COPYRIGHT NOTICE

FedUni ResearchOnline
<http://researchonline.federation.edu.au>

This is the author's version of a work that was accepted for publication in the AI Communication. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in the AI Communication, 29(1), 107-122 DOI: <http://doi.org/10.3233/AIC-150659>

Copyright © 2016, IOS Press and the authors. All rights reserved..

REPLOT: REtrieving Profile Links On Twitter for malicious campaign discovery

Perez Charles

University of Technology of Troyes, France

E-mail: charles.perez@utt.fr

Birregah Babiga

University of Technology of Troyes, France

E-mail: babiga.birregah@utt.fr

Layton Robert

Federation University, Australia

E-mail: r.layton@icsl.com.au

Lemercier Marc

University of Technology of Troyes, France

E-mail: marc.lemercier@utt.fr

Watters Paul

Massey University, New Zealand

E-mail: P.A.Watters@massey.ac.nz

Social networking sites are increasingly subject to malicious activities such as self-propagating worms, confidence scams and drive-by-download malware. The high number of users associated with the presence of sensitive data, such as personal or professional information, is certainly an unprecedented opportunity for attackers. These attackers are moving away from previous platforms of attack, such as emails, towards social networking websites. In this paper, we present a full stack methodology for the identification of campaigns of malicious profiles on social networking sites, composed of maliciousness classification, campaign discovery and attack profiling. The methodology named REPLOT, for REtrieving Profile Links On Twitter, contains three major phases. First, profiles are analysed to determine whether they are more likely to be malicious or benign. Second, connections between suspected malicious profiles are retrieved using a late data fusion approach consisting of temporal and authorship analysis based models to discover campaigns. Third, the analysis of the discovered campaigns is performed to understand the attacks. In this paper, we apply this methodology to a real world dataset, with a view to understanding the links between malicious profiles, their attack methods and their connections. Our analysis identifies a cluster of linked profiles focusing on propagating malicious links, as well as profiling two other major clusters of attacking campaigns.

Keywords: Online social networks, Suspicious profiles, Clustering, Malicious campaigns, Authorship attribution, Twitter

Introduction

The success of online social networks has lead to important challenges in the context of security. The survey provided in [10] regroups the security issues into the following four groups: privacy breaches, viral marketing, network structural attacks, and malware

attacks. In this paper, we focus on the detection of malicious content propagation (viral marketing) issued from a set of profiles that combines their effort (network structural attacks). The large number of users and the characteristics of the online social networks (e.g. small world properties) make them particularly vulnerable to malicious content propagation. Often per-

formed in the form of URLs contained in messages, malicious attacks can lead to a very large number of people being infected in a very short period of time. The Koobface worm has proved the real-life feasibility, on a very large scale, of malicious content propagation on online social networks [4]. From simple use of a profile for performing attacks, malicious actors have now moved on to a more collective and synchronised way of undertaking actions. This allows them to increase the impact of their attack by increasing artificially their reputation and by the same way the total number of targeted profiles. As a result, the detection of malicious profiles is often insufficient to eradicate malicious campaigns and a characterization of them is required.

One of the key measures to improving defences is to analyse and understand how these attacks occur and how they organise. Gathering intelligence on attacks allows for the creation of short term fixes, such as new signatures, and long term preventions, such as alterations to the policies of the organisation or system properties. As an example, a campaign spreading malicious links with random verses from a given book. A short term signature for this attack would be the detection of verses from the book. A longer term solution would be for the website to analyse the links to see if there is any correlation between the post and linked content.

The analysis of malicious campaigns can be a time consuming process, especially for attacks that obfuscate their actions. For this reason, techniques and methodologies that can automate parts of this process are needed for social media companies, security companies and law enforcement. Automation allows for both the breadth and depth of an analysis to be increased, without increasing the burden on analysts. The main contribution of this paper is therefore a methodology for discovering social media campaigns and providing intelligence in an automated method. This allows for an analyst to perform analysis on an entire campaign, rather than individual malicious profiles.

In this paper, we present a full stack methodology for the discovery of malicious campaigns on social networking websites. The methodology named REPlot, for REtrieving Profile Links On Twitter, contains three major phases. First, individual profiles are analysed to determine whether they are malicious or benign. These profiles are analysed using a set of behavioural features and a supervised machine learning algorithm is then used to classify them as malicious or benign. Second, connections between suspected malicious profiles

using a late data fusion approach consisting of temporal and authorship analysis based models. The temporal feature examines the correlation between posts using an entity based temporal similarity metric. The authorship analysis technique uses character n-grams to model authorship, linking disjoint profiles if they are likely to share a common author. Third, clustering is used to profile the different malicious campaigns. Clustering is performed by creating a graph based on the metrics analysed in phase two, with a view to finding sections of the graph with a high internal similarity.

By using this methodology, malicious profiles are not only discovered automatically, but also profiled. For an analyst needing to discover trends and patterns in attacking profiles, this level of automation reducing the number of profiles needing to be visited. Rather than analysing a very large number of individual profiles, a smaller number of clusters can instead be analysed.

