Social Networking, **2013, 2, 42-45**
http://dx.doi.org/10.4236/sn.2013.21005 Published Online January 2013 (http://www.scirp.org/journal/sn)

A Heuristic Reputation Based System to Detect Spam Activities in a Social Networking Platform, HRSSSNP

Manoj Rameshchandra Thakur[1*], Sugata Sanyal[2]
[1]Computer Science Department, Veermata Jijabai Technological Institute, Mumbai, India
[2]Corporate Technology Office, Tata Consultancy Services, Mumbai, India
Email: *manoj.thakur66@gmail.com, sugata.sanyal@tcs.com

Received November 12, 2012; revised December 27, 2012; accepted January 14, 2013

ABSTRACT

The introduction of the social networking platform has drastically affected the way individuals interact. Even though most of the effects have been positive, there exist some serious threats associated with the interactions on a social networking website. A considerable proportion of the crimes that occur are initiated through a social networking platform [1]. Almost 33% of the crimes on the internet are initiated through a social networking website [1]. Moreover activities like spam messages create unnecessary traffic and might affect the user base of a social networking platform. As a result preventing interactions with malicious intent and spam activities becomes crucial. This work attempts to detect the same in a social networking platform by considering a social network as a weighted graph wherein each node, which represents an individual in the social network, stores activities of other nodes with respect to itself in an optimized format which is referred to as localized data-set. The weights associated with the edges in the graph represent the trust relationship between profiles. The weights of the edges along with the localized data-set are used to infer whether nodes in the social network are compromised and are performing spam or malicious activities.

Keywords: Spam; Social Graph; Collaborative Filtering; Weighted Graph; Localized Data-Set; Trust Level

1. Introduction

A considerable amount of work has been done in the area of spam detection and trust based recommendation systems for social networking platforms. A brief overview of these is as follows: [2] suggests a dynamic personalized recommendation system that is based on the trust between agents. It uses the concept of feedback centrality and overcomes some of the limitations of earlier recommendation systems that use other trust metrics. In the model suggested in [3] an agent tries to filter interactions based on the information that it gains from its own social network. The model suggested in [3] identifies the impact of factors like preference heterogeneity of agents, network density among agents, and knowledge sparseness which are crucial factors for the performance of the model. The technique suggested in paper is however different from the earlier two works, in that it makes use of a weighted social graph [4] to view the relationship between profiles in a social networking platform . The technique suggested in [5,6] suggests a reputation based intrusion detection system to detect malicious and compromised nodes in a mobile ad-hoc network. Even though this work is not directly related to social networking platforms, the approach suggested in [5,6] is relevant to the problem of filtering malicious and spam conversations among agents in a social networking platform.

A number of works have been suggested in the recent past that address the issue of email spam activities and suggest techniques to combat them. Email spam, even thought not directly related to spam activities, is relevant to our work primarily because spam emails are analogous to spam and malicious activities in a social network. [7] presents a new email ranking and classification scheme that makes use the social email interactions to infer the spam/non-spam status of the sender of any given email. The rank assigned to an email address based on its interactions represents the reputation of the email address. The work suggests two level of ranking: global rank which is recipient email address agnostic and personalized rank that varies based on the email address which is receiving the email from a particular sender. [8] presents a email scoring mechanism which is again based on social interactions and assigns reputation ratings to email addresses. [9] attempts to discriminate spam activities from non-spam activities by applying various machine learning techniques. The classification is based on six distinct features that the work identifies. [10] attempts to assign the legitimacy to the sender of an email based on the features extracted from various email interaction.

[*]Corresponding author.

This work attempts to adopt a learning approach for detecting spam activities.

Intrusion detection, which involves trying to identify malicious and compromised nodes in a given network, is similar to the process of identifying compromised agents is a given social graph representing the way in which individuals are connected. [11] discusses some of the security issues associated with distributed computing infrastructures most of which apply to a social network as well. Approaches like the ones suggested in [12-16] are instrumental in not only addressing the problem of intrusion by malicious nodes in a network but are also indirectly helpful in devising similar approaches for spam and malicious agent detection in a social network.

2. The Social Graph

A social graph may be defined as a graph that represents the way individual are related to each other on the internet [17]. Even though it represents the relationship between individuals it doesn't manifest in any way the trust level among individuals. Two individuals might be related but might not have a high trust level. The ability to represent the trust level in a social graph can impart a powerful tool to detect and prevent unwanted interaction. For example, if A is not related to B but wants to interact with B. B will try to obtain relevant information from an individual C with whom B has a high trust level and based on the inputs B will decide whether to allow A to interact or not. The suggested approach attempts to derive this trust level among profiles based on previous interactions and the relationship type and represent the trust level as weights corresponding to the edges that represent the relationship between profiles.

