A Graph based Anomaly Node Detection in Socio-metric Networks

Bhuvaneswari Anbalagan[1*], Sariga Ilango[2]

[1]Teaching Fellow, [2]PG Student
Department of Computer Technology, Madras Institute of Technology,
Anna University, Chennai, India.

***Abstract*: The challenge of detecting anomaly node behavior is the most vital issue that complicate the social network graphs due to unstructured connectivity parameters in overlapped peer-peer online user communities. The socio-metric connectivity parameters are typically considered to determine the measurement error which is the divergence between the actual value of a node behavior and the observed value of that behavior in a social network graph. But it is practically difficult to identify the measurements of a particular concept as a summation of actual value plus error. In this paper, a graph based anomaly nodes are detected using the measurement scenarios namely False Positive nodes/edges, False Negative nodes/edges and False Aggregation/Disaggregation. The quantification of network measures such as Degree centrality, clustering coefficient, Network constraint and Eigenvector centrality exhibit the effectiveness of anomaly node behavior detection accuracy. The anomaly nodes are detected and erroneous nodes are removed.**

***Keywords:* Online Social network; Anomaly Node; Network Measurement; Behavior Detection;**

I. INTRODUCTION

The social network analysis is the study to analyze the online social networks. In terms of network theory, Social network analysis [1], [2] represents the social relationships in terms of nodes. They representing individual actors or users within the social network) and edges (which indicate relationships connecting the individuals). The social networks are often depicted in a social network illustration, where nodes are represented as points and edges are represented as lines. SNA is significant to scale up the recent research on measurement error in network data has typically focused on missing data. We embed missing data, which we term false negative nodes and edges, in a broader classification of error scenarios. It includes false positive nodes and edges and falsely aggregated and disaggregated nodes.

The mistakes that may occur while collecting or coding a network dataset is also a measurement error. For example, in a class room survey [3], if a respondent misspells the name of a contact, then the contact may erroneously be treated as two dissimilar persons. However, measurement error can also refer to the extent to which a network dataset represents the reality of the relationships within a group under study. For instance, even if all respondents report the correct spellings of their friends' names, the understanding of what qualifies as a friendship tie can vary by respondent. The major reasons for Measurement error to arise are: (a) missing data because of sampling method, (b) mis-specification of the limit in network, (c) top-coding of the number of edges, (d) errors due to miscoding and misreporting, (e) non-response nodes and (f) spurious nodes and edges. The first three categorize the Sampling-induced error and the next three categorize non-sampling induced error.

Initially in social network research, there exist three levels of empirical understanding. One, is the ideal network: the true set of relations among entities in which the edges represent actual, mutual friendships between individuals in the network. Two is the clean network: the set of relations among entities as coded in a network dataset with-out data entry mistakes in which the edges represent each respondent's own perception of friendship with others [4]. Three is the observed network: a network dataset, often suffering from coding errors in which the edges represent reported friendships, but some nodes and ties are erroneously coded that is actually available to a researcher. Secondly, we classify network measurement error into six basic types – missing nodes, spurious nodes, missing edges, spurious edges, falsely aggregated nodes, and falsely disaggregated nodes. The Social Network Users are categorized based on the Nodes and Edges [5].

- *Alpha Socializers* – online users who used sites in intense short bursts to flirt, meet new people, and be entertained.
- *Attention Seekers* – online users who craved attention and comments from others, often by posting photos and customizing their profiles.
- *Followers* – online users who joined sites to keep up with what their peers were doing.
- *Faithful's* – online users who typically used social networking sites to rekindle old friendships, often from school or university.
- *Functional* – online users who tended to be single-minded in using sites for a particular purpose.

The Social Networks give raise to errors such as flat file faults, out of data bound faults, data lost faults, spike faults [6]-[10]. Under the theme of the big data sets from real world complex networks, let's consider the social network data generated and exchanged within networks. However, the recent development in social networks fails to provide efficient support on fast detection of error and locating the errors [11], [12]. There exists large volume of error in form of text files and data logs generated by nodes in social network data sets. The Complex scale free social network generates various types of errors such as Omission error, Commission Errors, Edge/node attribution errors [13], [14]. Missing/omission edges and nodes can have huge impacts on errors in network variables, particularly for some centrality measures [15]. Like omission errors, the erroneous inclusion of nodes

and edges can affect the ultimate determination of node-level measures and the identification of key nodes called commission errors.

