2015 Second European Network Intelligence Conference

Evaluating Algorithms for Detection of Compromised Social Media User Accounts

David Trång
Uppsala University
Uppsala, Sweden
Email: david.a.trang@gmail.com

Fredrik Johansson
Swedish Defence Research Agency (FOI)
Stockholm, Sweden
Email: fredrik.johansson@foi.se

Magnus Rosell
Swedish Defence Research Agency (FOI)
Stockholm, Sweden
Email: magnus.rosell@foi.se

***Abstract*—Hijacking of user accounts on Online Social Networks (OSNs) such as Twitter is increasingly used for purposes such as large-scale spam campaigns and cyber crime-related scams and phishing attacks, but also for purposes such as spreading terrorism propaganda and as part of information operations in political and military conflicts. We present a novel method for evaluation of the performance of algorithms for detection of compromised social media accounts. In short, we create two artificially ”hijacked accounts” by randomly selecting two genuine user accounts and after a number of successive posts switch the remaining posts between them. This allows us to create large amounts of training and/or test data, something which is not easy to find for this problem. Further, the developed method is utilized to evaluate the performance of a modified version of the existing COMPA system.**

I. Introduction

Automated bot accounts (also known as Sybil accounts) created for the purpose of sending large amounts of spam have long been a problem for various online social networks (OSNs) such as Twitter. In more recent years we have seen how automated Sybil accounts have started to be used for various kinds of information operations in political and military conflicts, including use of Twitter for spreading terrorism-related propaganda and flooding of Twitter hashtags about political protests in e.g. Syria and Russia [1], [2]. Quite a lot of research has been devoted to detecting, analyzing, and characterizing spam and the accounts created and used for spreading such posts. Features that have been suggested in existing research literature include user account properties such as number of followers [3], content-based features measuring similarity among posts [4], regularity in posting behavior [5], [6], and network-based features such as degree centrality [7], clustering coefficients [8], and reciprocity [9].

However, arguably partially due to the relatively easiness with how automated classifiers can detect spam and Sybil accounts [10], cyber criminals, terrorist groups, and others who want to reach out to large audiences on social media with their spam, disinformation, phishing attacks etc. are nowadays also making use of more sophisticated methods. Instead of just simply creating Sybil accounts, it seems to become more popular [6] to hijack existing users' social media accounts and in this way leveraging the trust relationships which the legitimate owners of the accounts have established. The accounts are for example often compromised through phishing scams to steal login credentials and exploitation of cross-site scripting vulnerabilities [10]. One of the most famous hijackings of social media accounts is probably the hijacking in 2013 of a Twitter account belonging to Associated Press, tweeting out the message: ”Breaking: Two Explosions in the White House and Barack Obama is injured.” As a consequence of this message, there was a large instability in the financial market some time afterwards [11]. Another example of the use of compromised Twitter accounts to spread disinformation and propaganda is given in [12]. In the article it is described how Syrian hackers in support of the Assad regime have compromised accounts belonging to CBS News and Human Rights Watch. Other examples are cyber criminals' use of hijacked accounts to spread malware, make phishing attacks, sending botnet instructions, and to do various sorts of online scams. Also terrorist organizations are making use of other peoples' and organizations' social media accounts to gain credibility and to reach out to larger audiences, including ISIS' app ”The dawn of glad tidings” which on download gets permission to autonomosly publish tweets from the user's account, making it possible to perform large-scale social media campaigns. Finally, there are of course also many examples of how former girlfriends and boyfriends have compromised a person's social media accounts in order to send out messages that cause damage to the owner of the account. All in all, this means that hijacking of social media accounts is something which can be carried out by a large spectrum of attackers for a large number of purposes and is a relatively large problem. In fact, the threat of hijacking is according to [13] one of the largest challenges for OSNs.

Despite the popularity of hijacking social media accounts for propagating various types of illicit or unwanted information, there are surprisingly few research articles addressing this problem. Some OSNs such as Twitter have introduced more advanced and secure login procedures like the optional use of two-step verification in order to decrease the risk of hijacking of influential users' accounts. However, this causes extra burden on the user and is in practice often not used. In practice, the problem of compromised user accounts is still a very relevant problem. The lack of research on detection of compromised accounts is therefore something which we in this paper take a first step to address.

