SPOT 1.0: Scoring Suspicious Profiles On Twitter

Charles PEREZ, Marc LEMERCIER, Babiga BIRREGAH, Alain CORPEL

ICD (Charles Delaunay Institute)

UMR CNRS STMR 6279

University of Technology of Troyes, 12 rue Marie Curie, 10 010 Troyes Cedex

{charles.perez, marc.lemercier, babiga.birregah, alain.corpel}@utt.fr

***Abstract*—Everyday more than fifty million messages are generated by about two hundred million profiles on Twitter. Some users attempt to exploit the success of this microblogging platform and its relative freedom to perform malicious actions that can lead to identity or data theft. This work aims to propose a framework to assess suspicious behavior on Twitter. We present a tool developed for Scoring Suspicious Profiles On Twitter (SPOT 1.0) through a three-dimensional indicator that involves the degree of aggressiveness, the visibility and the level of danger.**

***Keywords*-social network analysis; Twitter; suspicious profiles; support vector machine; aggressiveness; visibility; level of danger;**

I. INTRODUCTION

During these last decades Twitter rapidly became the leader of microblogging platforms on the web. Each user of Twitter can send messages of less than one hundred forty characters, called tweets, to a list of volunteer contacts called followers. These users can also receive a list of messages produced by a set of individuals called followees. The global success of the platform and the spontaneity of the exchanges induce a new type of threats for users. An individual acting on this platform is highly vulnerable to malwares and malicious URLs. These vulnerabilities can have high impact since an increasing amount of online data is personal and sensible. In this work, we propose a tool to detect and classify the suspicious profiles and evaluate the danger that they represent. SPOT 1.0 focuses on evaluating the level of danger of malicious URLs that can be embedded into a tweet. Our choice is motivated by the fact that URL shortening services ([1]), is nowadays massively used on Twitter and allows an individual to hide a malicious link behind a short URL. These malicious URLs, quickly diffused on a platform such as Twitter, can now lead to large amount data or identity theft ([2]).

During many years one of the main diffusion vectors of suspicious websites was e-mail. In order to face this threat, several works propose solutions to detect an e-mail spam. The proposed analyses are only based on the e-mail content ([3], [4]) or can also be improved by websites characteristics ([5], [6]). As indicated in [2], social medias platforms will gradually replace emails as primary vector for distributing malicious codes and links. [7] and [8] propose a classification tool to detect profiles that generate spam. These analyses consider the behavior of a given user and the content of sent messages but do not take into account the real harmfulness of these profiles on the users.

The tool presented in this work analyzes a suspicious behavior through three axis of investigation: the user profile, the sent messages and the URLs contained in the tweets. Then we

978-0-7695-4375-8/11 $26.00 © 2011 IEEE
DOI 10.1109/ASONAM.2011.63

propose an indicator to reveal how aggressive is a profile and how effective can be an attack led by a malicious profile on Twitter.

II. Architecture of our Framework

A huge amount of tweets are published and stored each second on the main Twitter database. The density of exchanges on the platform is one of the major challenges when one wants to analyze the profiles and their tweets. This section presents an overview of the architecture of SPOT 1.0 as a set of successive modules numbered from I to VI in the figure 1.