We apply this methodology to a real world dataset, with a view to understanding the links between malicious profiles, their attack methods and their connections. Our analysis identifies a cluster of linked profiles focusing on propagating malicious links, as well as profiling two other major clusters of attacking campaigns. Our contributions are therefore twofold; first, we present a methodology consistent in approach for the detection and analysis of malicious profiles and second, we present the analysis of malicious profiles discovered on a social networking website.

State of the Art

Two main approaches can be used to address the problem of malicious action detection on online social networks: first, approaches that directly detect potential sources of malicious actions (i.e. messages or profiles); second, approaches that detect communities of malicious profiles. This section provides an overview of the state of the art of these two categories of approaches.

Malicious profile detection

The detection of malicious profiles or messages is often treated as a classification problem. Such approaches can be seen as an extension of traditional e-mail spam detection adapted to the characteristics of online social networks [9]. These techniques usually consider the existence of two classes of messages or

profiles: legitimate and spam. In some works, the level of danger of spam is also measured and is referred to as malicious.

We have presented a method for detecting malicious profiles on the Twitter platform in [22]. This approach, called SPOT (for Scoring Suspicious Profiles On Twitter), performs in two phases: first, suspicious profiles are detected based on content-based, graph-based and behaviour-based generated features; second, messages published by suspicious profiles are investigated to detect malicious URLs. Note that this work is a current extension of SPOT that integrates the similarity between suspicious profiles for characterizing suspicious social campaigns. The proposed methodology in the present paper can allow to bypass the scale limitation of SPOT which can arise when one consider the high number of candidate profiles to analyse. Indeed, at any time a profile that can be detected as suspicious by SPOT can be integrated to our methodology in order to identify whether it belongs or not to a malicious campaign set.

A method for identifying spam on Twitter is proposed in [27]. In the frame of this research, a set of behavioural and graph-based features extracted from the Twitter platform was generated. A Bayesian classifier is applied to the generated feature for identifying spammers on Twitter. One limitation of this approach is undoubtedly the large amount of data to handle. In this particular case, the rapid increasing of malicious actions can not be addressed in an exhaustive manner.

[24] have proposed a methodology for identifying spam on social networking sites. The methodology is based on a set of 'honey-profiles' that aim to raise the attention of malicious users. The data collected from the honey-profiles are used to perform spam detection. The features considered for the classification of a profile are the number of friends, the number of messages, the URL ratio, etc. The classification is performed using a random forest algorithm.

[2] has proposed to classify messages in the three following classes: legitimate, spam and malicious. The application of the approach is made on Facebook and remains on a Support Vector Machine (SVM) classification. In [18], the authors have handled the problem of collective attention spam. This problem refers to the possible use of a trending topic inside messages to artificially increase their visibility. A decision tree based random forests is proposed for addressing the problem. The set of features contains the number of hashtags, the number of URLs, a bag of words, etc.

It is important to note that, methodologies applied for detecting malicious profiles can hardly address the

large amount of data daily generated on social networking sites. Such approaches require the analysis of any profile while only a few represent a potential danger on the platform. To cover this problem, a priority must be fixed and a possible solution is to focus on the detection of networks of malicious profiles. These networks must be treated in priority since their action have a potential very high impact on users.

Social campaign detection

We present below some approaches that aim to detect connections between malicious profiles such as spam campaigns on online social networks. In the remainder of this section, we refer to suspicious campaigns as a set of suspicious profiles that perform similar actions and act in coordination. When the profiles of such campaigns propagate malicious URLs, the campaign will be referred as malicious.

[24] retrieves connections between malicious profiles on the basis of their published URLs. The assumption is that two profiles that publish messages containing the same malicious URL belong to the same malicious campaign. Note that the use of URL shortening services can complicate this task, as can dynamic URLs with randomised components.

Some approaches use a clustering-based algorithm for identifying communities of normal and spam profiles. As an example, [3] has proposed to apply a clustering algorithm on a large set of behavioural and content-based features. The results show that the approach can enable spam profiles to be identified with good accuracy. However, this approach does not enable the identification of tighter connections between spam profiles (i.e. spam campaigns).

[6] has introduced the concept of Socialbot Networks (SbN), a set of profiles that are owned and maintained by a unique human controller. Authors have proved the efficiency of such networks and have also highlighted that connecting bots together can artificially increase the attractiveness of a given malicious profile.

[11,12] have investigated the relationships existing between malicious profiles on online social networks. They have analysed the link-farming problem that refers to the creation of illegitimate connections to increase the visibility of profiles artificially. Their results have proved that not only do spammers generate link farming but also that many legitimate profiles (with high reputation) accept and generate non-legitimate connections to increase their audience. Such

observation highlights the potential difficulty of distinguishing malicious campaigns from legitimate ones while using the social graph of a platform.

[28] proposes a solution to identify worm propagation inside an online social network. The authors propose to create a decoy network of profiles and analyse the content received by these profiles. The detection rests on a local and network level using similarity measures between messages in order to determine whether there is evidence of propagation.

In [29], the authors have proposed an approach to detect promotion and spam campaigns on Twitter. Their approach is based on a URL-based similarity measure between two accounts. From this measure, they build a similarity graph from which sub-graphs are identified as campaigns. A classification of campaigns is performed by applying a support vector machine classification on campaign-based features.