One of very few attempts to detect hijacked accounts is the

978-1-4673-7592-4/15 $31.00 © 2015 IEEE
DOI 10.1109/ENIC.2015.19

COMPA system described by Egele et al. in [10]. Basically, their approach first checks for a set of similar messages and group such messages into clusters. For clusters in which a significant subset of the messages violate the behavioral profile of their corrresponding user accounts, the COMPA system flag these accounts as compromised. COMPA is shown to accurately detect compromised accounts with a low false positive rate, but due to the clustering part of the algorithm, only large-scale campaigns in which several compromised accounts are involved can be detected. In this paper we attempt to evaluate how well COMPA works for identification of single hijackings if the clustering part of the algorithm is removed. In order to make such an evaluation, we suggest a novel evaluation methodology in which we artificially create realistic well-controlled datasets in which ”hijackings” can occur at the points where we would like them to be. This overcomes the problem of a lack of real datasets containing data instances known to be compromised as well as uncompromised. Using our proposed evaluation methodology, we show that COMPA for some parameter settings can obtain a high recall, but that it in general has a too high false alarm rate to be used for most real-world applications. For this reason, we are suggesting modifications of COMPA that allow for higher precision, thereby being more suitable for real-world use.

The rest of this paper is structured as follows. In Section II, we present some related work. Next, we are in Section III summarizing the most important parts of the COMPA system introduced in [10] and explain how it can be adapted to allow for detection of single hijackings as well as more large-scale attacks. In Section IV we are presenting our novel method for evaluating algorithms for detection of compromised social media accounts. This method is in Section V used to evaluate the modifiication of COMPA for real-time detection of compromised accounts. A discussion of the obtained results and its implications is presented in Section VI, together with discussions about how the detection algorithms can be improved. Finally, we present some conclusions and ideas for future work in Section VII.

II. Related Work

As argued in Section I, quite a lot of research has been devoted to the problem of classifying OSN user accounts as being used for sending spam or not. Sybil accounts, i.e., accounts that have been created for the purpose of sending out spam or other types of unwanted messages are obviously of interest to detect, not least in relation to military and political information operations as discussed in more detail in [14]. Some examples of algorithms for detection of Sybil accounts are the semi-supervised learning framework SybilBelief [15] and the supervised learning approach presented and tested in [16]. While Sybil accounts easily can be created for reaching out to large audiences, their problem is that they quickly can be discovered and suspended, and that it is hard for them to gain enough credibility to really influence the receivers of the messages. For this reason it can be highly relevant to attempt to hijack existing user accounts when sending propaganda, attempting to do phishing attacks, etc.

The phenomenon of hijacked social media accounts has for example been studied in [13]. In their study, Thomas et al. focus on large-scale social contaigon attacks where hijacked accounts are utilized. They also investigate how criminals monetize on such accounts. Their approach relies on first clustering accounts based on content similarity and the presence of duplicate URLs. Next, labels are assigned to these clusters by classifiers that have been trained on a dataset consisting of tweets and accounts labeled as benign, compromised, and fraudulent. The labels of the training data are based on whether Twitter has deleted or disabled the tweets or accounts. This method is then used to estimate to which extent (large-scale) hijacking is a problem and how many users that are affected by it. In their study, 14 million users who have been victim of hijackings are identified. It is shown that 21% of the victims have not returned to Twitter after having their account compromised and that over 50% of the victims have lost online ”friends” in response to the spam sent from the victims' account. This is an impressive and well-performed study, but it is limited to large-scale attacks due to the clustering step of their system, thereby missing small-scale, more targeted, information operations. Moreover, their classifications are to a large degree dependent on Twitter's accuracy in identifying and suspending tweets from compromised accounts.

An unsupervised approach to detect anomalous user behavior is presented in [17]. They make use of principal component analysis (PCA) to model the behavior of ”normal” users using a small set of latent temporal, spatial, and spatio-temporal features and classify significant deviations from such normal behavior as anomalous. In theory, this makes it possible to detect various types of abnormal behavior, including both Sybil accounts and hijacked accounts. The main difference between their approach and the one suggested in this paper is that we create a normal model per user account, rather than a single normal model common to all users as in [17].

