Review of different types of Anomalies and Anomaly detection techniques in Social Networks based on Graphs

Sarbjeet kaur[1], Prabhjot Kaur[2]
[1] Research Scholar, [2] Assistant Professor,
Computer Science Engineering, Universal Institute of engineering and Technology, Lalru, Punjab, India

Abstract:

As today is a trend of social networking to communicate with each other, there is a possibility of anomalous users in online networks to steal other's personal information etc. It is necessary to understand the behavior of different users to find fake and genuine users from networks. To find anomalies in network we should understand social network analysis, different type of anomalies and different social network metrics. In this paper we have reviewed different type of social network metrics, types of anomalies and anomaly detection techniques based on graphs. This paper will help to understand social networking, social network metrics, anomaly types and anomaly detection techniques to find anomalous users from online social networks.

Keywords: *Anomaly, Anomaly detection, Metrics.*

1. INTRODUCTION

Social networking becomes very popular from last decade. As people prefer social networking websites to communicate with each other. With the help of social networks individuals connected with each other from large distance. While using social websites to interact with others there is possibility of fake users. Fake users enters into network by interacting with weaker node in the network to access information from other nodes. Fake users or fraudsters pretend to be someone and may obtain information from real users and hack data from network. There are number of social networks like Facebook, twitter, LinkedIn, WhatsApp etc. Some other social networks are available for interaction between two or more people. Where they share their personnel information directly or indirectly. Mostly users are unaware of fake users while connecting first time with others on social network.

Social network analysis is a technique to analyze different behaviors of online networks.

There is a variety of techniques to detect anomalous accounts from online social networks. Survey explain both traditional and modern ways to find fake accounts from social networks. Every online social network has its own structure and nature. These techniques are evaluated to find best technique to detect anomalous users from all online social networks like Facebook, twitter etc.

2. SOCIAL NETWORK METRICS

2.1 Closeness: This refers to the degree with which an individual is closer to all others in a network either directly or indirectly. Advance, it reflects the capacity to access information through the "grapevine" of network members. In this way, closeness is considered to be the inverse of the sum of the shortest distance (sometimes called geodesic distance) between every individual and all others accessible in the network.

2.2 Network Density: Network density is a measure of the connectedness in a network. Density is defined as the actual number of ties in a network, expressed as a proportion of the maximum possible number of ties. It is a number that varies in the vicinity of 0 and 1.0. At the point when density is close to 1.0, the network is said to be dense, otherwise it is sparse. When dealing with directed ties, the maximum possible number of pairs is used instead. The problem with the measure of density is that it is sensitive to the number of network nodes; accordingly, it can't be used for comparisons across networks that shift significantly in size [1].

2.3 Centrality: Local and Global: The idea of centrality comprises two levels: local and global. A hub is said to have local centrality, when it has a higher number of ties with different nodes, otherwise it is referred to as a global centrality. Whereas local centrality considers just direct ties (the ties directly connected to that hub), global centrality also considers indirect ties (which are not directly connected to that hub). For instance, in a network with a "star" structure, in which, all nodes have ties with one central hub, local centrality of the central hub is equal to 1.0. Whereas local centrality measures are expressed in terms of the number of nodes to which a hub is connected, global centrality is expressed in terms of the distances among the various nodes. Two nodes are connected by a path if there is a sequence of distinct ties connecting them, and the length of the path is simply the number of ties that make it up. The shortest distance between two points on the surface of the earth lies along the geodesic that connects them, and, by analogy, the shortest path between a particular

pair of nodes in a network is termed a geodesic. A hub is globally central in the event that it lies at a short distance from numerous different nodes. Such a hub is said to be "close" to a significant number of alternate nodes in the network, sometimes global centrality is also called closeness centrality. Local and global centrality depends mostly on the size of the network, and in this way they can't be compared when networks contrast significantly in size.

2.4 Betweenness: Betweenness is defined as the extent to which a hub lies between different nodes in the network. Here, the connectivity of the hub's neighbors is taken into account with a specific end goal to give a higher incentive to nodes which bridge clusters. This metric reflects the number of people who are connecting indirectly through direct links. The betweenness of a hub measures the extent to which an agent (represented by a hub) can fill the role of a broker or gatekeeper with a potential for control over others. Methodologically, betweenness is the most complex of the measures of centrality to calculate and furthermore suffers from the same disadvantages as local and global centrality [2].

2.5 Centralization: Centralization is calculated as the ratio between the numbers of links for every hub divided by the maximum possible sum of differences. Centralization provides a measure of the extent to which an entire network has a centralized structure. Whereas centralization describes the extent to which this connectedness is composed around particular focal nodes; density describes the general level of connectedness in a network. Centralization and density, in this way, are imperative complementary pair measures. While a centralized network will have large portions of its links dispersed around one or a couple of nodes, the decentralized network is one in which there is little variation between the number of links every hub possesses. The general procedure involved in any measure of network centralization is to take a gander at the differences between centrality scores of the most central hub and those of every single other hub [3].