A further method of linking profiles can be gained through analysis of the content, aiming to link profiles that share content generated by the same author. This is a problem in the field of authorship analysis, which aims to identify salient features within written documents that are hard to obfuscate. Authorship analysis techniques can be broadly categorised into two categories: vector space methods [30] and local profiles-based approaches [13].

Vector space methods take a set of documents and a set of features describing them, such as average sentence length or the frequency of a specific word. A matrix X is then generated such that $X_{i,j}$ is the value of feature j for document i . This feature matrix is then used as input into a standard machine learning algorithm, such as support vector machines [21].

Local profile methods commonly employ character n -grams as features, extracting the most descriptive n -grams from each author. An author is then profiled using the L most descriptive n -grams in their writing [13]. Documents of unknown authorship are profiled in the same way. These profiles are then compared to the author profiles using a distance (pseudo) metric, and the nearest is predicted as the likely author of the document. In their work on the Common N -Grams (CNG) model, [13] use the following equation to compare profiles P_1 and P_2 , where $P_i(x)$ is the frequency of n -gram x in profile P_i :

$$K(P_1, P_2) = \sum_{x \in X_{P_1} \cup X_{P_2}} \left(\frac{2 \cdot (P_1(x) - P_2(x))}{P_1(x) + P_2(x)} \right)^2 \quad (1)$$

Recent research has shown that this approach appears to have better results in a wide variety of contexts

[8], and more recently in online social media [14,15], despite the short content lengths. In addition, character n -grams are language independent and can be applied to other contexts, including malware detection.

Methodology

The proposed framework rests on a combination of behavioural and content-based analysis of Twitter profiles. The behavioural analysis mainly involves the SPOT tool and the content-based analysis is ensured by the NUANCE methodology [22,16]. Figure 1 gives a global overview of the proposed approach.

The first step (referred as step I in figure 1) aims to identify the suspicious users. For this purpose, a harvester module is connected to the online social network (i.e. Twitter). The module collects and stores the activities produced by Twitter profiles in a local database. A profile detection algorithm is performed and the selected profiles are represented by a set of behavioural features. A classification is performed to identify the suspicious and normal profiles.

The second step (referred as step II in figure 1) aims to identify important relationships between suspicious profiles. This work is performed using a content-based similarity measure. This similarity combines an authorship attribution method with an entity-based temporal measure.

Finally, a clustering algorithm is performed for highlighting campaigns (referred as step III in figure 1). The detected campaigns are characterised to evaluate their particularities (size, dynamic, etc.) and their level of danger.

This section provides explanation on these three key steps of our methodology.

Suspicious profile detection

The first step of our approach involves near real-time detection of suspicious profiles. The main objective of this step is to determine the class of a profile (i.e. suspicious or not). Two main constraints for this step exist: (1) the number of profiles treated and (2) the rapidity of the analysis. The number of users makes it impossible to propose an exhaustive approach. Thus, a priority strategy must be proposed. It is also obvious that one cannot perform content-based analysis in this step since the quantity of tweets to analyse is too large. We propose to define priority proportionally to the number of messages that appears in the public timeline of Twit-

Fig. 1. Overview of the proposed approach

ter for a given profile. The timeline of Twitter consists of a real-time random sampling of the tweets that are received on the platform. Logically, a profile that sends a very large number of messages increases its probability of being published in the timeline.

In this work, we focus on such active profiles because they are the most visible on the platform and thus may raise the largest audience when performing an attack. Note however that, once the malicious clusters are highlighted, it could be possible to test any profile and determine whether or not it belongs to one of these campaigns. We perform this step by connecting to the public timeline via the Twitter stream API and storing the received messages in a local database. A threshold K is set to determine whether a given profile p needs to be analysed. The classification of suspicious profiles is based on the SPOT 1.0 tool that performs a support vector machine algorithm on the set of behavioural features; the number of followers and friends, the reputation [26], the tweet frequency, the average number of URLs, hashtags, references in tweets, the rate of forwarded tweets (aka retweet) and the average of the Levenshtein distance applied between each couple of tweets of a profile.

Retrieving connections between suspicious profiles

The second step of the proposed approach concerns the investigation of the suspicious profiles obtained. A content-based analysis identifies connections between profiles. First, an authorship attribution algorithm is performed on the collected tweets of the suspicious profiles [15,14]. Then, a temporal similarity is measured on the entities contained in the messages. Both analyses contribute to the measure of the similarity of profiles in a late data fusion approach, where the results are combined after the models are applied.

The authorship analysis was performed using a trimmed version of the NUANCE algorithm [16], which was originally used in [14]. In the normal NUANCE algorithm, an ensemble is used to cluster documents by authorship. While it has been shown to be effective, the algorithm has a high complexity, making it infeasible to use in its current form for this problem. Instead, we used a single instance of the Recentred Local Profiles algorithm [17] to compare the profiles of each account.

The distance between all pairs of accounts was calculated and the likelihood values were calculated by comparing the distances to precomputed distances from a training dataset containing known account matches (see [14] for details). The distribution of the precomputed distances is used to calculate these empirical likelihoods, such that the likelihood of two accounts being matches is given as the percentage of pairs of accounts in the training dataset with a higher distance.

