SOME OBSERVATION OF ALGORITHMS DEVELOPED FOR ANOMALY DETECTION

Pallavi Raj
Department of Computer Science, Banaras Hindu University,
Varanasi, Uttar Pradesh, 221005, India
pallavi.blsitm@gmail.com

Rakhi Garg
Department of Computer Science,MMV, Banaras Hindu University,
Varanasi, Uttar Pradesh, 221005, India
rgarg@bhu.ac.in

Abstract - With the growth of social networks tremendous amount of data are generated in a regular interval of time that can be our search histories, likes, shares, posts, comments, etc. The generated data are publicly available which may become the prime target for the malicious users, who try to attack and harm the innocent users. Using anomaly detection techniques, we can identify the unusual behavior of such users. In social networks, anomalies can be detected by exploring the pattern hidden in the network. This paper mainly focuses on the graph mining techniques used for anomaly detection in social networks. The algorithm for anomaly detection using graph mining techniques has been categorized on the basis of different characteristics of anomalies, and the types of anomalies generated. In addition to this, the paper also emphasizes on the issues and challenges associated with each developed algorithm that will help the researchers and the scientists working in this area to find the solutions for problems associated with various algorithms.

***Keywords*:** Anomaly Detection; Graph Mining; Social network; Anomalies; Graph Anomaly Detection.

1. Introduction

Social networking sites are online platforms which let the users to build social relationships with other people who share similar interests regardless of their geographical locations. Over the past decades, social networking sites have become the most common and popular mode of communication. The openness of social networking sites lead to the exposure of personal information to the malicious users which in turn create security issues such as, cyber-attacks, profile cloning, spoofing, phishing, bullying, organized crimes, etc[Keyvanpour *et al.*, (2014); Liu and Chawla, (2015); Yu *et al.*, (2015)]. Also, spreading false or fake information on social media is on peak today. Hence, detection of these activities in the network is a big concern as their presence may lead to heavy losses.

According to Hawkins, anomaly is defined as " the observation which deviates so much from the other observation", or in other words, it is a deviation from a rule what is regarded as normal[Hawkins,(1980)]. Anomaly in the social networks can be defined as irregular or unexpected behavior which deviates from the normal pattern of user in a network [Savage *et al.*, (2014)]. For instance, it is considered to be normal when a user sends mails to a set of users who share connections among themselves but an anomalous user chooses its audience in a random manner which is unlikely to have a relation between them. One can identify this by analyzing the network pattern of both the users [Akoglu *et al.*, (2010)]. Another example that can be considered as anomaly is, by creating false identities and uses them to communicate with a random set of innocent users. Group review spamming, organized viral campaigns, etc are some of the other examples that can also be considered as anomalies.

Anomaly detection is a process of identifying rare occurrences in data set by analyzing the patterns that deviate from the normal patterns [Dang *et al.*, (2014)]. Anomalies can also be referred as outliers, exceptions, or peculiarities depending upon the application domain, in which it is being used. In social networks, anomalies can be detected by exploring the pattern hidden in the networks. Since, social networks can be best reflected in the form of graphs, anomalies in social networks can be detected by analyzing the pattern of graphs in the network using graph mining techniques.

Many data mining techniques have been used for detecting anomalies such as classification, clustering, etc. There were some issues associated with it such as handling structured data, high false alarm rates, which lead to the use of graph based anomaly detection. Graph based anomaly detection is the process of detecting anomalies from the data that are represented as a graph [Eberle and Holder, (2007)]. However, graph based anomaly detection have additional challenges as well such as interdependent objects, variety of definitions, and size of graph substructures. Therefore, graph based anomaly detection algorithms need to be designed by keeping in mind both efficiency and scalability.

This paper provides a comprehensive review of various existing anomaly detection methods in social networks with issues and challenges in each existing algorithm. The paper is organized as follows: section 2 provides brief details about social networks and graph mining. Section 3 contains classification of anomalies on the basis of different parameters. Section 4 describes different characteristics of anomalies in social network .Section 5 contains methods used for detecting anomalies in social network. Section 7 reviews the various anomaly detection algorithms and also highlights issues and challenges in each algorithm. Section 8 presents some of the examples of anomaly detection software that uses the concept of data mining algorithm. And finally section 9 concludes the paper with issues and challenges related to anomaly detection in social network and future direction of works for researchers and scientists.

2. Social network and Graph Mining

A social network is a network of people that helps in building personal as well as professional relationships. Social network is a platform that allows people to interact with each other in order to share common interests, posting information, messages, comments, etc. It is also referred as a virtual community, since people from any geographical location can communicate or interact with each other through various social networking sites. Social networks are represented in the form of graphs. A social network graph is a graph where the nodes represent people and their relationships or interactions represented by edges [Mislove *et al.*, (2007)]. There are different types of graphs that can be used to represent social networks such as directed, undirected, labeled and unlabeled as shown in figure 1.

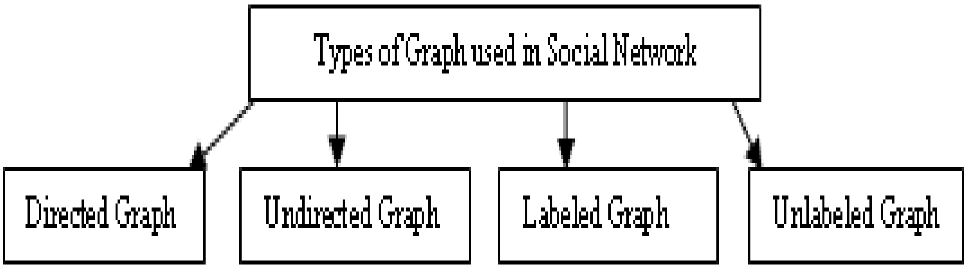
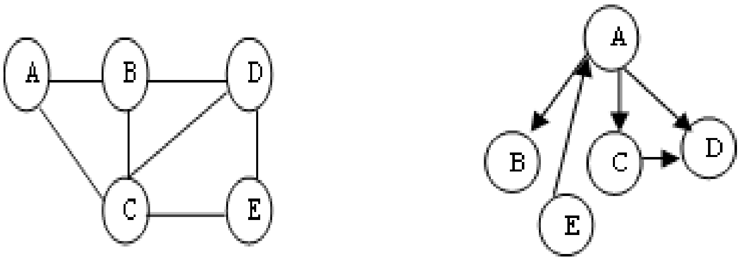


Fig.1. Types of graph used in a social networks.

Directed graphs have edges with directions whereas in undirected graphs edges do not have directions. For example, the friends graph of facebook are represented by undirected graph as shown in figure 2(*i*) whereas graphs of followers in twitter is represented by directed graph as shown in figure 2(*ii*).Here in case of facebook, it can be seen that A is a friend of B and C where as in case of twitter, the graph suggests the node A follows B, C and D and followed by the node E.



(i) *(ii)*

Fig 2. Undirected and directed graph of friends in facebook and twitter.

In labeled/attributed graph, properties are associated with the edges or nodes as shown in figure 3(*i*). Therefore, both the network structure and the information available from the nodes and edges are considered in order to detect anomalies. For example, consider a scenario in which we want to analyze the interactions of people, then the properties of the nodes such as age, gender, location were also be considered for detecting anomalies.On the other hand,in unlabeled/unattributed graph, details associated with the nodes or edges are ignored while detecting anomalies as shown in figure 3(*ii*).For example, consider the same scenario as given in the labeled graph, to detect anomalies in unlabeled graph only thing considered here is the total number of interaction by each person over a full period of time and details such as age, location, gender, etc. will be ignored.

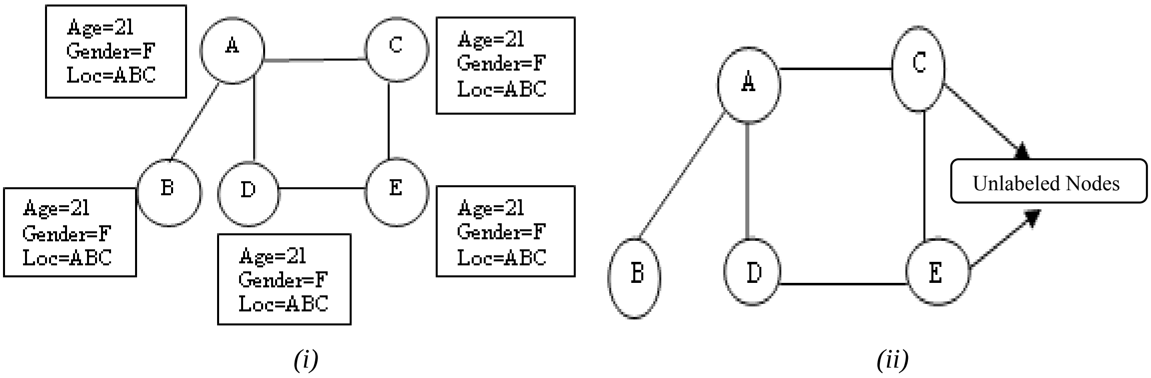


Fig.3. Labeled and unlabeled graph used to represent social networks.