II. Social Network Graph Metrics

Social network graph can be represented as nodes and edges. Each individual is a node. The connection that ties up the nodes in the network are called edges. Each node is the representation of data in the network. Network data can suffer from missing or spurious nodes, which we term false negative nodes and false positive nodes, missing or spurious edges, which are termed here as false negative edges and false positive edges, or the erroneous merging or splitting of nodes, which we call false aggregation and false disaggregation. The following figure 1 shows the conference citation network considered for our experimentation.

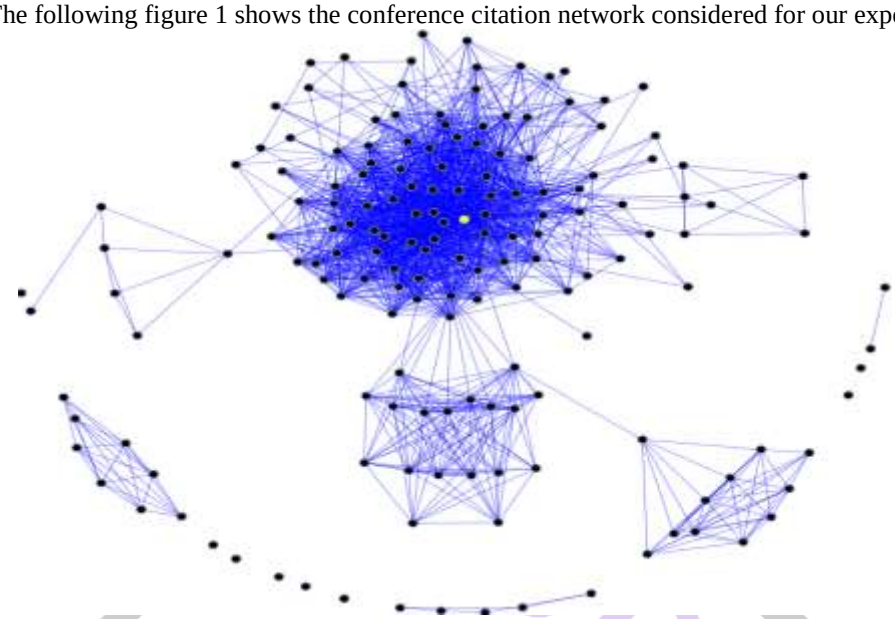

Fig.1. Conference Citation Network Graph

A. *Density and Degree*

Two most common evaluation metrics which is used in social networks are density and degree. The degree focuses on the individuals within the network. The density focuses on the entire network or communities within the network. Both these metrics degree and density signify connectivity. For example, Friendship network gathered through sociometric surveys

- *Density* represents the actual proportion present and indicates all possible relationships in a social network truly. The value ranges from 0 - 1 in which 0 represents the closer value, the sparser the network is represented by the value is to 1 the denser the network is. The number of possible relationships in a network is calculated using the formula: n!/(n-2)!

Where n = the number of nodes in the network and 2 is the maximum number of relationships possible between any two nodes in the network.

- *Nodal degree* is defined as the total number of relationships involving that node. The degree can be classified as In degree and Out degree.
- *Indegree* is the total number of relationships in which a particular node is the target where *Out degree* is the total number of relationships in which a particular node is the source.

B. *Miscoding and misreporting errors*

There is a possibility for the edges to be miscoded due to respondent or error of interviewer. Respondents may forget nodes or interview carelessly. It may lead them to misreport data on edges. In some time, there may be strategic reporting of edges, e.g., respondents may be reporting preferred rather than actual edges.

C. *Spurious Nodes*

Spelling mistakes in names of the nodes or multiple names for the same nodes. The first and last name of a particular node may be misinterupped which leads to a new name can also be reffered as spurious node. It can lead to the presence of duplication nodes in the network. There arises an issue when edges are inferred from existing data.

D. *Non-response nodes and edges*

There exist few nodes in the network that will never communicate in actively. Such edges or communications are missing as a result of non-response from nodes.

E. *False negative nodes*

The worst case of missing nodes in the network that are most expected to be. Such nodes are call as False negative nodes. It refers to the absence of nodes that should be present in a network.