Account matches with a likelihood above a given threshold were considered to be matches. While ‘normal’ thresholds, such as 0.9 or 0.95, could be used, we computed the optimal threshold using a training dataset that maximises the F-measure (i.e. the harmonic mean of precision and recall) and used that value, which was 0.989. Using higher thresholds resulted in lower recall values, while lower thresholds gave lower precision values.

Regarding the entity-based temporal similarity, the time of the experiment is divided into time frames T_i of equal length δt . Figure 2 presents the algorithm for measuring the entity-based temporal similarity between two profiles u and v at a given time frame T_i . Such similarity aims to identify profiles that act in co-ordination. The underlying assumption is that social campaigns that have a common goal should, at a given time frame, exhibit a minimum of similarity. The following three types of entities in Twitter are considered:

references (e.g. @screenname), hashtags (e.g. #hashtags) and URLs. For a given pair of profiles similar references can reveal that the targeted individuals are identical, sharing same hashtags can reveal a common approach to catch the attention of users and using same malicious URL can reveal the use of the same strategy.

Inputs:

- (u, v) a couple of Twitter profiles
- $E(u)$ the set of entities published by a profile u
- t_e^u the publication time of entity e by profile u

Output:

$B_T(u, v)$ the vector containing the similarity scores $B_{T_i}(u, v)$ of each time step i

```

1  $B_T(u, v) \leftarrow \emptyset$ 
2 foreach  $e \in E(u) \cap E(v)$ 
3   foreach  $T_i \in [0, T]$ 
4     if  $t_e^u \in [T_i - \frac{t}{2}, T_i + \frac{t}{2}]$  or  $t_e^v \in [T_i - \frac{t}{2}, T_i + \frac{t}{2}]$ 
5        $B_{T_i}(u, v) = B_{T_i}(u, v) + \log(\frac{|M|}{|\{m_j: e \in m_j\}|}) * e^{-\frac{|t_e^u - t_e^v|}{\delta}}$ 
6   endif
7 end foreach
8 end foreach
9 return  $B_T(u, v)$ 

```

Fig. 2. Algorithm for measuring the entity-based temporal similarity of profiles

The input of the algorithm is a couple of suspicious profiles denoted (u, v) and the output is a vector $B_T(u, v) = (B_{T_1}(u, v), B_{T_2}(u, v), \dots)$ where $B_{T_i}(u, v)$ is the temporal similarity at time step T_i . The temporal similarity vector is initialised at line 1. A loop is performed on each entity e that has been published by both profiles u and v (line 2). The moment of publication of such entities is investigated to identify the corresponding time step (lines 3 and 4). When the entity is identified as belonging to the current time frame, the similarity score is increased by a score such as indicated in line 5. The first component of the score is calculated based on the Inverse Document Frequency (IDF) feature [23]. In our case, a document refers to a message and a term corresponds to an entity e . m_j refers to the set of tweets that contain the entity and M is the number of tweets concerned in the experiment.

The second component of the score is a measure of the synchronization observed between the common entities that are published. The measure of the synchronization rests on an exponential decay. The mean life-

time δ , allows the similarity score to decrease more or less rapidly, as the synchronization delta ($|t_e^u - t_e^v|$) is important. Setting δ to an important value will enable the detection of strongly synchronised publication of profiles. Conversely, a small value will enable the detection of similar profiles even if they are not strictly simultaneous.

This entity-based temporal similarity score is combined with the result of the authorship attribution to provide a unique content-based similarity between a given pair of profiles. This global similarity is identified as the strength of a tie between profiles. We propose to define the strength $w_{T_i}(u, v)$ between a profile u and a profile v during a time step T_i as stated in equation 2.

$$w_{T_i}(u, v) = A(u, v) + B_{T_i}(u, v) \quad (2)$$

Where:

$A(u, v)$ refers to the result of the authorship attribution between u and v

$B_{T_i}(u, v)$ refers to the entity based temporal similarity between u and v

Note that we do not perform the authorship attribution dynamically. One reason is that the number of messages required to perform such analysis and obtain good performance is relatively important (approximately 200 tweets). In our case, this is made possible by collecting the last 200 tweets of each suspicious profile with the Twitter search API. However, we cannot ensure that a sufficient amount of messages is published at each time step for applying authorship attribution dynamically. The complexity of the algorithm is another reason for this choice. The alpha and beta coefficients must be adapted to the constraints of the experiment (time-response, number of messages, number of profiles, etc.).

Suspicious campaign detection

In the third step, we aim to reveal suspicious profiles that share significant similarities. For this purpose, we propose to build a suspicious graph denoted $G(N, E)$ that is composed by a set of nodes N (i.e. suspicious profiles) and a set of edges E (i.e. relationships). The weight on an edge between two profiles u and v is measured as the sum, over the time of the experiment, of the weights $w_{T_i}(u, v)$.

We apply a threshold to remove non-significant edges from the suspicious graph and then perform a

modularity-based clustering algorithm to this graph to detect suspicious campaigns [19]. The modularity of a graph is a measure of the strength of a particular division of a network into communities [20]. Modularity is high when a partition obtains many edges within communities and only a few between them. The modularity-based algorithm uses a greedy optimization to find partitions of the graphs that have a high modularity. It starts by considering each node as a community and successively merges communities that generate the highest increase in the modularity score.

