Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 78 (2016) 556 – 562

Procedia Computer Science

International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

A Novel Graph Centrality Based Approach to Analyze Anomalous Nodes with Negative Behavior

Ravneet Kaur[a*], Mankirat Kaur[a*], Sarbjeet Singh[a]

[a]*Computer Science Engineering, University Institute of Engineering and Technology, Panjab University, Chandigarh, India*

Abstract

Detection of different kinds of anomalous behaviors originating from negative ties among actors in online social networks is an unexplored area requiring extensive research. Due to increase in social crimes such as masquerading, bullying, etc., identification and analysis of these activities has become need of the hour. Approaches from two separate, yet, similar research areas, i.e. anomaly detection and negative tie analysis, can be clubbed together to identify negative anomalous nodes. Use of best measures from centrality based (negative ties) and structure based approaches (anomaly detection) can help us identify and analyze the negative ties more efficiently. A comparative analysis has been performed to detect the negative behaviors in online networks using different centrality measures and their relationship in curve fitting anomaly detection techniques. From results it is observed that curve fitting analysis of centrality measures relationship performs better than independent analysis of centrality measures for detecting negative anomalous nodes.

Keywords: Anomaly; Centrality; Negative Ties; Social Networks.

1. Introduction

In recent years, Online Social Networks have attracted a lot of population owing to its basic feature of bringing

[*] Corresponding author. Tel.: +91-977-999-1701;
E-mail address: ravneets48@gmail.com

[*] Corresponding author. Tel.: +91-946-354-3880;
E-mail address: mannsunshine09@gmail.com

doi:10.1016/j.procs.2016.02.102

together the people from different cultures at a common platform. But their growing popularity not only influenced different domains positively but also led to their extensive misuse like, sending spam emails, creating fake profiles, sending of unnecessary friend requests, etc. Similarly, the presence of negative ties among different users due to the existence of frequent set of negative social behavior towards other users is also on the rise. Nowadays, a lot of research is being focused on detection and analysis of these types of anomalous and negative activities[1-6].

Different centrality based measures like, degree, status[7], PN centrality[8], etc. were proposed by many researchers to identify the negative nodes and ties. The existence of negative nodes in online networks represents an unusual activity which as per the definition of anomaly[9,10] could be categorized as abnormal. Therefore, a number of anomaly detection techniques can also be applied for identifying such negative nodes. A better approach could also be developed by merging the techniques from two domains.

In this paper, a novel graph based approach has been proposed which uses degree and PN centrality as graph metrics in fitting curves to spot and rank the negative anomalous nodes.

The remaining paper is structured into different sections. Section 2 discusses the related work already performed in the respective domains of negative tie analysis and anomaly detection. Section 3 describes the novel approach followed to detect negative anomalous nodes followed by the Experimental analysis in Section 4. Finally, the paper has been concluded in Section 5 with few future directions.

2. Related Work

2.1. Centrality measures to Analyze negative ties

The term centrality refers to the central element or actor in any domain of knowledge. Likewise, in social networks, concept of centrality defines the most central and influential node in the network through which most of the information flows. Various centrality measures were proposed to analyze ties in networks such as, degree, betweenness and closeness centrality given by Freeman[11], and eigenvector centrality given by Bonacich[12].

Due to the difference in characteristics of flow in positive and negative tie networks, some of these measures are not applicable in identifying negative nodes. Hence, many new measures such as degree, status[7], PII[13] and PN[8] centrality were developed to analyze both types of ties simultaneously.

Degree measure calculates the overall degree of the node by summing the number of positive and negative ties of actor whereas status measure calculates the status of an actor by assigning the elements of eigenvector of adjacency matrix as the scores of actors[7]. PII measure calculates the power possessed by actor by analyzing its dependence of resources and information on other actors in political network of allies and adversaries[13]. PN measure calculates the centrality of actor contributed by direct as well as indirect positive and negative links. Out of these four measures PN centrality was able to identify maximum outsiders in different networks[8].

