Detection and Analysis of Hidden Activities in Social Networks

Divya Prakash
Student
SCT College of Engineering
Thiruvananthapuram

Subu Surendran
Associate Professor
SCT College of Engineering
Thiruvananthapuram

ABSTRACT
Social networks are flourishing because of fast growing Internet and the World Wide Web, and more research efforts have been put on Social Network Analysis (SNA). A social network can be modeled like a graph, where the nodes represent persons, and an edge between them represent direct relationship between the persons. One of the issues in SNA is to identifying criminals from groups of individuals. In a real social network, there must have various relationships between individuals, like friendships, business relationships, and common interest relationships etc. The internet itself is a huge social network. To model such a network, link analysis need to be proposed. A page in web may treat as a node, and hyperlink between them can be represented as relationships. After social network graph is constructed, link analysis and graph partitioning algorithms may be applied to identify the hidden links in that network. Most of the existing algorithms related to social network analysis assume that their existing only one single social network, with relatively multiple relationship like Web page linkage. In typical social networks, there always have various kinds of relations. Every relation can be identified as a relation network. These different types of relations play different tasks in different roles.

The work here attempts to find the problem of mining hidden relationships on social networks. Social network analysis (SNA) is a set of powerful techniques that can be used to identify clusters, patterns and hidden structures within social networks. Here the problem is identified with the following steps.

1. Analyzing information flow through the network using affected dataset,
2. Discovering non-obvious relations between actors, and
3. Identifying nodes that are directly or indirectly connected to most other nodes in the social network.

This is done with the help of mining algorithms like Min-cut and Regression.

Keywords
Graph Mining, Social network analysis, Community Detection.

1. INTRODUCTION
The Social Network is considered as social structure made with object as nodes and relation as edges. This allows users to meet others, keep relationships, share opinions. Social network analysis (SNA) finds the social relationships with nodes and edges. Social network analysis can be used to study hidden relationships. It can provide information about the unique characteristics of different relations.

Moreover, social network analysis can be used to understand criminal networks. Nowadays different criminal organizations are using some social networking sites for communication and recruitment purpose, such type of organisation will always hide their relationship. Hence socialnetworking information needs to be analyzed to track hidden relations. Social networks may be advantageous and they become great fun, but people really need to identify the complicated world of social network and wanted to move wisely [2, 3]. Social network analysis (SNA) is the best method to study criminal organizations. In a real human community, there may seen many relations such as community for peoples work at the same institute, sharing common interests, some are under the same religion etc. All this community can be represented by a big graph in which the vertex represents individuals, and the strength of relation can be evaluate by edges between them. The edges of this graph should be multiple because of the relations may become different. Hence we split the big graphs into multiple small graphs. Each graph should relate to one kind of relation. The existing relationships between people may not have an equivalent role. These relations can be modeled mathematically and extract the most dangerous relations. Hence each relation in the graph is represented with a weight matrix. Every element in this weight matrix represents the strength of relation between two objects.

Social network analysis (SNA) well suited for relation extraction. Amongst them community mining is one of the major directions. All the existing methods in community mining assume that there is only one kind of relation in the network. Actually, there should have numerous multiple related social networks, representing a particular kind of relationship with each other, and each relationship must act as a unique role for a particular task. Social network analysis (SNA) and mining community represents a great role in methodology from the traditional one.

In this work Relation extraction has been used to identify person, and relationship between them and Regression based algorithm associate different objects (such as persons, organizations) in social network, has been applied to find out the hidden relations, network intrusion detection, and other crime analyses that involve tracing abnormal activities. Social network analysis has been used to analyze criminal's associations and roles among entities in a criminal network.

Section 2 discusses related works about hidden relations in social network applications. Section 3 explores implementation methods in work, discusses implementation techniques using relation extraction from multiple relations in social network, discusses the Regression based algorithm trying to find a combined relation which makes the relationship between communities and also discusses a min cut-based algorithm for relation extraction in the case of user provides only one community details. Section IV discusses the Experiments and Result. Section V concludes the paper.