In [18], a Support Vector Machine (SVM) classifier is used to classify how users have reacted after having their accounts hijacked. The dataset used for training this classifier is based on the assumption that all accounts sending tweets containing the strings ”hacked” or ”compromised” in combination with the string ”account” have been compromised. Although such an approach is a first step towards a good dataset on which classifiers can be trained, it is in our opinion requiring more and better manual analysis and annotation since many of the tweets are likely to be false positives (since they e.g. can be messages mentioning other people having their accounts hijacked). Moreover, it is not clear from such a dataset exactly which tweets that have been sent during the time period when the user account have been compromised, making it hard to label the data instances correctly.

The COMPA system [10] is instead of relying on supervised learning classifiers based on an anomaly detection approach. COMPA builds a behavioral profile for each user account of interest, based on the user's previous activities in terms of features such as posting frequency, used language in the messages, and application usage. When new posts appear, they are compared to the user's previously built model, and if the new message is considered to be anomalous the account is flagged as potenitally hijacked. However, in order to reduce the number of false positives, accounts are not subject to calculation of anomaly scores until they have been clustered together with several other accounts sending similar content. This improves the precision of the system significantly, but also means that the original COMPA method cannot be applied as is for detection of single hijackings. As illustrated by the example of the hijacking of the Associated Press account, the hijacking of a single account can have large impact for information operation purposes. For this reason, we are in this paper using a slightly modified version of COMPA in which the clustering step is ignored. COMPA is described in more detail in Section III.

III. Detection of compromised accounts using COMPA

As explained in previous sections, COMPA was originally designed for detecting compromised user accounts in OSNs based on first discovering clusters of similar messages and then checking whether a significant subset of these messages are deviating strongly from the behavioral profiles of their senders [10]. Since the original version of COMPA is not able to detect cases in which an attacker posts just a few messages from a single compromised account, we are in this section presenting a straightforward modification of COMPA in which the clustering step of the algorithm is removed. We are here only giving sufficient details to explain the overall method and to allow for reimplementation of the modified version. For more in-depth explanations of the used features and the motivation to why the method works as it does, we refer the interested reader to [10].

COMPA is based on the concept of behavioral profiles, obtained by creating separate feature models in a separate training phase. Based on the messages received from the accounts of interest in the training phase, a behavioral profile is created for each user by extracting a set of feature values from each message, and then creating a statistical model for each feature. Different features can be extracted for different OSNs, but in this work the following features have been extracted: time (hour of day), message source, message language, message topics (hashtags), message links, and direct user interactions (mentions). Some of these are mandatory attributes present in each message, while others are optional (see Table I).

The mandatory features are those for which there is exactly one attribute value for each message, while optional features are those for which there can be zero to many attribute values in a single message (i.e., links, hashtags, and mentions in the case of Twitter). All features are stored as a list of tuples. Each tuple consists of an attribute value as a key, and the number of occurrences of that particular attribute value as its value, i.e.:

$$< \text{attribute value; nr of occurrences} >$$

.

TABLE I
Summary of used features and their corresponding weight α.

Feature	Mandatory?	Weight (α)
Time (hour of day)	Yes	0.88
Message source	Yes	3.3
Message language	Yes	0.58
Hashtags	No	0.39
Links	No	0.96
Mentions	No	1.4

To illustrate what such tuples could look like, a user who has written five messages in Norwegian and two in Spanish that end up in the observations used when creating the behavior profiles will for the message language features have two tuples:

$$< Norwegian; 5 > \text{ and } < Spanish; 2 > .$$

Since the mandatory features can have arbitrarily many attribute values, including zero, they always have the special key $null$ included, which encodes the number of messages in the training data from the current user which lack this particular feature. For the time of day model we smooth out the attribute values by also taking into account the values for the two adjacent hours. The final values for the keys h_i (where $i \in \{0, \ldots, 23\}$) are in our implementation of COMPA calculated as:

$$h_i = \frac{s_{i-1} + 2s_i + s_{i+1}}{4},$$

where $i \in \{0, \ldots, 23\}$ and s_i is the original number of occurrences for the i:th hour. This is a slight modification of how these values are calculated in the original COMPA implementation.