2.2. Structure based Techniques to detect anomalies

For detecting unusual activities in social networks, different anomaly detection approaches categorized under behavior based[14-16] or structure based scenario have been proposed by various researchers. Out of these, structure based approaches[17-19] seems to be more beneficial as they work on the structural characteristics of the network which a user cannot manipulate. The structural properties in a network like, the number and type of connections among nodes, centrality values etc. are those which cannot be fabricated and denied by a user hence, becomes important and significant to be worked upon.

A number of researchers have incorporated different metrics in curve fitting approaches using regression to detect various anomalies. For instance, Akoglu et. al.[18] proposed an Oddball algorithm which used various power laws to detect different type of anomalies. Likewise, Reza et. al.[19] introduced the use of average betweenness centrality (ABC) and community cohesiveness to predict the anomalous nodes and it was shown that the relation between ABC and number of edges was able to identify anomalies more accurately. Rezai et. al.[20] also evaluated the results of N vs. E on Twitter data set. Similarly, Henderson et al.[21] studied a number of already existing node-based and egonet-based characteristics recursively by calculating some aggregate values over the already existing features.

Both node and egonet-based features were used to retrieve neighborhood information along with the recursive features to extract behavioral information.

As already stated the negative behavior of nodes depicts an unusual activity that represents an anomalous behavior. Therefore, it seems intuitive to merge anomaly detection techniques with approaches of analyzing negative ties to tackle the problem in a better way.

3. Proposed Approach

From literature, it can be inferred that both the approaches are efficient enough to identify uncommon activities in their respective domains. Carefully analyzing the aspects of the approaches it is observed that both the approaches work towards identifying the unusual activities (anomalies) incorporating the structural properties of network. Therefore, the better approach could be developed by combining the best measures from both the methods.

The negative anomalous nodes representing the unusual activity can be detected by two approaches. The first approach for identifying negatively behaving nodes is by analyzing the positive and negative connections in egonet of nodes through centrality measures such as degree, PN centrality. The second approach is by analyzing the relationship between PN and Degree measure through curve fitting analysis for detecting negative nodes. Both the approaches are summarized as follows.

3.1. Independent Analysis of PN centrality and Degree measure

As compared to status, and PII measures, the PN centrality and degree measures are able to correctly identify most of the negatively behaving nodes in small societal dataset of Sampson monastery[22]. But scores assigned by PN measure is more precise and mapped small differences between centralities of nodes to considerably precise values that can be easily compared for analysis purpose[8].

- The Degree measure calculates the popularity of node in given network by analyzing the direct positive and negative links of node. The formula of degree is:

  $$D(x) = P(x) - N(x)$$

  where $P(x)$ is the total number of positive connections and $N(x)$ is the negative connections of node. It calculates the popularity of node gained from local neighborhood and cannot describe the global popularity due to the presence of all nodes of network.

- PN centrality calculates the popularity of a node based upon the popularity index of its neighboring nodes due to which it analyze both the direct as well as indirect ties among the different nodes. The PN measure is expressed as:

  $$PN = \left(I - \frac{1}{2n-2}A\right)^{-1} 1$$

  where A = P-2*N, P is the positive ties matrix and N is the negative ties matrix. 1/(2n-2) is the attenuation factor for normalizing positive and negative ties. PN measure does not consider the participation of all negative connections equally in imparting the centrality to a given node. According to this measure, the node which is having negative connections to the most popular node of network is considered more negative than the node which has negative connections to other negative nodes.

The score assigned by above measures is used to identify negative nodes of network i.e. the lowest score is assigned to the most negative node and highest score is given to the most positive node. From this analysis of scores the negative anomalous nodes can be detected.

3.2. Curve Fitting Analysis of PN Centrality and Degree Relationship

The use of distance based approaches in curve fitting can help to detect the presence of such negative activities in the network. Relationship between the centrality measures can be analyzed using fitting curves and the distance of various nodes from these curves can help us identify the unusual activities.

