J. Basic. Appl. Sci. Res., 3(8)614-626, 2013

© 2013, TextRoad Publication

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Anomaly Detection in Cliques of Online Social Networks Using Fuzzy Node-Fuzzy Graph

Mohammad Ali Doostari[1], Ramin Zeinali[2], Hamed Lashkari[3], Mehrana Ajamzamani[4]

[1,2] Department of Computer Science, Shahed University
[3]Faculty of Information Technology, Khaje Nasir University of Technology,e-
[4]Department of Computer Science, Islamic Azad University (Shahrood Science and Research Branch)

ABSTRACT

Analysis of social networks is a useful tool for providing information about social behaviors. One of the objectives achieved in analyzing social networks is to detect malicious behaviors and protect people privacy that results in providing security in social networks. There are different approaches for anomaly detection in social networks that can be categorized into two groups, parametric and non-parametric methods. In this paper, a new hybrid approach is presented which uses both parametric and nonparametric approaches. Features like continuous data variability and the inference which results in variety of conditions can be analyzed based on fuzzy logic. Furthermore the complexity of social networks needs a structured method that can simplify the analysis process, so cliques are evaluated as base structures for anticipating malicious intent in social networks. The proposed approach is used in empirical case study to detect malicious users and a comparison is made with previous approaches.

INDEX TERMS—online social network, anomaly detection, clique, malicious behavior, fuzzy node _ fuzzy graph

I. INTRODUCTION

Online social networks are the most popular events of the last decade. According to the news, Facebook as the biggest online social network has one billion users while more than half of them are active. Online social networks can be defined as the web services which provide sharing files and interactive relationship services between people in a virtual society. Such networks aggregate a lot of information which are stored as text content and links and consist of members' private information and relationships that must be protected. By growth of social networks, a lot of problems caused by users privacy violations have been reported [1], [2] and malicious users by different intents try to abuse user information.

First of all, malicious users must be member of online social network to fulfill his sabotage intent. We can consider malicious actions as abnormal behaviors which deviate from majority of users. According to this assumption, majority of users behave normal and rational. Therefore, users with different characteristic and behavioral model should be considered as candidate for malicious intents. But the problem of this assumption is the fact that malicious members usually try to simulate a normal behavior to hide from anomaly detection systems. For example, consider a malicious user in a social network who tries to join different groups and make friendship to communicate with others and simulate a usual behavior; but because of his unknown identity for members of groups, his requests and activities don't get response from other members of the groups and his relations will get one-way relation which cause the anomaly get more severe [7]. Although the malicious users try different methods to pretend a normal users' behavior, there must be a kind of anomaly in their relationships that can be identified by a suitable anomaly detection algorithm.

A lot of researches investigate malicious behaviors in online social network to fulfill the goal of this research which is malicious behavior detection and provide security protection. Typically, anomaly detection methods focus on the connections between graph's entities and develop different analysis. Examples include spectral decompositions excellently summarized in [3], scan statistics [4] and random walks [5], [6]. These methods generally need large scale computational capabilities to work properly in very large networks. They also require specifying type of anomaly detection method explicitly. The interest of this paper is anomaly detection in large dynamic networks in context where in principle malicious users to penetrate and fulfill their sabotage intent type of anomaly should be detected. Anomaly detection needs to analyze social network data that is stored as a large and complex graph. Therefore, for analyzing such data it's better to extract special patterns and then analyze anomaly according to these patterns. In this paper, a new approach is proposed based on fuzzy node _ fuzzy graph for detecting abnormal users in a clique of users in an online social network. The fuzzy node _ fuzzy graph method is used for aggregating parametric and nonparametric approaches and cliques are the structures which are evaluated for anomaly detection. Although analysis based on fuzzy node _ fuzzy graph can be used in any level in a social network, but as online social network cliques are suitable places for malicious users to penetrate and fulfill their sabotage intent [8], and also because of the problems in evaluating whole the social network with all the users, we use cliques for fuzzy node _ fuzzy graph analysis.

The scientific contributions of this paper are: 1) we use anomaly detection algorithm that is hybrid approach from parametric and non-parametric methods based on fuzzy logic. 2) We use cliques as base structure and by parallel computing,

[*]**Corresponding Author:** Mohammad Ali Doostari, Department of computer science, shahed University,
E-Mail: doostari@shahed.ac.ir

decrease the complexity of analysis operations to get a better conclusion.

The remainder of the paper was organized as follows: In section II, previous works on this field is reviewed and related works will be categorized. In section III, relevancy of fuzzy logic and social network is investigated. In section IV, the purpose of detecting anomaly and extracting user with the most anomalies will be discussed. In section V, the result of a practical implementation for this approach on social network of Iranian students1 and the simulated data come from the VAST Challenge 20082 will be described. Finally, in section VI, conclusion and possible future works of anomaly detection in online social network will be discussed.

II. Related works

Anomaly detection has attracted lots of interests and although it seems simple but has many difficulties. In Hawkins [9] anomaly is defined as “an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.” Also, Barnett and Lewis [10], and Johnson [11] represent another similar definition that has the same semantic.

There are two classes for anomaly detection methods: parametric and non-parametric. Parametric methods assume that there is a standard distribution of observations fit the data [9], [10]. The other class includes distance-based and density-based data mining methods. These methods imply that the n-D point is too far from other points and exist in low-density area [7], [12], [13]. Typical methods include LOF [14] and LOCI [15]. These methods assign an anomaly value to each point. So the points can be sorted and the most abnormal ones can be collected. The density based anomaly detection methods are used in [16],[17], [18] for large data set analysis. Also, outputs of clustering algorithms can be used as inputs for anomaly detection algorithms [19], [20]. In [21], [22] anomaly detection is done according to centrality of graph and analysis of neighborhood graph appearance.

As mentioned before, anomaly detection in social networks uses wide range of methods. In this paper, we try to integrate previous methods and propose a hybrid solution based on the fuzzy node _ fuzzy graph idea to combine previous methods.

