Malicious Users' Circle Detection in Social Network Based on Spatio-Temporal Co-Occurrence

Zahid Halim, Mian Maqsood Gul, Najam ul Hassan, Rauf Baig, Shafiq Ur Rehman and Farhat Naz
FAST-National University of Computer and Emerging Science, Islamabad, Pakistan.
{zahid.halim , i060225,i060430, rauf.baig, shafiq.rehman, fartat.naz}@nu.edu.pk

Abstract -- Online social networks have witnessed massive increase from the point of view of users during last decade. However, it is also becoming center of attraction for spammers. It is a complex problem to trace spammers on a large scale. Since spammers communicate covertly so by analyzing simple graph of social network, they cannot be identified. In order to find the circle of people involved in the malicious messaging, we associate people on the basis of their spatio-temporal co-occurrence i.e. people frequently communicating with each other. In this paper, we associate people on the basis of their spatio-temporal co-occurrence and find the users involved in malicious communications.

Keywords-- Social Networks, Prediction, Spatio-Temporal Analysis, Web mining

I. INTRODUCTION

Early social networking websites primarily focused on bringing group of people together in order to interact through chat rooms and share information through personal homepage. However, in 2002 these social networking sites emerged as the most popular sites, by allowing users to publicize and share contents. Social network sites may become at risk to different types of malicious action such as message spamming. These spammers are encouraged to spam in order to broadcast pornography or for promotion of certain content.

These social network websites have become the center of internet users' attention, by creating a versatile, common platform and allowing people to connect with other people- who share their common interests; activities, political views, language, cultural and religious values or nationality. Apart from internet users, these websites also got researchers attention who tries to understand the reason behind users' engagement with them and how to extract useful information from these.

The reason behind this introduction is to provide enough background that forces us to do our research on these social network websites.

A. SOCIAL NETWORK

A social network is group of social entities i.e. comprising of individuals or organizations called "nodes or actor", which are linked by certain relationship called "tie, edge or link". So social network sites are basically sites that allow to publish them self and allow to form relation with other people, it also allow individuals to eloquent and make visible their social network, which make it different from other computer mediated communications.

Social network analysis has many useful insights as one can identify important nodes, nodes with many connections or those with greatest effect on the network [1, 2] or to find the subset of network that show interesting patterns [3, 4]. Moreover, it also helps us in identifying nodes i.e. who is core of network and who is on the periphery or where are the clusters and who is in it.

Knowledge of social network is important. For example, viral marketing take advantage of relation between consumers to increase business output [1, 2] or one can identify group of people involved in malicious communication. Social network can also help in expanding social circle [4].

Despite of many uses, analyzing social network to find required pattern is a difficult job. In this paper, we are computationally mining social network for finding group of users involved in malicious activities from spatio-temporal data i.e. data is associated with location and time. Moreover, we will be focusing on the spatio-temporal co-occurrence to establish ties between users in order to detect the group of users involved in malicious activities.

B. OUR CONTRIBUTION

Currently there are many techniques used for filtering malicious content. However, what we are proposing here is not only identifying malicious content but also the circle of people involved in malicious communication. We first of all associate people upon spatio-temporal co-occurrence i.e. associate people that frequently communicate with each other. Now with the help of this we can identify that relation which covertly exists between users involved in malicious activities. Once we associate people, we apply latent semantic indexing (LSI) to identify malicious content in the user's messages. Once a user causing malicious activity is identified, we then start searching users, who are connected to him/her and this way we are able to identify whole circle of people involved in malicious communication.

II. PREVIOUS WORKS

Existing spam/malicious filter currently are typically stationary such as white listing [5]. One of the techniques to detect malicious content is *Bayesian classifier*. Illustration of *Bayesian classifier* is given in [6], which describe how it can be applicable to filter malicious content. One more method that is connected with technique used in our work is proposed by Gee [6], he uses *latent semantic analysis* (LSI) to filter malicious content. However in Gee's [6] work textual feature of email is used. In addition to this work the literature in [20-23] address spatio-temporal mining in other domains too.