Accordingly, it was theoretically examined that the plotting of PN vs. degree curve helps us to analyze negative anomalous nodes. The implication used is that the nodes with high negative degree i.e. having more number of negative connections will have the lowest PN score and similarly, the nodes with high positive degree will attain the high PN score. This corresponds to a normal behavior. But the negative anomalous nodes are those which have high positive degree but low PN score. Hence, such nodes will lie far away from the curve depicting negative behavior. The relationship between PN and degree could be represented by a linear or a power law as:

- **Linear Law:** $y = Cx + \theta$ (1)

In Equation 1, y represents PN centrality score and x denotes the degree measure. C and θ defines the gradient or the slope of fitting curve and a constant factor respectively.

- **Power Law:** $y = Cx^{\theta}$ (2)

Similarly, Equation 2 defines a power law where again x and y represent degree and PN centrality respectively. Besides this, C defines the slope and θ the power law exponent.

The distance of nodes from fitting curve helps to determine the required anomalies. One of the well sought out distance based method to detect as well as score the encountered anomalies proposed by Akoglu et. al. [3] in Equation 3, helps us analyze the nodes.

$$\mathrm{a}_{\mathrm{score}}(i) = \frac{\max\,(y_i, Cx_i}{\min(y_i, Cx_i)} * \log(|y_i - Cx_i| + 1) \qquad (3)$$

The formula described in Equation 3 calculates the distance of each node i, from the fitting curve. The common and obvious trend states that anomalous nodes would lie far away from the curve (i.e. their distance from the curve will be larger) whereas the normal nodes would either lie on the curve or near to it.

Moreover, anomalous nodes as compared to other nodes have high anomalous score ordered as per their degree of abnormality, i.e. more the abnormality higher will be the score of that node.

4. Experimental Analysis

The experiments have been performed on Epinions network data set collected from the SNAP repository. This is the online social trust network of who trusts whom, of a common consumer review website Epinions.com, where users create signed relations of trust/distrust with each other[23]. Members of site can give positive or negative ratings to products of website as well as rate the reviews given by other members. The visitors of site can check the new and old reviews of product and then decide which product to purchase. The links between nodes or persons who give reviews about products are explicitly labelled as positive or negative[24]. All the relationships formed by trust interact with each other to form Web of Trust, which is then combined with review ratings to choose which review to be shown to user. The dataset contains large number of nodes from which a subset of 122 nodes is identified which contains the most influential nodes by using Random Walk algorithm of Scale down sampling[25]. The actual negative anomalous nodes from this subset are detected as 45 by analysing the structural properties of network. The description of dataset is summarized in Table 1.

Table 1. Dataset Description

Dataset	Epinions
Type of network	Trust network
Type of connections	Trust (user trusts user)
Type of anomalies found	Negative Anomalous nodes
Number of most influential nodes analyzed	122
Mode of selecting the analyzed sample	Random Walk
Number of actual anomalies present	45

The behavior of nodes have been analyzed using three principle ways:

- By plotting the power law fitting curves between centrality measures. (PN centrality and Degree measure)
- By plotting the linear law fitting curves between PN and Degree centrality measures
- By analyzing the scores assigned by the PN centrality and Degree measure.

In order to compare the outputs of different methods, statistical parameters like Precision, Recall and finally F-score is calculated and studied. F-score for different measures is computed as a result of precision and recall and the ultimate goal is to have an F-score value that incorporated minimum number of false positives and negatives. F-score is calculated as described in Equation 4.

$$\mathrm{F}_{\mathrm{score}} = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

where, $Precsion = \frac{TP(i)}{[TP(i)+FP(i)]} * 100$

$$Recall = \frac{TP(i)}{[TP(i)+FN(i)]} * 100$$

Here, with respect to negative anomalous nodes,

TP(i) indicates the True positives i.e. number of negative anomalous node which are correctly classified as negative

FP(i), False positives, indicates the number of actual positive nodes which are incorrectly classified as negative.

FN(i) indicates False negatives stating the number of actually negative nodes misclassified as positive.