III. Anomaly detection in social networks base on fuzzy node _ fuzzy graph

There are different parametric and non parametric approaches for anomaly detection in social networks. However it’s obvious that malicious members try to deceive anomaly detection systems [7]. Both the parametric and nonparametric methods have specific advantages. In this paper, a hybrid approach is proposed that use both parametric and non parametric approaches and aggregate them with fuzzy node _ fuzzy graph method.

A. Adaptation of fuzzy logic with social networks

Content of online social network changes rapidly and members of each network or sub network usually have different behaviors based on their culture and thoughts. Therefore definition of anomaly in each network is different from others. For example, an international social network by different cultures and nationalities has much more variety in user behaviors in comparison with a university social network. This uncertainty in social network environment and dynamic situation is in adaption with fuzzy logic principles.

B. Fuzzy graph and structure for anomaly detection

Definition of fuzzy graph is stated as follow [23].

$$G = (V, Y), V = \{v_i\}, F = (f_{ij}) \\ 0 \le f_{ij} \le 1 \text{ , } f_{ii} = 0, 1 \le i \le n \qquad (1)$$

In (1) statement f_{ij} is the edge fuzzy value between nodes i and j. Each edge fuzzy value is equal to the distance of that node from its neighbors and is counted based on anomaly detection algorithm. Considering our security purposes (preventing malware distribution, preventing spying), all the members of a network may be examined or just some special patterns may be focused.

A clique is the pattern which is used in this paper. In a clique, density of links in neighborhood graph is equal to one and exists just in a group of friends and relatives. A clique of people is a suitable place for malicious users to interfere and extract people information and relations [8]. If a malicious user wants to inference in a group, s/he must make relationships with some of the people in that group and use chain of trust between group members to make other relationships.

In such situation, the people of a group may make relation with unknown user based on trust and in the worse scenario the malicious user can extract all the information and relations in a group.

Therefore, in the first step, we extract cliques of users in the network to check anomaly according to this structure.

C. Fuzzy non parametric anomaly assignment to edges

Non parametric anomaly in a clique can be calculated base on behavioral distance in semantic graph [7]. Behavioral distance of neighbor nodes is used as fuzzy edge value in a clique of users. Figure 1(a) shows a sample of a clique neighborhood graph. For calculating the semantic neighborhood graph of each node, the neighbors with distance equal to one would be considered. These nodes with all of their relations with each other is called ego. Figure 1(b) shows ego of a node in a clique. In the last step for creating semantic graph, independent activities of users would be calculated. Independent activities are those which are related to a solo user independent of other users in a social network. Figure 1(c) shows posting and membership of users as independent activities in our case study. Dependent activities are any interactive between users like commenting and rating in

[1]http://www.jdir.ir

[2] http://www.cs.umd.edu/hcil/VASTchallenge08

other users' posts. Figure 1-d shows dependent activities on any independent activity in our case study.

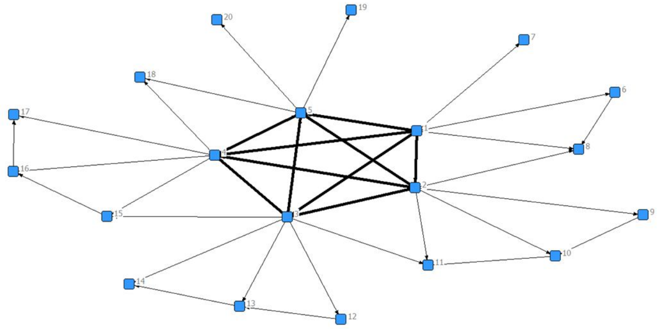

Fig. 1(a). Neighborhood graph of a five-node-clique in our case study

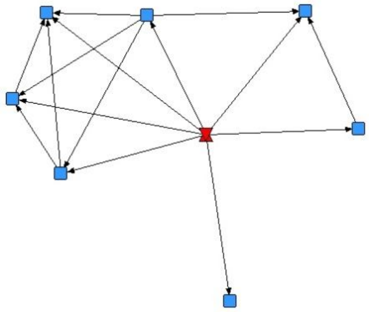

Fig. 1(b). Neighborhood graph of node u1

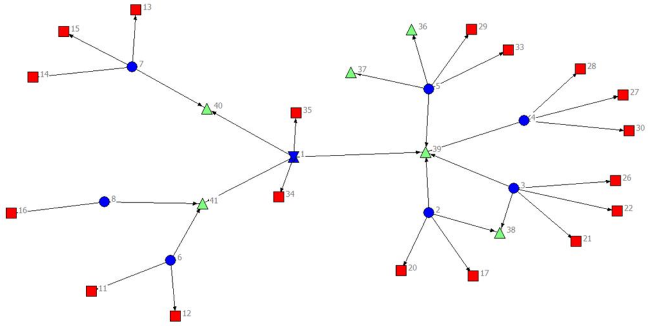

Fig. 1(c). Semantic graph for independent activities for u1 (circular nodes show network users, rectangular nodes show posts of a user and triangular nodes shows groups that a user belongs to)

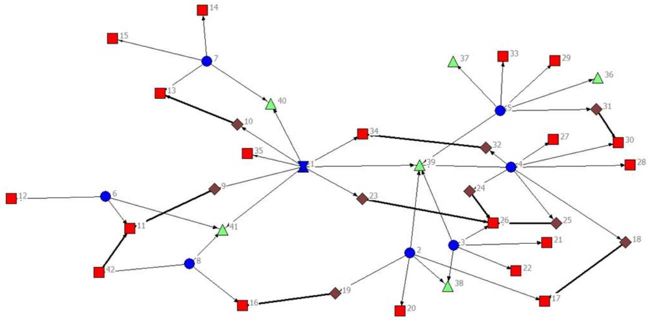

Fig. 1(d). Semantic graph for dependent activities for u1 (diamond nodes show comments or rating on a post)

It's assumed that u1, u2, u3, u4, and u5 are in a five node clique, so their behavior should be almost similar. For example if u1 and u2 are university friends, they should have common friends and close interests with behavior similarities. Membership, posting and commenting are considered as three usual paths in semantic graph to specify nonparametric anomaly value for a user in clique.

