International Journal of Engineering & Technology, 7 (4.5) (2018) 518-521

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Centrality measure based approach for detection of malicious nodes in twitter social network

Krishna Das [1] *, Smriti Kumar Sinha [1]

[1] Department of Computer Science & Engineering, Tezpur University, Assam, India
**Corresponding author E-mail: krishnadas@tezu.ernet.in*

Abstract

In this short paper, network structural measure called centrality measure based mathematical approach is used for detection of malicious nodes in twitter social network. One of the objectives in analysing social networks is to detect malicious nodes which show anomaly behaviours in social networks. There are different approaches for anomaly detection in social networks such as opinion mining methods, behavioural methods, network structural approach etc. Centrality measure, a graph theoretical method related to social network structure, can be used to categorize a node either as popular and influential or as non-influential and anomalous node. Using this approach, we have analyzed twitter social network to remove anomalous nodes from the nodes-edges twitter data set. Thus removal of these kinds of nodes which are not important for information diffusion in the social network, makes the social network clean & speedy in fast information propagation.

Keywords: *Malicious Nodes; Centrality Measures; Clustering Coefficient; Anomaly Behavior; Information Diffusion*

1. Introduction

Social network is an interdependent relationship among the users which connects each other nodes via various forms of connections such as friendship, follower, followee etc. Some nodes are highly influential, some do not have any role in the network to disseminate information. Such nodes unnecessarily creates long network path and delay in the information diffusion process. These nodes are considered as malicious or fraud nodes in the network. This is only a partial work that we are investigating to identify such malicious nodes in twitter social networks for smooth information diffusion. Information diffusion is a process in social network by which a piece of information moves ahead through other nodes and reaches to every other node via interactions among the nodes in the network. Identifying abnormal users and events in social networks [13],[12] such as days with an abnormally high number of messages or network with less number of computations but extensive slow in information diffusion process, are some of the application of anomaly node detection in social network. The aim of this paper is to show how centrality measure based graph theoretical approach can be used to identify malicious nodes in social networks. For the purpose of various centrality value computation of nodes, the network is assumed as undirected and un-weighted for our analysis.

2. Related work

Different anomaly detection approaches categorized under behavior based or structure based scenario have been proposed in literature. We have considered here only structure based anomaly detection.
Approach in this research paper. Bimal Viswanath et. al. [3] in their research paper proposed unsupervised anomaly detection techniques over user behavior to distinguish potentially bad behavior from nor- mal behavior of social network users. Landherr A., Friedl B. and Heidemann J. [11] have shown the application of Centrality mea- sures in a social network to find out popular nodes in the network. David Savage ET. al. [7] in their survey paper have mentioned network feature based classification methods for detection of anomaly nodes in social network. Ravneet Kaur, Mankirat Kaur and Sarbjeet Singh
[15] in their research work used curve fitting anomaly detection method to identify anomalous nodes in social network.Stephen Ran- shous et. al [17] in their survey paper have provided comprehensive overview of anomaly detection in dynamic networks which includes graph based methods for anomaly detection. Nicholas A. Heard et. al. [18] have used Bayesian model to find out the nodes in a social network exhibiting malicious behavior.

3. Problem definition & motivation

Anomaly can be defined as deviation from some expected and necessary behaviour. In literature sometimes a general definition is "patterns in data that do not conform to a well defined notion of normal behaviour" [6]. Another recent review defines anomalies as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" [1] [9].
Anomalies in online social networks can signify irregular, and often illegal behaviour . Malicious nodes may undermine effectiveness by disrupting or spamming the network. Network analysts must cope with complex structures, disconnected components, well-connected clusters, and multiple attributes for nodes and links. They often deal with networks of dynamically evolving structures, where links act as pathways for volatile information or commodity flows. Hence it is necessary to find out important as well as non-functional unwanted malicious nodes for better information diffusion and strong cohesive structures in the social network. Anomalies in so-

cial networks are often representative of illegal and unwanted behaviour. So these must be removed to have a fast information diffusion social network.