2. RELATED WORKS

Nowadays the importance of social networks are increasing, with law enforcement and intelligence agencies have come to realize the value of detailed knowledge of hidden relations, or details about the network which offenders about the criminals who are involved in crimes together [8,9]. Different crimes activities are operating in a hidden fashion for conceal their illegal. For investigating such activities; it will not only concentrate on individuals and also attempts to determine criminal groups. Thus, it should be most valuable to understand networks of criminals from data sources which are available to investigators, and need to understand these data with the help of social network analysis methods. Social network analysis may have provided important information about persons in criminal history. Hence, investigators can understand the players in data and point them to closer look. In general, knowledge from the network structures dataset provides a basic idea of hidden related agencies to make strategic or tactical decisions.

2.1 Community Mining

With the increasing growth of World Wide Web, community mining is also growing. Much research work has been developed for mining the unique communities of web pages, in general, community can be said to be collection of objects they share some common properties. Community mining has many related properties to the graph-cut problem. In general social network analysis and community mining can be seen under graph mining [25]. The sub-graph identification is very useful in community mining. Most of the previous techniques related to graph mining and community mining are based on a single graph, in which there is only one kind of relationship between the objects. But, in actual social networks, the objects always have different kinds of relations.

2.2 Relational Mining

Relational mining especially multi-relational data mining has received a lot of attentions in recent years. Multi-relational data mining aims at dealing with knowledge discovery from relational databases consisting of multiple tables. It tries to analyze data from a multi-relational database directly, without the need to transfer the data into a single table first. Thus the relations mined can reside in a relational or deductive database. Intuitively, the different tables in relational database can be thought of as different relation networks. Thus, the relation extraction techniques described in this paper has potential applications to multi-relational data mining.

2.3 Feature Extraction

Feature extraction is very much useful in machine learning and data mining at classification and clustering. Feature extraction can be used for finding a linear combination of the original features that can better describe the internal structure of the data set. Actual feature extraction methods include Locality Preserving Projection (LPP), Linear Discriminate Analysis (LDA), and Principle Component Analysis (PCA). LDA is performed under supervised mode, LPP can be performing under either supervised mode or unsupervised mode and PCA is performed under unsupervised. From the different relationships of objects, find a linear combination of relations, such as intrinsic relationship between the objects with the query from the user.

3. METHODOLOGY

This chapter comprises of the methods which are adopted for the thesis and these mainly consist of algorithms. The algorithms are Regression-Based algorithm and Min Cut-Based Algorithm.

3.1 A Regression_Based Algorithm

Linear regression is widely used in biological, behavioral and social sciences to describe possible relationships between variables. In general, regression analysis helps one to understand how the typical value of the dependent variable changes when any one of the independent variables is varied, during the other independent variables are held fixed. Regression analysis is commonly used for forecasting and prediction. Regression analysis is also used to identify which independent variables are associated to the dependent variable, and to diagnose the forms of these association or relationships.

The key idea of this algorithm is trying to find a combined relation which makes the relationship between the members within the community and at the same time the relationship between the same members with in different communities with examples provided by the user. All the relation can be normalize to make the biggest strength weight on the edge to become 1. The target relation between the labeled objects can be constructed as follows:

$$\widetilde{M_{ij}} = \begin{cases} 1, & \text{example } i \text{ and example } j \text{ have the same label;} \\ 0, & \text{otherwise.} \end{cases}$$

Where M is a $m \times m$ matrix and M_{ij} represent the relationship between node i and j. After building the target relation matrix identify a linear combination from the existing relations which is very much approximate the target relation in the sense of L2 norm. Sometimes, if two objects belong to the same community and can only provide the possibility that two objects belong to the same community then the user will become uncertain. In that case, M can be defined as follows.