In the first step, dependency of any path to node is extracted base on membership of users in different groups and dependency probability value of each node is calculated. For example all membership paths in figure 1(c) are fourteen and the dependency probability value of each node is presented in table 1. The dependency probability value of other activities like posting and commenting also should be calculated. In table 2 and 3 the dependency probability value of posting and commenting is calculated for the proposed model.

TABLE I

DEPENDENCY PROBABILITY VALUES OF USER MEMBERSHIP

Node	U1	U2	U3	U4	U5	V1	V2	V3
N	3	2	2	1	3	1	1	1
P	21%	14%	14%	7%	21%	7%	7%	7%

TABLE II

DEPENDENCY PROBABILITY VALUES OF USER POSTING

Node	U1	U2	U3	U4	U5	V1	V2	V3
N	5	3	2	7	3	2	3	2
P	19%	11%	7%	26%	11%	7%	11%	7%

TABLE III

DEPENDENCY PROBABILITY VALUES OF USER COMMENTING

Node	U1	U2	U3	U4	U5	V1	V2	V3
N	3	1	0	4	1	0	0	1
P	30%	10%	0	40%	10%	0	0	10%

The same calculation should be done for all the other nodes of the clique. Therefore, the neighbor semantic graph and dependency probability value of each node in the sample clique should be calculated. The sum of behavioral distance differences in three paths with each neighbor in semantic graph is considered as behavioral distance with neighbors. For example in the proposed sample, the neighbor semantic graph of u1,as mentioned in table 2, imply that 19% of posting paths relate to u1 and 11% relates to u2 so the behavioral distance of u1 and u2 will be 0.08 in this path. As mentioned in tables 1, 2, and 3, behavioral distance between u1 and u2 is 0.35 which is the fuzzy value of the edge from u1 to u2.

The fuzzy distance graph can be presented in an N*N matrix where N is the number of nodes in the clique. Weight of each edge is equal to fuzzy anomaly value calculated with nonparametric method. The fuzzy nonparametric matrix and graph of the sample clique is presented in figure (2). Each row of fuzzy distance matrix represents behavioral distance of that node from its neighbors. For example, the first row represents behavioral distance of node u1 from its neighbors.

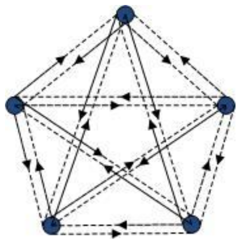

Fig. 2(a). graph of fuzzy distances

$$\begin{array}{c|ccccc|}
 & \boldsymbol{u1} & \mathbf{u2} & \mathbf{u3} & \mathbf{u4} & \mathbf{u5} \\
\mathbf{u1} & \mathbf{0} & \mathbf{0.35} & \mathbf{0.49} & \mathbf{0.31} & \mathbf{0.28} \\
\mathbf{u2} & \mathbf{0.12} & \mathbf{0} & \mathbf{0.03} & \mathbf{0.43} & \mathbf{0.06} \\
\mathbf{u3} & \mathbf{0.14} & \mathbf{0.13} & \mathbf{0} & \mathbf{0.48} & \mathbf{0.1} \\
\mathbf{u4} & \mathbf{0.29} & \mathbf{0.50} & \mathbf{0.53} & \mathbf{0} & \mathbf{0.47} \\
\mathbf{u5} & \mathbf{0.05} & \mathbf{0.18} & \mathbf{0.21} & \mathbf{0.46} & \mathbf{0}
\end{array}$$

Fig. 2(b). Matrix of fuzzy distances

D. Assign a fuzzy parametric anomaly value to nodes

Besides the fuzzy distance of each edge that is calculated based on nonparametric features, we can use parametric features of social networks and assign a value base on the distance from normal patterns .This method is proposed in [24]. We can specify some features based on the objectives of our analysis. Here the objective is to find users who behave abnormal and interfere in the clique by malicious intents. The neighbor graph (ego net) for such a user will be so crowded, because he doesn't use his real identity and also try to make relationship with all the members of the different groups to observe their activities. We define "e/n ratio" for specifying abnormal users who make a lot of relationships. In this ratio, 'e' is the total number of edges in one ego and 'n' is the number of nodes in that ego. Figure 3 shows the difference in this ratio for different users in our case study. For each user in studied dataset, the ego net is made and the ratio of present edges to number of nodes is considered as e/n ratio. The e/n ratio of each user is a point in figure 3 which shows the deviation from normal behavior that is shown by the red line. The number of nodes is Ni and the number of edges is Ei. So the Ego net density power law is as follow [25]:

$$\boldsymbol{E_i \propto N_i^{\propto},\ 1 \leq \propto \leq 2} \qquad (2)$$

A common way to identify density power law parameter is least squares method. In this method, density of the data is considered. And the logarithms of horizontal and vertical axis are calculated. In our experiments the ego net density power law exponent α equaled to 1.57.

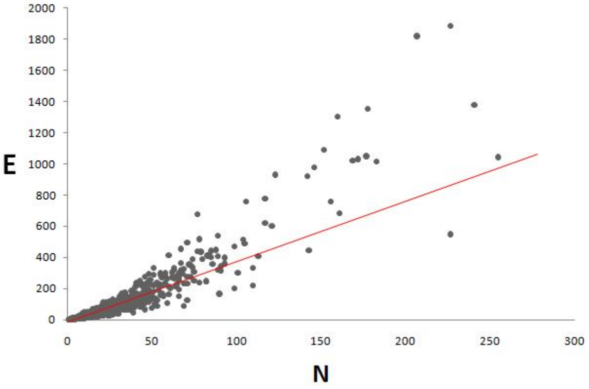

Fig. 3. e/n ratio for our case study. The line shows normal behavior and the red dots are the sample we explained in this article

As shown before, the fuzzy value of each edge is assigned base on behavioral distances. Now, another fuzzy value would be assigned to nodes to make a graph which is called fuzzy node _ fuzzy graph [26]. The value of each node is assigned in proportion of anomaly fuzzy value in parametric methods. The e/n normal ratio is considered to be usual behavior in social network and distance from this normal value will be assigned to each node [27].