A graph is a powerful tool for representing relationships between entities in a social network. Another important factor about graph representation is its robustness, which makes it hard for an opponent to alter it. Anomalies are basically relational in nature. For example, organized frauds usually take place by closed collaborations of a fraud related group. So, it can be easily detected by using graph mining techniques. Graph mining is a tool used to extract useful patterns from the graph data that can be further used for classification or clustering purposes. Graph mining finds its applications in many different areas, for example, web browsing patterns, biological networks, social networks, chemical compounds, etc [Jiang,(2011)]. Graph mining techniques have been categorized into graph clustering, graph classification, and subgraph mining as shown in figure 4.

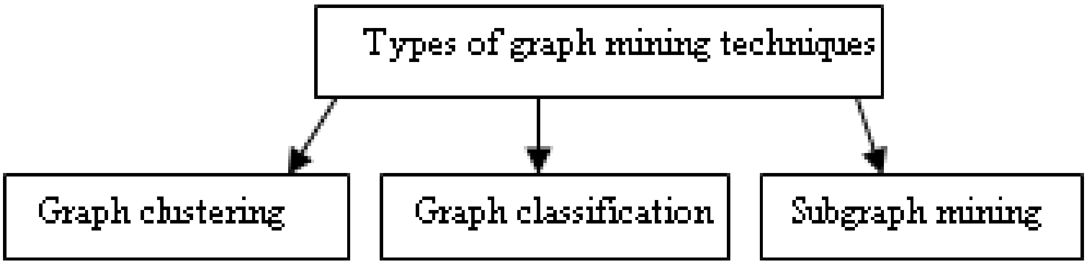


Fig.4. Classifications of graph mining techniques.

Graph Clustering is a process for identifying hidden community structure in a network by grouping the nodes of a graph into cluster. It groups the node in such a way that dense edges connections are belonging to one cluster and few edges among vertices are partitioned into different cluster. Graph clustering is based on unsupervised learning [Newman, (2012); Xu *et al*., (2012); Clauset, (2005)]. The main aim of graph classification is to classify individual graph in a graph database into two or more classes. Graph classification is based on supervised as well as unsupervised learning. Graph classification is a term used to denote two different tasks. The first one is to build a model that predict the class label of a whole graph and the second one is to predict the class label of a node in a single graph. Nodes are classified on the basis of interests, beliefs or other characteristics [Saigo *et al*., (2009)]. Subgraph is a graph whose vertices and edges are subsets of another graph. Frequent subgraphs are the pattern that occurs frequently in a large set of graph. The process of finding frequent subgraph is called frequent subgraph mining [Chen *et al*., (2007); Cook and Holder, (1994)]. In other words, we can say, a subgraph is a frequent if its frequency in a given dataset is no less than a minimum support threshold. Support of a graph is the number of times it appears in the given graph.

3. Types of anomalies

In general, anomaly is defined as deviation from a rule what is regarded as normal. In case of social network it is an unexpected behavior of a user. Anomalies are classified into different types depending upon the nature and behavior of user as shown in figure 5.

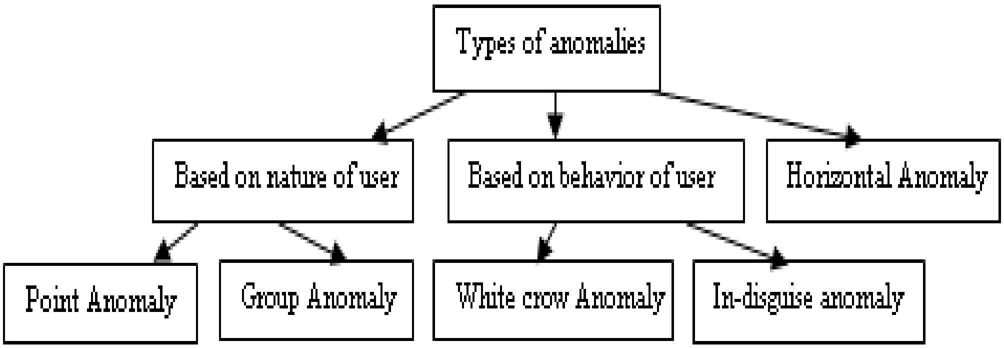


Fig.5. Types of anomalies on the basis of different parameters.

Point anomaly is the identification of the abnormal behavior of individual user. For example, unusual file access, and abnormal network communications. Group anomaly denotes the unusual patterns of group of people. For example, a group of users creating false product reviews [Chandola *et al*., (2009)]

In white crow anomaly, the data is deviated very much from the normal observations. For example, if we are analyzing records of a person and in the age field it is entered as 250, then this kind of deviation is white crow anomaly, because practically this is not possible. It can be detected by analyzing nodes, edges or subgraphs.

In case of in-disguise anomaly, only a minor deviation is observed from the normal pattern. For example, when someone tries to peep into someone's social account, in this case user pretends to behave like a normal user. It is detected through uncommon nodes or entity alterations [Chandola *et al*., (2009)].

In horizontal anomalies, a user completely depends on the different sources interacted by them. For example, a user might be in different social networks and may have completely different types of friends in different social networks. This type of behavior can be considered as anomalous [Gao *et al.,* (2013)].

4. Characteristics of anomalies in social network

In social network, anomalies are identified by analyzing the pattern of interaction between individuals, timestamp of interactions between individuals and the properties of individual involved. Therefore, it requires analyzing different aspects of data such as nature of input network, type of graph, and the type of anomalies generated in order to detect anomalies. The following subsections describe different characteristics of anomalies in social networks as shown in figure 6.

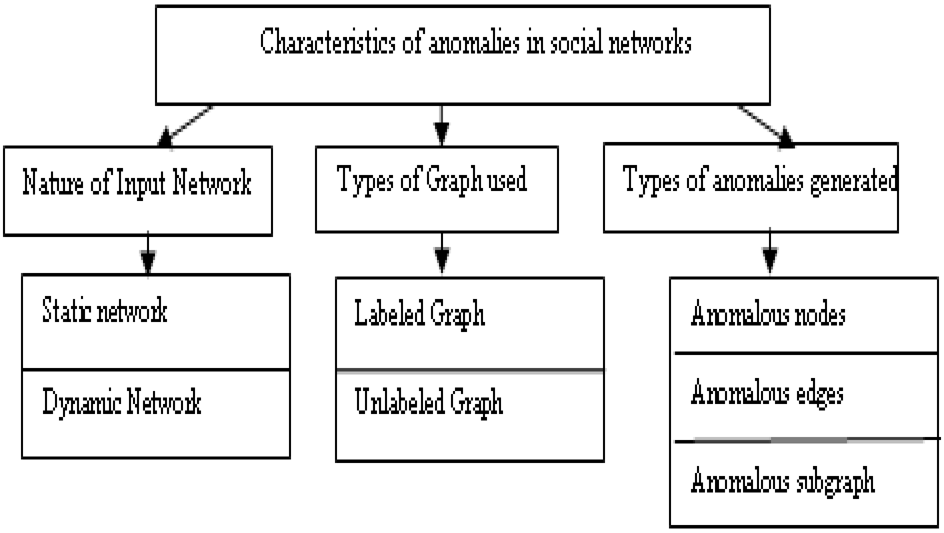


Fig.6. Different characteristics of anomalies in social network.

4.1 Nature of input network

Based upon the network structure being used, network is divided into static and dynamic networks. Static network are those network in which connectivity between nodes is defined by the existence of physical links. In static network, only the current behavior of the nodes is analyzed. The only fact considered is, that interaction has occur. Static network, allows the changes to happen slowly after some time [Chandola *et al.*, (2012); Kriegel *et al.*, (2010)]. For example, in an online social network such as facebook, the behavior of normal people or we can say, the normal people follows some common pattern which deviates significantly from the pattern followed by malicious or anomalous user. Therefore, one can easily extract the hidden pattern by looking at the relationship of user. In spite of providing false information, user cannot hide link they have established with other users. This case is also available to other online platform such as online banking, online examination, online courses, etc.

In dynamic network, changes occur continuously over time. Social network are dynamic in nature. If we consider the example of social networking sites say facebook, we may add new friends, or delete existing friends; accordingly edges must be added or deleted. Therefore anomalies in a dynamic network can be detected by analyzing the pattern by considering different timestamps [Aggarwal and Subbian, (2014)].

4.2 Types of graph used

Based on the information available in a graph, the graph is divided into labeled/attributed graph and unlabeled/unattributed graph.In labeled/attributed grapd, anomalies are detected by considering not only the network structure but also the properties associated with the nodes/edges. However, in unlabeled/unattributed grapd, anomalies are detected by considering only the network structures.

4.3 Types of anomalies generated

Since social network is represented in the form of a graph and therefore anomalies in social network are categorized into anomalous node, anomalous edge, and anomalous subgraph as shown in figure 7. The categorization has been done on the basis of output produced by the different existing anomaly detection method.