4. Methods for anomaly detection

Centrality measures helps in detection of either the most important node or the least important in the social network. Using this measurement, we are to find out centrality measure values for every nodes in the network to detect whose centrality measure is very poor, which will imply that a node having very less centrality measure values are not essential nodes in the network for carrying informa- tion. Hence these nodes are regarded as malicious nodes present in the network. There are many types of centrality measures that are used to find out malicious nodes. We have computed only three prominent centralities for detection of malicious nodes in twitter social network data. We first formally defined these centrality mea- sures and illustrate them by means of some examples. Then we analyzed them with respect to the previously formulated properties in a twitter social network. A sample social network graph having 10 number of nodes v1, v2 upto v10 is shown in fig 1. Computation of various centrality measures for every node in fig 1 is listed in table 1. Accordingly ranks of every nodes according to their position in the sample network is also shown in the table along with their centrality values.

Degree Centrality (DC): It represents the simplest centrality measure and determines the number of direct contacts as an indicator of the quality of a node's interconnectedness [14]. Degree Centrality is represented using following formula:

$$DC(v) = \sum_{i=1}^{n} a_{iv}$$

A general overview to understand the degree centrality is the Kth-path centrality which counts the number of paths less than or equal to K that originate from a node.

Closeness Centrality (CC): This centrality measure is based on the idea that nodes with a short distance to other nodes can spread information very productively through the network [2]. In order to calculate the closeness centrality of a node v, the distances between the node v and all other nodes of the network are summed up [16]. By using the reciprocal value we achieve that the CC value increases when the distance to another node is reduced, i.e. when the integration into the network is improved. Closeness Centrality is represented using following formula:

$$CC(v) = \frac{1}{\sum_{i=1} d_G(v,i)}$$

This is a diameter based measure which counts the total lenght of the

Walks in a graph. It computes the average of the shortest distances to all other nodes from the source node whose closenes centrality is to be counted.

Betweenness Centrality (BC): Betweenness centrality measure of a network node is considered to be well connected if it is located on as many of the shortest paths as possible between pairs of other nodes. The underlying assumption of this centrality measure is that the interaction between two non-directly connected nodes u and v depends on the nodes between u and v. According to Freeman [8] the BC of a node v is therefore calculated as

$$BC_j(v) = \sum_{i=1, i \neq v}^{n} \sum_{j=1, j \neq v}^{N} \frac{g_{ij}(v)}{G_{ij}}$$

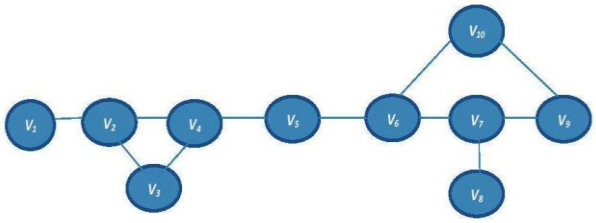

Fig. 1: A Network Example to Explain Centralities.

to the own centrality to a greater extent than a relationship to a less well interconnected node. For a node v, the EC is therefore defined as [4] a function of its neighbours in terms eigen vectors.

Katz's Centrality Measure (KC): According to Katz not only the number of direct connections but also the further interconnectedness of actors plays an important role for the overall interconnectedness in a social network [2]. Therefore, Katz includes all paths of arbitrary length from the considered node v to the other nodes of the network in the calculation of this centrality measure.

5. Experimental results

We have collected twitter data set from UCI net data repository. It consists of near about 2, 00,000 nodes & 10,00,000 edges. We have normalized the data set discarding those nodes having more than 100 edges & less than 10 edges to maintain the bias free data set in our study. Finally it consists of number of nodes: 26,491 number of edges: 32,688 and with no missing values.

A node is a entity that exists in a network graph. Each Twitter user is a node in this network. Each node can have a set of attributes and related network metrics that measure their position within the larger network. We have applied NodeXL simulator for the exper- iments. It combines network metrics it calculates about each node in the network with data extracted from Twitter that describe it. For example, the number of people the user follows, the number of users following that user, the number of Tweets the user has created to date, the number of tweets that person made etc. We have calculated the following network metric as described below:

Unique Edges 32688 Edges with Duplicates 0 Total Edges 32688 Self-Loops 0

Maximum Nodes in a Connected Component 26491 Maximum Edges in a Connected Component 32688 Maximum Geodesic Distance (Diameter) 6

Average Geodesic Distance 3.734586 Graph Density 9.31618E-05 Modularity 0.738719

Self-loop in graph theory represent an edge, which connects a vertex to itself. In our data set, nodes having self-loop has been discarded to produce a error free network analysis result. Again, geodesic distance between two nodes in a network is defined as a path with the minimum number of edges. Average geodesic distance in our graph is 3.73. Graph density is the sum of all the ties divided by the number of all possible ties. Graph density provides the idea about the speed of information diffusion among the nodes. Graph density of our data set is found 9.31. Network modularity measures the strength of divisions or modules of a network nodes into clusters. High modularity network has dense connections among the nodes

Within the cluster but very sparse connections among the nodes in Different clusters. Accordinh to M.E.J. Newman, social network modularity is given This is the network flow related measure which gives the idea about

How much a given node lies in the shortest paths of other nodes.

