Egyptian Informatics Journal (2016) 17, 199–216

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

FULL-LENGTH ARTICLE

A survey of data mining and social network analysis based anomaly detection techniques

Ravneet Kaur *, **Sarbjeet Singh**

University Institute of Engineering and Technology, Panjab University, Chandigarh, UT, India

Received 20 February 2015; revised 26 October 2015; accepted 13 November 2015
Available online 28 December 2015

KEYWORDS

Anomaly detection;
Online social networks;
Social network analysis;
Data mining;
Graph based anomaly detection

Abstract With the increasing trend of online social networks in different domains, social network analysis has recently become the center of research. Online Social Networks (OSNs) have fetched the interest of researchers for their analysis of usage as well as detection of abnormal activities. Anomalous activities in social networks represent unusual and illegal activities exhibiting different behaviors than others present in the same structure. This paper discusses different types of anomalies and their novel categorization based on various characteristics. A review of number of techniques for preventing and detecting anomalies along with underlying assumptions and reasons for the presence of such anomalies is covered in this paper. The paper presents a review of number of data mining approaches used to detect anomalies. A special reference is made to the analysis of social network centric anomaly detection techniques which are broadly classified as behavior based, structure based and spectral based. Each one of this classification further incorporates number of techniques which are discussed in the paper. The paper has been concluded with different future directions and areas of research that could be addressed and worked upon.

1. Introduction

Online Social Networks (OSNs) have gained much attention in recent years in terms of their analysis for usage as well as detection of abnormal activities. The term has been defined differently by different authors. Like, Schneider et al. [1] formally defined OSN as "OSNs form online communities among people with common interests, activities, backgrounds, and friendships. Most OSNs are Web-based and allow users to upload profiles (text, images, and videos) and interact with others in numerous ways". Adamic and Adar [2] used the term social networking instead of Online social networks and defined it as "Social networking services gather information on users' social contacts, construct a large interconnected social network, and reveal to users how they are connected to others in the network". Regardless of the terminology used for defining it, social networks have become a communication platform where different users with a personalized user profile interact and share information with each other. Starting with Six

* Corresponding author. Tel.: +91 9779991701.
E-mail addresses: ravneets48@gmail.com (R. Kaur), sarbjeet@pu.ac.in (S. Singh).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.11.004

Degrees in 1997 [3], Online Social Networks such as Twitter, LinkedIn and Facebook have attracted large number of people. At present, almost every domain is linked in one form or the other with the social networks. Be it entertainment, education, trading, business, communication etc., OSN has made an influence on each of them. For example, mostly companies have started promoting their brands and products on social networking sites to increase the popularity of their products which in turn enhances their sales [4].

Contrary, to the positive side of social networking sites, its increasing popularity and open and free use have also led to their extensive misuse [5]. Malicious users are using it in a different way by behaving and obeying patterns differently from their peers. For example, a normal user often send emails to set of users which usually have connection among themselves but an anomalous user chooses its audience at random which are unlikely to have a relation in between them. Similarly, in the social networks such as Facebook and Google+ people who add friends indiscriminately, in "popularity contests" can be considered anomalous [6]. A new set of social network attacks may include unnecessary friend requests on Facebook, spam emails etc. "Millions of people fell for Facebook scams in 2014. They lost money, reputation and even their jobs after simply clicking on the wrong social media link", claimed the Online security firm BitDefender [7].

An anomaly is defined as an unusual activity exhibiting a different behavior than others present in the same structure. The term also called an outlier, abnormality or exception, has been defined in numerous ways by different authors. Some of the most popular and commonly used definitions are presented in Table 1.

Table 1 Various definitions of anomaly.

Defined by	Defined in	Defined as
Grubbs [8]	1969	An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs
Barnett and Lewis [9]	1994	An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data
John [10]	1995	An outlier can also be considered as a *surprising veridical data*, a situation in which a point otherwise belonging to class A but in actual is placed in class B, thereby making the true (veridical) classification of that point surprising to the observer
Aggarwal and Yu [11]	2001	Outliers may be considered as noise points lying outside a set of defined clusters or alternatively outliers may be defined as the points that lie outside of the set of clusters but are also separated from the noise
Chandola et al. [12]	2009	Patterns in data that do not conform to a well defined notion of normal behavior
Savage et al. [13]	2014	Regions of the network whose structure differs from that expected under the normal model

There is usually confusion between certain terms relating to anomalies which are otherwise different from it. For example, as indicated in the definition proposed by Aggarwal and Yu [11], the presence of anomalies is considered different from noisy data as noise is often viewed as a random error or a variance depicted in a variable and has no relevance during data analysis. As an example, while detecting credit card faults randomness in the behavior can be analyzed in terms of a person's purchase activities. Consider a scenario in which if one day a person buys a bigger lunch than he normally do, or have an extra cup of coffee than usual, it may seem like "random errors" or "variance" but it is actually the "noisy transactions". And hence, it must not be considered as anomalous; otherwise, it will be highly expensive for the company to verify so many transactions or lose the consumers by troubling them with several false alarms [14]. What is usually practiced is to remove noise before performing anomaly detection. Similarly, anomaly detection is also considered analogous to novelty detection [15,16] in which previously unobserved novel patterns in the data are detected. They may initially appear to be same but in novelty detection upon the confirmation of new topics they are generally incorporated into the model of normal behavior.