Once the training phase has been completed, COMPA calculates anomaly scores for new messages obtained from user accounts for which it has existing behavior profiles. An anomaly score Φ is calculated for each new message. This overall anomaly score is calculated as a weighted sum of the individual anomaly scores computed for each feature, as illustrated in Equation 1:

$$\Phi = \sum_i \alpha_i \phi_i. \quad (1)$$

Here, α_i is the weight of the current feature (shown in Table I) and ϕ_i is the anomaly score for the current feature. The total anomaly score is a value in the interval $[0, \sum_i \alpha_i]$, since each ϕ_i is a value between 0 and 1, where 0 is fully normal and 1 is fully anomalous.

The individual feature anomaly scores are computed differently depending on whether the feature is mandatory or not. For mandatory features, new attribute values which are not present in the user's behavior profile are considered fully anomalous and are assigned an anomaly score of 1. If the observed attribute value exists, we compare the number of occurrences for the current attribute value to $\bar{M}$, where $\bar{M}$ is defined as the total number of occurrences summed over all attribute values for the current feature, divided by the number of different attribute values for the current feature. If the number of occurrences for the present attribute value is greater than or equal to $\bar{M}$, an anomaly score of 0 is returned. Otherwise the anomaly score for the current feature is calculated as the number of occurrences of the current attribute value divided by the total number of occurrences for all attribute values for the user's current feature.

For optional features, the anomaly score is instead calculated as follows. If the attribute value is present in the user's behavior profile created from the training phase, an anomaly score of 0 is returned. Otherwise the anomaly score is calculated by dividing the observed number of $null$ values for the current feature in the user's behavior profile with the number of different observed attribute values for the current feature. In case there are no $null$ values in the user's behavior profile for the current feature, an anomaly score of 0 is returned.

IV. Evaluation Methodology

As there is a lack of research on algorithms and systems for detecting compromised accounts, there is also, quite naturally, a lack of research on how to evaluate such algorithms and systems. In the original evaluation of the COMPA system [10], Egele et al. make an impressive effort to assess the system's false positive and false negative rates on real data by using a combination of manual evaluation, automatic comparisons against what Twitter a few months later had removed or suspended, and comparisons of posted URLs against known blacklists. Although extensive effort has been put into this evaluation, it is relying on that OSNs such as Twitter and Facebook are successful in detecting and deleting or suspending messages sent from hijacked accounts. Moreover, the manual assessment may work for evaluating groups of accounts participating in large-scale campaigns (which is the scenario to which COMPA is applied and evaluated in [10]), but it does not scale to the situation of interest here since a single message from a single hijacked account can be highly interesting in a military and political information operation context. For this reason, we are requiring a more fine-grained evaluation methodology in which we:

1) are not dependent on the performance of external services, and
2) have full control over the ground truth so that we know exactly which messages that have been the result of a hijacked account or not, allowing for more precise and reliable evaluation.

Before presenting our suggested evaluation methodology in more detail, we will first describe the expected setting on which to apply the evaluation method. In general, we would like to have a set of interesting accounts to follow in real-time, and for each new post from any of these accounts we want to judge the new post to be from a compromised or uncompromised account. Now, this can be accomplished using supervised classifiers as in [13], as well as anomaly detection algorithms as in [10]. The set of interesting accounts can be very different for different applications. It can for example consist of a single company Twitter account we would like to protect against hijackings, all major newspapers' social media accounts, or all active accounts on Twitter.

Irrespectively of the application and the method used to come up with a judgement, we can treat all posts judged to be from a compromised account when it in reality has not been hijacked as false positives. Similarly, all posts being sent from compromised accounts without being judged to be from such an account can be counted as false negatives. A perfect system or algorithm will judge all posts from an uncompromised account to be normal/uncompromised and switch over to abnormal/compromised judgements as soon as the account has been hijacked. Now, assume that all accounts are hijacked simultaneously (which is unrealistic in a real-world scenario but which can be assured in simulated data). In such a case, the perfect behavior of an algorithm for detection of compromised accounts would be to have zero detections of compromised accounts up to the point where the hijackings take place, and then judge all the following posts to be the result of hijackings.