A. LATENT SEMANTIC INDEXING

Latent Semantic Indexing (LSI) analysis as discussed in [7] is a statistical technique that calculates association among entities of corpus, in an attempt to conquer issues of conventional matching.

In general, it involves creating a weighted *term-document matrix*, upon which we perform *Singular Value Decomposition* and using this matrix to find concepts contained in the text.

Support vector machine (SVM) is created using as a term-document matrix. Singular value decomposition (SVD) is applied estimating usages. Then SVD-derived matrices are summarized to k dimensions [7, 8, 9, 10] provides an excellent background on kernel machines and how they are not susceptible to local minima like other methods.

The end result is a condensed vector for each term [8]. Studies have shown that these vectors are robust, effective indicators of meaning and enjoy a higher recall than searching only with individual terms [11, 8, and 9].

One issue with LSI is that it does not support the ad-hoc addition of new documents once the semantic set has been generated.

Work in [7] provides an exact summary of how comparison of two vectors is generated. A cosine of 1 signifies that the two vectors (be they term or document) are considered to be exactly similar which is different from identical. New test documents not previously included in the semantic set can be used for comparison as well, by combining the vectors of their composite terms.

B. MINING SOCIAL NETWORK

Apart from *co-occurrence*, there are also some other criteria, on which ties between actors can be formed i.e. *similarity*, *communication* etc and using this information, we can analyze network.

Similarity is basically degree of likeness and symmetry between two or more concepts or objects i.e. sharing same properties or features. Generally, friends tends to be alike [12], from this we conclude that people who share common features tends to be associated to one another in some way. For example, websites with similar text and links represents group of related individuals [13].

Communication, the exchange of information or resources is normally observed between related people. This means, communication can infer association between people. For example emails [14], instant messaging [5] can be used to trace association, since such communications are directed i.e. between sender and receiver.

Our current work mainly focus on time series approach i.e. data associated with each node is considered as the collection of time series and using this time series, we calculate the distance between time series with the help of Euclidean [15] or LCSS [16] distance function. If the distance is less (based upon a threshold) then the resulting nodes are considered to be co-occurring.

III. MINING SOCIAL NETWORKS FROM SPATIO-TEMPORAL EVENTS

Event is basically an occurrence of co-occurring nodes in social network. In order to identify the spatio-temporal behaviors of nodes, common actions of social network is studied i.e. actors (nodes) group together, when they interact. By this, we mean that social event will result in spatio-temporal co-occurrence. Spatio-temporal event can be defined as follows.

Definition: for event to be spatio-temporal event it must satisfy following condition:

- *Actors involved in event must have same location.*
- *The difference in time of participation of actors in particular event must be less then certain threshold* ***(δ_t)***.
- *Number of actors in event must be at least two.*

The first two conditions specify the constraints for spatio-temporal co-occurrence. The value of both can be as restraining or as lenient, depending upon how you associate people. However, it is neither possible nor practical to have exact co-occurrence. Furthermore, by allowing this tolerance to be diverse, an appropriate tolerance can be selected depending upon need.

Third condition is general, which is true for all events i.e. for event to occur; at least two actors must interact.

Once we create spatio-temporal events, we give them weight between 0 and 1. With the help of these weights, we will be able to predict association between actors. In order to be more precise, we set threshold and events whose value is greater than threshold and can be selected only. Same algorithmic approach is followed in [17]

After joining, we will find actors involved in malicious activities. In order to do so, we will be using LSI.

IV. LATENT SEMANTIC INDEXING (LSI)

In order to apply on malicious content filtering, the posts i.e. text message (in our case) is considered as document and each post is considered as vector in space. In order to create vector space model, we represent each post according to its feature value. This leads us to an important question i.e. which features of posts are significant for classification into malicious text or not. In our case, we use *text based feature set.*

A. DIMENSIONALITY REDUCTION

When tokenizing text, one frequently faces very high dimensional data, which is difficult to handle. Document frequency thresholding plays important role in reducing dimensions. An overview of different term selection technique is given in [18].

Document Frequency Thresholding

The basic idea is that, the frequent items play no significant role in distinguishing between malicious and non-malicious text.
Once dimensions are reduced using *document frequency thresholding*, we can also apply *Information Gain* technique to increase performance.