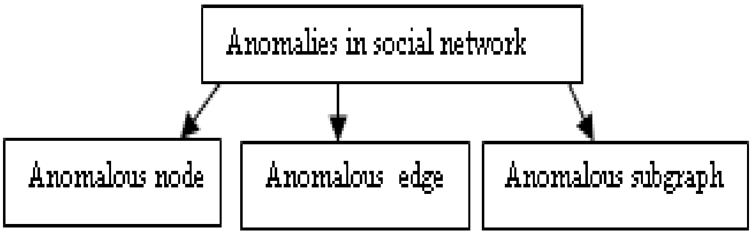


Fig. 7. Types of anomalies in social network.

4.3.1. *Anomalous Nodes*

Anomalies in a node aim to detect subset of vertices whose behavior significantly deviates from the usual behavior in a network. For example, a user who sends messages to a random set of users generally represents a star or near star like pattern, since messages is sent from the single source.This type of behavior can be categorized under anomalous nodes[Hassanzadeh *et al.*, (2012)].

4.3.2. *Anomalous Edges*

Anomalies in edges are used to identify subset of edges that has irregular interaction or dense connections among them. For example, some users communicating more or less frequently than usual in a network [Heard *et al.*, (2010)].

4.3.3. *Anomalous Subgraph*

Anomalous subgraph aims to detect sub network whose pattern of interactions among the nodes is somehow irregular as compared to other nodes in a network. For example, a group of users who try to give fake reviews for some products [Pandit *et al.*, (2007)].

5. Methods used for anomaly detection in social network

Many researchers have proposed different methods for anomaly detection in social networks such as as community based, structure based, probability based, etc as shown in figure 8. The methods are implemented according to the type of network , type of graph and the type of anomalies being generated.

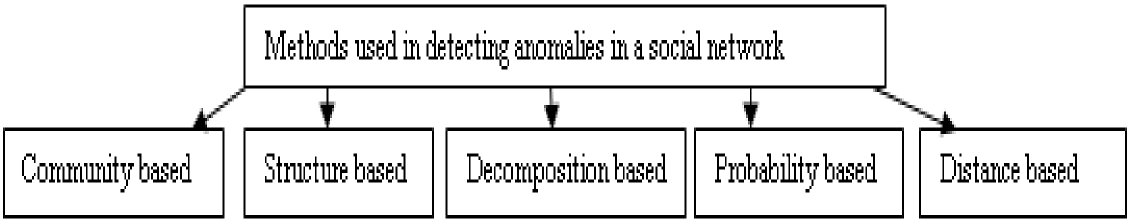


Fig.8. Types of method used in detecting anomalies in social network.

5.1. Community based method

The aim of community based approach is to find the densely connected nodes which form cluster or communities and detect the nodes or edges that have interconnectivity among different communities. Many of the researchers tries to find communities as well as anomalies simultaneously while other first identify communities form and at the second step anomalies is detected. Community based approach in static labeled graph aims to detect those nodes in the graph whose attribute value deviate significantly from other member of the specific group they belongs to, also called as outliers or community outliers. In dynamic network, community based method aims to detetct communities based anomalies between suceding graph.Chen et al. identifies six types of communities based anomalies : grown community, shrunken community, merged community, split community, born community and vanished community[Chen *et al.*, (2012)]. Community based approach used to detect node that do not follow the common trend of the communities[Sun *et al.*, (2005)].

5.2. Structure based method

In structure based method, the given network structure is analyzed by computing the graph specific features, such as node degree, centrality, ego-net, etc [Henderson *et al.*, (2011)]. There are two different structure based approach used for detecting anomalies they are feature based approach and proximity based approach. In feature based approach, the graph representation is used to extract graph centric features to compute various measures associated with the nodes, dyads, triads, egonet, etc. The node level features used to compute centrality measures such as eigenvector, betweeness, closeness, etc. Proximity based approach exploits the graph structure to measure closeness of objects in the graph. In this approach, the importance of node is measured, using page rank algorithm. Page rank is an algorithm used by Google search engine results. It works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. In static labeled graph, structure based approach are used to identify substructures that are rare in network with respect to attribute and network structure.

5.3. Probability based method

In probability based method, previous graph snapshots are used to model the “normal” behavior of a graph constructed on the probability theory,scan statistics. Then all the incoming graphs are compared against the model graph to specify normal or anomalous behavior [Priebe *et al.*, (2005)].Probability based approach uses link prediction techniques for detecting anomalous edge. Link prediction is a method for predicting links between network structure and attributes of nodes and connections. It provides a similarity score for each of the none existing link and higher scores means high possibility of appearing that link.

5.4. Distance based method

In distance based method, similarity or difference is measured. Two objects are similar if they have minimum difference. The main idea of this approach is to calculate the similarity/difference of the evolving entities such as, edge weight, if the similarity/difference is large as compared to the normal evolution, then the entity is considered as anomalous [Mongiovi *et al.*, (2013)].

5.5. Decomposition based method

The decomposition based method detects temporal anomalies by resorting to matrix or decomposition of the time evolving graph. The method can be divided into two categories based on the representation of the graph: matrices vs. tensor [Koutra *et al.*, (2012)].

6. Issues and challenges with various anomaly detection algorithms

Anomaly detection algorithm has been categorized on the basis of different characteristics of anomalies such as nature of input network, types of graph, and types of anomalies generated.Anomaly detection is a very complex process and therefore various issues and challenge are encountered while detecting it.The following sub section describes various algorithms used for detecting anomalies and there related issues and challenges.

6.1. Algorithms for anomaly detection in static unlabeled graph

In static unlabeled graph, anomalies are identified by analyzing the network pattern while ignoring all details. The aim of static unlabeled graph is to uncover anomalous node, edge, or sub graph as shown in figure 9. The methods used by static unlabeled graph to detect anomalies are community based and structure based. Following are the algorithms developed by various researchers to detect anomalies in static unlabeled graph:

- **SCAN** - Structural clustering algorithm for network (SCAN) is an algorithm proposed by Xu et al. to detect node anomalies [Xu *et al*., (2007)]. It is a density based network clustering algorithm to identify cluster, hubs, and outliers. It group vertices based on a structural similarity measures. Instead of considering only direct connections it also considers the way they share neighbors. Because neighborhood around two connected vertices is also important. Vertices u and v are structural similar if (u,v)>=E and u is a core vertex if it has at least μ structurally similar neighbors, where μ is the minimum cluster size threshold and E is the minimum similarity threshold. SCAN is fast as compared to other graph clustering algorithm.

  Issues and challenges

  1.It is difficult to apply SCAN to a large graph because of its high computational complexity.
  2.It is deppendent on minimum threshold parameters, which is difficult to select.

- **GSkeletonClu** – To overcome the difficulty of selecting minimum threshold parameter existed in SCAN, Xu et al. and Sun et al. proposes a framework called GSkeletonClu. GSkeletonClu also aims to find hubs and outliers using graph clustering technique but it automates the way of selecting minimum threshold parameters[Xu *et al*., (2007); Sun *et al*., (2010)].

  Issues and challenges

  1.It is slower than SCAN.
  2.It cannot be applied to a larger network.

- **AUTOPART**– In case of large, sparse graph such as web graphs or social graphs where new nodes/edges are frequently changing i.e. sometimes new nodes/edges are added whereas sometimes existing edges/nodes are deleted from the graph. In such situation, we need some automatic process that should be able to figure out the number of node groups by itself and should allow incremental update so that anomalies can be detected efficiently. To overcome the situation, Chakrabarti proposed a method AUTOPART, which is a parameter free algorithm and allows incremental update to detect anomalies in a social network [Chakrabarti, (2004); Chakrabarti *et al*., (2004)]. This algorithm automatically partitions the network into clustered by reorganizing the rows and columns of the adjacency matrix. The edges that do not belong to any cluster as well as the edges that have interconnectivity between different clusters are reported as anomalies. To cluster the network, this algorithm uses the concept of Minimum Description length (MDL) principle. MDL is used for reorganizing the adjacency matrix of the graph into homogeneous blocks. The homogeneous block consists of “similar nodes”. The main advantage of this algorithm is user does not need to set any parameter.

  Issues and challenges

  1.It is computationally expensive for the large graph.
  2.This algorithm consider reassigning only one node at a time, if there exists a situation where moving one node only increases the cost. Although moving two nodes together might decreases the cost but there is no guarantee of algorithm to work properly.

- **Non Negative Residual matrix factorization** - Non negative residual matrix factorization (NrMF) is an algorithm proposed by Tong and Lin to spot anomalous nodes or edges in a large graph based on graph communities [Tong and Lin, (2011)]. Matrix factorization is a tool to find patterns such as communities, anomalies, etc from the large set of a graph represented by adjacency matrix A. The adjacency matrix is decomposed by multiplying the two low rank matrices F&G and some residual matrix R. The low rank factor F&G shows the communities structure and residual matrix R indicates anomalies in a network.

$$\underbrace{A}_{\text{Adjacency matrix}} = \underbrace{F\ X\ G}_{\substack{\text{low rank matrices}\\ \text{(Represents communities)}}} + R \longrightarrow \text{Residual matrix (Represents anomalies)} \tag{1}$$

Residual matrix factorization is improved by interpreting non-negativity. Non-negativity matrix factorization is carefully designed by keeping in mind two things. The first one is to achieve accuracy in anomaly detection results. Since, one of the main challenges in detecting anomalies is lack of ground truth, which is achieved by non-negative residual matrix factorization. The second one is scale up with the growing data with respect to the size of the graph.