Based on the above intuition, we are now ready to describe our proposed evaluation methodology. Up to a randomly selected point in time, or a point in time of our choice, we want to have genuine messages posted by the user accounts. Each time such a genuine message is classified as originating from a compromised account, this will increase the count of false positives. After this point in time, we would like to replace genuine messages with posts that could have been sent from a hijacked account. This can be accomplished in several ways. One way could be to have a large collection of previous messages from compromised account from which content and its associated metadata can be randomly selected. The main problem with such an approach is that it, as argued previously, is hard to collect social media posts which we can be sure are originating from hijacked accounts without demanding a great deal of manual work. Moreover, a too limited set of posts from hijacked accounts would not necessarily give a good estimate of how well the detection algorithm would work in practice. Another possible alternative would be to randomly generate the social media posts and their associated metadata (such as language and source), but this would probably not yield realistic posts. What we instead have aimed for is to:

1) Randomly select a pair $< U_1, U_2 >$ of social media user accounts.
2) Select N consecutive messages from each user.
3) At an arbitrarily selected m, where $1 < m \leq N$, swap all messages $M_m, \ldots, M_N$ between the selected pair of users U_1 and U_2.

This process is illustrated in Figure 1. Using this approach, we can create large datasets containing synthetically compromised user accounts in which we have full control over when the hijackings have taken place. These synthetically generated hijackings are probably more challenging to detect than many real-world hijackings since the posts will not contain keywords and URLs typical for many large-scale spam and scam attacks in which hijacked accounts are used. However, this will arguably give better indications of the precision with which more well-targeted small-scale information operations can be detected.

All posts originating from an account which not yet have been compromised and which the evaluated algorithms classify as uncompromised will hence be counted as true negatives. If they instead are classified as compromised, they will be counted as false positives. In the same way, posts from a compromised account which are classified as uncompromised will be treated as false negatives, while those classified as compromised will count as true positives. In this way it becomes possible to apply standard metrics such as precision (see Equation 2), recall (see Equation 3), and F_1-score (see Equation 4), where the F_1-score is the harmonic mean of precision and recall. In these equations, TP refers to the number of true positives, FP to the number of false positives, and FN to the number of false negatives.

$$Precision = TP/(TP + FP) \tag{2}$$

$$Recall = TP/(TP + FN) \tag{3}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

These standard metrics can then be used to compare various algorithms on the same datasets. It also becomes possible to compare algorithms using visual inspection as shown in Section V.

V. Experiments

In this section we present an experiment in which we have evaluated the modified COMPA algorithm on synthetic data using the evaluation methodology described in Section IV. The data used for building the behavior profiles and for generating the synthetic evaluation dataset is based on tweets relating to the Russia-Ukraine conflict as explained below in Section V-A. COMPA's success rate in detecting the synthetically generated compromised accounts is then summarized in Section V-B.

A. Collection of a Twitter dataset

At the Swedish Defence Research Agency (FOI) we have collected a large number of tweets related to the Russia-Ukraine conflict, even before the Russian annexation of Crimea took place. From this very large collection of tweets we have together with domain experts selected a time period of 45 days, ranging from 2014-04-01 to 2014-05-15, to be of extra interest from an information operation's perspective. During this time period more than 4,000,000 tweets matching our selected keywords from over 700,000 different Twitter accounts were collected.

B. Evaluation of COMPA

From the dataset described in Section V-A, we have randomly selected 1000 user accounts which have posted at least 100 messages. The first 60 messages from each account have been used in the training phase for creating COMPA's behavior profiles, while the next 40 messages have been set aside for the evaluation phase. The evaluation methdology described in Section IV have been used to create artificial hijackings of all the 1000 user accounts. For all randomly selected pairs of user accounts we have switched the remaining messages after tweet number 20 in the evaluation phase, so that tweet number 21 to tweet number 40 in the evaluation phase are from the ”compromised” account.