The basic idea of *Information Gain* is to identify how well every feature separates the given data. *Information Gain* along with other techniques for feature selection is given in [18].

B. TRAINING AND CLASSIFICATION

The training consists of indexing the malicious and non malicious text along with the computation of *singular value decomposition* and classification consists of query indexing and retrieval of closest message from the training set. Every new message is classified based on the feature set used and comparing it with the query vector, if the closest message from training set is malicious then the new message is malicious else it is non malicious. The distance is measured on the basis of angle between the query and training vector [19].

V. EXPERIMENTAL EVALUATIONS

For experiment we use data from Facebook. The dataset consist information about user's personal information and posts. We first create spatio-temporal database from this data i.e. store only information of user regarding its location and time of interaction. Once we have spatio-temporal database, we start associating people on the basis of spatio-temporal co-occurrences for different values of threshold ***(δ_t)***. Figure 1 shows when the value of threshold ***(δ_t)*** is very small (e.g. 60 sec).

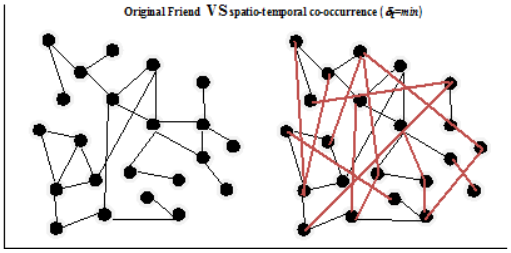

Figure 1: Original friend graph VS Spatio temporal graph (δt =min)

The resulting graph shows a large deviation from the original graph because it considers those interactions which occur accidently. These wrong associations are showed with red lines Figure 2 shows when the value of threshold ***(δ_t)*** is very large (e.g. 28800 sec or 8 hrs).

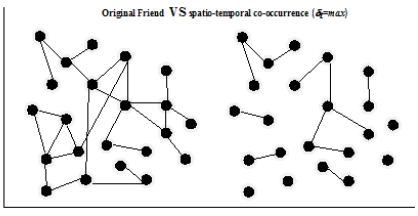

Figure 2: Original friend graph VS Spatio temporal graph (δt =max)

The resulting graph also shows deviation from original graph because it misses interaction between friends who communicate for short time. So we have to select an appropriate value, so that

it neither selects accidental values nor misses important data, only then we will be able to have most appropriate graph that resemble the correct association as shown in Figure 3.

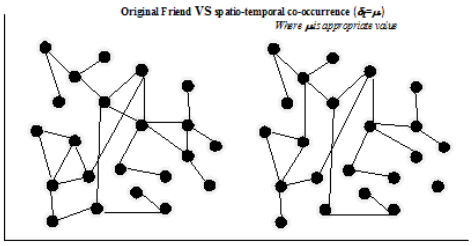

Figure 3: Original friend graph VS Spatio temporal graph (δt =μ)

Once we set the appropriate value, we increase our dataset size in order to better find the circle of people involved in malicious communication. After associating people on the basis of spatio-temporal co-occurrences, we find malicious text send by the particular individual. For this, we apply latent semantic indexing and after finding the user involved in malicious activities, as shown Figure 4 with larger circle, we are able to find other users who are involved in these activities as well.

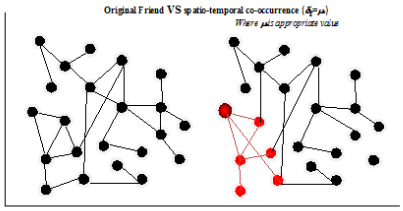

Figure 4: Users' involved in malicious Communication

VI. CONCLUSION

In this paper we propose spatio-temporal mining on social network to identify circle of users involved in malicious activities with the help of latent semantic analysis. We then compare the results produced from spatio temporal co-occurrence with that of original association/ties with in social network, which is very encouraging as the association generated by spatio-temporal co-occurrence and real one are very close to each other. Once we set the value of threshold to appropriate level, we increase the number of nodes i.e. actor so that we can get better picture. Overall, experiment indicate that Latent Semantic Indexing perform very well for identifying malicious contents, if the feature set is properly chosen. One obvious limitation of this approach is how you select your feature set and how rich it is. If the feature set is very small then most of the malicious content will not be traced. However, the greater your feature set, better the performance gained. The results generated are very encouraging and creates many opportunities.