Node fuzzy value is calculated base on following.

$$\boldsymbol{\mu = (\frac{e}{n} - N)\big/(maxd - N)} \\ \boldsymbol{if \frac{e}{n} < N\ then\ \mu = 0} \qquad (3)$$

In (3) N is the e/n normal value (line in figure 3) and Maxd is the maximum value of e/n ratio in n-th point. If the node has the maximum value of e/n ratio, then the numerator and denominator of (3) would be equal and anomaly in that point has its maximum value. Nodes with maximum complexity (e/n ratio) are those who have the most anomalies and should be examined. If the e/n ratio of a node is less than N, the parametric anomaly value for that node would be considered zero because low values shows people with normal behaviors and aren't useful to consider. Figure (4) shows the matrix of fuzzy node _ fuzzy graph for the sample graph and the corresponding graph. Column f_u shows the parametric anomaly value assigned to each node.

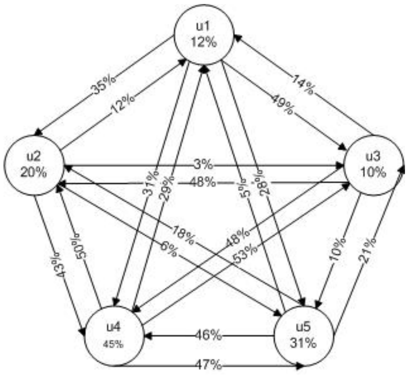

Fig. 4(a). graph of anomaly fuzzy node _ fuzzy graph for the sample data

$$\begin{array}{c|cccccc} & u1 & u2 & u3 & u4 & u5 & f_u \\ \hline u1 & 0 & 35\% & 49\% & 31\% & 28\% & 12\% \\ u2 & 12\% & 0 & 3\% & 43\% & 6\% & 20\% \\ u3 & 14\% & 13\% & 0 & 48\% & 10\% & 10\% \\ u4 & 29\% & 50\% & 53\% & 0 & 47\% & 45\% \\ u5 & 5\% & 18\% & 21\% & 46\% & 0 & 31\% \end{array}$$

Fig. 4(b). graph of anomaly fuzzy node _ fuzzy graph for the sample data

Both of fuzzy values from nodes and edges should be considered to detect anomaly in one place. Therefore instead of analyzing fuzzy node _ fuzzy graph which is so hard, we use the following conversion in (4) that is considered as one of the basic conversion methods for converting fuzzy node _ fuzzy graph to a fuzzy graph. So, parametric and nonparametric anomaly values can be aggregated in one graph or matrix [28].

$$T_\lambda(p,q) = \begin{cases} 0\,, & p\vee q < 1-\lambda \\ p\wedge q, & p\vee q \geq 1-\lambda \end{cases}, \lambda\epsilon[0,1] \qquad (4)$$

In (4), q is edge fuzzy value and p is node fuzzy value, λ is a value between 0 and 1 and can have different values. Best value for λ for creating resultant matrix of fuzzy node _ fuzzy graph can use Measure the Quality of Communication Function and Distance function to specify values. As the value of λ change, communication function should increase and distance function should decrease [28]. The value of λ in our case study is equal to 0.8. The graph that is created as the consequence of merging node and edge values of the graph is called resultant graph.

IV. ABNORMAL NODE(S) EXTRACTION

Nodes are assigned by anomaly values and node or nodes with the most anomalies are considered to be candidate for sabotage. In this process and in the first step, clustering is used to help specifying sabotage candidates.

Clustering is a method to categorize objects base on their similarities. We can use this concept to separate users with most anomalies with each other in one cluster. Clustering can be operated on symmetric matrix; therefore the resultant graph from previous section will be converted to symmetric matrix with the following formula.

$$\begin{array}{l} \boldsymbol{S = (S_{ij}), \frac{2}{S_{ij}} = \frac{1}{f_{ij}} + \frac{1}{f_{ji}}} \\ \boldsymbol{where\ S_{ij} = 0\ if\ f_{ij}.f_{ji} = 0} \end{array} \qquad (5)$$

If the relation is one-way relation, that relation will not be considered in symmetric matrix. Symmetric matrix for anomaly fuzzy node _ fuzzy graph of our case study is shown in figure (5). The value of cut value can change to make different clustering trees for the anomaly fuzzy graph. The analysis of clustering trees result in finding nodes with the most anomaly (figure 6). The cut value for clustering shouldn't be so high to make separate islands and shouldn't be so low to put all the nodes in one cluster, though the tradeoff between two values must be chosen [20].

$$\begin{array}{c|ccccc} & u1 & u2 & u3 & u4 & u5 \\ \hline u1 & 0 & 0.12 & 0 & 0.16 & 0.09 \\ u2 & 0.12 & 0 & 0.0 & 0.33 & 0.12 \\ u3 & 0 & 0 & 0 & 0.28 & 0 \\ u4 & 0.16 & 0.33 & 0.28 & 0 & 0.38 \\ u5 & 0.09 & 0.12 & 0 & 0.38 & 0 \end{array}$$

Fig. 5. Symmetric matrix for anomaly fuzzy node _ fuzzy graph

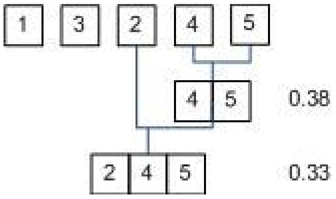

Fig. 6. Clustering with values of symmetric anomaly fuzzy node _ fuzzy graph

In the last step the objective is to find the node with maximum anomaly. We cluster nodes with optimized cut value and use optimized cut value for resultant matrix to draw the result graph. The relations with values less than cut value can be omitted because they almost don't have much effect on the result of detecting abnormal nodes (figure 7).