The presence of anomalies in our data poses many problems which need to be tackled carefully. For example, some sort of malicious users may construct a set of false identities and use them to communicate with a large random set of innocent users [17]. Hence, detection of these anomalous activities in a network is a big concern as their presence may lead to heavy losses. For example, in a computer network an anomalous traffic pattern could mean that a hacked computer is sending out sensitive data to an unauthorized destination [12]. Nowadays, not only the detection but the reason why these activities took place along with the methods to prevent these behaviors is on the rise. Here in this paper, various techniques used to detect and handle the anomalous behavior are covered. At first, a generalized view of various data mining techniques applicable to multiple domains and applications is given and then a special reference is given to some of the popular anomaly detection methods applicable to social networks.

The paper is organized into different sections. Section 2 contains the novel categorization of anomalies on the basis of number of parameters. The major data mining and social network techniques for anomaly detection have been discussed in Sections 3 and 4 respectively. Finally, Section 5 presents conclusion along with some future directions that could be addressed.

2. Types of anomalies

Anomalies or the abnormal activities can be classified into different categories based upon number of parameters. This section discusses some of these categories.

2.1. Based on nature of anomalies

Chandola et al. [12] classified anomalies into mainly three categories based upon the nature and scope of anomalies:

2.1.1. Point anomalies

Point anomalies, also referred to as global anomalies are found if a data object (i.e. a point) shows a different behavior than that of the rest of the data. Although being the simplest kind of anomaly to be detected yet major problem associated with detecting point anomalies is finding a suitable measurement in deviation of the object from other objects. Let us assume that for a normal network every node must have at least two neighbors linked to it. As shown in Fig. 1, nodes in Group V2 form such type of network and thus represent a normal behavior but group V1 contains isolated points. Because of their dissimilar behavior to other nodes they are predicted to be representing an anomalous behavior.

Similarly we may also have *local anomalies* which are studied relative to their local neighborhood only. For example, if we group a set of individuals based on their links in the network as friends and check their income (some parameter), a particular individual, let say A, might be having a fairly low income compared to his friends suspecting a local anomaly while overall in the global context his income might be insignificant as many people may have similar income representing a normal behavior. This behavior is depicted in Figs. 2(a) and 2(b).

2.1.2. Contextual anomalies

Frequently, referred to as conditional anomalies, these are present in the data set if the data object deviates significantly with respect to a specific context. For example, temperature may be considered as a contextual anomaly. If for example, today's temperature is 28 °C. Whether it is anomalous or not depends upon the time and location. It is viewed as an anomaly when considered in winters in Toronto. But in summers in Toronto this much temperature is normal and hence no abnormality is seen.

While detecting contextual anomalies two attributes of the data object define the data set:

- *Contextual attributes*: These attributes define the context of the object. For example, in temperature example, contextual attributes are date and location.
- *Behavior attributes*: Characteristics of an object are defined using these attributes and in a way help to identify the anomalous behavior of an object with respect to its context. In the temperature example, temperature, humidity etc. can be considered as behavior attributes.

When dealing with behavior attributes, a data object may be considered anomalous with respect to one context whereas normal in a different context. Usually proximity based methods, discussed in Section 3.1 are used for contextual anomaly detection.

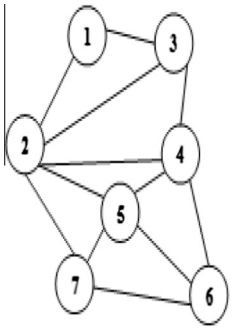

Figure 2a Groups on the basis of friendship links.

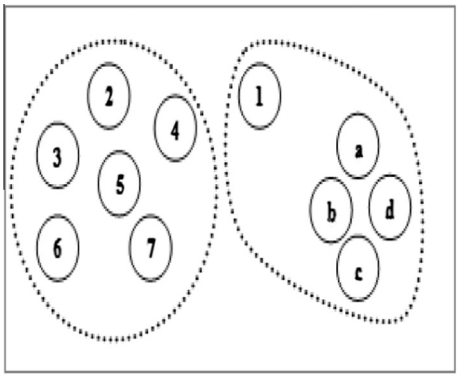

Figure 2b Groups according to income.