Issues and challenges

1. NrMF can only be applied to non-negative data.
2. It requires initializations of parameters.

- **OddBall** - is a technique proposed by Akoglu et al. that aims to detect anomalous node based on the analysis of ego network [Akoglu *et al*., (2010)]. This technique mainly focuses on the nodes which are densely connected(near clique) or the nodes which are sparsely connected(near star) that are particular in social network, the previous indicates a regular and strong interaction, the later suggests a person in a central position, who is capable of reaching to a wide independent audience. This technique works by extracting all ego networks from the input graph. Then selecting features of ego network such as number of triads, total weight of edges, etc that could be a sign of anomalies. Triad is a social group consisting of three peoples. After that, pointing out anomalies by using any outlier detection method. Mainly power law is used to detect deviation from the normal pattern. To predict near star/clique in twitter data set, Rezaei et al. uses the same approach of oddball to analyzed number of nodes and edges [Rezaei *et al*., (2017)].

  Issues and challenges

  1. OddBall is used only under static network, future work could derive Oddball to consider dynamic network also.

- **Random walk with restart –**Random walk with restart is an algorithm proposed by Sun et al. [Sun *et al*., (2005)]. They uses bipartite network to identify anomalous node and addresses two primary problems:-

  1. Neighborhood formation
  2. Anomaly Detection

     To solve the problem of neighborhood formation, sun et al. proposed an algorithm called random walk with restart and graph partitioning algorithm to find the neighborhood of each node in a bipartite network and uses this algorithm to detect anomalous nodes in a network. Random walk with restart is widely recognized as one of the most important node proximity measure for a graph, as it capture the whole graph structures and is robust to noise in the graph. Graph partitioning is a technique to divide the graph into number of disjoint sub graph called clusters.

     Issues and challenges

     1. It is highly dependent on the choice of similarity measures.
     2. It does not scale well to a large graph.

- Using graph properties of a node's ego net such as number of nodes, number of edges, betweeness centrality, etc Hassanzadeh et al. proposed a framework to identify anomalous nodes in social network[Hassazadeh *et al*., 2012]. Neighborhood information was retrieved using both node and ego net based features and behavioral information was extracted using recursive features.

- **Page rank algorithm** – Page rank is an algorithm proposed by Brin and Page, they uses the concept of page rank algorithm in order to detect anomalies in a static network[Brin and Page, (1998)]. Page rank algorithm is one of the most popular algorithms based on random walk. Random walks on the graph start from any random node and then from each node pick a random link to follow. This algorithm work by taking input a set of unlabeled data points and determines a ranking of node which are anomalous.

Issues and challenges

1. It cannot be applied to a large set of graph
2. Determining ranking of a node is a difficult task.

- **RLA(Random Link Attack)** - In some of the fraudulent activities, such as emails, viral marketing in a social networks, a group of users create a set of fake identities and uses these identities to interact with a large random set of innocents users. To detect these types of activities, Shrivastava et al. defined Random Link Attack (RLA) to model such malicious activities and proposed an algorithm to mine subgraph satisfying the RLA property [Shrivastava *et al.*, (2008)]. For detecting random link attack, Greedy and TRWalk technique is used .The main idea of this approach is to count external triangle around each node. A triangle is a clique containing three nodes. Neighbor of a regular users have many triangles, but random victims do not. This algorithm works in two steps: In first step, it identifies the suspect by conducting two tests i.e. clustering and neighborhood independence test on each individual node. In order to detect the suspect nodes in the network, the triangle in each ego net are counted, with a lower triangle count indicating an attackers. In the second step, the attackers set are identified by growing the neighborhood of the suspect nodes.

Issues and challenges

1. It requires too many parameters.
2.It invlove high computational costs.
3.This algorithm does not work properly when there exists multiple RLA's.

In figure 9 we have showed the various algorithms developed for static unlabeled graph and there categorization on the basis of type of anomalies and the type of methods used for implementation.

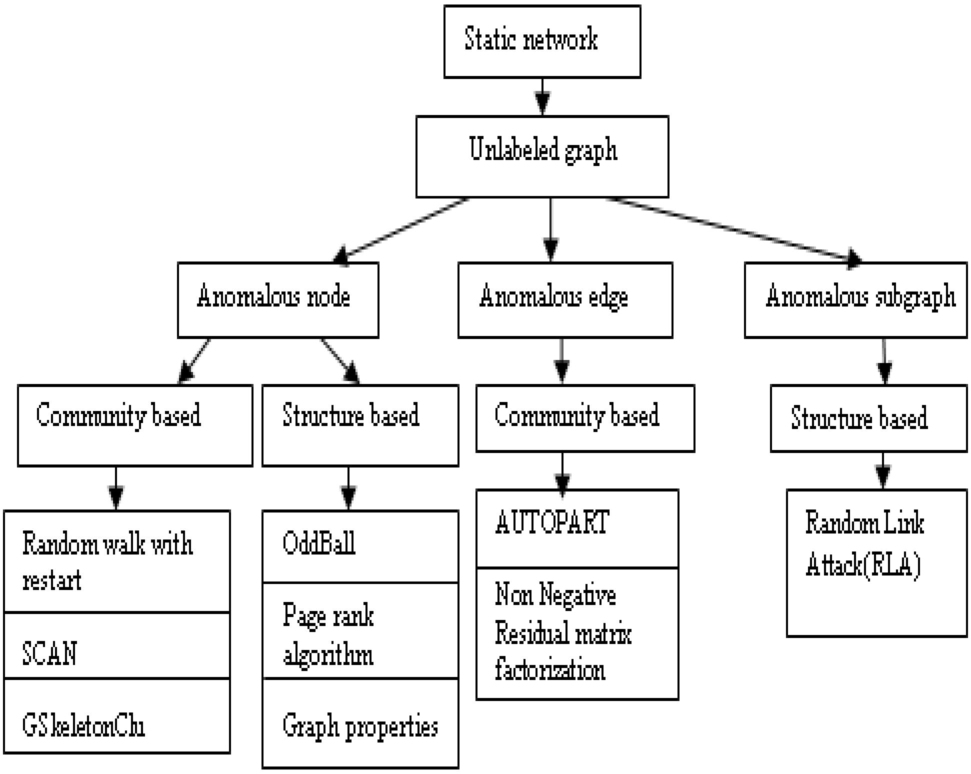


Fig.9. Algorithms developed for static unlabeled graph.

6.2. Algorithms for anomaly detection in static labeled graph

In static labeled graph, anomalies are detected by analyzing the network structure as well as the information gathered from nodes/edges. The aim of static labeled graph is to uncover anomalous nodes and sub graph as shown in figure 10. The methods used by static labeled graph to detect anomalies are community based and structure based. Following are the algorithms developed by various researchers to detect anomalies in static labeled graph:

- **CODA**- Community outlier detection algorithm (CODA) is an algorithm proposed by Gao et al. that integrated the two steps process of identifying communities and detecting anomalies to find more meaningful outliers [Gao *et al.*, (2010)]. This algorithm uses the concept of probabilistic model to detect normal data and outliers. The information at each object is collected by generating a mixture model and network information is generated by hidden markov random field (HMRF) model. Hidden variables are random variables whose values are unobservable. This approach uses two types of data i.e. continuous data and test data. For continuous data, gausian mixture is used and for text data, multinomial distribution is used.

Issues and challenges

1. CODA has poor scalability and high sensitivity to hyper-parameter choice and initialization.

2. CODA discovers communities by using only node and data linkage while ignoring the edge attributes. If we consider the edge details too along with node it might leads to meaningful communities.

3. CODA does not work well with increasing number of attributes.

- **GoutRank**- is a node outlier ranking technique proposed by Miller et al. [Miller *et al.*, 2013]. The aim of this algorithm is to reveal complex anomalies, which is possible through a subset of relevant attributes. It take nodes attributes and graph structures as input and output ranking of all nodes ordered by deviation. The nodes having low score are considered as outliers and nodes having high scores are considered to be normal objects.

  Issues and challenges

  1. Calculating the score of each node so that we can rank it accordingly is a very complex task.

  2. Selection of relevant subspaces and subgraph is tedious.

- **Non-negative matrix tri-factorization–**It is a framework proposed by Yang et al. for detecting anomalies based on bipartite graph and co clustering algorithm. It mainly focuses on detecting anomalous behavior in micro blogging [Yang *et al.*, (2015)]. Co-clustering is based on non negative matrix tri-factorization that can able to detect anomalous user and messages simultaneously. This algorithm works in three steps: In the First step, a bipartite graph between user and messages are created to model homogeneous and heterogeneous interaction. In the second step, heterogeneous interaction and homogeneous interaction are integrated using distances metric learning. At the final step, users and messages are co-clustered based on co-clustering algorithm non negative matrix tri-factorization.