2.6 Social Network Performance: Once the network analysis is completed, the network dynamics anticipate the performance of the network which can be assessed as a combination of: (1) the network's robustness to the removal of ties as well as nodes, (2) network efficiency in terms of the distance to traverse starting with one hub then onto the next and its non-redundant size, (3) effectiveness of the network in terms of information benefits allocated to central nodes lastly, (4) network diversity in terms of the history of each of the nodes.

2.7 Robustness: Social network analysts have highlighted the importance of network structure with relation to the network's robustness. The robustness can be assessed based on how it becomes fragmented when an increasing fraction of nodes is evacuated. Robustness is measured as an estimate of the tendency of individuals in networks to shape local groups or clusters of individuals with whom they share similar characteristics, i.e., clustering [4].

2.8 Efficiency: Network efficiency can be measured by considering the number of nodes that can instantly access a large number of various nodes – sources of knowledge, status, and so on., through a relatively small number of ties. These nodes are treated as non-redundant contacts. For instance, with two networks of equal size, the one with more non-redundant contacts provides a larger number of benefits than the others. Also, it is very obvious that the gain from another contact redundant with existing contacts will be minimal. Notwithstanding, it is wise to consume time and energy in cultivating another contact to un-reached people. Henceforth, social network analysts measure efficiency by the number of non-redundant contacts and the average number of ties a sense of self has to traverse to reach any adjust, this number is referred to as the average path length. The shorter the average path length relative to the size of the network and the lower the number of redundant contacts, the more efficient is the network.

2.9 Effectiveness: Effectiveness targets the cluster of nodes that can be reached through non-redundant contacts. In contrast, efficiency aims at the reduction of the time and energy spent on redundant contacts. Every cluster of contacts is an independent source of information. One cluster around this non-redundant hub, regardless of how numerous its members are, is just a single source of information, because people connected to each other tend to think about the same things at about the same time.

2.10 Diversity: While efficiency is about getting a large number of (non-redundant) nodes, a hub's diversity, conversely suggests a critical performance point of view where those nodes are diverse in nature, i.e., the history of every individual hub inside the network is imperative. It is particularly this aspect that can be explored through case studies, which is a matter of intense discussion among social network analysts. It seems to suggest that social scientists should prefer and use network analysis as per the first strand of thought developed by social network analysts instead of actor-attribute-oriented accounts based on the diversity of each the nodes [5].

3. TYPES OF ANOMALIES

Anomalies or the abnormal activities can be classified into various categories endless supply of parameters. This segment talks about some of these categories [6].

3.1 Based on nature of anomalies

The anomalies are classified into chiefly three categories based upon the nature and scope of anomalies:

3.1.1 Point anomalies: Point anomalies, additionally alluded to as global anomalies are found if a data data. Despite the fact that being the simplest kind of anomaly to be detected yet major problem related with detecting point anomalies is finding a suitable measurement in deviation of the object from different objects. Give us a chance to assume that for a normal network each hub must have no less than two neighbors linked to it. Hubs in Group V2 shape such kind of network and in this manner represent a normal behavior however group V1 contains separated points. As a result of their different behavior to different hubs they are predicted to represent an anomalous behavior. Additionally, we may likewise have local anomalies which are studied relative to their local neighborhood as it were. For instance, on the off chance that we group a set of individuals in view of their links in the network as friends and check their income (some parameter), a particular individual, let say A, may have a genuinely low income compared to his friends suspecting a local anomaly while by and large in the global context his income may be insignificant the same number of individuals may have comparative income representing a normal behavior.

3.1.2 Contextual anomalies: Frequently, alluded to as conditional anomalies, these are available in the data set if the data object deviates essentially with respect to a particular context. For instance, temperature might be considered as a contextual anomaly. On the off chance that for instance, today's temperature is 28 C. Regardless of whether it is anomalous or not relies on the time and location. It is seen as an anomaly when considered in winters in Toronto. Yet, in summers in Toronto this much temperature is normal and henceforth no abnormality is seen. While detecting contextual anomalies two attributes of the data object define the data set [7]:

- Contextual attributes: These attributes define the context of the object. For instance, in temperature illustration, contextual attributes are date and location.
- Behavior attributes: Characteristics of an object are defined utilizing these attributes and in a route help to identify the anomalous behavior of an object with respect to its context. In the temperature case, temperature, humidity and so on can be considered as behavior attributes.

3.1.3 Collective anomalies: Collective anomalies are encountered at whatever point a collection of data objects all in all depicts an alternate behavior than others, though the individual data objects may not be anomalous. One of the fine principles received to detect collective anomalies is to consider the behavior of the group of objects alongside the background information about the relationship among those data objects.