Let $a = [a_1, a_2, \ldots a_{n1}]^T \in R^n$ denote the combination coefficients for different relations. The approximation problem can be characterized by solving the following optimization problem:

$$a^{opt} = \arg \min_a ||M - \sum_{i=1}^{n} a_i M_i||^2 \qquad (1)$$

This can be written as a vector form. Since the matrix $M_{m \times m}$ is symmetric, we can use a $\frac{m(m-1)}{2}$ dimensional vector v to represent it. The problem (1) is equivalent to:

$$a^{opt} = \arg \min_a ||v - \sum_{i=1}^{n} a_i v_i||^2 \qquad (2)$$

Equation (2) is actually a linear regression problem. From this point of view, the relation extraction problem is interpreted as a prediction problem. Once the combination coefficients are computed, the hidden relation strength between any object pair can be predicted. In real applications, the user does not need to specify the relationships between any pair of objects. That is, the vector v needs not to be $\frac{m(m-1)}{2}$ dimensional. Assume that v is a k-dimensional vector in the following.

First consider the simplest case that

$$\sum_{i=1}^{n} a_i v_i = v \qquad (3)$$

define:

$$V = [v_1 , v_2 , \dots v_n] \quad (4)$$

Thus, equation (3) can be rewritten as follows:

$$V_a = v \quad (5)$$

Suppose the rank of V is min(k, n). We have the following facts:

• when k < n, the set of solutions to equation (5) forms a (n − k) dimensional vector subspace;

• when k = n, there is a unique solution to equation (5);

• when k > n, there is no solution to equation (5).

In the first two cases, we get a solution with perfect match (The minimization that, the value of k reflects the quantity of the user's information needs. k is small when the query submitted by the user is simple. With a complex query, k can be larger than n. In this case, the optimal solution to (2) is obtained when the derivative of this objective function with respect to a is zero.

By some algebraic steps, we have:

$$\Rightarrow \quad V^T(v - Va) = 0 \quad (6)$$

$$\Rightarrow \quad V^T Va = V^T v \quad (7)$$

Since the matrix V has full rank as we assumed, i.e., rank *(V) = min(k, n)*, the matrix $V^T v$ is
invertible and the optimal solution to (4.2) is

$$a^{opt} = (V^T v) - 1 \quad (9)$$

When the matrix V is rank deficiency, i.e., rank $(V) < min(k, n)$, there will be multiple solutions with the same minimization value. In such case, choose the a with minimum norm as the solution. The objective function (2) models the relation extraction problem as an unconstrained linear regression problem.

3.2 A Min Cut-Based Algorithm

In previous algorithm exhibit a widespread idea for exacting the hidden relation from the social network. This method cannot give a better result when the example provided by the user is from only one community. Then regression model should fail in such a case. For dealing with the single community issue, need to identify the weakest connection in the extracted relation.In graph theory, tightness of the graph can be evaluating the minimum cut value on the graph. Let M be the weight matrix of a weighted graph G. And the number of vertices are indicated as m. A cut on the graph is which, separates the vertices into two disconnected groups defined as A and B in which A ∩ B = {} and A ∪ B = G.

Graph connectivity is one of the classical subjects in graph theory, and has many practical applications, for example, in chip and circuit design, reliability of communication networks, transportation planning, and cluster analysis. Finding the minimum cut of an undirected edge-weighted graph is a fundamental algorithmically problem. Precisely, it consists in finding a nontrivial partition of the graphs vertex set V into two parts such that the cut weight, the sum of the weights of the edges connecting the two parts, is minimum. The usual approach to solve this problem is to use its close relationship to the maximum flow problem. The famous Max-Flow-Min-Cut-Theorem by Ford and Fulkerson [1956] showed the duality of the maximum flow and the so-called minimum s-t-cut. There, s and t are two vertices that are the source and the sink in the flow problem and have to be separated by the cut, that is, they have to lie in different parts of the partition. Until recently all cut algorithms were essentially flow algorithms using this duality. Finding a minimum cut without specified vertices to be separated can be done by finding minimum s-t-cuts for a fixed vertex s and all |V|-1 possible choices of t € V \ {s} and then selecting the lightest one. The thesis throughout deals with an ordinary undirected graph G with vertex set V and edge set E. Every edge e has nonnegative real weight w(e). The simple key observation is that, knowledge of how to find two vertices s and t, and the weight of a minimum s-t-cut.