  Issues and challenges

  1.Selecting number of clusters is a big issues because high values lead to ovelapping communities and low values leads to number of smaller communities.

- **FocusCo**-is a method proposed by Peroozi et al. to detect outliers in large attributed graph with node attributes [Peroozi *et al.*, (2014)]. Given an initial set of nodes provided by a user as an example of the kind of similarity they are interested in ,then the algorithm identifies subset of attributes, that the given nodes agree on which is called" focus attributes" and then find clusters of densely connected node which agree with the focus attribute called "focused cluster". Based on the focused cluster, nodes deviating in focus attributes values are identified and called as "focused outliers".

  Issues and challenges

  1. It focuses on detecting complete anomalies, ignoring the fact that nodes can also be partially anomalous.
  2. FocusCo is a semi-supervised method therefore it needs labeled data as examples of similar nodes.

- **SUBDUE-** is an algorithm proposed by Noble and Cook [Noble and Cook, (2003)].They defined anomaly as "anomalous substructure occurring frequently when compared to normative substructures." They address two problems while detecting anomalous substructure. The first problem is "finding unusual substructure" and the second problem is "finding unusual subgraph". They solved the above problem by using an approach called SUBDUE System and introduced two methods for identifying unusual pattern in a network. In the first method, unusual substructures are identified and in the second method, networks are partitioned into unique separate subgraph to detect anomalous subgraph by comparing each of the subgraph against other in order to find unusual occurrence. The main aim of their approach is to find substructure that occurs infrequently. SUBDUE is used to discover frequent substructures, or subgraph that compresses well the input graph. It compresses the graph by replacing subgraph with pointers. SUBDUE uses the minimum description Length (MDL) principle to identify pattern minimizing the description length of the entire graph compressed with the pattern.

  Issues and challenges

  1. It is limited to only categorical attributes.

- **GBAD**-Eberle and Holder in 2007 proposed an algorithm called GBAD (graph based anomaly detection) which is a suite of three anomaly detection algorithm based on the concept of SUBDUE system [Eberle and Holder, (2007)]. GBAD-MDL aims to detect anomalous modification; GBAD-Probability (P) aims to detect anomalous insertions; GBAD-Maximum Partial Substructure (MPS) aims to detect anomalous deletions. This algorithms work on unweighted graph and discrete vertex/nodes and edge label. It ignores numeric attributes in a graph. Limitations of this algorithm are overcome by Davis et al. who proposed YAGADA(yet another graph based anomaly detection algorithm)[Davis *et al.*, (2011)]. It detects anomalies by considering both, structural information, and numeric attributes.

Issues and challenges

1. It ignores numeric attributes.

- **SODA**-Various work has been done in the area of detecting subgraph outliers, but they only focus on detecting outliers for the whole network or in a community. To overcome the limitation, Gupta et al. in 2014 proposed a query based algorithm called as subgraph outlier detection algorithm (SODA)[Gupta *et al.*, (2012); Gupta *et al.*, (2014)]. Instead of taking the global perspectives, this query based algorithm gives user a flexibility to find outliers in the form of a query. This algorithm takes input a query and returns the matching anomalous subgraph sorted on their outlinerness score from the original graph. This algorithm uses SPath to compute all the matches for a query. A subgraph can be anomalous based on connectivity structure within itself and with neighborhood. An anomalous subgraph may have many unexpected edges and missing many expected edges. To capture the features of link existence, the author developed an optimization model. It uses the information within one-hop neighborhood to infer feature weight. The optimization is also used to calculate the outlier score apart from learning feature weights.

Issues and challenges

1. It involves high computational cost.

In figure 10 we have showed the various algorithms developed for static labeled graph and there categorization on the basis of type of anomalies and the type of methods used for implementation.

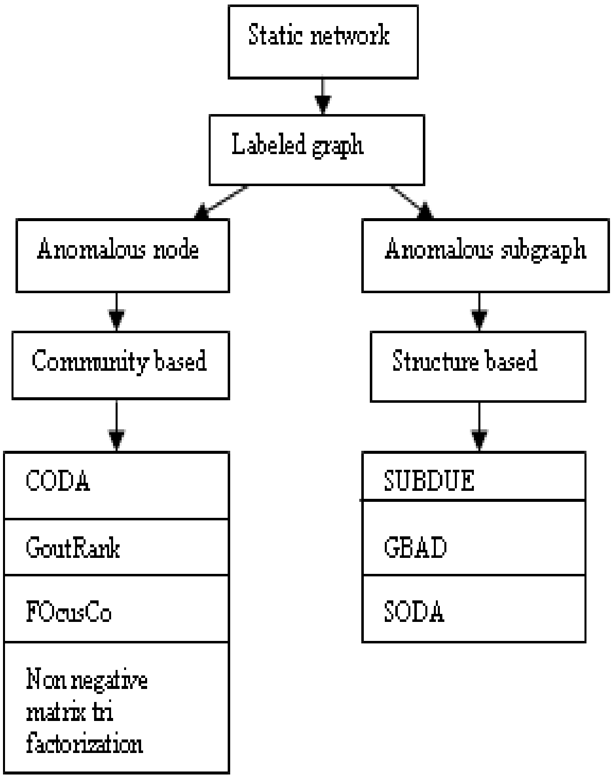


Fig.10. Algorithms developed for static labeled graph.

6.3. Algorithms for anomaly detection in dynamic unlabeled graph

In dynamic unlabeled graph, anomalies are detected by considering the changes occur in the structure of the network. The aim of dynamic unlabeled graph is to uncover anomalous nodes, edges and sub graph as shown in figure 11. The methods used by dynamic unlabeled graph to detect anomalies are community based, decomposition based, probability based and distance based. Following are the algorithms developed by different researchers to detect anomalies in dynamic unlabeled graph:

- **ECOutliers(evolutionary community outliers**) – is a concept proposed by Gupta et al. to detect temporal ouliers[Gupta *et al.*, (2012a); Gupta *et al.*, (2012b)].Temporal outliers are detected by identifying objects that have different evolutionary behavior compared to other. Evolutionary community outliers are the outliers that evolve differently with respect to other communities members.Evolutionary outliers are detected by considering multiple snapshot of the given discovered communities.Example of ECoutlier could be a author changing his research areas across time while other community member continues in same area. One of the way to identify ECoutlier is to detect node that changes their community without following trends across time.But it may suffer from the ECoutlier itself. Outlier detection and community matching cannot be seperated. Therefore, the author proposed an algorithm to integrate community matching and outlier detection. It generates outlierness score for each node after matching communities from sequential snapshot.

Issues and challenges

1. Community trend pattern and outliers are not known before.
2. Another challenge is to define outlier score, after the pattern are discovered.

- **COM2**- is a novel and fast tensor analysis method proposed by Araujo [Araujo, (2014)]. The advantages of this method is : (i)it is scalable and (ii)needs no user defined parameters. The main focus of this method is to find time varying communities , also refer as a 'comet'. It is so named because this type of communities comes and goes like a comet. The method COM2 works in three steps :

  Step 1: In this step, candidate communities are spoted by using PARAFAC decomposition. The output of this step represents score of source, destination, and timestamp.

  Step 2: In this step, the scores obtained from step1 is used to find for important communities. Then , the correct size of the communities is determined by using minimum description length (MDL) principle.

  Step 3:In this step, based on the communities detected, author uses the tensor deflation principle to find novel communities.

  Issues and challenges

  1.COM2 can be applied only to edge-labeld graph, future work can be extended to exploit side information like node attributes also.

- **DBMM**–Dynamic behavior mixed-membership model is an algorithm proposed by Rossi et al. which is fully automatic, non parametric and flexible to detect anomalous node and role of node in the graph[Rossi *et al*., (2013)]. 'Role' of a node refers to whether the node is forming star like patter, or clique. This model focus on detecting behavior of a node rather than finding communities. DBMM uses minimun description length(MDL) and non matrix factorization(NMF) to automatically select the structural role of a node. Transitional model is generaterd to predict changes in next time stamp.

  Issues and challenges

  1.This algorithm is not scalable,i.e. not suitable for analysis of large graph.

- **Scan Statistics** - also known as moving window analysis is an algorithm proposed by priebe et al. to detect anomalous nodes [Priebe *et al*., (2005)]. The idea of scan statistics is to scan a small window over data, calculating local statistic for each window. The maximum value of the local statistics, calculated is known as the scan statistics, denoted by M(x). Enron database uses the method scan statistics on weekly data to detect network instance with an unusually high connectivity compared to the past. If the maximum value of local statistic is larger than a threshold value, then the network instance is considered as anomalous.