The clusters are characterised both by content-based and behaviour-based attributes in the aim to clearly identify their purpose. The characterization applied to clusters has a double aim; first it enables better understanding of the strategies used by malicious actors and, second, it can enable identification of potential undetected profiles that belong to these clusters by applying features such as the affinity to cluster.

Results

We tested our methodology on a set of profiles that produced an important number of messages that were published in the public timeline of Twitter during the time of the experiment (from 01/03/2012 to 10/04/2012). For this experiment, the threshold parameter K was set to 20 messages (such a threshold enables identification of only very active profiles). A set of 1,000 suspicious profiles identified by the module I was selected for deeper investigation. The suspicious profile classification finally obtained an accuracy of 81% with a precision equals to 96% and a recall equals to 82%. Figure 3 represents the suspicious graph $G(N, E)$ which contains the connections between profiles with regard to their similarity (parameters α and β was set to equal weights). Graphs are displayed using Gephi software [5].

In sub-figure (a) the graph contains all the connections identified during the experiment. The number of nodes equals 680, that is approximately 70% of the suspicious individuals analysed. This first observation indicates that less than 30% of suspicious profiles from our dataset appear to behave completely independently of any kind of campaign. The measured clustering coefficient is equal to 0.505, the modularity is equal to 0.681 and the average path length is equal to 2.87. In sub-figure (b) we have highlighted the main clusters by deleting non-significant connections (edges with a low weight).

Fig. 3. Suspicious graph that contains detected campaigns

We present below the results of the behavioural characterization for the three major campaigns identified. Figure 4 illustrates the number of nodes belonging to each of the three main clusters during the time of the experiment.

Cluster 1 has approximately 100 participants with a maximum number of 140 at day 20. Cluster 2 has a relatively stable number of participants of 40. Cluster 3 has a decreasing number of participants: 140 at day one and only 25 at day 20. The first observations show that the suspicious networks are dynamic and evolve significantly over time. A decrease in the size of a network can be explained by various factors such as the deletion of malicious profiles by Twitter. It is also plausible that a malicious user managing a network at a given time decides to stop the campaign, either temporarily or permanently. Conversely, an increase of the

Fig. 4. Number of profiles belonging to the clusters over time

number of participants in a suspicious network can be explained by creation of new profiles, increasing the number of suspicious profiles acting in coordination at a given time.

We set forth in table 2 to 4 the behavioural characteristics of the clusters obtained. For comparison purpose, we have displayed in table 1 the average values for the same characteristics over a large set of normal profiles. We observe that among the most relevant parameters, the average number of hashtags (*Avg #*) used in tweets is much higher for suspicious profiles than for normal profiles. This observation confirms the fact that a suspicious profile aims to gain in visibility by including hashtags in its messages, as hashtags enable a community of interest to receive tweets.

On the other hand, normal profiles use more references (*Avg @* = 0.68) than suspicious profiles (*Avg @* = 0.30). Remember that tweets containing references (e.g. @screenname) are automatically received by the referenced person. The scores reveal that malicious campaigns are not targeting particular individuals but rather aim to reach larger communities identified by hashtags. The average URL number per tweet is also much higher for profiles belonging to suspicious communities (*Avg URL* = 0.70) than for normal profiles (*Avg URL* = 0.20). This reveals that these profiles act on the platform mostly to increase the visibility of a given set of websites.

The behavioural characteristics of suspicious clusters indicate that there is no significant distinction between the three suspicious campaigns. To better characterise them, we propose to investigate the quality of the URLs that they have promoted.

Figures 5 to 7 illustrate the box plots of the scores obtained for the set of URLs produced by each campaign over a set of four indicators. On this plot the dots represent the outliers of the data sample. The minimum

Table 1
Average features scores for normal profiles

	Averages
Account age	513
Freq tweet	94.0
Avg #	0.18
Avg @	0.68
Avg URL	0.20
Freq RT	0.16
#Friends	1049
#Followers	2836
#Tweets	110

Table 2
Behavioral characteristics of the spam campaign 1

	Avg	Min	Max	Med
Account age	668	312	1323	560
Freq tweet	106	0.11	429	91.9
Avg #	2.00	0.10	5.50	1.7
Avg @	0.60	0.00	3.9	0.00
Avg URL	0.69	0.00	1.23	0.94
Freq RT	0.20	0.00	3.30	0.00
#Friends	5086	0.00	22127	3234
#Followers	6473	3.00	22680	4328
#Tweets	321	229	485	297

Table 3
Behavioral characteristics of the spam campaign 2

	Avg	Min	Max	Med
Account age	462	308	1363	365
Freq tweet	88.8	0.07	402	40.5
Avg #	2.03	0.33	6.53	1.61
Avg @	0.01	0.00	0.09	0.00
Avg URL	0.81	0.01	1.07	0.96
Freq RT	0.00	0.00	0.00	0.00
#Friends	1349.1	0.00	9819	191
#Followers	1297	0.00	8949	460
#Tweets	268	232	314	268

Table 4
Behavioral characteristics of the spam campaign 3

	Avg	Min	Max	Med
Age account	615	322	989	575
Freq tweet	139	11.0	370	87.2
Avg #	2.01	1.01	3.48	1.78
Avg @	0.04	0.00	0.15	0.01
Avg URL	0.81	0.43	1.00	0.9
Freq RT	0.02	0.00	0.08	0.01
#Friends	999	6.00	2210	890
#Followers	1404	155	2882	1289
#Tweets	293	237	381	278

and maximum (excluding outliers) are represented by the lower and upper horizontal lines. The bounds of the main rectangle are defined by the values of the lower and upper quartiles. The median is displayed as a bold line.