Let s and t be two vertices of a graph G. Let G/{s, t} be the graph obtained by merging s and t. Then a minimum cut of G can be obtained by taking the smaller of a minimum s-t-cut of G and a minimum cut of G/{s, t}.This holds since either there is a minimum cut of G that separates s and t, then a minimum s-t-cut of G is a minimum cut of G; or there is none, then a minimum cut of G/{s, t} does the job.So a procedure finding an arbitrary minimum s-t-cut can be used to construct a recursive algorithm to find a minimum cut of a graph.

The following algorithm, known in the literature as maximum adjacency search or maximum cardinality search, yields the desired s-t-cut.

Minimumcutphase(G, w, a)

A ← {a}

while A ≠ V

add to A the most tightly connected vertex.

store the cut-of-the-phase and shrink G by merging the two vertices added last.

A subset A of the graphs vertices grows starting with an arbitrary single vertex until A is equal to V. In each step, the vertex outside of A most tightly connected with A is added. Formally,

add a vertex z ∉ y A such that w(A, z) = max{w(A, y)|y ∉ y A},

where w(A, y) is the sum of the weights of all the edges between A and y. At the end of each such phase, the two vertices added last are merged, that is, the two vertices are replaced by a new vertex, and any edges from the two vertices to a remaining vertex are replaced by an edge weighted by the sum of the weights of the previous two edges. Edges joining the merged nodes are removed.

The cut of V that separates the vertex added last from the rest of the graph is called the cut-of-the-phase. The lightest of these cuts-of-the-phase is the result of the algorithm, the desired minimum cut:

minimumcut(G, w, a)

while |V|> 1

minimumcutphase(G, w, a)

if the cut-of-the-phase is lighter than the current minimum cut then store the cut-of-the-phase as the current minimum cut.

Notice that the starting vertex a stays the same throughout the whole algorithm. It can be selected arbitrarily in each phase instead.

4. IMPLIMENTATION

As shown in the figure 1 relation extraction process is done using regression based algorithm and min cut based algorithm. From the dataset the hidden relations from social network can be extracted. Thus the experimental results show that the proposed methods more better, more reliable and scalable system for complex applications.

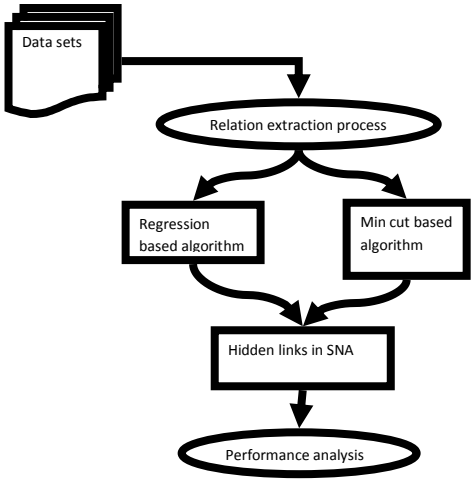

Figure 1 Extracting hidden links in Social Networks

4.1 Threshold Cut

In some cases, the user expects the mined community has a reasonable size, and relation strength in the mined community is strong enough. A simple method based on predefined threshold can be used for this purpose. This method is named as Threshold Cut. There are two ways to perform this algorithm.

This work the threshold cut value of two algorithms is determined separately. In the first algorithm ie Regression based algorithm the output with a dataset as input is a $k \times k$ metrics with different strength between nodes, from that min value is determined and set as threshold for first algorithm. In the second algorithm min cut based algorithm the output is minimum cut value. This minimum cut value is set as Threshold for this algorithm. The relation strength which is higher than this defined threshold is calculated as hidden relations. There for two set of output are listed.

5. RESULT

Figure 2 shows experimental result of the system, the graph represented by the social network with the nodes plotted as circles represents the individual persons in the social network and edges between them represented the relationship between the persons. The social network dataset DBLP are used for the experiment. The DBLP dataset is a computer science bibliography provides a comprehensive list of research papers in computer science. From this dataset hidden relationships are traced out using regression and min cut algorithms. In figure 2 edges highlighted by green color indicate people who have hidden links.