Table 2. Comparison of different analysis methods

Dataset	Method	Formula	Precision (%)	Recall (%)	F-score
Epinions	PN Centrality	$PN = [I - \frac{1}{2n-2} A]^{-1}$	100	82.22	90.24
	Degree Measure	*P(x)+N(x)*	100	75.56	86.073
	D vs PN (Power Law)	PN= 2.682*D$^{0.01064}$	100	93.33	96.55
	D vs PN (Linear Law)	PN= 5.151 - 2.515N	89.13	91.11	90.11

From Table 2, it can be seen that the use of inter-dependence relationship between Degree measure and PN measure through curve fitting approach helps to analyze the negative anomalous nodes with higher accuracy as compared to independent PN and Degree measure. Both power law and linear law curves were plotted but power law relationship help better predict the required nodes than the linear laws. Fitting curve for the power law relationship between the two measures is shown in Fig. 1. As the PN centrality measure is found to be an efficient measure for the detection of negative tie behavior[8], but its efficiency is found to be much less than Degree and PN relationship (power law).

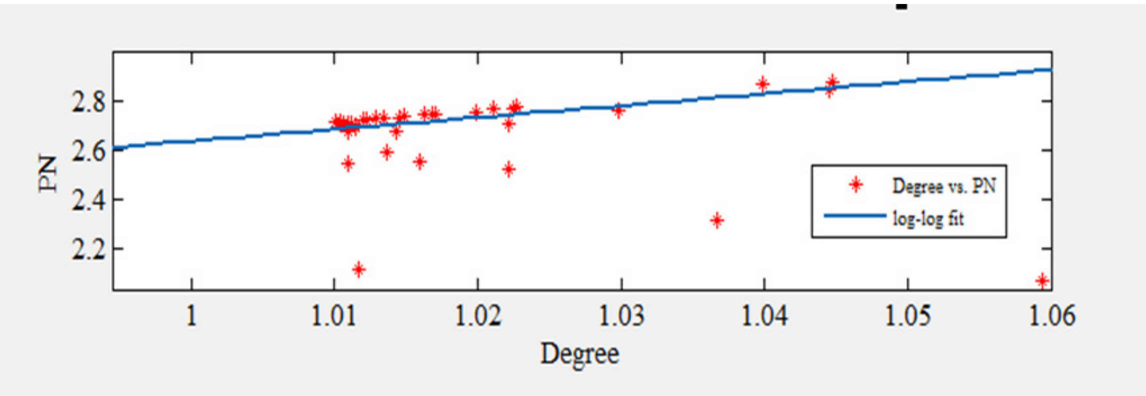

Fig. 1. Power law fitting curve between Degree and PN measure for Epinions Data set.

5. Conclusion

Analyzing the negative behavior using different structure based centrality measures in curve fitting approaches is an interesting area to be worked upon. An effort has been made to provide an idea of an efficient approach by combining best possible measures from the domains of negative tie analysis and anomaly detection. From the discussion in the paper it can be concluded that the use of proposed approach forms an effective baseline for analyzing negative anomalous nodes in online social networks. The proposed approach is an initial step towards analyzing the negative tie behavior using curve fitting approach and in order to check the behavior it has initially been implemented on a small subset of Epinions signed social network. It could be further extended to real world large social network data sets like Twitter, Facebook containing millions of nodes.