3. Collaborative Filtering and Unwanted/Malicious Activity Detection

In a given social networking platform, the following holds, "If Profile A is victimized by a malicious interaction by C then the chances of profile B being victimized by profile C is high". It is this relationship that has inspired the use of collaborative filtering [18] for the suggested approach. The suggested approach adds an additional constraint that in order for profile B to detect whether profile C is trying to initiate a malicious interaction it will only try to take recommendation from profiles which it trusts *i.e.* profiles with high trust level.

4. Weighted Social Graph

The suggested approach views the social networking plat-form as a weighted social graph [4]. Each node represents a profile (an individual), an edge represents a relationship between profiles and the weight corresponding to the edge represents the trust level among profiles. Each profile has a localized dataset associated with it that holds a table with the following format:

Profile Id (*say X*) → *< incoming activity with X >* : *< outgoing activity with X >*

We refer to the localized dataset as *LD*, such that *LD* (*X*) represents the localized dataset entry of profile *Y* for profile X, such that the incoming activity with *X* is represented as *I*(*X*) and the outgoing activity with *X* as *O*(*X*). Thus,

$$LD(X) = \{I(X), O(X)\} \tag{1}$$

The overall view of the weighted social graph can be depicted as shown in **Figure 1**.

5. Localized Data-Set and EDGE Weight Calculation

Corresponding to each profile, the localized data-set refers to a table with the following entries:

- The first column represents profile Ids to which the given profile (*Y*) is connected *i.e.* has a relationship (*X*).
- The second column has entries in the following format (*LD*(*X*)):

$$LD(X) \to \{I(X), O(X)\} \tag{2}$$

Incoming activity, *I*(*X*): it represents the activities which were initiated by *X* in which the considered profile was the destination. These interactions include activities like message sent from *X* to the considered profile, friend request from *X* to the considered profile, comments on a photo of the considered profile etc. It must be noted that the type of interaction as mentioned earlier may vary based on the social networking website considered.

Outgoing activity, *O*(*X*): it represents the activities which were initiated by the considered profile wherein profile *X* was the destination. These interactions include activities like message sent to *X*, friend request sent to *X*,

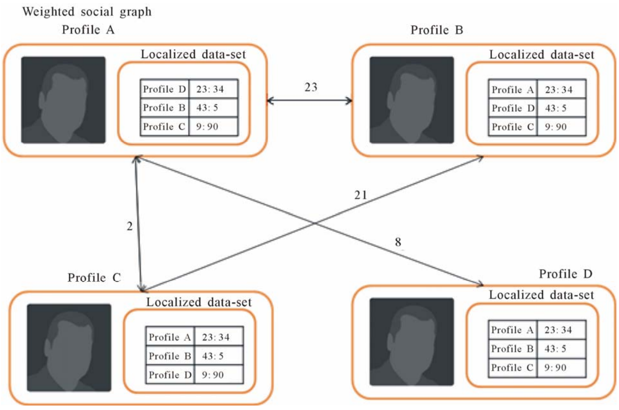

Figure 1. Weighted social graph.

comments on a photo of profile *X* etc. It must be noted that the type of interaction as mentioned earlier may vary based on the social networking website considered.

In real life interactions, the trust among individuals increases over time as the interactions among the individuals increase. These interactions are however bi-directional *i.e.* they include interactions which are initiated by both the involved individuals. The suggested approach employs this concept to calculate the trust levels among profiles. The trust level of profile X for a connected profile *Y* is represented by *T* (*X*, *Y*) and *T* (*X*, *Y*) vice versa.

$$T(Y,X)=\left[I(X)/O(X)\right] \quad (3)$$

$$T(X,Y)=\left[I(Y)/O(Y)\right] \quad (4)$$

In order to calculate *O*(*Y*) and *I*(*Y*), the entries in the localized data-set of X are considered corresponding to profile Y. A similar procedure is used to calculate *O*(*X*) and *I*(*X*). A value for [*I*(*X*)/*O*(*X*)] that is close to 1 represents a high trust level of profile *X* with respect to profile Y. The value of *I*(*X*) increasing at a rate higher than that of *O*(*X*) represents a spam activity initiated by profile *X* and vice versa. It must be noted that the suggested technique will allow first few spam messages, if any, after which as the value *T*(*X*, *Y*) increases the possibility of a spam increases and after the upper threshold the interaction will be blocked and marked as spam interactions. Consider the scenario where say a fake profile A is created which sends out a friend request to a legitimate profile B. Now in such a scenario for B, *O*(*A*) and *I*(*A*) are both 1, since A initiated an outgoing interaction and B replied to it. However after the first few spam messages the value [*I*(*A*)/*O*(*A*)] will increase thus preventing A form initiating any further spam messages.