Anomaly value for each node is equal to sum of input and output edges in the resultant anomaly fuzzy graph. This value for each node is based on parametric and non parametric anomaly distances from its neighbors. Besides the anomaly value, the number of effective edges (edges with values more than optimized cut value) in input and output can be considered for detecting anomaly. It means that the candidate node for abnormal behavior should have maximum number of edges in the resultant anomaly fuzzy graph [22]. Indeed, centrality in resultant fuzzy graph is an important factor in specifying abnormal candidate nodes.

In our sample, node u4 has centrality in resultant anomaly graph, because it has the most edges with its neighbors. It also has the highest value in resultant anomaly fuzzy graph so it can be a candidate for sabotage.

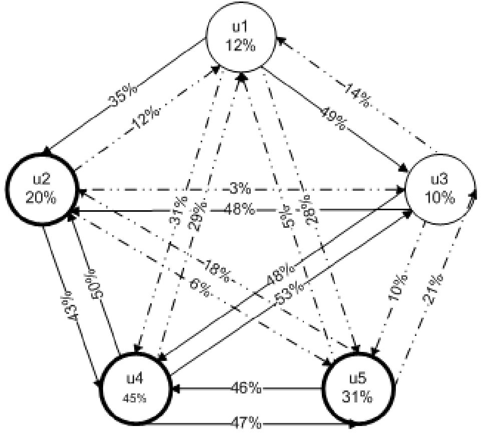

Fig. 7. Resultant fuzzy graph-heavy weighted edges (edges with values less than cut value won't be considered)

V. EMPIRICAL RESULTS

Evaluation of proposed approach was examined in two datasets which are collected from two different sources. The first dataset was from Iranian student social network and the second was a simulated dataset which was examined in other researches. The first implementation showed that 47% of detected abnormal users were malicious which shows the hopefulness of the approach. In second implementation, our approach was compared with the concluded results from similar researches and showed that the approach is comparatively better in this dataset. In the following parts, implementations of the approach on these two dataset are discussed.

A. Empirical results on real data

The proposed method is evaluated on a dataset from social network of Iranian students which in addition to users' neighborhood graph, has user relationships and behaviors like posting and commenting. Therefore we could analyze the datasets base on users' behavior.

Concern about using this method in larger social networks might be addressed but the result of this research on the Iranian students social network can be generalized to larger social networks because the result of small social network can be generalized to larger networks [29]. So our research can be generalized to larger social networks.

UCINET [30] is a tool used by social network analysts to visualize and understand social networks. This tool is used for some part of calculations implemented in our study.

In this research, we concentrated on cliques as basic structure for evaluating user activities. Therefore we examined the dataset and extract six hundred cliques with five users in each clique. The number of users in this sample of social network was 2374 and some users were in more than one clique. In the evaluated case study, number of users in each clique wasn't so much and we evaluated five member cliques to fulfill our purpose. The proposed anomaly detection algorithm runs on all the cliques with different cut values and λ values for each clique. Among these six hundred cliques that we assessed, 509 cliques had the features we proposed on section 4 and the number of candidate users with different behavior was 420 because some users specified as candidate node in more than one clique.

According to the proposed method assumption, abnormal users in cliques can be malicious users. Therefore validity of user identity for candidate users should be examined. It's assumed that 5% of users in the social network are malicious users [31]. So we find 118 users with high anomaly values as final candidates. Calculated anomaly value for 2374 members of the evaluated network is shown in a chart in figure 8.

Fig. 8. Anomaly values for all users in evaluated social network

The validity process involves other users in the social network as trustworthy identities and asks them about candidate user identities. The result was 55 correct detection as none of the users in the clique knows the candidate users identity. Therefore s/he should be fake identity. In 34 cliques, there was at least one user who knew the identity of the candidate user and in 29 cliques the result was unknown because there was at least one person in clique who didn’t answer the question. So we couldn’t conclude that the user is malicious. The overall result of our study is presented in figure (9).

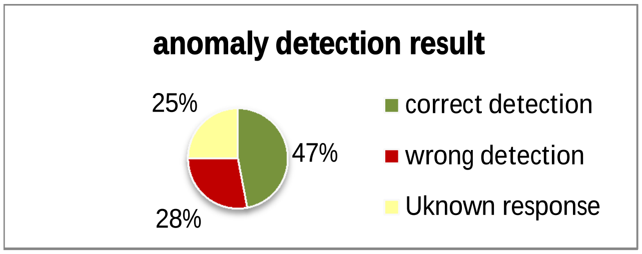

Fig. 9. Empirical result for the research

B. empirical results on the simulated data

In the second step, we compare our approach with other approaches which are tested on the simulated data come from the vast challenge 2008. As used in other approaches, we consider the simulated cell phone data from the Mini Challenge focused in the area of social network analysis. The cell phone call records cover a fictional ten-day period on an island, narrowed to 400 unique cell phones during this period. As well as the time of each phone call and details of who phoned whom, an identifier of the cell tower from which the call originated is also given. The records should provide critical information about an important social network structure. From the results of award winning published work on this challenge [32], work which used a combination of PageRank [33] and visual analytic methods. This dataset is also used in [34] base on Bayesian methods for anomaly detection. In this approach, the analysis is based on all the links between two nodes which are assessed in different time slices. In first step, the relationship between two nodes is assessed by Bayesian probability model and nodes with greater differences in relationship values from threshold value in two sequential time slices would be selected for the second step which uses standard network tools such as spectral clustering for anomaly detection.

Heard et al in [34] tried to enhance the approach represented in [32] and propose an approach which is dynamic and real time in compare to previous approach used in [32] that is static, but yet the previous approach has found 11 abnormal nodes whereas the latter approach detected 8 abnormal nodes.

The result of these approaches is drawn in a spectral cluster plot using two components of the symmetric laplacian of the historical adjacency matrix [3] and is given in figure (10).