F. False positive nodes

In difference to false negatives, nodes that are erroneously present in a network are called false positive nodes. For example, various spamming scripts are generated to post or tweet various content that generate false user activity in online communities. These spam scripts mimic the online activity like a human. The issue is very hard to filter completely when the amount of node data is very large in the online communities. It is proved that almost 27% of all Facebook accounts are not real

G. False negative edges

The relationships between nodes should be available and reported, but not observed in a network is called false negative edges. In socio-metric surveys, the high risk of the false negative edges come on or after respondents' imperfect recall of their ego-networks

H. False positive edges

False positive edges happen when relationships between nodes are erroneously available in a network. In socio-metric surveys, respondents sometimes description or relations that are not actually present. In online community survey data, many contacts listed by users and the random respondents do not mean to represent real world relationships. The kind of error relationships due to error edges can be avoided by filtering.

I. False aggregation and disaggregation

False aggregation refers to the error scenario in which two different nodes, A and B, are mistakenly treated as one single node. False disaggregation is the opposite problem, in which one node A, is erroneously treated as two separate nodes, edges to different nodes A and B. During data cleaning, the false aggregation and disaggregation of nodes typically occur very commonly.

III. Graph Based Anomaly Node Detection

The proposed work conduct the computation required in the whole process of abnormal error detection and localization in online Social network sites particularly on nodes and edges. The proposed framework is shown in Figure 2.

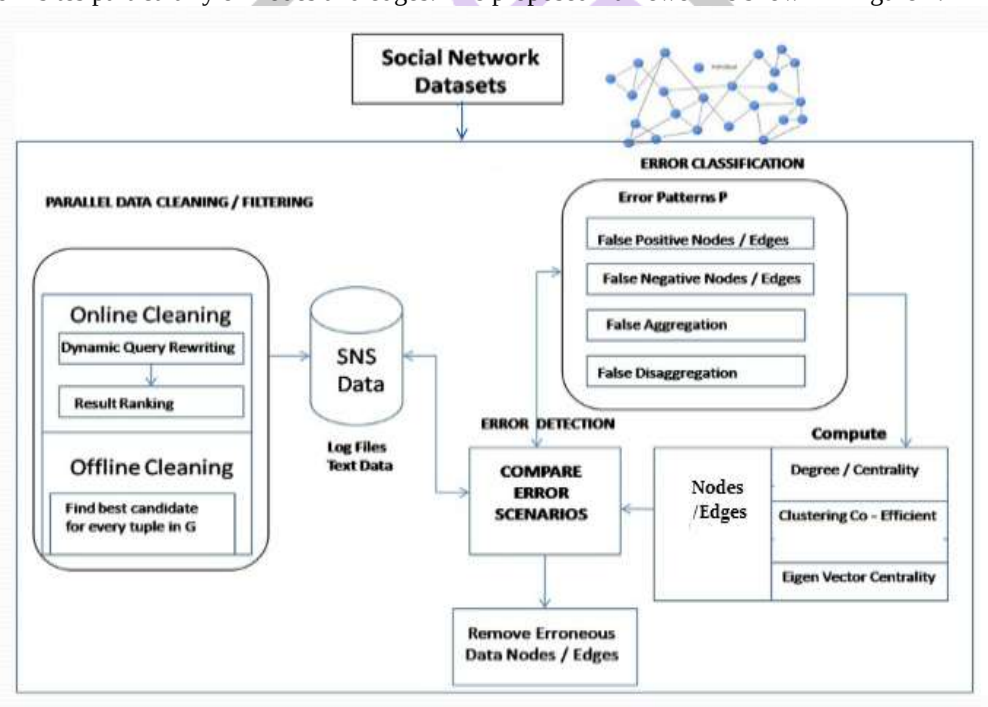

Fig. 2. Proposed Framework

Social network data can suffer from missing or spurious nodes, which can be termed false negative nodes and false positive nodes, missing or spurious edges, which are termed here as false negative edges, and false positive edges, or the erroneous merging or splitting of nodes, which can be called as false aggregation and false disaggregation. In the phase of anomaly node detection, there are three inputs for the error detection algorithm. First, data is cleaned either online or offline. Secondly, the error pattern template P is created which detect the false positive nodes and edges, false negative nodes and edges, false aggregation and disaggregation. Then, the graph of social network SNS G(V,E) of a specific group collected and stored in form of text files and log files. The following metrics are computed on the Conference Citation network.

A. Degree centrality

The degree of a node is the number of edges incident to the node. The diameter of a network is the largest distance between any two nodes in the network. *Average degree* of a network refers to average of the degrees over all nodes in the network. However, it might not be representative, since the distribution of degrees might be skewed. The nodes with a large number of neighbors (i.e., edges) have high centrality (Figure 3) with sample sub-graph. Therefore, is given by $C_d(v)=deg(v)$.