In order to illustrate the variation in *T*(*Y*, *X*) and hence the reputation of profile *X* with respect to profile *Y* we present a graph showing the cumulative number of incomeing and outgoing messages by profile *Y* with respect to profile *X*. **Figure 2** represents a non-spam normal interaction and **Figure 3** represents a spam interaction.

It must be noted that the time unit used for storing the number of interactions starts with the granularity level of seconds and as and when the time proceeds the granularity level of the data to be stored increases. For example one configuration that can be used is to store the counts on a per second basis for one minute after which store the counts on a per minute basis for one hour and after which store the counts on a per hour basis. The flat sections in the line graphs above represent no interaction. In **Figure 2** throughout the time span represented in the graph the count of *I*(*X*) and *O*(*X*) is almost the same. As a result the value of *T*(*X*, *Y*) is close to 1 throughout the time span, which represents a valid non-spam interaction. In **Figure 3** however it can be seen that after point P the difference between *I*(*X*) and *O*(*X*) increases at a higher rate. Thus after point P the chances of the interaction between *X* and *Y* being a spam increases. It must be noted that the graphs presented in **Figures 2** and **3** are not based on any actual data. The graphs have been used to merely illustrate the way in which the suggested technique operates.

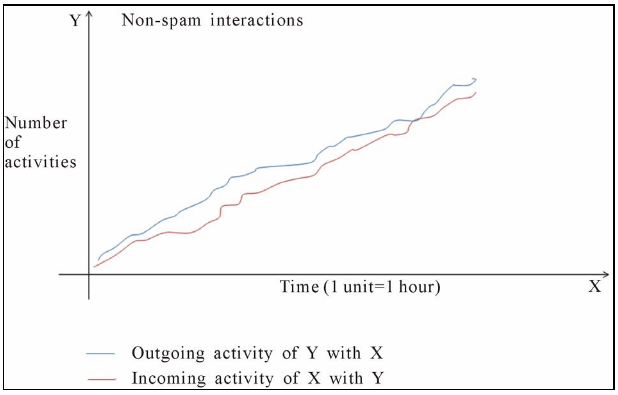

Figure 2. Graph representing non-spam interaction.

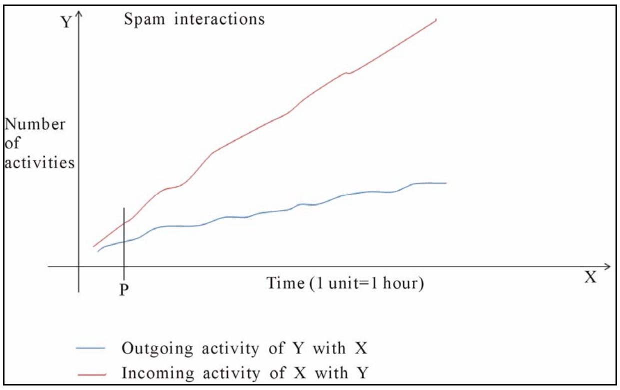

Figure 3. Graph representing spam interaction.

6. Collaborative Filtering of Interactions

For a given profile if an incoming interaction is initiated from a profile then the profile first checks if the profile is connected. If the source profile is connected then the interaction is accepted only if the trust level between the two profiles is higher than a predefined threshold and the localized data-set of each of the profiles is updated. If however the source profile is not connected then the destination profile tries to find the trust level of the source profile with a third profile with which the destination profile has a high trust level. For example if A tries to interact with B then B will accept the interaction only if the trust level between A and B is greater than a particular threshold. However if A is not connected to B then B tries to derive or infer the trust level from a third profile C such that the trust level between B and C is higher than the threshold and C is connected to A.

7. Conclusion

The suggested technique thus addresses the issue of malicious and spam interactions among profiles in a social networking platform in an effective way by correlating the scenario with the interactions in the society. The use of the weighted social graph imparts the suggested technique the ability to not only view and understand the way individuals are connected in a social networking platform but also reflects the trust level among individuals which helps to filter out malicious and unwanted spam interactions. It must be noted that the suggested technique will be unable to prevent spam and malicious interaction if already existing legitimate profiles with high trust level are compromised. The solution to this problem is outside the scope of this work however a potential solution to this problem is the N/R one time password system suggested in [19]. The problems of passwords of legitimate profiles being disclosed by means of attacks like password guessing attacks can be addresses by the approach suggested in [19].

REFERENCES

[1] Social Networking Statistics, URL (last checked 14 Dec 2012) http://www.internetsafety101.org/Socialnetworkingstats.htm.

[2] E. F. Walter, S. Battiston and F. Schweitzer, “Personalized and Dynamic Trust in Social Networks,” *Proceedings of The Third ACM Conference on Recommender Systems*, Association for Computing Machinery, New York, pp. 197-204.