3.1.4 Horizontal anomalies: Recently, another sort of anomaly, called horizontal anomaly has evolved in social networks which depict the presence of object (i.e. a point) demonstrates an alternate behavior than that of the rest of the anomalies based upon the diverse sources of data available. For instance, a similar user might be available in various communities on various social networks. Essentially, a user may have comparable kinds of friends on number of social networks (e.g. Facebook, Google+) however totally various types of friends for another social network (e.g. Twitter). This depicts an unusual action which can be considered as anomalous.

3.2 Based on static/dynamic nature of network/graph structure

Further classification of anomalies based upon the network structure being used distinguishes them as being static or dynamic. Static networks, for example, bibliographic networks, allow the changes to happen slowly after some time while dynamic networks, for example, mobile applications, allow the faster communications and continuous changes in the networks.

3.2.1 Dynamic anomalies: A dynamic anomaly exists with respect to past network behavior in which changes happen in the network with the passage of time. For instance, it might involve changes in the way interactions occur in the network.

3.2.2 Static anomalies: A static anomaly occurs with respect to remainder of the network ignoring the time factor. Just the present behavior of a node is examined with respect to others in the network.

3.3 Based on information available in network/graph structure

Depending upon the sort of information available at a node or an edge, anomalies can be arranged as labeled or unlabeled [8].

3.3.1 Labeled anomalies: Labeled anomalies are identified with both structure of the network and the information gathered from vertex or edge attributes. For instance, labels on nodes may indicate the attributes of individuals involved in the communication movement and that on the edges represent their interaction behavior.

3.3.2 Unlabeled anomalies: Unlabeled anomalies are connected just to the network structure. No trait of a node or an edge is contemplated. Their classification is generally studied as follows and distinctive procedures have been developed and deployed to detect these types of anomalies.

3.3.3 Static unlabeled anomalies: This sort of anomaly occurs when behavior of an individual remains static and the attributes, for example, time of individuals involved, kind of interactions, and its duration are ignored because of unlabeled nature of the network in which labels on nodes and edges are

ignored. Just the way that interaction occurred is important.

3.3.4 Static labeled anomalies: When alongside the network structure labels on the vertices and edges are additionally viewed as, then the anomalous substructures found are referred to as static labeled anomalies. Static labeled anomalies are used in spam detection, for instance, to detect opinion spam (which involves the fake product reviews). A set of hidden labels are normally assigned to the vertices and edges ones they will give negative reviews though fraudulent users are comprehended to do the reverse.

3.3.5 Dynamic unlabeled anomalies: This kind of anomaly emerges when we have dynamic networks that change with time. Behavior of the data object is diverse with respect to past day and age relative to the network structure. For instance, while considering just the pattern of interactions, there are greatest of six courses in which a maximal clique can evolve: shrinking, growing, splitting, merging, appearing or vanishing. These involve studying the network structure with respect to the network structure prevalent at some past day and age. Now and again, the normal behavior does not result in any network change; then, any area changes may likewise predict an anomalous behavior [9].

3.3.6 Dynamic labeled anomalies: In a dynamic network when anomalous behavior is observed by considering labels of the vertices and edges likewise; then, anomalies observed are classified as dynamic labeled anomalies. Dynamic networks are worked upon by considering the structure of the network at fixed time intervals and treating them in an indistinguishable path from for a static network.

3.4 Based on behavior

Different classes of anomalies to be specific, "white crow anomalies" and "in-disguise anomalies" are displayed here.

3.4.1 White crow anomaly: It emerges when one data object deviates fundamentally from different observations resembling the essential anomaly definition. For instance, while examining the student record, if a record is found where stature of a student is entered as 56 ft, which is impossible, then it is taken as a white crow anomaly. These anomalies are generally detected as particular nodes, edges, or subgraphs representing the abnormal behavior.

3.4.2 In-disguise anomaly: It is considered as a little deviation from the normal pattern. For instance, anybody attempting to peep into somebody's social network account would not have any desire to get caught; along these lines, he will attempt to act in an indistinguishable way from a normal user. Such anomalies are perceived through peculiar patterns, which additionally include uncommon nodes or entity alterations. These are hard to be detected as they are hidden inside the network.

which are iteratively updated. In the product review framework, a bipartite graph with one subset of vertices as users and different as products is taken in which the edges between the subsets represent the product reviews. Hidden labels are assigned to both users and products. For users the name can be in the type of honest or fraudulent and for the products it could be either great or awful. A normal honest user will give accurate results i.e. for good products they give positive reaction and for awful

4. ANOMALY DETECTION METHODS:

Following three categories of data mining approaches are used to detect anomalous users from online social networks [10].