  Issues and challenges

  1. The result of scan statistics are sensitive to the parameter settings.

- **NetProbe**–is a method proposed by Pandit et al. is a fast and scalable emethod for detecting fraud in online auction network [Pandit *et al*., (2007)]. NetProbe is a system that analyzes transaction within user of auction sites to spot doubtful fraudster. It imitates auction user and transaction as a Markov Random Field to detect suspicious pattern and uses belief propagation mechanism to detect fraudsters. Fraudsters in an auction network can be identified by detecting bipartite cores or near bipartite cores. Fraudsters can create two types of identities: fraud and accomplice. Fraud identities are those who actually carried out the fraud, whereas accomplices are those who only help the fraudster to carry out their job. Accomplice themselves behaves like a normal user but also interact with the fraudster creating a bipartite core, to help fraudster to gain high feedback rating. Therefore, fraudster can be detected by spotting bipartite cores.

  Issues and challenges

  1. It is difficult to differentiate between normal users and anomalous users because fraudsters generally behave like normal users.

- **CMD**–By using matrix decomposition Sun et al. proposed compact matrix decomposition (CMD) to calculate sparse low rank matrix approximation [Sun *et al*., (2007)].

Issues and challenges

1.It suffers from high computational cost and memory requirements.

- An algorithm based on Eigen behavior analysis to detect events and the equivalent anomalous nodes in a network is proposed by Akoglu and faloutsos [Akoglu and faloustos, (2010)]. Anomalous node can be detected by comparing their behavior with their previous behavior.

  Issues and challenges

  1.It is difficult to maintain the history of dynamic updates.
- **PCA**- Yu et.al designed a localized principal component analysis (PCA) algorithm that can continuously maintain information about the changes in different neighborhoods of the network [Yu *et al.*, (2013)].
- **Issues and challenges**

  1. PCA suffers from problem of over fitting of datasets.
- **Link prediction -** Link prediction model is proposed by Huang and Zeng to detect anomalous email[Huang and Zeng, (2005)]. The detection framework uses the following data as input: set of email accounts and a list of tuples, capturing a specific email message with its sender, recipients, and its delivery time. The output produced is an anomaly score denoted by A(θ , t) for each distinct sender-recipient tuple θ. The probability of the maximumlikelihood of the possible link is calculated by Expectation Maximization(EM). Based on the probability potetntial score is given. The email with low score is considered as anomalous.

  Issues and challenges

  1. It requires too many parameters.
  2. It involves high computational costs.
- **Stream based algorithm** -is proposed by Aggarwal et al..Stream based outlier detection algorithm is used to identify unusual bridging edges[Aggarwal *et al.*, (2011)]. The author proposed a probabilistic algorithm that are used to maintain structural summary of the network.In order to capture the structural summary, node is partitioned and to dynamically maintain the partitioning , structural reservoir sampling approach is used. To define anomalous behavior of an edge, likelihood fit of an edge is calculated using the structural generation model. The likelihood probabilities of new incoming graph is calclated using the stored probability.Then the likelihood probability is compared, if the probability is below taverage of all the graph recieved, the network is considered as anomalous.

  Issues and challenges

  1.It is difficult to maintain real time structural summary of the network.
- **NetSpot**–is developed by Mongiovi et al. to discover anomalous subgraph [Mongiovi *et al.*, (2013)]. The method first identifies all the significant anomalous regions (SAR), in different network areas and time span.SAR is an NP hard problems also known as heaviest dynamic subgraph. Heaviest dynamic subgraph aims to find connected subgraph whose sum of its edge weights is highest. Anomalous outcome of each edge is calculated by observing its statistical p-value, if its p-value is lower than the edge is considered as anomalous. NetSpot alternates between graph space and time domain by computing heaviest subgraph and maximum score. The algorithm then produces output, the area of a network whose anomaly score is higher than a given threshold value.

  Issues and challenges

  1. It suffers from computational complexity i.e. computing the heaviest subgraph and maximum subsequence for each edge.
- **Thompson and Eliassi-Rad** – uses a probability based approach to identify anomalous subgraph [Thompson and Eliassi-Rad, (2009)]. They first construct a dynamic graph based on time stamped edges and called this graph as “cumulative graph”, that confine all past edges but gives more weightage to the recent edges.. Then using this graph they try to find persistent pattern. Persistent pattern is created by extracting connected components of edges that interact regularly and whose weights are above threshold value. These persistent patterns act as a basis for normal behavior and are utilized to detect anomalous behavior in a network. Anomalous behavior is detected by comparing the current activity at a particular time and expected normal activity based on recent behavioral pattern. An event is considered as anomalous if differs remarkably from the expected activity.

  Issues and challenges

  1**.** Selecting minimum threshold values is tedious, as it depends upon various components.
- **chen et al.** identify six types of commuities based anomalies that can arises in a dynamic network[Chen *et al.*, (2012)]. The six types of communities based anomalies are : grown community, shrunken community, merged community, split community, born community and vanished community. The author proposed a techniques, graph representative and community representative, to detect these types of communities based anomalies. Graph representative helps to reduce the computational cost and communities representaive is used to identify communities based anomalies. Community are

maximal clique in a network. Graph representative is a representatives of nodes that are present in preceding and succeeding graph. Anomalies are detected by comparing graph of different timestamp.

- **Issues and challenges**
  1. This methods needs to know beforehand the number and size of communities before graph partitioning.

In figure 11 we have showed the various algorithms developed for dynamic unlabeled graph and there categorization on the basis of type of anomalies and the type of methods used for implementation.

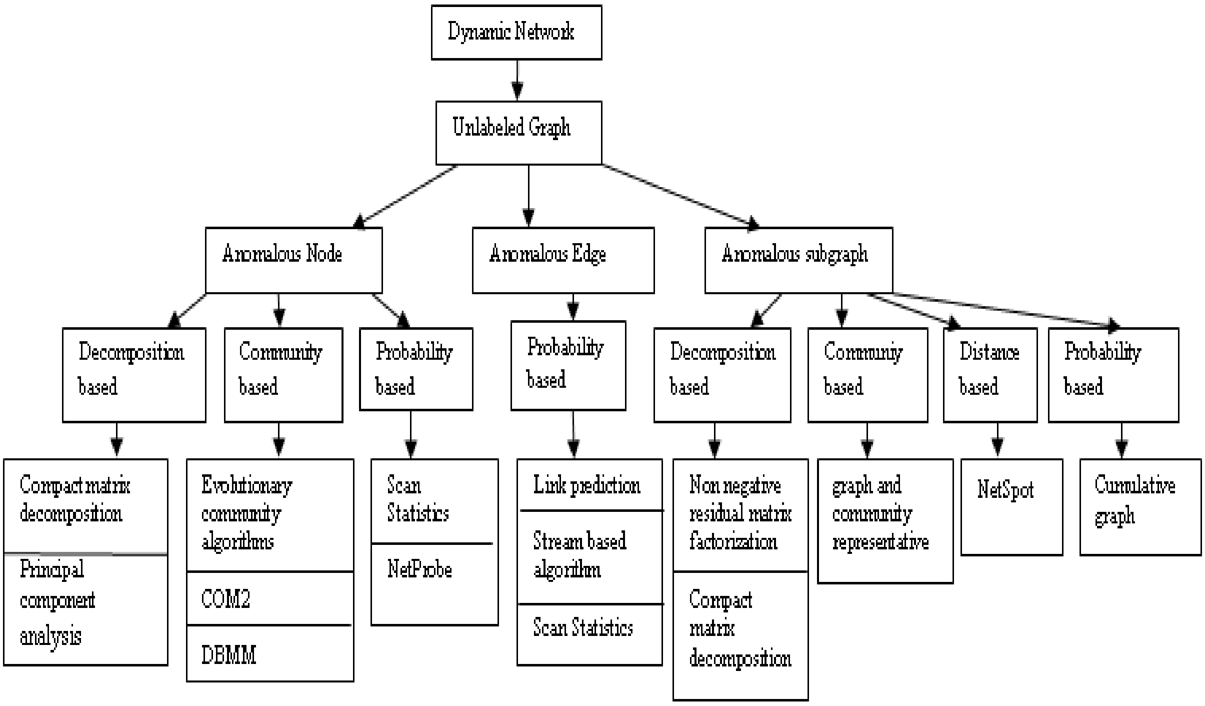


Fig. 11. Algorithms developed for dynamic unlabeled graph.

6.4. Algorithms for anomaly detection in dynamic labeled graph

In dynamic labeled graph, properties associated with the nodes or edges are considered to detect anomalies in addition to the changes in network structure over time. Not much work has been done in the field of dynamic labeled network, only few papers are available demonstrating the work as shown in figure 12.

- ParCube- Paplexakis et al. proposed a method called Parcube which is decomposition based approach [Paplexakis *et al.*, (2012)]. ParCube uses a tensor decomposition to represent the data using sparser latent factors. ParCube is a method that can handle any types of graph and can process graph that do not fit in memory. The method works by first selecting the original tensor and then performs tensor decomposition on the slices selected by original tensor. After that they merge the decomposition to obtain the outcome.This method detects the anomalous node by following the unusual pattern recorded by factor of tensor decomposition.
- Bayseian method- proposed by Heard et.al is a probability based method used to detect anomalous nodes, edges, and subgraph[Heard *et al.*, (2010)]. This method works in two stages. In the first stage, anomalous node is identified and in the second stage a subgraph is established. The subgraph is erected in a way that it includes the set of nodes identified in the first stage as well as the nodes which have recently communicated with this set of nodes. After identifying the reduced subset of interesting nodes, a network tool such as spectral clustering are used to identify the cluster rather than identifying changes in the structure.