[3] E. F. Walter, S. Battiston and F. Schweitzer, “A Model of a Trust-Based Recommendation System on a Social Network,” *Autonomous Agents and Multi-Agent Systems*, Vol. 16, No. 1, 2008, pp. 57-74. doi:10.1007/s10458-007-9021-x

[4] Uniform Resource Locator, “Weighted Graphs.” (last checked 02 Dec 2012) http://courses.cs.vt.edu/~cs3114/Fall10/Notes/T22.WeightedGraphs.pdf.

[5] A. K. Trivedi, R. Arora, R. Kapoor, Sudip Sanyal and Sugata Sanyal, “A Semi-Distributed Reputation-Based Intrusion Detection System for Mobile Ad hoc Networks,” *Journal of Information Assurance and Security*, Vol. 1, No. 4, 2006, pp. 265-274.

[6] A. K. Trivedi, R. Kapoor, R. Arora, Sudip Sanyal and Sugata Sanyal, “RISM—Reputation Based Intrusion Detection System for Mobile Ad hoc Networks,” *Third International Conference on Computers and Devices for Communications*, Institute of Radio Physics and Electronics, University of Calcutta, Kolkata, 18-20 December 2006, pp. 234-237.

[7] P. A. Chirita, J. Diederich and W. Nejdl, “MailRank: Using Ranking for Spam Detection,” *Proceedings of the 14th ACM international conference on Information and Knowledge Management*, Bremen, 31 October-5 November 2005.

[8] J. Golbeck and J. Hendler, “Reputation Network Analysis for Email Filtering,” *Proceedings of the 1st Conference on Email and Anti-Spam*, Mountain View, 2004.

[9] B. Markines, C. Cattuto and F. Menczer, “Social Spam Detection,” *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, Madrid, 21-21 April 2009. doi:10.1145/1531914.1531924

[10] H. Y. Lam and D. Y. Yeung, “A Learning Approach to Spam Detection based on Social Networks,” *Proceedings of the Fourth Conference on Email and Anti-Spam*, Mountain View, 2007.

[11] R. Bhadauria and S. Sanyal, “Survey on Security Issues in Cloud Computing and Associated Mitigation Techniques,” *International Journal of Computer Applications*, Vol. 47, No. 18, Foundation of Computer Science, New York, 2012, pp. 47-66. doi:10.5120/7292-0578

[12] R. A. Vasudevan, A. Abraham, S. Sanyal and D. P. Agrawal, “Jigsaw-Based Secure Data Transfer over Computer Networks,” *IEEE International Conference on Information Technology*: *Coding and Computing*, Las Vegas, Vol. 1, 2004, pp. 2-6.

[13] A. Abraham, R. Jain, S. Sanyal and S. Y. Han, “SCIDS: A Soft Computing Intrusion Detection System,” In: A. Sen, *et al.*, Eds., *6th International Workshop on Distributed Computing*, *Lecture Notes in Computer Science*, Vol. 3326, Springer Verlag, Berlin, pp. 252-257.

[14] S. Sanyal, D. Gada, R. Gogri, P. Rathod, Z. Dedhia and N. Mody, “Security Scheme for Distributed DoS in Mobile Ad Hoc Networks,” *Technical Report*, School of Technology & Computer Science, Tata Institute Of Fundamental Research 2004.

[15] S. Pal, S. Khatua, N. Chaki and S. Sanyal, “A New Trusted and Collaborative Agent Based Approach for Ensuring Cloud Security,” *Annals of Faculty Engineering Hunedoara International Journal of Engineering*, Vol. 10, No. 1, 2012, pp. 71-78.

[16] P. Rathod, N. Mody, D. G., Rajat G., Z. Dedhia, S. Sanyal and A. Abraham, “Security Scheme for Malicious Node Detection in Mobile Ad Hoc Networks,” In: A. Sen *et al.*, Eds., *6th International Workshop on Distributed Computing*, *Lecture Notes in Computer Science*, Vol. 3326, Springer Verlag, Berlin, 2004, pp. 541-542.

[17] Uniform Resource Locator, “Social Graph.” (last checked on 11 Dec 2012) http://en.wikipedia.org/wiki/Social_graph.

[18] Uniform Resource Locator, “Recommender System.” (last checked 10 Dec 2012) http://en.wikipedia.org/wiki/Recommender_system.

[19] V. Goyal, V. Kumar, M. Singh, A. Abraham and S. Sanyal, “CompChall: Addressing Password Guessing Attacks,” *IEEE International Conference on Information Technology*: *Coding and Computing*, Vol. 1, 4-6 April 2005, pp 739-744.