1. Supervised Learning Techniques.

2. Unsupervised Learning Techniques.

3. Semi-Supervised Learning Techniques.

4.1 Supervised anomaly techniques:

Supervised anomaly Techniques are used to model both normal and abnormal behaviors. These techniques require pre-labelled data for anomaly detection classified as normal or abnormal. Different training models are used to identify the normal or abnormal data from dataset. Supervised techniques work on two approaches:

1 Training model is compared with dataset to find analogues data from data set that is classified as normal data.

2. In opposite to above method some anomalous data is compared against training model to find abnormal data from dataset.

4.2 Unsupervised anomaly detection Techniques:

Unsupervised techniques work on clustering mechanism. These Techniques have no pre- labelled data normal or abnormal. These techniques find clusters of nodes whose behavior is similar to group. sometimes this assumption becomes wrong as many anomalies also make clusters with similar pattern. So unsupervised techniques are inefficient to find accurate results.

4.3 Semi Supervised anomaly detection Techniques:

In semi supervised techniques data set is only labeled with one label as normal. Training model detect abnormal class by itself from dataset.[11].

Table 1- Anomaly Detection Techniques based on Graphs

Sr. No.	Method	Type of Network	Conclusion
1.	Matched filtering for subgraph detection in dynamic networks	Dynamic Networks	Detected weak subgraph anomalies.
2.	Query-based Graph Cuboid Outlier Detection	Heterogeneous networks	Detection of graph cuboid outliers
3.	Anomaly Detection using Scan Statistics on Time Series Hypergraphs	Hypergraphs	Scan statistics on hypergraphs can detect certain anomalies that are not apparent by using scan statistics on graphs
4.	Facebook Immune System	Social network (Facebook)	Overview of various threats to Facebook and different techniques to protect Facebook from threats
5.	Uncovering Large Groups of Active Malicious Accounts in Online Social Networks	Facebook and Instagram	Unveil more than two million malicious accounts and 1156 large attack campaigns within one month.
6.	Discovering Important Nodes through Graph Entropy The Case of Enron Email Database	Enron GRAPH	Defined and addressed the problem of important nodes and finding closed group around them
7.	Strangers Intrusion Detection	Social networks	This method is effective in detecting various types of malicious profiles
8.	Opinion Fraud Detection in Online Reviews by Network Effects (FRAUDEAGLE)	Synthetic as well as real networks	FRAUDEAGLE successfully reveals fraud-bots in a large online app review database.
9.	Subgraph Detection Using Eigen vector L1 Norms	Network graphs	Detect small, dense anomalous subgraphs embedded in a background
10.	Community Trend Outlier Detection using Soft Temporal Pattern Mining	Temporal datasets	Highly effective and efficient in detecting meaningful community trend outliers

5. CONCLUSION

This paper reviewed different types of anomalies in social networks, social network metrics and various possible anomaly detection techniques based on graphs. In future for better anomaly detection in online social networks probabilistic structures can be used with social network metrics.

5. REFERENCES

1. G. E. Hinton and S. T. Roweis," Stochastic neighbor embedding", 2002, NIPS
2. J. Huang, H. Sun, J. Han, H. Deng, Y. Sun, and Y. Liu," SHRINK: a structural clustering algorithm for detecting hierarchical communities in networks", 2010, CIKM
3. G. Karypis and V. Kumar," A fast and high quality multilevel scheme for partitioning irregular graphs", 1998, SIAM J. Sci. Comput., 20(1):359–392,

4. A. Lancichinetti, S. Fortunato, and F. Radicchi," Benchmark graphs for testing community detection algorithms", 2008, Phys. Rev. E, 78(4):046110
5. T. Lou and J. Tang," Mining structural hole spanners through information diffusion in social networks", 2013, WWW
6. A. Agovic, A. Banerjee, A. R. Ganguly, and V. Protopopescu," Anomaly detection using manifold embedding and its applications in transportation corridors", 2009, Intelligent Data Anal., 13(3):435–455,
7. L. Akoglu, H. Tong, and D. Koutra," Graph based anomaly detection and description: a survey", 2015, Data Min. Knowl. Discov., 29(3):626–688
8. L. Armijo," Minimization of functions having lipschitz continuous first partial derivatives", 1966, Pacific J. Math, 16(1):1–3,
9. P. Bogdanov, C. Faloutsos, M. Mongiovi, E. E. Papalexakis, R. Ranca, and A. K. Singh," Netspot: Spotting significant anomalous regions on dynamic networks", 2013, SDM
10. Hodge, Victoria J., and Jim Austin. "A survey of outlier detection methodologies." Artificial Intelligence Review 22.2 (2004): 85-126.
11. Ravneet kaur, Sarbjeet singh," A survey of data mining", Egyptian Informatics Journal(2016)17,199-216.