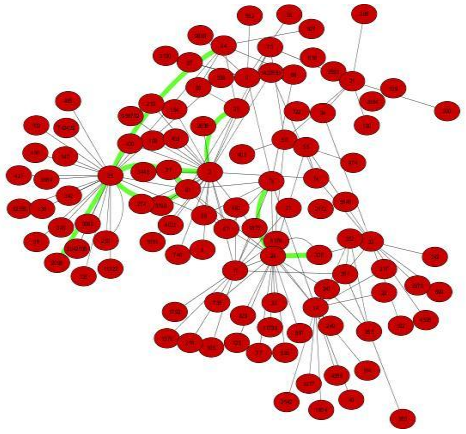

Figure 2 Result of hidden relation Detection in social network

Table 1 shows experimental result of the system, contains number of hidden relations traced from different datasets using Regression method and Min cut method.

Table 1 Finding hidden relations with different datasets.

Datasets	No of Relations	Using Regression Method		Using Min Cut Method	
		No of hidden relations	No of genuine relations	No of hidden relations	No of genuine relations
DBLP Dataset	118	69	49	69	49
ego-Facebook	109	105	4	105	4
com-Orkut	250	225	25	225	25
com-Live Journal	200	128	72	128	72
Email-Enron	129	95	34	95	34

6. CONCLUSION

Social network analysis is the same which applies to any innovative technology. It is just one among other tools useful in understanding hidden activities and becomes only piece of the puzzle. Role of subject matter experts are essential for providing a context for the research. Another fact is that, the primary assumption of network analysis of hidden activity may not be fully valid. Untiring efforts must be made in social network analysis to trace out hidden root cause of criminals. But most pathetic phase is that the research to solve the troubles in this field by using advanced computational ways is still in its beginning stage. More strengthened and coordinated attempts are inevitable for gaining momentum to this. Due to sensitive nature of real crime datasets, they are not easily available for academic research ends and involve problems and difficulties. In this work a conceptual framework for detecting and analyzing hidden groups and their activities in social network is presented as a foundation towards framing advanced computational methods for detecting organized criminals in social networks.

7. REFERENCES

[1] G. Aggarwal, T. Feder, K. Kenthapadi, Samir Khuller, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving Anonymity via Clustering," Proceedings of PODS, June 26-28, Chicago, Illinois, US, 2006.

[2] Z. Baird, J. Barksdale, and M. Vatis, "Creating a Trusted Network for Homeland Security", Markle Foundation, 2003.

[3] Z. Baird and J. Barksdale, "Mobilizing Information to Prevent Terrorism: Accelerating Development of a Trusted Information Sharing Environment", Markle Foundation, 2006.

[4] C. Best, J. Piskorski, B. Pouliquen, R. Steinberger and H. Tanev, "Chapter 2. Automating Event Extraction for the Security Domain, Intelligence and Security Informatics: Applications and Technique", Editors: H. Chen and C. C. Yang, Springer-Verlag, to appear in 2008.

[5] . Caruson, S. A. Macmanus, M. Khoen, and T. A. Watson, "Homeland Security Preparedness: The Rebirth of Regionalism," Publius, 35(1), 2005, pp.143-189.

[6] R. R. Friedmann and W. J. Cannon, "Homeland Security and Community Policing: Competing or Complementing Public Safety Policies," Journal of Homeland Security and Emergency Management, 4(4), 2005, pp.1-20.

[7] Arquilla, J., Ronfeld, D.: "Networks and Netwars: The Future of Terror, Crime, and Militancy". RAND Corporation, Santa Monica, CA (2001)

[8] Xuning Tang and Christopher C. Yang, " Terrorist and Criminal Social Network Data Sharing and Integration," College of Information Science and Technology Drexel University, Philadelphia 2009.