In figure 12 we have showed the various algorithms developed for dynamic labeled graph and there categorization on the basis of type of anomalies and the type of methods used for implementation.

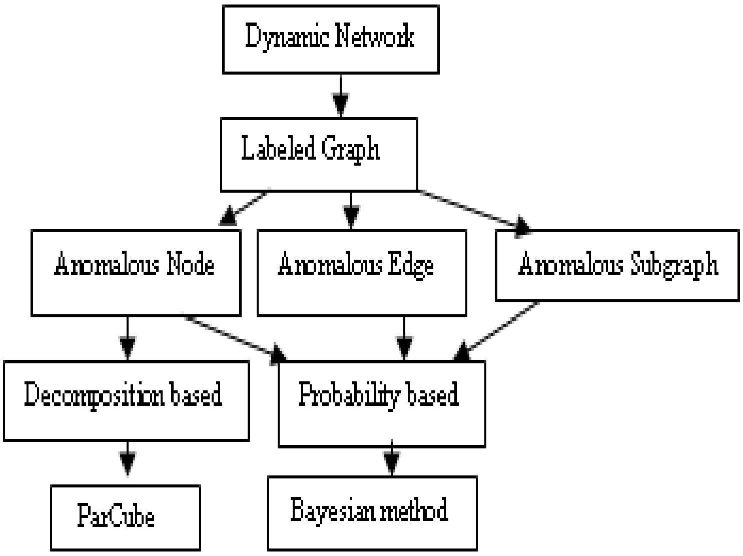


Fig.12. Algorithms developed for dynamic labeled graph.

7. Anomaly Detection Software

Anomaly detection software are used to identify, monitor,and correct anomalies.Following are some of the anomaly detection software that uses the techniques of data mining and machine learning to detect anomalies.

1. Weka Data Mining – It is written in java, developed at the University of Waikato, New Zealand. It includes features like machine learning, data mining, classification, clustering, regression, etc. It is freely available to the user.

2. ELKI –is also written in java and open source data mining software. It included algorithm belong to clustering, outlier detection and database indexes.

3. Scikit-learn – It features various classification, regression, and clustering algorithms including DBSCAN and K-means.

4.Kepler - Kepler a data mining suite, which includes methods for deviation detection.

5. Anodot provides real-time analytics and automated anomaly-detection systems to find outliers in Big Data and transform them into valuable business insights.

8. Conclusion

In this paper, we have reviewed basic concepts of anomalies, types of anomalies, social network,graph mining and various existing techniques to detect anomalies in a social network with issues and challenges in each existing techniques and also provided some example of anomaly detection software that are used in real life to identify anomalies. However, it is not posible to review each and every techniques,we have tried our best to cover the most important techniques. The existing techniques has been categorized on the basis of different characteristics of anomalies in social network as shown in figure 13. Detecting anomaly is a very tedious task, especially when the world is genearting large amount of data in few seconds.In spite of various work done in the area of anomaly detction in a social network, there remains a number of flaws that could be addressed for future work. For example, there are very few work done in the area of anomaly detection in dynamic labeled graph so, there are huge opportunity to develop new techniques/methods under this domain.Another area that can be explored for detecting anomalies in social network is structure based approach that can be use to detect some new types of anomalies in social network in near future.

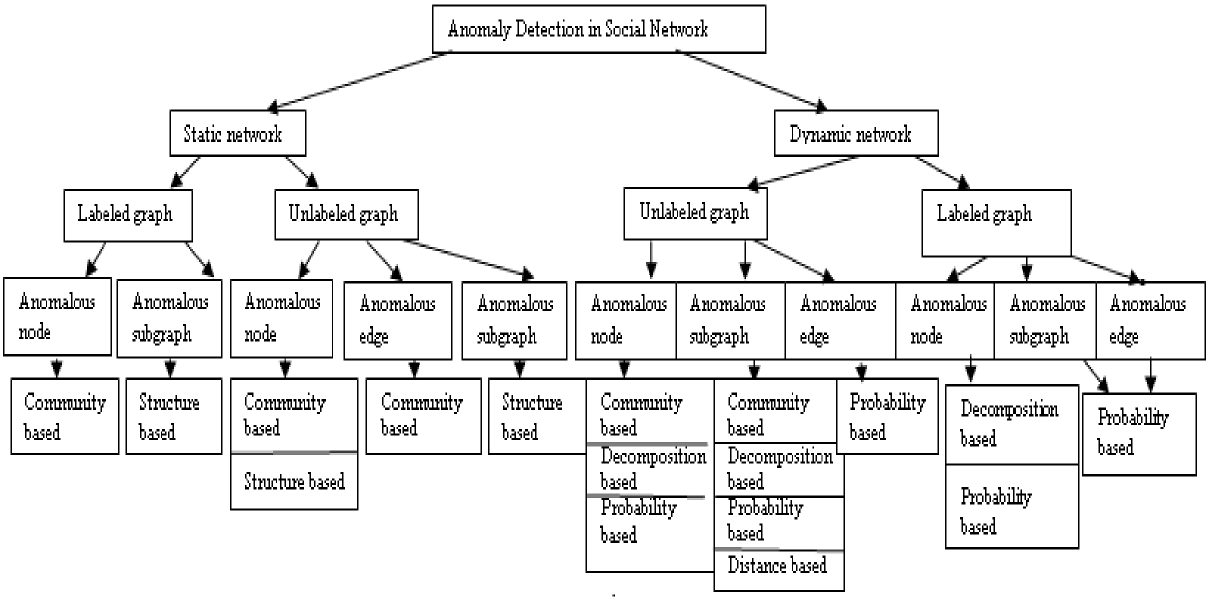


Fig.13. Summary of anomaly detection techniques in social networks.

References

[1] Aggarwal, C.C., Zhao, Y., Yu, P.S., 2011.Outlier detection in graph streams. In: Proceedings of the IEEE 27th International Conference on Data Engineering(ICDE), IEEE, pp.399–409.

[2] Aggarwal, C., Subbian, K., 2014. Evolutionary network analysis: a survey. ACM Comput. Surv. (CSUR) 47 (1), 10.

[3] Akoglu, L., Faloutsos, C., 2010. Event detection in time series of mobile communication graphs. In: Proceedings of the Army Science Conference, Springer, Berlin, Heidelberg.

[4] Akoglu, L., McGlohon, M., Faloutsos, C., 2010. Oddball: spotting anomalies in weighted graphs. Advances in Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, pp.410–421.

[5] Araujo, M.,Papadimitriou,S.,Gnnemann,S.,Faloutsos,C.,.,Swami,A.,Papalexakis,E.E.,Koutra,D.,2014.Com2:fast automatic discovery of temporal (‘comet’).

[6] Chandola, V., Banerjee, A., Kumar, V. Anomaly detection: a survey. ACM Comput. Surv. 2009; 41(3):15.

[7] Chandola,V., Banerjee,A., Kumar,V. Anomaly detection for discrete sequences: A survey. IEEE Transactions on Knowledge and Data Engineering (TKDE), 24(5):823{839, 2012.

[8] Chen, C.,Yan,X.,Zhu,F.,Han,J.,2007.gapprox:mining frequent approximate patterns from a massive network. In: Proceedings of the Seventh IEEE International Conference on Data Mining(ICDM),IEEE,pp.445–450.

[9] Chen, Z., Hendrix,W.,Samatova,N.F.,2012.Community-based anomaly detection in evolutionary networks. J. Intell.Inf.Syst.39 (1), 59–85.

[10] Clauset, A., 2005. Finding local community structure in networks. Phys. Rev. E 72, 026132.

[11] Cook, D.J., Holder,L.B.,1994.Substructure discovery using minimum description length and background knowledge.J.Artif.Intell.Res.,231–255.

[12] Chakrabarti, D., 2004.Autopart: Parameter-Free Graph Partitioning and Outlier Detection. Knowledge Discovery in Databases: PKDD.Springer, pp.112–124.

[13] Chakrabarti, D., Zhan, Y., Faloutsos, C., 2004. R-MAT: a recursive model for graph mining. In: Proceedings of the SIAM International Conference on Data Mining, Vol. 4, pp.442–446.

[14] Dang, X.H.,Assent,I.,Ng,R.T.,Zimek,A.,Schubert,E.,2014.Discriminative features for identifying and interpreting outliers.In:Proceedings of the IEEE 30th International Conference on Data Engineering(ICDE),IEEE,pp.88–99.

[15] D. M. Hawkins. Identifcation of outliers, volume 11.Springer, 1980.

[16] Eberle, W., Holder, L., 2007.Anomaly detection in data represented as graphs. Intell. Data Anal.11 (6), 663–689.