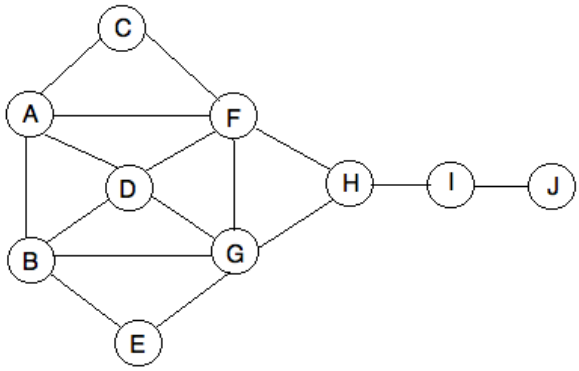

Fig. 3 Sample Network Graph model

In social networks, nodes with high degree centrality are considered to be important."*Closeness centrality*, $C_c(v)$: nodes with short paths to *all* other nodes in the network have high closeness centrality. *Betweenness centrality*, $C_b(v)$: Nodes (or edges) which occur in many of the shortest paths have high betweeness centrality.

Table 1. Degree, Closeness, Betweeness

High -Low	Degree	Closeness	Betweeness
From Highest	D	F,G	H
	F,G	D,H	F,G
to	A,B	A,B	I
	C,E,H	C,E	D
lowest	I	I	A,B
	J	J	C,D,J

B. Clustering coefficient

It measures the extent to which my friends are friends with one another (Figure 4).

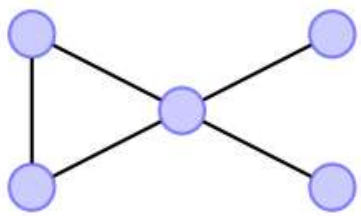

Fig. 4. Connected graph

The clustering measure is represented by the overall clustering coefficient Cl(g), given by

$$Cl(g) = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes}}$$

where a "connected triple" refers to a node with edges to an unordered pair of nodes.
Another measure of clustering is defined on an individual node basis. The individual clustering for a node i is

$$Cli(g) = \frac{\text{number of triangles connected vertex i}}{\text{number of triples centered at i}}.$$

The average clustering coefficient is

$$Cl\mathbf{Avg}(g) = \frac{1}{n} \sum_i Cli(g).$$

The overall clustering coefficient for the network is 3/8. The individual clustering for the nodes are 1, 1, 1/6, 0, and 0.

C. Eigenvector centrality

A node's eigenvector centrality is the unit-normalized sum of its edges or connections to its neighbors, wherein each tie to a neighbor is weighted by the neighbor's ties, and each of the neighbors ties are weighted, and so forth. To facilitate calculation, lets consider graph G's representation as an adjacency matrix A, the eigenvector centrality of node i is given by the ith element of A's unit-normalized principal eigenvector.

IV. Experiments & Results

Researchers need the information of the nodes and edges in the network for various research purposes. Based on the interaction and relationship in the social network, it is possible to obtain the information using the directionality of the nodes and edge connectivity. It allows the researchers to construct the directed and weighted graphs. Information about nodes can be obtained by direct elicitation method. In the method, asking all nodes to report to neighboring nodes that they communicate frequently restricted for specific dimension and specified boundary in network. e.g., the users who carry the citation towards a conference papers. It describe one simulation run of citation network for each error scenario below (Figure 5).