The results of this experiment are summarized in Figure 2. The figure is showing the cumulative sum of the users that has been flagged as compromised at each point in time (i.e., each sent message) for various threshold settings. The threshold setting specify how large the total anomaly score is allowed to be before COMPA flag the user account as compromised. An optimal algorithm would here not flag any user between tweet 1 and tweet 20 in the evaluation phase, and then flag all users as compromised from tweet 21 since that is the first tweet originating from a compromised account.

As can be seen in Figure 2, very different results can be obtained depending on the choice of threshold settings. If setting this to a low value, nearly all compromised accounts are detected soon after the hijacking has taken place. However, this comes with the price of a very high false positive rate. If instead increasing the threshold value, a much lower false positive rate can be obtained, but then a much lower recall is reached. Clearly, there is a trade-off between precision and recall for this modified version of COMPA, either we have a too low precision to be used for real-world application, or we obtain a low recall. Possible further modifications of COMPA to get around this problem is discussed in more detail in Section VI.

VI. Discussion

The results obtained in Section V are not unexpected. COMPA was originally intended for finding groups of hijackings taking place in parallell and although the modification of COMPA to allow for detection of single hijackings is straightforward, it is quite obvious that not as good results will be obtained for a single account. There are just too many things which can cause false alarms in the single account setting if the anomaly threshold is not set high, including changes in when the user is active on Twitter or which topic one is tweeting about. Although this probably can be countered partially by increasing the number of messages from which to create the feature models in the training phase, it is not unproblematic since it also reduces the usability of the method for user accounts which are sending messages rarely. For this reason we also suggest to modify the COMPA algorithm by taking into account individual differences in anomaly scores. We propose to refine the algorithm by calculating anomaly scores for each message used for creating

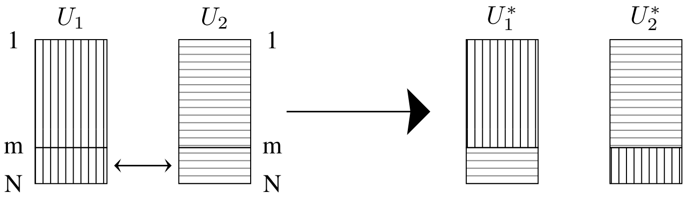

Fig. 1. Illustration of how two artificial hijackings are created by swapping posts between two randomly selected users U_1 and U_2. N consecutive posts from each are selected. The artificially hijacked account U_1^* is created by taking the $m-1$ first posts from U_1 and the last $N-m+1$ posts from U_2. Similarly for U_2^*.