[17] Gao J, Du N, Fan W, Turaga D, Parthasarathy S, Han J. A multi-graph spectral framework for mining multi-source anomalies. In: Graph embedding for pattern analysis. Springer; 2013. P.205–27.

[18] Gao, J.,Liang,F.,Fan,W.,Wang,C.,Sun,Y.,Han,J.,2010.On community outliers and their efficient detection in information networks.In:Proceedingsofthe16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM,NY,USA,pp.813–822.

[19] Gupta, M.,Gao,J.,Sun,Y.,Han,J.,2012a.Community trend outlier detection using soft temporal pattern mining. In: Machine Learning and Knowledge Discovery in Databases,Springer,Berlin,Heidelberg,pp.692–708.

[20] Gupta, M.,Gao,J.,Sun,Y.,Han,J.,2012b.Integrating community matching and outlier detection formining evolutionary community outliers.In :Proceedings of the18th ACM SIGKDD international conference on Knowledge discovery and data mining,ACM,NewYork,NY,USA,pp.859–867.

[21] Gupta, M., Gao, J., Aggarwal, C., Han, J., 2014.Outlier detection for temporal data. Synth. Lect. Data Min. Knowl.Discov.5 (1), 1–129.

[22] Gupta, M., Mallya,A.,Roy,S.,Cho,J.H.,Han,J.,2014.Local learning for mining outlier subgraphs from network datasets. In: Proceedings of the 2014 SIAM International.

[23] Hassanzadeh, R., Nayak,R.,Stebila,D.,2012.Analyzing the effectiveness of graph metrics for anomaly detection in online social networks. In : Web Information Systems Engineering—WISE 2012,vol.7651.Springer,Berlin,Heidelberg,pp. 624–630.

[24] Heard, N.A.,Weston,D.J.,Platanioti,K.,Hand,D.J.,etal.,2010.Bayesiananomaly detection methods for sociall networks.Ann.Appl.Stat.4(2),645–662.

[25] Henderson,K.,Gallagher,B.,Li,L.,Akoglu,L.,EliassiRad,T.,Tong,H.,Faloutsos,C.,2011.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, ACM, New York,NY,USA,pp.663–671.

[26] Jiang, C., 2011. Frequent Subgraph Mining Algorithms on Weighted Graphs (Ph.D. thesis), University of Liverpool.

[27] Keyvanpour,M.,Moradi,M.,Hasanzadeh,F.,2014.Digitalforensics2.0.In:ComputationalIntelligenceinDigitalForensics:Forensic Investigation and Applications, SpringerInternationalPublishing,Switzerland,pp.1746.

[28] Koutra,D.,Papalexakis,E.E.,Faloutsos,C.,2012.Tensorsplat:spotting latent anomalies in time. In: 2012 16th Pan-Hellenic Conference on Informatics (PCI), IEEE, pp.144–149.

[29] Kriegel, H. P., Kroger, P., & Zimek, A. (2010). Outlier detection techniques. *Tutorial at KDD*, *10*.

[30] L. Page, S. Brin, R. Motwani, and T.Winograd, “The Page Rank citation ranking : Bringing order to the web,” Stanford In for Lab, Tech. Rep. 1999-66, 1999.

[31] Liu, Y., Chawla, S., 2015.Social media anomaly detection: challenges and solutions. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp.2317–2318.

[32] Miller,B.A.,Arcolano,N.,Bliss,N.T.,2013.Efficient anomaly detection in dynamic, attributed graphs: Emerging phenomena and big data.In:2013 IEEE International Conference on Intelligence and Security Informatics (ISI),IEEE,pp.179– 184.

[33] Mislove,A.,Marcon,M.,Gummadi,K.P.,Druschel,P.,Bhattacharjee,B.,2007.Mea- surement and analysis of online social networks. In : Proceedings of the7th ACM SIGCOMM Conference on Internet Measurement , ACM , New York , NY, USA, pp.29–42.

[34] Mongiovi, M., Bogdanov, P., Ranca, R., Papalexakis, E. E., Faloutsos, C., & Singh, A. K. (2013, May). Netspot: Spotting significant anomalous regions on dynamic networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 28-36). Society for Industrial and Applied Mathematics.

[35] Newman, M. E. (2012). Communities, modules and large-scale structure in networks. *Nature physics*, *8*(1), 25-31.

[36] Noble,C.C.,Cook,D.J.,2003.Graph-basedanomalydetection.In:Proceedingsofthe Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM,NewYork,NY,USA,pp.631–636.

[37] Pandit, S.,Chau,D.H.,Wang,S.,Faloutsos,C.,2007.Netprobe:a fast and scalable system for fraud detection in online auction networks. In : Proceedings of the 16thInternational Conference on World Wide Web, ACM, NewYork, NY,USA, pp. 201–210.

[38] Papalexakis, E.E., Faloutsos,C.,Sidiropoulos,N.D.,2012.Parcube:sparse parallelizable tensor decompositions , Machine Learning and Knowledge Discovery in Databases. Springer, Berlin,Heidelberg,pp.521–536.

[39] Perozzi,B., Akoglu,L., Sanchez,P.L.,Muller,E.. Focused clustering and outlier detection in large attributed graphs. In ACM Special Interest Group on Knowledge Discovery and Data Mining (SIG-KDD), 2014.

[40] Priebe, C.E., Conroy, J.M., Marchette, D.J., Park, Y., 2005. Scan statistics on Enron graphs. Comput.Math.Organ.Theory11 (3), 229–247.

[41] Rezaei, Z., & Jalali, M. (2017, October). Sentiment analysis on Twitter using McDiarmid tree algorithm. In *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)* (pp. 33-36). IEEE.

[42] Rossi, R.A., Gallagher,B.,Neville,J.,Henderson,K.,2013.Modeling dynamic behavior in large evolving graphs. In : Proceedings of the Sixth ACM International Conference on Web Search and Data Mining , ACM , New York , NY ,USA , pp.667– 676.

[43] Saigo, H.,Nowozin,S.,Kadowaki,T.,Kudo,T.,Tsuda,K.,2009.gBoost:a mathematical programming approach to graph classication and regression.Mach. Learn. 75(1),69–89.

[44] Savage,D.,Zhang,X.,Yu,X.,Chou,P.,Wang,Q.,2014.Anomaly detection in online social networks.Soc.Netw.39,62–70.

[45] Shrivastava, N., Majumder, A., Rastogi, R., 2008. Mining (social) network graphs to detect random link attacks. In: Proceedings of the IEEE 24th International Conference pp.on Data Engineering (ICDE), IEEE, 486–495.

[46] Sun, H.,Huang,J.,Han,J.,Deng,H.,Zhao,P.,Feng,B.,2010.gskeletonclu:density- based network clustering via structure-connected tree division or agglomeration. In: Proceedings of the IEEE 10th International Conference pp.on Data Mining (ICDM), IEEE, 481–490.

[47] Sun, J., Qu, H., Chakrabarti, D., Faloutsos, C., 2005. Neighborhood formation and anomaly detection in bipartite graphs. In: Proceedings of the Fifth IEEE Inter- national Conference on Data Mining (ICDM), IEEE, p. 8.

[48] Sun, J., Xie, Y., Zhang, H., Faloutsos, C., 2007. Less is more: compact matrix de- composition for large sparse graphs. In: Proceedings of the SIAM International Conference on Data Mining (SDM), SIAM, Philadelphia, USA, pp. 366-377.

[49] Thompson, B., Eliassi-Rad, T., 2009. Dapa-v10: discovery and analysis of patterns and anomalies in volatile time-evolving networks. In: Notes of the 1st Work- shop on Information in Networks (WIN).

[50] Tong, H., Lin, C.Y., 2011. Non-negative residual matrix factorization with application to graph anomaly detection. In : Proceedings of the International Conference on Data Mining (SDM),SIAM,Philadelphia,USA,pp.143–153.

[51] Xu, X.,Yuruk,N.,Feng,Z.,Schweiger,T.A.,2007.Scan:a structural clustering algorithm for networks. In Proceedings of the13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,ACM,pp.824–833.

[52] Xu,Z.,Ke,Y.,Wang,Y.,Cheng,H.,Cheng,J.,2012.A model-based approach to attributed graph clustering. In : Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data,ACM,pp.505–516.

[53] Yang,W.,Shen,G.W.,Wang,W.,Gong,L.Y.,Yu,M.,Dong,G.Z.,2015.Anomalydetection in micro blogging viaco-clustering .J. Comput. Sci. Technol. 30(5), 1097–1108.

[54] Yu, R., He, X., Liu, Y., 2015. Glad: group anomaly detection in social media analysis. ACM Trans. Knowl. Discov. Data (TKDD) 10 (2), 18.

[55] Yu,W.,Aggarwal,C.C.,Ma,S.,Wang,H.,2013.On anomalous hotspot discovery in graph streams. In: Proceedings of the IEEE 13th International Conference on Data Mining(ICDM),IEEE,pp.1271–1276.