[9] K.Liu and E.Terzi, "Towards Identity Anonymization on Graphs," in ACM SIGMOD'08 Vancouver, BC, Canada: ACM Press, 2008.

[10] C. C. Yang, "Information Sharing and Privacy Protection of Terrorist or Criminal Social Networks," Proceeding of IEEE International Conference on Intelligence and Security Informatics, Taipei, Taiwan, 2008.

[11] B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks," in IEEE International Conference on Data Engineering, 2008.

[12] D. Thacher, "The Local Role in Homeland Security," *Law &Society*, 39(3), 2005, pp.557-570.

[13] B. Thuraisingham, "Security Issues for Federated Databases Systems," Computers and Security, North Holland, December,1994.

[14] B. Thuraisingham, "Chapter 1. Assured Information Sharing:Technologies: Challenges and Directions, Intelligence and Security Informatics: Applications and Technique", Editors:H.Chen and C. C. Yang, Springer-Verlag, to appear in 2008.

[15] R. C. Wong, J. Li, A. Fu, and K. Wang, "k-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing," Proceedings of SIGKDD, August 20-23, Philadelphia, Pennsylvania, US, 2006.

[16] X. Xiao and Y. Tao, "Personalized Privacy Preservation, "Proceedings of SIGMOD, June 27-29, Chicago, Illinois, 2006.

[17] C. C. Yang, N. Liu, and M. Sageman, "Analyzing the Terrorist Social Networks with Visualization Tools," Proceedings of the IEEE International Conference on Intelligence and Security Informatics, San Diego, CA, US, May 23 – 24, 2006.

[18] C. C. Yang and T. D. Ng, "Terrorism and Crime Related Weblog Social Network: Link, Content Analysis and Information Visualization," Proceedings of the IEEE International Conference on Intelligence and Security Informatics, New Brunswick, New Jersey, May 23-24, 2007.

[19] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. "Mining hidden community in heterogeneous social networks. Technical report", Computer Science Department,UIUC, UIUCDCS-R-2005-2538, May, 2005.

[20] Nasrullah M, Henrik L, D. M. Akbar Hussain**, "**Constructing Hierarchy of Non-hierarchical Terrorist Networks, Study from Theory to Implementation for Analyzing and Destabilizing Terrorist Networks", DCMMC - 2006, Washington DC,USA, September 28 – 29 2006.

[21] Newman M. E. J.; "A measure of betweenness centrality based on random walks", cond-mat/0309045,2003.

[22] Latora V, Marchiori M.; "A measure of centrality based on network efficiency", arxiv.org preprint condmat/0402050, 2004.

[23] J.J. Xu, and H. Chen, "Untangling Criminal Networks: A Case Study", ISI 2003 pp. 232-248, 2003.

[24] R. V. Hauck, H. Atabakhsh, P. Ongvasith, H. Gupta, H.Chen, Using Coplink to analyze criminal-justice data. IEEE Computer, Vol. 35, No. 3: 3037, 2002.

[25] J. M. McGloin et al., "Investigating the stability of cooffending and co-offenders among a sample of youthful offenders Criminology", 2008.

[26] A. J. Reiss, "Co-offending and criminal careers, Crime and Justice: A Review of Research", IEEE 1988.

[27] Hsinchun Chen Michael Chau,Shu-hsing Li Shalini Urs,Srinath Srinivasa G.Alan Wang(Eds.), "Intelligence and Security Informatics",Springer Verlag Berlin Heidelberg 2010.

[28] Xiaowei Ying , Xintao Wu ; Barbara, D. ,"Spectrum based fraud detection in social networks",, IEEE 27th International Conference on Data Engineering (ICDE),2011,Page 912 – 923.

[29] Abhishek Sachan, Devshri Roy, "TGPM: Terrorist Group Prediction Model for Counter Terrorism",International Journal of Computer Applications (0975 – 8887) Volume 44– No10, April 2012.

[30] Tayebi, M.A. ,Glasser U," Organized Crime Structures in Co-offending Networks ",IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC), 2011,Pages 846 – 853.