Eigenvector Centrality (EC): Eigenvector centrality is based on the idea that a relation-ship to a more interconnected node contributes

By-Q = (edges within cluster) - (edges within clusters in a random graph with similar node degrees) Lets us consider, ki =degree of node i M

$= \sum ki = 2 \mid E$

Table 1: Results of the Example Network Figure 1

Degree centrality results										
Nodes	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}
DC	1	3	2	3	2	3	3	1	2	2
Rank	9	1	5	1	5	1	1	9	5	5
Closeness centrality										
CC	1/34	1/26	1/27	1/21	1/19	1/19	1/23	1/31	1/29	1/25
Rank	10	6	7	3	1	1	4	9	8	5
Betweenness centrality										
BC	0	8	0	18	20	21	11	0	1	6
Rank	8	5	8	3	2	1	4	8	7	6

Accordinh to Adjacency matrix, $A(i, j) = 1, if (i, j) \in E$, Otherwise 0
Then the modularity is given by,

$$Q = \sum A(i, j)(k_i k_j)/M \quad i, j$$

Belongs to the same group.
Various centrality measures are computed using NodeXL Version 1.0.1.245 as follows from the experiments:
Degree Centrality represents the number of connections a particular node has. The degree centrality of a vertex, for a given graph G = (V, E) with V nodes and E edges, is defined as DC(v) = deg(v) From our analysis we have found out following degree centrality measure value of the network:

Minimum Degree= 1

Maximum Degree= 8571

Average Degree= 2.468

Median Degree= 1.000

So, a node having degree less than the below average degree centrality value may be considered as malicious nodes which can be discarded from the network.
Betweenness Centrality represents how many short paths a particu-Lar node makes. It makes an node powerful to control the flow. We have obtained the following betweenness centrality measures: Maximum Betweenness Centrality=186548573.796
Average Betweenness Centrality= 36221.464
So, a node having Betweenness Centrality less than the below average value or sharp 0 value may be considered as malicious nodes which can be discarded from the network.
Closeness Centrality represents the mean of the geodesic distances between some particular node and all other nodes connected with it. In our network, we have found out the closeness centrality values as given below:

Minimum Closeness Centrality =0.003 Maximum Closeness Centrality =0.005 Average Closeness Centrality= 0.004 Median Closeness Centrality= 0.004

So, a node having Closeness Centrality less than the below average value or sharp 0 value may be considered as malicious nodes which can be discarded from the network.
Clustering coefficient: The cluster coefficient represents the number of existing connections that a particular node can have from all possible connections in its neighborhood. This measure describes the relative strength of connectivity. The clustering coefficient of a node v is 1 if every neighbor connected to v is also connected to every other node within the neighbor-hood of v, and 0 if no node that is connected to v connects to any other node that is connected to A. Hence those clusters or connected component in the network whose clustering coefficients is 0 is discarded from the network as these lusters are irrelevant in the social network in case of information diffusion task.
Network-Level Analysis: From our analysis we have find out following for the network that we have studied. From our data analysis, total number of connected components of the nodes is 20.

- G1, G2, G3 are the users of elite groups who are well connected among themselves & take parts in all the tweeter network activities such as follows, replies, mentions and tweets.
- G4, G5, G6,G7,G8,G9,G10 & G11 are the groups of users having less centrality measures and take part in some of the activities of twitter networks.
- G12 to G20 are the groups of less number of members and they either take part in at least any one of activities or some time don't play any role in the network. Hence G12 to G20 consisting of a small fraction of nodes, can be considered as anomaly group members which do not have any big role in the network as their centrality values are very very less.