References

1. Yoo S, Kim S, Choudhary A, Roy OP, Tuithung T. Two-Phase Malicious Web Page Detection Scheme Using Misuse and Anomaly Detection. *Int J Reliab Inf Assur*. 2014;2(1).
2. Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: a survey. *Data Min Knowl Discov.* Springer; 2014;1–63.
3. Ranshous S, Shen S, Koutra D, Harenberg S, Faloutsos C, Samatova NF. Anomaly detection in dynamic networks: a survey. *Wiley Interdiscip Rev Comput Stat.* Wiley Online Library; 2015;7(3):223–47.
4. Vigliotti MG, Hankin C. Discovery of anomalous behaviour in temporal networks. *Soc Networks*. Elsevier; 2015;41:18–25.
5. Harenberg S, Bello G, Gjeltema L, Ranshous S, Harlalka J, Seay R, et al. Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdiscip Rev Comput Stat*. Wiley Online Library; 2014;6(6):426–39.
6. Miller B, Arcolano N, Bliss NT, others. Efficient anomaly detection in dynamic, attributed graphs: Emerging phenomena and big data. *In: 2013 IEEE International Conference on Intelligence and Security Informatics (ISI),*. 2013. p. 179–84.
7. Bonacich P, Lloyd P. Calculating status with negative relations. *Soc Networks*. Elsevier; 2004;26(4):331–8.
8. Everett MG, Borgatti SP. Networks containing negative ties. *Soc Networks*. Elsevier; 2014;38:111–20.
9. Barnett V, Lewis T. Outliers in statistical data. Wiley New York; 1994.

10. Savage D, Zhang X, Yu X, Chou P, Wang Q. Anomaly detection in online social networks. *Soc Networks*, Elsevier B.V.; 2014;39:62–70.
11. Freeman LC. Centrality in social networks conceptual clarification. *Soc Networks*. Elsevier; 1979;1(3):215–39.
12. Bonacich P. Factoring and weighting approaches to status scores and clique identification. *J Math Sociol. Taylor & Francis*; 1972;2(1):113–20.
13. Smith JM, Halgin DS, Kidwell-Lopez V, Labianca G, Brass DJ, Borgatti SP. Power in politically charged networks. *Soc Networks.* Elsevier; 2014;36:162–76.
14. Vanetti M, Binaghi E, Carminati B, Carullo M, Ferrari E. Content-based filtering in on-line social networks. *Priv Secur Issues Data Min Mach Learn*. Springer Berlin Heidelberg; 2011;127–40.
15. Viswanath B, Bashir MA, Crovella M, Guha S, Gummadi KP, Krishnamurthy B, et al. Towards detecting anomalous user behavior in online social networks. In: *Proceedings of the 23rd USENIX Security Symposium (USENIX Security)*. 2014.
16. Xiao C, Freeman DM, Hwa T. Detecting Clusters of Fake Accounts in Online Social Networks. In: *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*. 2015. p. 91–101.
17. Shrivastava N, Majumder A, Rastogi R. Mining (social) network graphs to detect random link attacks. In: *2008 IEEE 24th International Conference on Data Engineering, 2008 ICDE*. 2008. p. 486–95.
18. Akoglu L, McGlohon M, Faloutsos C. Oddball: Spotting anomalies in weighted graphs. *Adv Knowl Discov Data Min*. Springer Berlin Heidelberg; 2010;410–21.
19. Hassanzadeh R, Nayak R, Stebila D. Analyzing the effectiveness of graph metrics for anomaly detection in online social networks. *Web Inf Syst Eng* 2012. Springer Berlin Heidelberg; 2012;624–30.
20. Rezaei A, Kasirun ZM, Rohani VA, Khodadadi T. Anomaly detection in Online Social Networks using structure-based technique. In: *2013 8th International Conference for Internet Technology and Secured Transactions (ICITST)*. 2013. p. 619–22.
21. Henderson K, Gallagher B, Li L, Akoglu L, Eliassi-Rad T, Tong H, et al. It’s who you know: graph mining using recursive structural features. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011. p. 663–71.
22. Sampson SF. Crisis in a cloister. Ph. D. Thesis. Cornell University, Ithaca; 1969.
23. Guha R, Kumar R, Raghavan P, Tomkins A. Propagation of trust and distrust. In: *Proceedings of the 13th international conference on World Wide Web*. 2004. p. 403–12.
24. Massa P, Avesani P. Controversial users demand local trust metrics: An experimental study on epinions. com community. In: *Proceedings of the National Conference on artificial Intelligence*. 2005. p. 121.
25. Leskovec J, Faloutsos C. Sampling from large graphs. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006. p. 631–6.