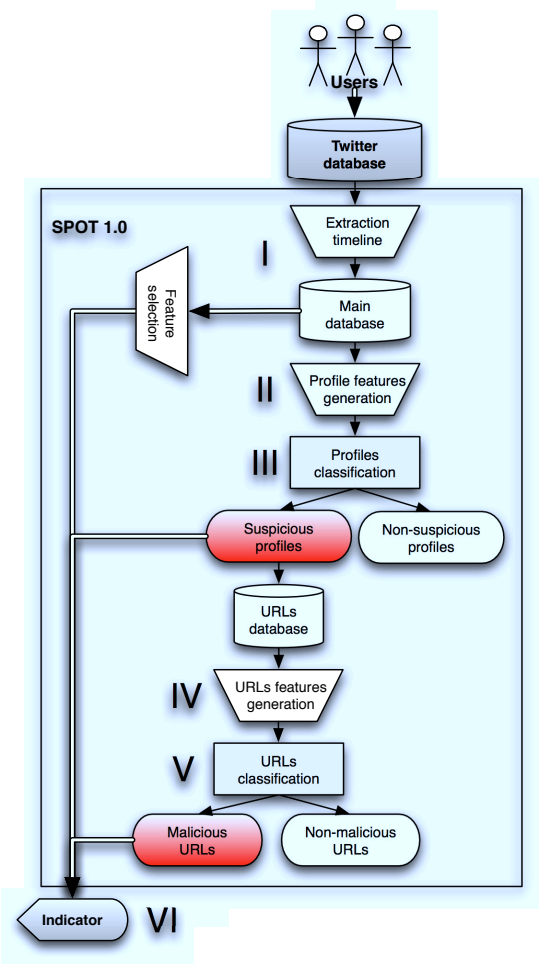

Figure 1. Overview of the architecture of SPOT 1.0

The *first module* of SPOT 1.0 is dedicated to collect on line tweets and profiles of users through the streaming API of Twitter. Simultaneously, to increase the amount of collected data, a webcrawler retrieves the lasts tweets produced by profiles. Although all profiles cannot be retrieved, the streaming API permits to collect an important subset of the most interesting active profiles. An active profile on Twitter naturally increases his likelihood to appear in the public timeline and by the same way increases his probability to be analyzed by our tool. The collected data is stored in SPOT's main database. It contains the information of Twitter profiles, the tweets sent by users and the different entities that appear in these tweets. From the collected data, the *second module* generates automatically a set of indicators that describes the behavior of a profile. The set of features contains: the number of followers and followees, the reputation, the frequency of tweet, the average number of URLs, hashtag, trends and references in tweets, the rate of replied tweets and the average distance ([9]) between any pair of tweets. Based on these features, incoming profiles are partitioned by the *third module* into suspicious or non-suspicious groups by a machine-learning algorithm. The classification method trains on an initial partition including suspicious profiles reported by users and the non-suspicious profiles reported by Twitter as verified. These two kinds of profiles have been collected using the Twitter streaming and search APIs**.** The current version of SPOT is based on the support vector classification ([10]). The parameters of the machine are selected with a cross-validation process associated with a grid search ([11]). We used the LIBSVM library for the implementation of the project ([12]). The *fourth module* investigates deeper the suspicious profiles. The URLs contained in their tweets are analyzed. When an URL is a short one, the module retrieves the corresponding long URL by extracting the location value from the HTTP header while connecting to the shortening service. This method allows to retrieving quickly the long URL without being redirected. The *fifth module* classifies each URL from the database into malicious and non-malicious. This classifica-

tion is done using the URL itself but also characteristics of the domain and content based features. Finally the *sixth module* combines the different results of the platform in order to evaluate the profiles in a relevant way.

In the next section we present the three-dimensional indicator built to evaluate the behavior of all the profiles.

III. Evaluation of suspicious profiles

We propose to use the result of SPOT 1.0 in order to analyze the severity of suspicious profiles on Twitter. We consider three aspects: (1) the degree of aggressiveness of a profile (2) the visibility of its action and (3) the level of danger of its tweets.

Degree of aggressiveness

The aggressiveness dimension reveals the speed of actions per profile on the platform. Twitter allows its users to achieve two main actions on the site. The first one is to publish tweets and the second one is to get new friends (i.e. followees). We built our first dimension as the frequency of tweet published per hour (f_{tweets}) plus the frequency of generated outgoing links per hour ($f_{friends}$) divided by the maximum limit of actions (equation 1).

$$A_p = \frac{f_{tweets} + f_{friends}}{f_{max}} \tag{1}$$

The maximum number of actions that can be performed through the API is currently 350 actions $f_{max}(API) = 350$. Typically a malicious action will be more efficient if the actor of this action is highly active (i.e. aggressive). When some conditions are met, the likelihood to succeed a malicious action increases with frequency of tweets per hour. One must notice that the act of following somebody can increase the number of followers and enrich a database of screen names. These latter can be used for directed attacks. A low score of the degree of aggressiveness reveals a dormant profile while a high score reveals a hyperactive profile.