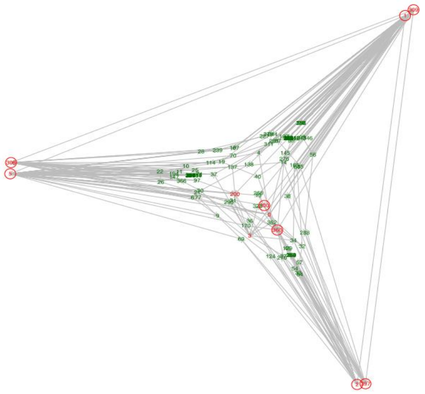

Fig. 10. circular nodes (309,1,306,5,2,397,360,300) are detected in Heard et al method and red nodes(309,1,306,5,2,397,360,300,0,200,3) are detected in ye et al method[3]

Heard et al imply that the three undetected nodes can be detected by analyzing call events, for example node 200 has the most relationship frequency and can be identified as the network leader. Node 3 is one of the six nodes in relationship with network leader and node 0 contacts all nodes that are in relationship with node 200. But such analysis is not possible in large scale networks, Therefore a detection algorithm must be designed to find anomalies in a step-by-step process.

In this part we compare our method with these two methods and tried our method on simulated data comes from the VAST Challenge 2008. First of all the e/n ratio for ego of each node is calculated and its' distance from normal value which is taken from (3) assigns to node fuzzy anomaly value. Our method tries to detect nodes with a lot of relationships in different groups that has a crowded neighborhood graph. The e/n ratio and normal behavior graph for nodes of vast challenge 2008 is shown in figure (11). As it's shown all the detected nodes in [34] and [32] are above the normal behavior value and has crowded egos.

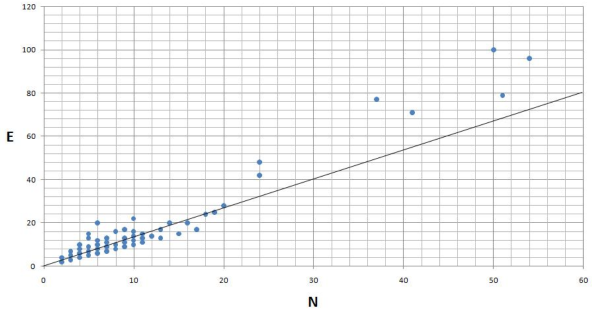

Fig. 11. e/n ratio for nodes in vast challenge 2008 dataset and normal behavior line

In the next step, cliques in cell phone social network are extracted. Dependent and independent behaviors in semantic neighborhood graph are calculated based on proposed method in section 3.3. The dependent value of selected semantic path to each node is calculated in semantic graph of that node. Behavioral distance of each node from its neighbors in clique is assigned to the connecting edge as fuzzy anomaly value. Then the fuzzy anomaly values of nodes and edges are converted to one value on each edge that is calculated based on formula 3. For example node 200 is part of three cliques. One of these cliques is a four node clique that all of its neighbors with distance equal to one are shown in figure (12).

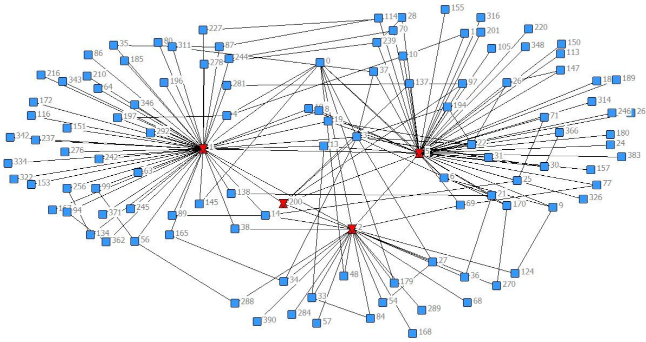

Fig. 12. four node clique including 200,5,2,1 and its neighbors with distance equal to one

Semantic neighborhood graph of node 200 is shown in figure (13). The connection path from tower cells is the only path in ego semantic graph. As figured in 13(a) dependency of this path to node 200 in neighborhood semantic graph of node 200 is equal to 0.28. Neighborhood semantic graph of other three nodes is sown in figure (13).