This evaluation has been performed based on the Web Of Trust (WOT) API that uses a crowdsourcing approach to evaluate the quality of a website [1]. The four criteria are 1) trustworthiness 2) reliability, 3) privacy and 4) child safety. Trustworthiness refers to the global trust that one gives to the website (service, safety, etc.). Reliability (vendor reliability) refers to trust that one gives to the website for performing business transactions (e.g. buying, selling, etc.). Privacy refers to the trust regarding handling of personal information. Finally, child safety refers to the existence of age-inappropriate material on the site. Each criterion is given a score between 0 and 100 that reflect the trust of the website and that should be interpreted as follows: ≥ 80 Excellent; ≥ 60 Good; ≥ 40 Unsatisfactory; ≥ 20 Poor; ≥ 0 Very poor.

The results highlight that clusters one, two and three seem to quote trustworthy websites overall. However, a few anomalies for clusters two and three reveal the fact that some of the websites have a very poor score regarding child safety. Cluster one reveals a significant set of abnormal websites with very poor characteristics. It is observed that a set of quoted websites is not completely trusted for any criterion analysed. This clearly highlights the fact that a set of websites is malicious and that the campaign has malicious (e.g. phishing) intent.

Fig. 5. Characterisation of URLs produced by spam campaign 1

Fig. 6. Characterisation of URLs produced by spam campaign 2

Fig. 7. Characterisation of URLs produced by spam campaign 3

Performances

We measure the performance of our approach with Purity, Inverse Purity and F measures [25]. These measures aim to evaluate the differences between the test data sample (i.e. profiles with verified labels) denoted

L with the obtained data sample (i.e. profiles with labels issued from the algorithm) denoted C .

Given a set of manually labelled spam campaigns $L = L_1, \dots, L_m$ and a set of detected campaigns $C = C_1, \dots, C_k$, the purity of a clustering approach is defined as follows:

$$Purity = \frac{k}{|C|} \max_{j \in 1, \dots, m} Precision(C_i, L_j) \quad (3)$$

Where the precision of a detected campaign C_i , for a given category L_j is defined as follows:

$$Precision(C_i, L_j) = \frac{|C_i \cap L_j|}{|C_i|} \quad (4)$$

The basic relationship between precision and recall is obtained by the following formula:

$$Recall(C_i, L_j) = Precision(L_i, C_j) \quad (5)$$

The inverse purity is defined by:

$$Inverse Purity = \frac{m}{n} \max_{j \in 1, \dots, k} Recall(C_j, L_i) \quad (6)$$

Note that for each detected campaign, the precision is evaluated by identifying the labelled campaign that shares the highest number of common profiles. Similarly, the recall of a labelled campaign is calculated based on the detected campaign that contains the highest number of common profiles.

Finally, the performances are evaluated with the F-measure, which is defined in equation 7.

$$F = \frac{m}{n} \max_{j \in 1, \dots, k} F(L_i, C_j) \quad (7)$$

Where:

$$F(L_i, C_j) = \frac{2 * Recall(L_i, C_j) * Precision(L_i, C_j)}{Recall(L_i, C_j) + Precision(L_i, C_j)} \quad (8)$$

Figure 8 presents the precision and recall obtained for the three main campaigns detected in our data sample. We obtained a purity equal to 0.81 and an inverse purity equal to 0.74. The final F-measure obtained is equal to 0.76.

Note that these scores are calculated based on the full time of the experiment. The dynamic aspect of the campaign can generate variation of the results that depends on the selected time-frame.

	Cluster_1	Cluster_2	Cluster_3
Precision	0.79	0.74	0.81
Recall	0.98	0.67	0.74

Fig. 8. Precision and recall for the three main campaigns of the data sample.

Discovered Campaigns

We have evaluated the reputation related to an average benign profile such as the one of a profile belonging to detected campaigns. The reputation of a Twitter profile denoted u represented by a node in the social graph is defined in equation 9 as proposed by [26]. This value is equal to the number of followers divided by the number of friends and followers. This measure allows to calculate the level of attractiveness of a profile (captured by the number of incoming connections) moderated by its own proportion to create connections.

$$R(u) = \frac{d_{in}(u)}{d_{in}(u) + d_{out}(u)} \quad (9)$$

Based on our results, a given benign profile has an overall reputation of 0.73 which reveals the fact that it attracts a number of followers that is roughly equals to double the number of their friends. Remember that the values of average number of friends and followers are relatively high because of the selection process of SPOT, that only analyses the most active profiles of the Twitter platform.