6. Future works & conclusions

There are various types of anomaly detection techniques in social networks for different purposes. For example to find out fraud & malicious users in social networks, one may use text processing methods to find out spammers and harmful sensitive messages using Natural Language Processing Techniques. For a stable, fast information diffusion and secure social networks, network structure based methods like centrality measure methods are more fruitful. Different social network anomaly detection approaches can be usefully categorized based on characterization of anomalies as being static or dynamic and labelled or unlabelled. Depending on this characterization, different features of the network may be examined. In our study, we have considered only static unlabelled, undirected & unweighted social network nodes for finding out malicious nodes. Both supervised & unsupervised methods must be simultaneously used in dynamic environment for detection of all kind of anomaly such as behavioural anomaly and network structural anomaly. Another promising task which is in line is to compute information diffusion of certain nodes situated in various network position to provide a information diffusion capacity of those nodes for a comparative study of malicious unwanted nodes and node positions with high priority in the network.
In this way anomaly nodes in social network can be identified to make the network faster information diffusion model. Our model of application of centrality measure for detection of malicious nodes in social network can address only network structural anomaly. So behavioural methods & structural methods both must be used as hybrid approach for detection of all kinds of malicious nodes in social networks. We are working towards this to include all possible measure in hybrid approach to detect malicious nodes in various social network platform to design an efficient anomaly method as our future work. Moreover, for practical implementation for detection of malicious user, the said approach must be applied in various social network data to validate the results. These are future works under- taken by us in our further continous studies to design an efficient cross platform social network anomaly detection method to find out less important users in the network.

References

[1] Barnett V. and Lewis T.,"Outliers in Statistical Data", third ed. John Wiley & Sons, 1994, Chichester, UK.
[2] Beauchamp MA.,"An improved index of centrality", Behav Sci, 10(2), 161–163, 1965.
[3] Bimal Viswanath, M. Ahmad Bashir, Mark Crovella and Saikat Guha, "Towards Detecting Anomalous User Behavior in Online Social Networks", 23rd USENIX Security Symposium, August 20–22, 2014, San Diego, CA.
[4] Bonacich P. and Lloyd P.,"Eigenvector-like measures of centrality for asymmetric relations", textitSocial Networks, 23(3), 191–201, 2001.
[5] Chandola V., Banerjee A. and Kumar V.,"Anomaly detection for discrete sequences: a survey", IEEE Trans. Knowl. Data Eng., 24 (5), 823–839, 2012.
[6] Chandola V., Banerjee A. and Kumar V.,"Anomaly detection: A survey", ACM Computing Survey, (CSUR) 41 (3), 15, 2009.

[7] David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou and Qingmai Wang, "Anomaly detection in online social networks", Social Networks, 39, 62–70, 2014.
[8] Freeman L. C.,"Centrality in social networks: conceptual clarification", Social Networks, 1(3), 215–239, 1979.
[9] Hodge V.J. and Austin J.,"A survey of outlier detection methodologies", Artif. Intell. Rev., 22(2), 85 –126, 2004.
[10] Katz L.,"A new status index derived from sociometric analysis", Psychometrika, 18(1), 39 –43, 1953.
[11] Landherr A., Friedl B. and Heidemann J., (2010),"A Critical Review of Centrality Measures in Social Networks", Business & Information Systems Engineering, 2(6), 371–385, 2010.
[12] Li H., Cui J.T. and Ma J.F.,"Social influence study in online networks: a three-level review", J. Comput Sci Technol., 30(1), 184 – 99, 2015.
[13] Li K-L, Huang H-K, Tian S-F, and Xu W., "Improving one-class SVM for anomaly detection", Int Conf Mach Learn Cybernetics,30(5), 77 81, 2003.
[14] Nieminen J.,"On the centrality in a graph", Scand J Psychol, 15(1), 332 –336, 1974.
[15] Ravneet Kaur, Mankirat Kaur and Sarbjeet Singh,"A Novel Graph Centrality Based Approach to Analyze Anomalous Nodes with Negative Behavior", International Conference on Information Security & Privacy (ICISP2015), 11 –12 December 2015, Nagpur, INDIA.
[16] Sabidussi G.,"The centrality index of a graph", Psychometrika, 31(4), 581–603, 1966.
[17] WIREs Comput Stat 2015, 7:223–247. doi: 10.1002/wics.1347.
[18] Nicholas A. Heard ET. al., "Bayesian anomaly detection methods for social networks", The Annals of Applied Statistics, 2010, Vol. 4, No. 2, 645 – 662.