Visibility

It is broadly accepted that the efficiency of a malicious content depend directly on the number of individuals that can access its messages. It is important to notice that the tweet is not only received by followers but may also be received or viewed by other members of the network. Indeed, the references and hashtags in a tweet extend the set of possible recipient of this tweet. References appear in a tweet as an "@" sign followed by the screen name of a user (e.g. @screen_name). The referenced profile(s) will automatically receive the message. On the other hand a user can place one or several hashtags in a tweet. Hashtags appear as a "#" symbol followed by a tag (e.g. #tag). This action tags the tweet with a specific keyword and can be found by a user while searching tweets on a specific topic. We built the second dimension as shown on the following equation:

$$V_p = \frac{\sum_{E\epsilon\{@,\#\}} Avg(E) * C(E)}{140} \tag{2}$$

This dimension reveals the efficiency a tweet of 140 characters has been built in order to perform a visible action. The score depends on the average use of the elements ($E\epsilon\{@, \#\}$) in a tweet. We use the main database in our tool SPOT 1.0 to retrieve the average cost of characters needed for a reference ($C(@) \approx 11.4$) and for a hashtag ($C(\#) \approx 11.6$). A user whose action is optimized will manage to use strategically the two entities and therefore will get a higher value of visibility than a user that does not use entities.

Level of Danger

We evaluate the level of danger of a profile based on the number of malicious tweets that it sent. In this work, one assumes that a tweet is considered malicious if it contains at least one malicious (short or long) URL. We propose to compute this dimension as the

ratio of malicious tweets (M_{tweets}) by the total number of tweets analyzed by our tool (T_{tweets}) as shown in the following equation.

$$D_p = \frac{M_{tweets}}{T_{tweets}} \quad (3)$$

A user who produces only suspicious tweets has a Dp value equal to one while a user who does not produce any malicious contents will have his Dp score equal to zero.

These three dimensions permit to represent profiles on a 3D plot that reveal the real efficiency of suspicious and normal profiles. This three-dimensional indicator is frequently updated and allows to detect malicious profiles that was previously asleep.

IV. Results

We present in this section the results of the implementation of SPOT on a sample of collected profiles. We then introduce some interesting feedbacks and statistical observations to better understand the behavior and the strategy of a malicious attacker on Twitter.

Global observation

SPOT collects approximately 500,000 tweets everyday. Every tweet is collected with the corresponding profile of its sender. These information help to generate in real-time the features that are necessary to classify suspicious profile and their corresponding URLs. The figure 2 plots 5,000 automatically classified profiles based on the three dimensions presented. This visualization tool aims to give a fast representation of the situation to help decision making on a global or a local point of view.

According to this representation the profiles with high values in all three dimensions will be considered as more malicious. In figure 2(c), suspicious profiles appear in red while non suspicious one appear in blue. Since the risk studied in this paper is linked to the URLs, all the profiles located in the plane defined by the aggressiveness and visibility dimensions can be considered as "non malicious".

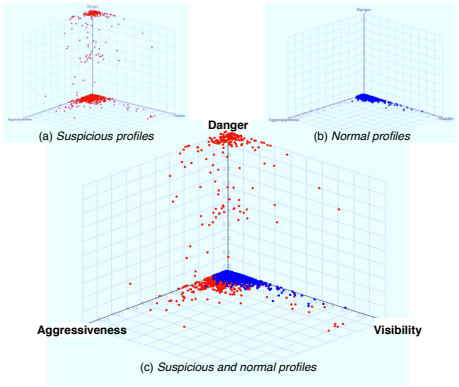

Figure 2. Three dimensional scoring of profiles by SPOT

The positions of suspicious profiles (figure 2(a)) are more distributed and so more heterogeneous in their strategies and behavior. SPOT can also help to identify normal profiles that have a high visibility on the network (figure 2(b)).

Observations about malicious behaviors

One can investigate the strategy of malicious actors of the platform using the collected features with respect to the results of our tool. Table I indicates the average values of some features used for classification on our platform. This table distinguishes the averages of profiles classified as suspicious and normal.