REFERENCES

[1] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network", KDD,(2003), pp. 137–146.

[2] M. Richardson and P. Domingo, "Mining knowledge-sharing sites for viral marketing", KDD, (2002), pp. 61–70.

[3] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, "Mining newsgroups using networks arising from social behavior", WWW, (2003), pp. 688–703.

[4] M. Mukherjee and L. B. Holder, "Graph-based data mining on social networks", in Workshop on Link Analysis and Group Detection (in conj. with KDD), (2004).

[5] J. Resig, S. Dawara, C. M. Homan, and A. Teredesai, Extracting social networks from instant messaging populations", in Workshop on Link Analysis and Group Detection (in conj. with KDD), (2004).

[6]Androutsopoulos, J. Koutsias, K. Chandrinos, and C. D. Spyropoulos, "An experimental comparison of naive Bayesian and keyword based anti-spam filtering with personal e-mail messages", in SIGIR '00: proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval,2000, pp. 160–167.

[7] N. Cristiani and B. Scholkopf, "Support Vector Machines and Kernel Methods: The New Generation of Learning Machines". AI Magazine. Fall 2002 (Vol 23, no. 3), pp. 31-41, 2002.

[8]T.K. Landauer. and S.T. Dumais, "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge",Psychological Review, 104, 211-240, 1997.

[9] T.K. Landauer, P.W. Foltz, D. Laham, "An Introduction to Latent Semantic Analysis", Discourse Processes, 25, 259-284, 1998.

[10] S. Deerwester, S.T. Dumais, , G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, 41:391-407, 1990.

[11] K. Gee, "Text Classification Using Latent Semantic Indexing", Master's thesis. The University of Texas at Arlington, 2001.

[12]K. Carley, "A theory of group stability", American Sociological Review, 56 (1991), pp. 331-354.

[13] L. A. Adamic and E. Adar, "Friends and neighbors on the web, Social Networks", 25 (2003), pp. 211–230.

[14] M. F. Schwartz and D. C. M. Wood, "Discovering shared interests using graph analysis", CACM, 36 1993),pp. 78–89.

[15] Y. Wang, E. Lim, and S. Hwang, "On mining group patterns of mobile users", DEXA, (2003), pp. 287–296.

[16] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories", ICDE, (2002),pp. 673-684.

[17] H. W. Lauw, E. P. Lim, T. T. Tan and H. H. Pang, "Mining Social Network from Spatio-Temporal Events", 2009, pp. 88-89.

[18] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization", in Proceedings of ICML-97, 14th International Conference on Machine Learning, D. H. Fisher, ed., Nashville, US, 1997, Morgan Kaufmann Publishers, San Francisco, US, pp. 412–420.

[19] A. N. Langville, "The linear algebra behind search engines", in Journal of Online Mathematics and its Applications (JOMA), 2005, Online Module,2005.

[20] Masaaki Ishikawa and Takayuki Tanabe, Simulation Analyses of Spatio-temporal Patterns formed by Chemotactic Bacteria under Random Fluctuations, International Journal of Innovative Computing, Information and Control, vol.5, no.1, pp.57-66, 2009.

[21] Muhammad Aziz Muslim, Masumi Ishikawa and Tetsuo Furukawa, Task Segmentation in a Mobile Robot by mnSOM and Clustering with Spatio-temporal Contiguity, International Journal of Innovative Computing, Information and Control, vol.5, no.4, pp.865-876, 2009.

[22] Chung C. Chang and Li-Wei Sung, A Computer Maintenance Expert System Based on Web Services, ICIC Express Letters, vol.3, no.4 (B), pp.1209 - 1214, 2009.

[23] Youwei Wang, Wenjing Zhang, Cheng Zhang, Hong Ling and Fuchun Zhao, Establishing Website Navigation Support Systems by Mining Sequential Patterns, ICIC Express Letters, vol.4, no.2, pp.395-400, 2010.