtain the profiles. In their work, they used models related to time, source, language, topic, location, URLs and mentions.

Zangerle and Specht (2014) analysed the behavior of the compromised accounts after they have been hacked. A supervised classification approach was realized to classify the tweets as a mean to identify hacked accounts. A dataset involving 1,231,468 tweets from December 2012 to July 2013 was assembled via the Twitter Streaming API. Support Vector Machines was employed as a classifier. The non-English tweets and retweets were filtered. For training, they took a sample of 2,500 tweets and manually labeled them. Once classified, they obtained an accuracy of 82.51%. After manually analysing the tweets and applying the classification to the whole dataset, they observed :

— 27% of hacked users move to another account,
— 14% apologize for posted tweets and an other 14% for directed messages,
— 10% have been hacked by a relative or friend,
— 4% change their password.

All the detected classes, except the false positives, state that they have been hacked, the only difference is in their behavior.

Nauta et al. (2017) conducted a similar work to Egele et al. (2013). However, they took interest on Dutch users, and they proposed other models. The classification yielded a lower false positve ratio than Egele et al. (2013) by using the J48 classifier. However, they acknowledged the different sizes of the datasets and their languages.

4 Conclusion

This paper reviews the studies that have been conducted to resolve the spam issue in Twitter. The review involves studies analysing the behavior of spammers individually and in groups. The strategies they respect to stay hidden and efficient approaches to detect them. The cybercrime knowledge that can be extracted by processing the tweets. This survey studies all the aspects of Twitter spammers, even the studies that have been interested in compromized accounts are stated. This state of the art can be leveraged to get past researchers knowledge; then, propose a unified framework not restricted to Twitter involving the detection of spammers while considering all their behaviors.

References

Bindu, P., R. Mishra, and P. S. Thilagam (2018). Discovering spammer communities in twitter. *Journal of Intelligent Information Systems*, 1–25.

Campbell Jr, J., A. C. Mensch, G. Zeno, W. M. Campbell, R. P. Lippmann, and D. J. Weller-Fahy (2015). Finding malicious cyber discussions in social media. Technical report, MIT Lincoln Laboratory Lexington United States.

Chen, Z., R. S. Tanash, R. Stoll, and D. Subramanian (2017). Hunting malicious bots on twitter: An unsupervised approach. In *SocInfo*.

Egele, M., G. Stringhini, C. Krügel, and G. Vigna (2013). Compa: Detecting compromised accounts on social networks. In *NDSS*. The Internet Society.

Feldman, B. (2017). Twitter has 450,000 'suspicious log-ins' a day. Retrieved from `http://nymag.com/selectall/2017/09/twitter-has-450-000-suspicious-log-ins-a-day.html`. Last visited: May 15th 2018.

Katpatal, R. and A. Junnarkar (2018). Spam detection techniques for twitter.

Kaur, P., A. Singhal, and J. Kaur (2016). Spam detection on twitter: A survey. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 2570–2573.

Lee, K., J. Caverlee, K. Y. Kamath, and Z. Cheng (2012). Detecting collective attention spam. In *Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality*, pp. 48–55. ACM.

Lee, K., J. Caverlee, and S. Webb (2010a). The social honeypot project: Protecting online communities from spammers. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, New York, NY, USA, pp. 1139–1140. ACM.

Lee, K., J. Caverlee, and S. Webb (2010b). Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 435–442. ACM.

Lee, K., B. D. Eoff, and J. Caverlee (2010). Devils, angels, and robots: Tempting destructive users in social media.

Lee, K., B. D. Eoff, and J. Caverlee (2011). Seven months with the devils: A long-term study of content polluters on twitter.

Lee, K., K. Y. Kamath, and J. Caverlee (2013). Combating threats to collective attention in social media: An evaluation. In *ICWSM*.

Mittal, S., P. K. Das, V. Mulwad, A. Joshi, and T. Finin (2016). Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pp. 860–867. IEEE.

Nauta, M., M. B. Habib, and M. van Keulen (2017). Detecting hacked twitter accounts based on behavioural change. In *13th International Conference on Web Information Systems and Technologies, WEBIST 2017*. INSTICC Institute for Systems and Technologies of Information, Control and Communication.

Raghav, J. (2014). Fighting spam with botmaker. Retrieved from `https://blog.twitter.com/engineering/en_us/a/2014/fighting-spam-with-botmaker.html`. Last visited: May 15th 2018.

Rao, P. R., A. Katib, C. A. Kamhoua, K. A. Kwiat, and L. Njilla (2016). Probabilistic inference on twitter data to discover suspicious users and malicious content. *2016 IEEE International Conference on Computer and Information Technology (CIT)*, 407–414.

Sikandar G, M. (2018). 100 social media statistics you must know [2018] + infographic. Retrieved from https://blog.statusbrew.com/social-media-statistics-2018-for-business/. Last visited: May 15th 2018.

Vo, N., K. Lee, C. Cao, T. Tran, and H. Choi (2017). Revealing and detecting malicious retweeter groups. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 363–368. ACM.

Whittaker, Z. (2016). A hacker claims to be selling millions of twitter accounts. zd-net. Retrieved from https://www.zdnet.com/article/twitter-32-million-credentials-accounts-selling-online/. Last visited: May 15th 2018.

Wu, T., S. Wen, Y. Xiang, and W. Zhou (2017). Twitter spam detection: Survey of new approaches and comparative study. *Computers & Security*.

Yang, C., R. Harkreader, and G. Gu (2013). Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security 8*(8), 1280–1293.

Yang, C., R. Harkreader, J. Zhang, S. Shin, and G. Gu (2012). Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, pp. 71–80. ACM.

Yang, C., R. C. Harkreader, and G. Gu (2011). Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *International Workshop on Recent Advances in Intrusion Detection*, pp. 318–337. Springer.

Yang, C., J. Zhang, and G. Gu (2014). A taste of tweets: reverse engineering twitter spammers. In *Proceedings of the 30th annual computer security applications conference*, pp. 86–95. ACM.

Zangerle, E. and G. Specht (2014). "sorry, i was hacked": a classification of compromised twitter accounts. In *SAC*.

Résumé

Les cyber-criminels exploitent les réseaux sociaux afin d'affecter plus de victimes. Les spammers de Twitter créent plusieurs bots et se comportent de différentes manière selon leurs attentions. En particulier, les spammers émergent le réseau par des liens malveillants amenant vers des malware ou liens d'hameçonnage. Achévant ça par s'engager avec d'autres utilisateurs ou réagir en observant les tendances. Ces criminels se manifestent individuellement, ou au sein d'un ou plusieurs groupes coordonnées avec des objectives assez précis. Décidément, le renforcement de la cyber-sécurité dans Twitter est indispensable afin d'éviter les pertes. Plusieurs chercheurs ont étudié les différentes aspects du spamming dans Twitter. Ce papier inclut un background sur les informations traitées dans Twitter, et un survey détaillé sur les papiers discutant le problème de spam. Les papiers ont été publiés de 2010 à 2018. Ce survey n'est pas limité à la détection des spammers mais il discute également les approches de la détection des communités de spam, les comptes piratés, *collective attention spam*, et l'extraction des connaissances sur le cybercrime. Par conséquant, cette étude peut être considérée essentielle afin de pouvoir designer un framework unifié de la détection du spam.