	Suspicious	Normal
Age in days	190	**465**
Number of Friends	**941**	753
Number of Followers	1,600	**1,724**
URLs per tweets	**0.22**	0.07
Références per tweet	0.80	0.84
Hashtags per tweet	**0.25**	0.14
Retweet per tweet	0.09	**0.2**
Frequency tweets per day	**131**	35
Distance between tweets	**0.68**	0.59
Reputation	0.58	**0.59**

Table I
Comparison of Averages of features for suspicious and normal profiles

It is readily seen that the age of a suspicious profile is much less important than the age of the normal profiles. This could be explained

by the efforts of Twitter in the detection and the deletion of some suspicious profiles. The suspicious profiles have more friends and fewer followers than "normal" users ([7], [13]) and are usually less followed by other profiles. One of the main reasons could be that their tweets do not attract many users. A suspicious profile generates more URLs and hashtags in his tweets than "normal" ones; often with the intention of dispatching advertisements or phishings. Finally suspicious profiles send more tweets per day and these tweets are more similar than the one sent by normal users.

The standard deviation for the same features is presented in table II.

	Suspicious	Normal
Age in days	**164**	175
Number of Friends	3,642	**2,679**
Number of Followers	9,562	**9,000**
URLs per tweets	0.38	**0.16**
Références per tweet	0.75	**0.58**
Hashtags per tweet	0.76	**0.27**
Retweet per tweet	**0.19**	0.25
Frequency tweets per day	151	**26**
Distance between tweets	0.18	**0.16**
Reputation	0.21	**0.18**

Table II
COMPARISON OF STANDARD DEVIATION FOR SUSPICIOUS AND NORMAL PROFILES

The standard deviation for most of the features is less important for normal profiles than for suspicious ones. This means that suspicious profiles are more heterogeneous than normal profiles and that a normal attitude is more predictable than a suspicious one. Two characteristics remain important to discriminate the suspicious profiles: a low operating lifetime and a low number of responses to received tweets. This observation confirms that most of the malicious actions are managed automatically.

V. CONCLUSION

We have presented a tool developed for Scoring Suspicious Profiles On Twitter through a tri-dimensional indicator. Our tool allows firstly to classifying a suspicious behavior from a normal one and then analyzes more deeply the suspicious profiles through a danger dimension. In this work, we choose the malicious URLs as the danger dimension. However, SPOT is designed to manage other dimensions of danger such as targeted keywords. Further works will be devoted to the topological analysis of the malicious profiles detected by SPOT.

REFERENCES

[1] D. Sullivan. (2009, April) Url shorteners: Which shortening service should you use? [Online]. Available: http://searchengineland.com/analysis-which-url-shortening-service-should-you-use-17204.

[2] D. Alperovitch, T. Dirro, P. Greve, R. Kashyap, D. Marcus, S. Masiello, F. Paget, and C. Schmugar, "2011 threats predictions," Mc Afee, Tech. Rep., 2010.

[3] A. Wang, "Don't follow me: Spam detection in twitter," in *Int'l Conference on Security and Cryptography (SECRYPT)*, 2010.

[4] R. R. S. Appavu alias Balamurugan, "Data mining techniques for suspicious email detection: a comparative study," in *European Conference on Data mining*, Portugal, 2007.

[5] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," Retrieved Sep, Tech. Rep., 2006.

[6] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM Press, 2007, pp. 639–648.

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[8] A. H. Wang, "Detecting spam bots in online social networking sites: a machine learning approach," in *Proceedings of the 24th annual IFIP WG 11.3 working conference on Data and applications security and privacy*, ser. DBSec'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 335–342.

[9] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady.*, vol. 10, no. 8, pp. 707–710, February 1966.

[10] V. Vapnik and A. Lerner, "Pattern recognition using generalized portrait method," *Automation and Remote Control*, vol. 24, 1963.

[11] C. W. Hsu, C. C. Chang, and C. J. Lin, *A practical guide to support vector classification*, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2003.

[12] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001.

[13] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?" in *Proceedings of the 26th Annual Computer Security Applications Conference*, ser. ACSAC '10. New York, NY, USA: ACM, 2010, pp. 21–30.