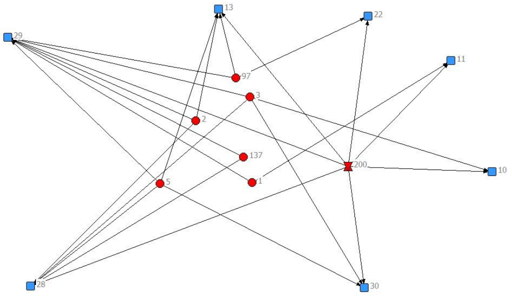

Fig. 13 (a). Semantic graph for node 200

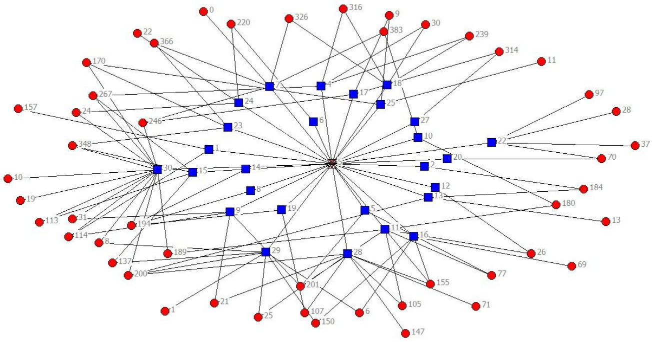

Fig. 13 (b) Semantic graph for node 1

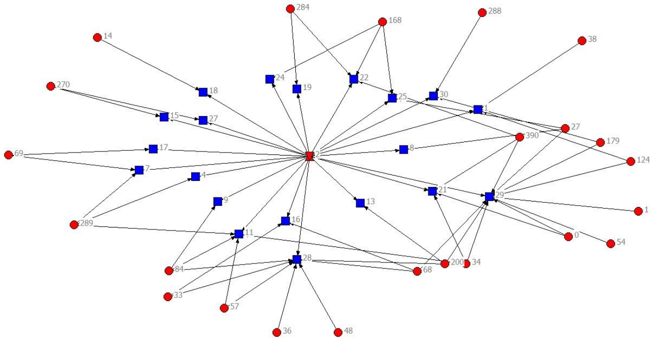

Fig. 13 (c) Semantic graph for node 2

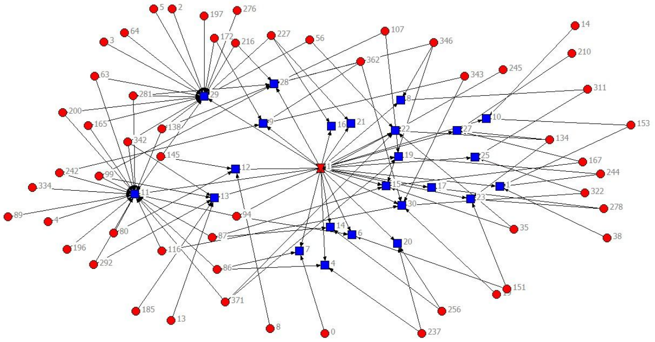

Fig. 13 (d).Semantic graph for node 5

Behavioral dependency differences are considered as fuzzy anomaly value (non parametric anomaly) on edges is shown in figure 14(a). Fuzzy anomaly value of each node (parametric anomaly) which is the deviation from normal behavior is also displayed inside each node. For example dependency path of linking to cell tower or linking from cell tower for node 1 in its semantic graph is 0.19 and this value for node 2 is 0.29, so fuzzy edge between these two nodes is 0.1. Then the fuzzy node-fuzzy graph will be converted to fuzzy graph by conversion formula 3 (figure 14 (b)).

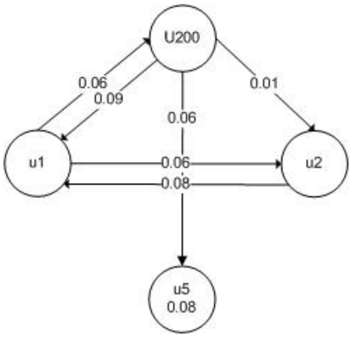

Fig. 14 (a). Anomaly fuzzy node - fuzzy graph for the sample clique

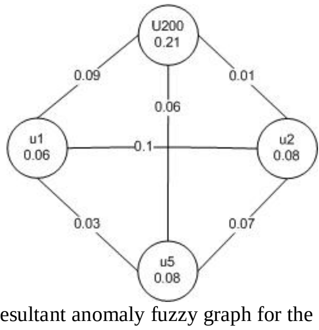

Fig. 14 (b). Resultant anomaly fuzzy graph for the sample clique

As shown in figure 14 (b), node 200 has the most anomaly value and anomaly edges (centrality of anomaly fuzzy graph) in this clique. So node 200 is the best candidate for sabotage. Node 200 is involved in two other cliques and also has the most abnormality in these two cliques.

In proposed method, 5% of nodes with the most anomalies will be considered as sabotage candidates. So 16 nodes were detected with maximum abnormal behavior and are shown in figure (15). All the 11 nodes that were detected in [32] are also detected by our proposed algorithm. The last step in our approach was identity approval which is not practical for this dataset.

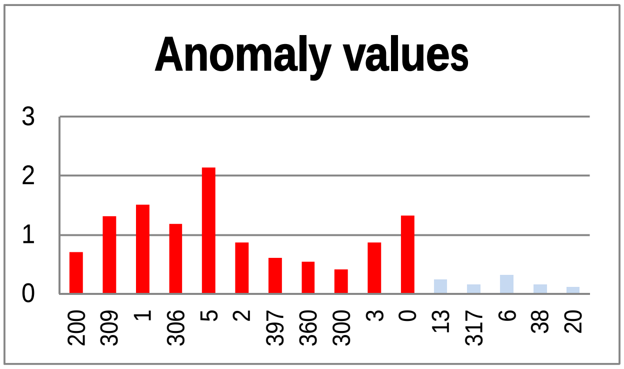

Fig. 15. Abnormal detected nodes that 11 nodes have the most anomaly values in simulated data come from the vast challenge 2008

VI. Result and Future works

As discussed before, different kinds of privacy violation was reported in online social networks. The purpose of this research was to prevent malicious users from interfering in groups and misuse the information. We use anomaly detection algorithm that is hybrid approach from parametric and non-parametric methods based on fuzzy logic. Because social networks include lots of users and information, operating any analytical theory on such a big information repository takes a lot of cost and time. Therefore in this research, we use cliques as base structure and by parallel computing, decrease the complexity of analysis operations to get a better conclusion. Semantic neighborhood graph of each user is used for nonparametric anomaly and e/n ratio in ego net is used for parametric anomaly. Then the anomaly value of each user is calculated based on proposed approach to specify most abnormal users.

In our case study in Iranian student social network, 55 (47%) of the candidate abnormal users were validated by trusty users. They were validated by asking other normal users that were assumed to be trusty. In addition our empirical result showed that implementation of our approach on the simulated data come from the VAST Challenge 2008, detected abnormal nodes which were detected in previous approaches but this approach was more successful as mentioned. The proposed approach has integrated different methods for anomaly detection so has some advantages on previous ones. Also social network graph is sparse and the computational overhead is not a limitation in this approach [35]. The dynamicity of our approach makes it practical for real time anomaly detections.

The other structures like star, near star and heavy vicinity or etc can be used to analyze users' behaviors for goal-based anomaly detection. So by defining other purposes, any other structure can be used for future work. Also in future research, other methods besides questioning and trust to normal users can be used to assure that users with abnormal activity are malicious or not.