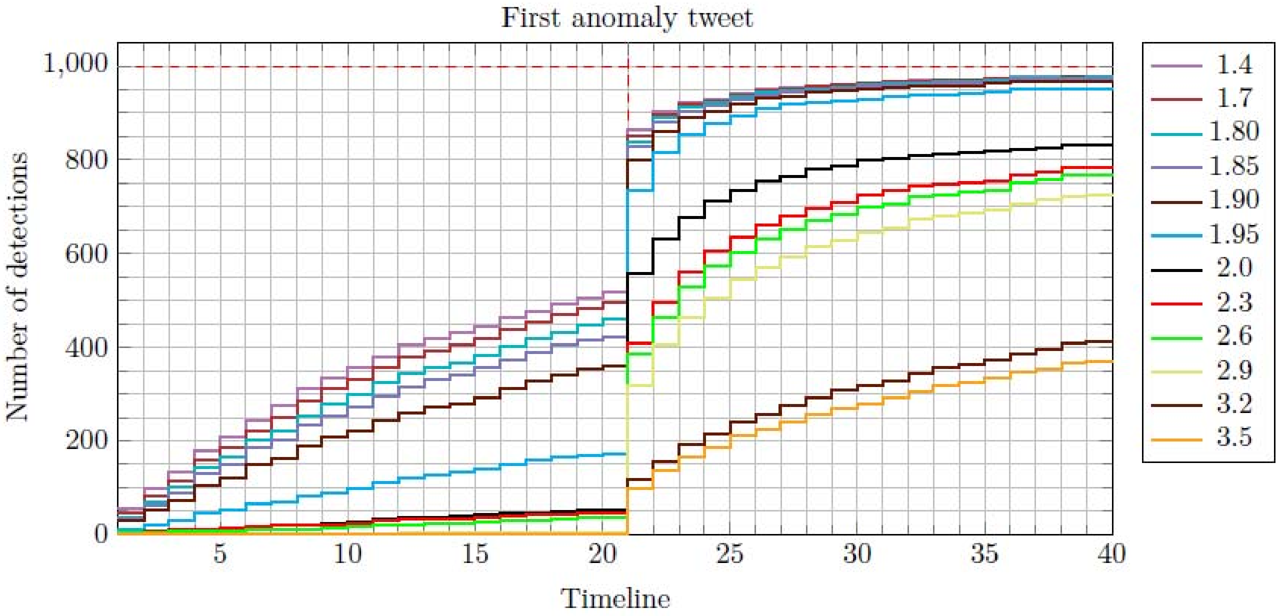

Fig. 2. Cumulative sum of users flagged as compromised by COMPA at various time steps for different anomaly threshold settings. All accounts are compromised from time step 21 to 40 in the evaluation phase.

the behavior profiles in the training phase (except for the first message since no normal model exists at this point). Then these anomaly scores obtained on messages that are known to be from uncompromised accounts can be used to adjust the anomaly scores calculated for new observations from the same user account. One way to do this is to calculate the mean and standard deviation from the user account's anomaly scores obtained in the training phase, and when used on new unlabeled messages flag all new observations which receive higher anomaly scores than the corresponding account's mean plus x standard deviations as anomalous. If the anomaly score is lower than this threshold, the new observation is flagged as normal. In this way the number of false positives is likely to be decreased. A complementary solution to this is to add more features to the model. Two such sets of features we would like to try to add to the COMPA method in the future is to make use of stylometric features and more advanced time-based features such as those utilized in [19] and [20]. Such features have in preliminary experiments been shown to discriminate among different authors surprisingly well, also on Twitter.

VII. Conclusions

Despite the widespread use of hijacking of social media accounts for various purposes such as large-scale spam campaigns, cyber crime scams and phishing attacks, as well as more targeted information operations, algorithms for detection of compromised accounts is an underresearched area, especially compared to research on bot and spam detection. An exception to this is the COMPA system, proposed in [10]. Although this system has been shown to obtain high precision for detection of large-scale spam campaigns involving several compromised accounts, it is not applicable to detecting single hijacked accounts in its original version.

In this paper we have presented a novel methodology for evaluation of algorithms for detecting compromised accounts.

It allows for easy creation of large training and test sets in which we have full knowledge of when the hijacking occurs. The artificially created hijacked accounts are probably sometimes even harder to detect than real ones as they are created from "normal" accounts.

As an example we have used our evaluation methodology to evaluate a modified version of COMPA which allows for detection of single compromised user accounts. The results show that the modified version of COMPA can yield either good recall or precision depending on the parameter settings (the value used for the anomaly threshold), but that there is a very clear trade-off between the two. The behavior is also quite unstable, i.e., the detection behavior is strongly dependent on the used threshold. In order to overcome these problems, we have suggested to implement a modified version of COMPA in which additional features are used.

A. Future Work

As future work we intend to implement an improved version of COMPA into our framework for analyzing information operations on Twitter. In this way we hope to be able to get a better understanding for how compromised accounts are used for information operation purposes in military and political conflicts. Detection of hijacked accounts is just one, but an important, piece of the complex puzzle when trying to understand how information operations are carried out in various social media.

By implementing a new version of COMPA in which the mean and standard deviation of the anomaly scores obtained on the training data are used to adjust the threshold for when a user account should be flagged as anomalous, we hope to be able to improve the precision of the algorithm. We also aim at adding new features to our modified version of COMPA, including stylometric and time-based features. Finally, we would like to evaluate alternative methods using the same evaluation methodology as developed in this work, including the supervised learning algorithm developed in [13].

In future work we would also like to develop the evaluation method further. There are several ways to vary the generation of artificially created compromised accounts. For instance: instead of pairing two ordinary accounts we could create a "hijacking" by adding posts from a known bot[1] to the end of a sequence of posts from an ordinary account. It would probably be much easier to detect the hijacking in this particular variation than the more interesting artificial hijacking in the original version. However, as it is rather easy to create the data we could potentially study the methods from different perspectives.