The three campaigns possess an average reputation of 0.56, 0.49 and 0.58 respectively for campaign I, II and III. Although these scores are lower than for a benign profile these are however relatively high. These scores highlight the fact that surprisingly the malicious campaigns do succeed to attract a quantity of followers that is higher or at least as important as their number of friends. This observation may rise from three different reasons : (1) the organisation of profiles into campaigns allows to artificially increase the reputation by connecting together (this is known as link farming) [12], (2) many profiles may find into malicious profiles an opportunity to increase their audience and reputation (they usually follow people that follow them), (3) they fully exploit hashtags or references in the aim to increase their reputation. It is important to note that, as the reputation of malicious campaign increases, the number of vulnerable profiles also increases.

Over the three campaign, campaign number II exhibits the lowest score of reputation. A possible reason is its lower number of participants (about 40 profiles).

This makes the artificial increase of reputation less efficient than for campaign I and II that owns a higher number of profiles (respectively 100 and 140). Also campaign II has a lower average number of tweets, retweets and references. These three factors may also be taken into account as a result of this lower reputation score.

We have manually investigated the three identified campaigns and have highlighted a set of techniques that they apply in order to reach their goal. We present below the four main identified techniques and present in table 5 the level of usage of these techniques by the three main campaigns.

(1) The use of hashtags that encourage users to follow profiles of a campaign (denoted EF for encourage followers). Among the most commonly used hashtags for these purpose, we have identified : #500aday, #1000ADAY, #f4f, #Follow4Follow, #followme, #InstantFollow, #MustFollow, #OpenFollow, #OpenFollowPro, #TeamAutoFollow, #TeamFollowBack.

(2) The use of words and hashtags that encourage readers to download, view a media or visit a given webpage (denoted ED for encourage downloads). Among the most common words and hashtags, we can quote : #Download, #freedownload, #freedownloads.

(3) The use of pseudo-automatic tweets (denoted PA for pseudo-automatic). Such tweets have a common root and only a few words, numbers or punctuation marks are modified from tweets to tweets. Usually, the dynamic part is either the final characters, the references or hashtags used in the tweets. Note that, entirely similar tweets cannot be sent on the Twitter platform by the same profile.

(4) The use of external media or tools to feed the tweets (denoted EM for external media). Some profiles send every period of time a tweet that is extracted from a RSS feeds but also some sentences fetched from books. Note that for this purpose, some social media management tools such as TweetAdder, Twitterfeed, HootSuite may be used.

We evaluated each technique for each campaign on a three level scale : ('++') for an extensive usage, ('+') for a usage and (' ') for a non usage.

The results provided in Table 5 indicates that campaign 3, that has the higher reputation score has an extensive use of hashtags for encouraging followers. In fact, we have observed that most of their tweets only contains hashtags that refer to the follower/following processes. A representation of the most common words and hashtags are represented in figure 9.

Campaign I, as already shown in the previous section contributes to the propagation of many malicious

Table 5
Techniques used by the different campaigns

	EF	ED	PA	EM
Campaign 1	+	++	+	
Campaign 2				++
Campaign 3	++

Fig. 9. Cloud of the top words produced by spam campaign III

URLs. This observation is confirmed by the results provided in table 5 and figure 10 that highlight the fact that the tweets published by this campaign strongly encourage users to download or visit a given webpage. Note that, such campaign also uses pseudo-automatic messages such as hashtags that encourage followers. These observations may explain the fact that despite the malicious propagation of URLs, this campaign has a quite high level of reputation.

Fig. 10. Cloud of the top words produced by spam campaign I

Finally, we have observe that the campaign II only remains on the use of external feed. This means that it catches parts of existing books or websites to publish tweets. Such campaign does not encourage followers nor downloading but however provides from time to time URLs that represent a threat for users.

Discussion

There are no silver bullet approaches to security or attack analysis, and therefore any methodology that solves a problem in this space must be considered in

the broader context of the global problem. Given this, the main limitation of the provided approach is that it discovers malicious campaigns on social media websites, but does not necessarily discover all such campaigns. This limitation occurs for two reasons. Firstly, malicious campaigns by their nature disobey normal rules around how platforms are used. This means that discovering patterns of usage are likely to have a high precision by low recall, as different attacks work in different ways. Second, it is often forgotten that the actors behind these attacks are often intelligent, motivated and persistent. Attackers often keep up to date with technological advances and patterns of analysis, and can alter their attack to circumvent new protection measures. This ‘cat and mouse game’ has been an ongoing trait of security research and practice and is likely to continue to be [7].

The results of the previous section do highlight that our method does provide a high recall value for the discovered clusters. These clusters represent approximately 70% of the overall set of malicious profiles, and therefore we can approximate the lower bound of the overall recall value at 0.55, which is the recall of the known clusters as a proportion of the overall population. This recall value indicates a significantly reduced workload for an analyst needing to examine new trends of attacks on social media websites.

Conclusion

In this paper we have proposed a novel approach to detecting relationships between suspicious profiles on social networking sites. Our approach is mainly based on a combination of behavioural and content-based algorithms. As a first step, the suspicious profiles are identified based on a support vector machine algorithm applied to synthesised behavioural features of the profiles. The similarity between the detected suspicious profiles is performed with content-based analyses. This similarity is measured by applying authorship attribution to the messages published by the suspicious profiles. Additionally, we measure a temporal similarity based on the identification of suspicious profiles that send messages simultaneously with the same entities. We build a suspicious graph from the similarity measure and perform a clustering algorithm to identify suspicious campaign. Malicious campaigns detection problem, as discussed in this work, can be used as an entry point to assess the exposure of users to threats and thus provide counter-measures.