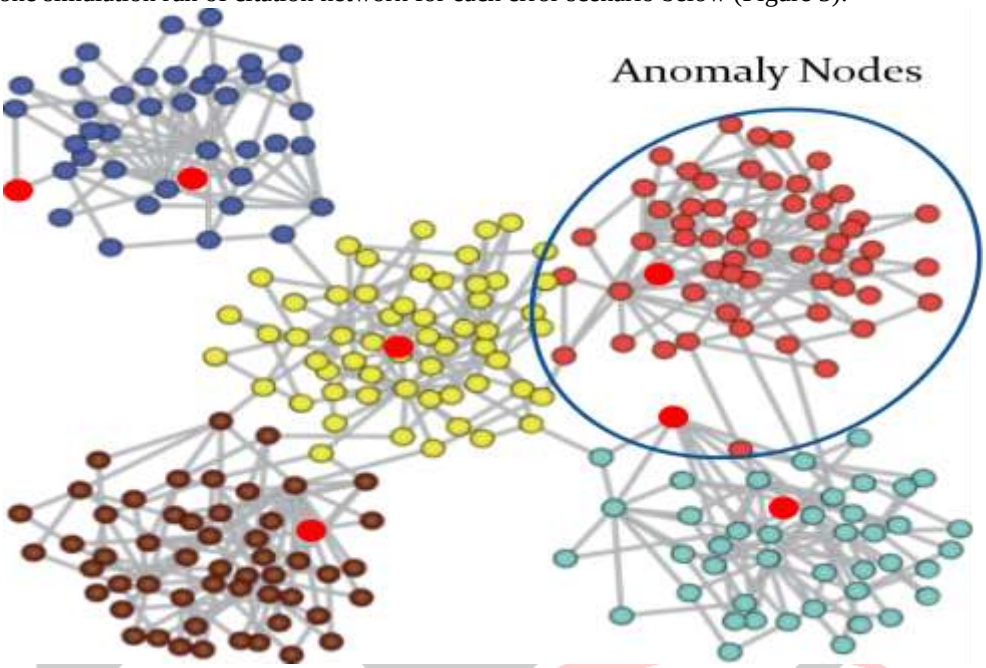

Fig. 5. Anomaly Nodes in Conference Citation Network

Each and every node has to report to their neighboring node whether they potentially interacted with other nodes in network that reduce recall errors. Requesting all nodes to report their mutual participation in familiar groups and activities and evaluating the interactions inside the groups and activities. Dataset can be collected from Existing Data Sources. The initial simulation approach involves simulating an error scenario on what we take to be a unstructured social network, G(V, E). The degree centrality, clustering co efficient, Eigen vector centrality is computed for the citation network using the nodes and edges. The framework exception is the effect of false positive edges on clustering coefficients in the citation network, which appears to be just as damaging as false negative edges. Finally, the effects of false aggregation and false disaggregation lie between the effects of false negatives and false positives.

Table 2. Error Scenario measure of Conference Citation Network

Conference Citation Network	Degree centrality	Clustering coefficient	Network constraint	Eigenvector centrality
False Negative nodes	0.45	0.70	0.33	0.09
False Positive nodes	0.56	0.78	0.21	0.01
False Negative edges	0.12	0.92	0.63	0.18
False Positive edges	0.44	0.51	0.04	0.56
False aggregation	0.14	0.21	0.17	0.64
False disaggregation	0.11	0.21	0.40	0.52

It is likely because false aggregation introduces the combined effects of false negative nodes and false positive edges, whereas false disaggregation reflects the effects of false negative edges and false positive nodes. Finally the output of the error detection algorithm is the error set D'. It calculate the Spearman's rank correlation, clustering co-efficient, Eigen vector Centrality of M and M' to compute reference ranges of error detection. Using the detected error node patterns, the proposed framework detect anomaly and irrelevant node connectivity using the reference ranges obtained from Spearman's rank correlation. The dynamic

query processing over streaming data flows is done while data cleaning so that the error detection complexity is reduced. In order to accomplish the proposed work, the social network data will be stored where the nodes and edges connectivity can be detected as graph. The proposed anomaly nodes are detected and its node distribution is shown in Figure 6.