References

[1] R. Gross and A. Acquisiti, "Information relevation and privacy in online social networks," In Workshop on Privacy in the Electronic Society, 2005.

[2] T. N. Jagatic, N. A. Johnson, M. Jakobsson and F. Menczer, "Social phishing," Communications of the ACM, 2007.

[3] U. Luxburg, "A tutorial on spectral clustering." Statistics and Computing, 2007, pp.395–416

[4] C. E. Priebe, J. M. Conroy, D. J. Marchette, Y. Park, "Scan statistics on Enron graphs", Computational & Mathematical Organization Theory, 2005,pp 229–247.

[5] J.-Y. Pan, H.-J. Yang, C. Faloutsos, P. Duygulu, "Automatic multimedia cross-modal correlation discovery", Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2004,pp 653–658

[6] H. Tong, C. Faloutsos, J. Y. Pan, "Fast random walk with restart and its applications", Sixth IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, 2006, pp 613–622

[7] S. Lin, and H. Chalupsky, "Discovering and Explaining Abnormal Nodes in Semantic Graphs," IEEE Transactions On Knowledge And Data Engineering, 2008, pp1039-1052

[8] S. Whitsitt, A. Gopalan, S.Cho, J. Sprinkle, S. Ramasubramanian, L. Suantak and J. Rozenblit, "On the Extraction and Analysis of a Social Network with Partial Organizational Observation," Engineering of Computer Based Systems (ECBS), IEEE, 2012.

[9] D. Hawkins,: "Identification of Outliers". Taylor & Francis, 1980

[10] V. Barnett, T. Lewis.: “Outliers in Statistical Data. John Wiley and Sons” , 1994

[11] R. A. Johnson , D. W. Wichern: “Applied Multivariate Statistical Analysis”. Prentice Hall, 1998

[12] K. Bhaduri, B. L. Matthews, C. R. Giannella,: “Algorithms for speeding up distance-based outlier detection”. International conference on Knowledge discovery and data mining, 2011

[13] X. Wang, I. Davidson: “Discovering contexts and contextual outliers using random walks in graphs”, international conference on data mining ,2009

[14] M.B. Markus , H. Kriegel, R. T. Ng, J. Sander: LOF: “Identifying density-based local outliers”. In conference on Management of Data, 2000, pp.93–104

[15] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, C. Faloutsos: “LOCI: Fast outlier detection using the local correlation integral”. In conference on Engineering , 2003

[16] C. C. Aggarwal, P. S. Yu: “Outlier detection for high dimensional data. In ACM international conference on Management of data”, 2001, pp37 – 46

[17] Z. Bakar, R. Mohemad, A. Ahmad, M. M. Deris: “A Comparative Study for Outlier detection Techniques in Data Mining”, IEEE Conference on Cybernetics and Intelligent Systems, 2006

[18] C. Chaudhary, A. S. Szalay, A. W. Moore: “Very fast outlier detection in large multidimensional data sets”. In DMKD , 2002

[19] S. Basu, M. Bilenko, R. J. Mooney: “A probabilistic framework for semi-supervised clustering”. In International Conference on Knowledge Discovery and Data Mining, 2004

[20] S. Guha, R. Rastogi, K. Shim: “An efficient clustering algorithm for large databases” .In conference on Management of data, 1998

[21] P. Sun, S. Chawla: “On local spatial outliers”, In IEEE International Conference on Data Mining , 2004

[22] S. Wasserman, K. Faust: “Social Network Analysis: Methods & Applications”. Cambridge University Press, 1994

[23] A. Rosenfeld, L.A. Zadeh, K. S. Fu, M. Shimura, “Fuzzy Sets and Their Applications”, Academic Press, 1975, pp 77-95

[24] McGlohon, M., Akoglu, L., Faloutsos, C.: “Weighted graphs and disconnected components Patterns and a model., In ACM international conference on Knowledge discovery and data mining , 2008

[25] M. McGlohon, L. Akoglu, and C. Faloutsos. “Weighted graphs and disconnected components”: Patterns and a model. In ACM SIG-KDD, Las Vegas, 2008.

[26] H. Uesu, S. Kimiaki, H. Yamashita: “Analysis of Fuzzy Node Fuzzy Graph and its Application”, In Conference on Soft Computing and Human Science, 2007

[27] V. Chandola, A. Banerjee, V. Kumar: “Anomaly Detection: A Survey”. ACM Computing Surveys, 2009

[28] H. Uesu, H. Yamashita, T. Takizawa, M. Yanai: “Optimal Fuzzy Graph Based on Fuzzy Node Fuzzy Graph Analysis”. In International Conference on Soft Computing and Intelligent Systems, 2006

[29] G. Yan, G. Chen, S. Eidenbenz, N. Li: “Malware Propagation in online Social Networks”. In ASIACCS ACM , 2011

[30] B. MacEvoy, L. Freeman: “Ucinet: A Microcomputer Package for Network Analysis”, Mathematical Social Science Group, 1987

[31] J. Gaoy, F. Liang, W. Fan, C. Wangy, Y. Sun, J. Han: “On Community Outliers and their Efficient Detection in Information Networks”. In ACM international conference on Knowledge discovery and data mining , 2010

[32] Q. Ye, T. Zhu, D. Hu, B. Wu, N. Du, B. Wang, “Cell phone mini challenge award: Social network accuracy—exploring temporal communication in mobile call graphs”. In IEEE International Symposium on Visual Analytics Science and Technology, 2008, pp 207–208.

[33] S. Brin, L. Page: “The anatomy of a large-scale hypertextual web search engine”. Computer Networks and ISDN Systems, 1998, pp107–117.

[34] N. A. Heard, D. J. Weston, K. Platanioti, D. J. Hand: “Bayesian anomaly detection methods for social networks.” Institute of Mathematical Statistics in Annals of Applied Statistics, 2010, pp 645–662.

[35] C. Faloutsos, K. S. McCurley, A. Tomkins: “Connection subgraphs in social networks”. In Proceeding of SIAM International Conference on Data Mining, 2004