Acknowledgments

This work is part of the CyNIC (Cybercrime, Nomadism and IntelligenCe) CPER project supported by the Champagne-Ardenne region and European Regional Development Fund (ERDF). Dr Layton’s work on this project was funded by the Internet Commerce Security Laboratory, a joint research venture between Federation University, Westpac, the Australian Federal Police and IBM.

References

- [1] Web of trust (wot). <https://www.mywot.com/>. Accessed: 2014-01-23.
- [2] S. Abu-Nimeh, T. Chen, and O. Alzubi. Malicious and Spam Posts in Online Social Networks. *Computer*, 44(9):23–28, 2011.
- [3] F. Ahmed and M. Abulaish. An mcl-based approach for spam profile detection in online social networks. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pages 602–608. IEEE, 2012.
- [4] J. Baltazar, J. Costoya, and R. Flores. The real face of koobface: The largest web 2.0 botnet explained. *Trend Micro Research*, 5(9):10, 2009.
- [5] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI conference on weblogs and social media*, volume 2. AAAI Press Menlo Park, CA, 2009.
- [6] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network : When Bots Socialize for Fame and Money . In *the 27th Annual Computer Security Applications Conference*, page 93, New York, New York, USA, 2011. ACM Press.
- [7] R. Broadhurst, P. Grabosky, M. Alazab, B. Bouhours, S. Chon, and C. Da. Crime in cyberspace: Offenders and the role of organized crime groups. *Available at SSRN 2211842*, 2013.
- [8] H. Escalante, M. Montes-y Gómez, and T. Solorio. A weighted profile intersection measure for profile-based authorship attribution. In *10th Mexican international conference on Advances in Artificial Intelligence*, pages 232–243, 2011.
- [9] I. Fette, N. Sadeh, and A. Tomasic. Learning to Detect Phishing Emails. Technical report, School of Computer Science Carnegie Mellon University, 2006.
- [10] H. Gao, J. Hu, T. Huang, J. Wang, and Y. Chen. Security Issues in Online Social Networks. *IEEE Internet Computing*, 15(4):56–63, 2011.
- [11] S. Ghosh, G. Korlam, and N. Ganguly. The effects of restrictions on number of connections in OSNs: a case-study on twitter. In *WOSN 10: Proceedings of the 3rd conference on Online social networks*, June 2010.
- [12] S. Ghosh, B. Viswanath, F. Kooti, N. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummedi. Understanding and combating link farming in the twitter social network. *WWW 12: Proceedings of the 21st international conference on World Wide Web*, 2012.

- [13] V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Pacific Association for Computational Linguistics*, 2003.
- [14] R. Layton, C. Perez, B. Birregah, P. Watters, and M. Lemercier. Indirect information linkage for OSINT through authorship analysis of aliases. *The International Workshop on Data Mining Applications in Industry & Government (DMAApps 2013)*, pages 1–12, Jan. 2013.
- [15] R. Layton, P. Watters, and R. Dazeley. Authorship Attribution for Twitter in 140 Characters or Less. In *2010 Second Cybercrime and Trustworthy Computing Workshop*, pages 1–8. IEEE, July 2010.
- [16] R. Layton, P. Watters, and R. Dazeley. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, Available on CJO, 2011.
- [17] R. Layton, P. Watters, and R. Dazeley. Recentred local profiles for authorship attribution. *Journal of Natural Language Engineering*. Available on CJO, 2011.
- [18] K. Lee, J. Caverlee, K. Y. Kamath, and Z. Cheng. Detecting collective attention spam. In *WebQuality 12: Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, page 48, New York, USA, 2012. ACM Press.
- [19] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69(6 Pt 2), May 2004.
- [20] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69, Feb 2004.
- [21] J. Novak, P. Raghavan, and A. Tomkins. Anti-aliasing on the web. *Proceedings of the 13th conference on World Wide Web - WWW 04*, page 30, 2004.
- [22] C. Perez, M. Lemercier, B. Birregah, and A. Corpel. Spot 1.0: Scoring suspicious profiles on twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 377–381, July 2011.
- [23] K. Sparck Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 1972.
- [24] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *the 26th Annual Computer Security Applications Conference*, page 1, New York, USA, 2010. ACM Press.
- [25] C. J. Van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 1974.
- [26] A. Wang. Don't follow me: Spam detection in twitter. In *Int l Conference on Security and Cryptography (SECRYPT)*, 2010.
- [27] A. H. Wang. Detecting spam bots in online social networking sites: a machine learning approach. In *DBSec 10: Proceedings of the 24th annual IFIP WG 11.3 working conference on Data and applications security and privacy*, pages 335–342, Berlin, Heidelberg, June 2010. Springer-Verlag.
- [28] W. Xu, F. Zhang, and S. Zhu. Toward worm detection in online social networks. In *ACSAC 10: Proceedings of the 26th Annual Computer Security Applications Conference*. ACM Request Permissions, Dec. 2010.
- [29] X. Zhang, S. Zhu, and W. Liang. Detecting spam and promoting campaigns in the twitter social network. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1194–1199. IEEE, 2012.
- [30] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57:378–393, 2005.