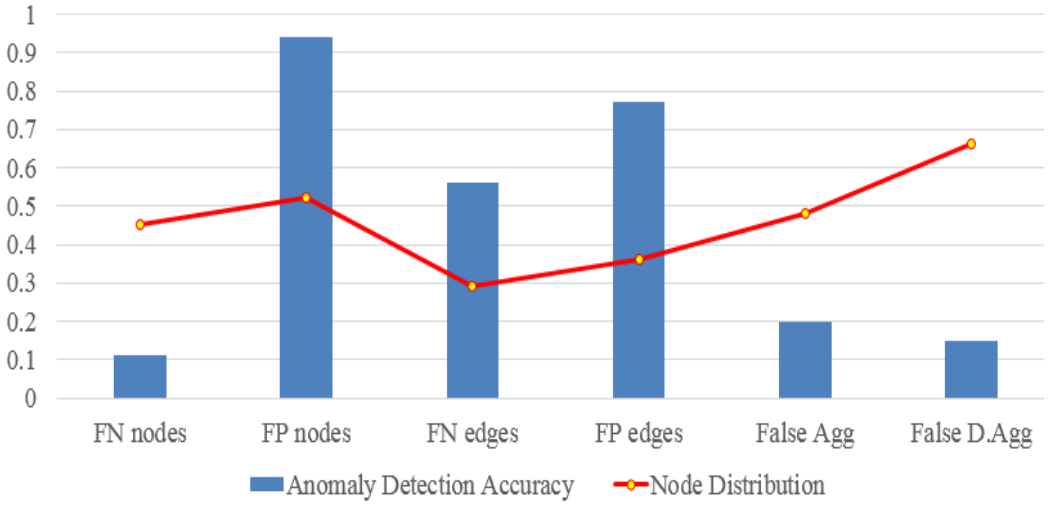

Fig. 6. Comparison of Anomaly Node Detection and distribution

The proposed framework will improve the time efficiency and minimize storage space of Social network data. Thus, when investigative the top 70% of nodes, numerous nodes that prior had just two neighbor have significantly more to pick up as far as positioning by degree centrality than the effectively top-positioned hubs. All in all, these outcomes propose that deciding in favor of speaking to an excessive number of powerless ties as genuine ties makes for more solid system measures than including just more grounded ties. The impacts of false collection and false disaggregation additionally differ with the diagram auxiliary highlights.

V. Conclusion

In the paper, a systematic comparison of six different measurement error scenarios for a conference citation network is discussed. The experimental results show a little perception about how the size of a network affects and influence the robustness and measurements to error scenarios. Finally, given the plenty of online community network data available, researchers are obliged to also be insightful to false positive nodes and edges in their dataset. False negative edges pose the biggest problem in almost every plot, whereas false negative nodes are not just about as unfavorable. The abnormality is identified in the online community and the erroneous nodes are removed.

References

[1] Wei peng, feng li, xukai zou, and jie wu,"A Two-Stage deanonymization attack against Anonymized social networks", IEEE Transactions on Vomputers, vol. 63, no. 2, February 2014

[2] G. Hackeling, Mastering Machine Learning with scikit-learn, Packt Publishing Ltd, 2017.

[3] Kim, M., Leskovec, J., The network completion problem: inferring missing nodes and edges in networks. In: SDM, pp. 47–58. 2011.

[4] Sadikov, E., Medina, M., Leskovec, J., Garcia-Molina, H., Correcting for missing data in information cascades. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM'11. ACM, New York, NY, USA, pp. 55–64, 2011.

[5] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[6] P. McFedries, Twitter tips, tricks, and tweets, Wiley Publishing, 2010.

[7] J. Scott, Social Network Analysis: A Handbook. SAGE Publications, 2000.

[8] B. Krishnamurthy and C.E. Wills, "Characterizing Privacy in Online Social Networks," Proc. First Workshop Online Social Networks (WOSN), 2008.

[9] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08), pp. 93-106, 2008.

[10] Salas, A., P. Georgakis, C. Nwagboso, A. Ammari, and I. Petalas. (2017) "Traffic event detection framework using social media." In Smart Grid and Smart Cities (ICSGSC), IEEE International Conference on, pp. 303-307.

[11] Abhishek Sharma, Yuan Tian and David Lo, "NIRMAL: Automatic Identification of Software Relevant Tweets Leveraging Language Model", IEEE 22nd International Conference in Software Analysis, Evolution and Reengineering, pp. 449- 458, March 2015.

[12] Po-Ching Lin and Po-Min Huang, "A Study of Effective Features for Detecting long-Surviving Twitter spam Accounts", 15th IEEE International Conference in Advanced Communication Technology (ICACT), pp. 841-846, January 2013.

[13] Lin Yao, Yanmao Man, Zhong Huang, Jing Deng, and Xin Wang, "Secure Routing based on Social Similarity in Opportunistic Networks", IEEE Transactions on Wireless communications, vol. 15, pp. 594-605, January 2016.

[14] S. Justin Samuel, B. Dhivya, "An Efficient Technique To Detect and Prevent Sybil Attacks in Social Network Applications", IEEE International Conference in Electrical, Computer and Communication Technologies, pp. 1-3, March 2015.

[15] Michael Fire, Roy Goldschmidt, and Yuval Elovici, "Online Social Networks: Threats and Solutions", IEEE Communication Surveys, vol. 16, pp. 2019-2036, May 2014.

[16] Xia Hu, Jiliang Tang, Huiji Gao , and Huan Liu, "Social Spammer Detection with Sentiment Information", IEEE Conference in Data Mining, pp. 180-189, December 2014.