Acknowledgements

This work was supported by the R&D programme of the Swedish Armed Forces. We would like to thank the other members of the SIA project for fruitful discussions about various types of information operations in social media in general and Twitter in particular. We also would like to thank Christopher Teljstedt for implementing parts of the code used for extracting the features used in the COMPA implementation. Thanks also to Elliot, 2.5 years, for letting his father work with the paper rather than playing with dinosaurs. Lastly, we express our gratitude to personnel in the Swedish Armed Forces for your support and good discussions.

[1] The bots could for instance be found by using an automatic method for bot detection.

REFERENCES

[1] J. C. York. (2011, April) Syria's twitter spambots. The Guardian. [Online]. Available: http://www.theguardian.com/commentisfree/2011/apr/21/syria-twitter-spambots-pro-revolution

[2] U. Franke and C. V. Pallin, "Russian politics and the internet in 2012," FOI, Tech. Rep. FOI-R–3590–SE, 2013.

[3] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *Dependable and Secure Computing, IEEE Transactions on*, vol. 9, no. 6, pp. 811–824, 2012.

[4] A. A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, and C. Yang, "Cats: Characterizing automation of twitter spammers," in *Communication Systems and Networks (COMSNETS), 2013 Fifth International Conference on*. IEEE, 2013, pp. 1–10.

[5] C. M. Zhang and V. Paxson, "Detecting and analyzing automated activity on twitter," in *Passive and Active Measurement*. Springer, 2011, pp. 102–111.

[6] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," in *Proceedings of the 17th ACM conference on Computer and communications security*. ACM, 2010, pp. 27–37.

[7] D. DeBarr and H. Wechsler, "Using social network analysis for spam detection," in *Proceedings of the Third International Conference on Social Computing, Behavioral Modeling, and Prediction*, ser. SBP'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 62–69.

[8] P. Boykin and V. Roychowdhury, "Leveraging social networks to fight spam," *Computer*, vol. 38, no. 4, pp. 61–68, April 2005.

[9] M. Fire, G. Katz, and Y. Elovici, "Strangers intrusion detection - detecting spammers and fake profiles in social networks based on topology anomalies," *Human*, 2012.

[10] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting compromised accounts on social networks," in *Proceedings of the Network and Distributed System Security Symposium*, 2013.

[11] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on twitter," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, ser. SNAKDD '13, 2013.

[12] M. Fisher. Syria's pro-Assad hackers are hi-jacking high-profile twitter feeds. The Washington Post. [Online]. Available: http://www.washingtonpost.com/blogs/worldviews/wp/2013/04/22/syrias-pro-assad-hackers-are-hijacking-high-profile-twitter-feeds/

[13] K. Thomas, F. Li, C. Grier, and V. Paxson, "Consequences of connectivity: Characterizing account hijacking on twitter," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. New York, NY, USA: ACM, 2014.

[14] C. Teljstedt, M. Rosell, and F. Johansson, "A semi-automatic approach for labeling large amounts of automated and non-automated social media user accounts," Accepted for publication and oral presentation at the Second European Network Intelligence Conference (ENIC2015).

[15] N. Z. Gong, M. Frank, and P. Mittal, "Sybilbelief: A semi-supervised learning approach for structure-based sybil detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, 2014.

[16] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network sybils in the wild," in *Proceedings of the 2011 ACM SIGCOMM Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 259–268.

[17] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Towards detecting anomalous user behavior in online social networks," in *Proceedings of the 23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, 2014, pp. 223–238.

[18] E. Zangerle and G. Specht, ""sorry, i was hacked": A classification of compromised twitter accounts," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2014, pp. 587–593.

[19] F. Johansson, L. Kaati, and A. Shrestha, "Detecting multiple aliases in social media," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. ACM, 2013, pp. 1004–1011.

[20] ——, "Time profiles for identifying users in online environments," in *Proc. 1st Joint Intelligence and Security Informatics Conference, IEEE Computer Society*, 2014, pp. 83–90.