2.1.3. Collective anomalies

Collective anomalies are encountered whenever a collection of data objects as a whole depicts a different behavior than others, whereas the individual data objects may not be anomalous.

Objects in group G in Fig. 3, represent a collective anomaly on the basis of let us say, density parameter. Density of this group G is very high as compared to others. But on the other hand, each individual black data object in G is not an anomaly with respect to others. Similarly, in a real-life scenario we may assume that the figure shows the set of students who reserved a seat for a particular course and if one of them leaves a course, it may be considered as normal but if multiple students start leaving the course then they as a complete group represented by G are considered as anomalous (see. Fig. 3).

One of the fine principles adopted to detect collective anomalies is to consider the behavior of the *group of objects* along with the background information about the relationship among those data objects.

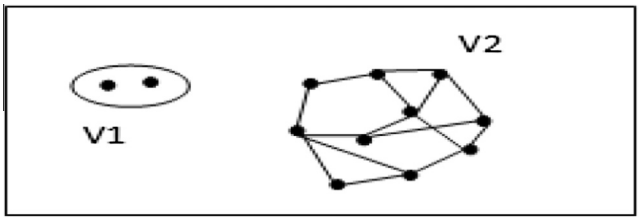

Figure 1 Point anomalies.

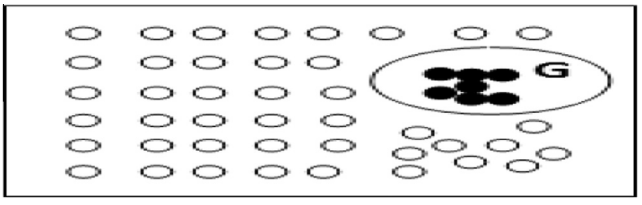

Figure 3 Collective anomalies.

2.1.4. Horizontal anomalies

Recently, another type of anomaly, called *horizontal anomaly* [18] has evolved in social networks which depict the presence of anomalies based upon the different sources of data available. For example, the same user may be present in different communities on different social networks. Similarly, a user may have similar kinds of friends on number of social networks (e.g. Facebook, Google+) but completely different kinds of friends for another social network (e.g. Twitter). This depicts an unusual activity which can be considered as anomalous.

2.2. Based on static/dynamic nature of network/graph structure [13]

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

2.2.1. Dynamic anomalies

A dynamic anomaly exists with respect to previous network behavior in which changes occur in the network with the passage of time. For example, it may involve changes in the way interactions take place in the network.

2.2.2. Static anomalies

A static anomaly occurs with respect to remainder of the network ignoring the time factor. Only the current behavior of a node is analyzed with respect to others in the network.

2.3. Based on information available in network/graph structure [13]

Depending upon the type of information available at a node or an edge, anomalies can be categorized as labeled or unlabeled.

2.3.1. Labeled anomalies

Labeled anomalies are related to both structure of the network and the information gathered from vertex or edge attributes. For example, labels on nodes may specify the attributes of individuals involved in the communication activity and that on the edges represent their interaction behavior.

2.3.2. Unlabeled anomalies

Unlabeled anomalies are related only to the network structure. No attribute of a node or an edge is taken into consideration.

Their classification is mostly studied as follows and different techniques have been developed and deployed to detect these types of anomalies. A number of such techniques have been discussed in Section 4.

- *Static unlabeled anomalies*

This type of anomaly occurs when behavior of an individual remains static and the attributes such as age of individuals involved, type of interactions, and its duration are ignored due to unlabeled nature of the network in which labels on nodes and edges are ignored. Only the fact that interaction took place is important.

- *Static labeled anomalies*

When along with the network structure labels on the vertices and edges are also considered, then the anomalous substructures found are referred to as static labeled anomalies.

Static labeled anomalies are used in spam detection, for example, to detect opinion spam (which involves the fake product reviews). A set of hidden labels are usually assigned to the vertices and edges which are iteratively updated. In the product review system, a bipartite graph with one subset of vertices as users and other 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 label can be in the form of honest or fraudulent and for the products it could be either good or bad. A normal honest user will give accurate results i.e. for good products they give positive response and for bad ones they will give negative reviews whereas fraudulent users are understood to do the reverse.

- *Dynamic unlabeled anomalies*

This type of anomaly arises when we have dynamic networks that change with time. Behavior of the data object is different with respect to previous time period relative to the network structure. For example while considering only the pattern of interactions, there are maximum of six ways in which a maximal clique can evolve: shrinking, growing, splitting, merging, appearing or vanishing [19]. All of these involve studying the network structure with respect to the network structure prevalent at some previous time period. Sometimes, the normal behavior does not result in any network change; then, any neighborhood changes may also predict an anomalous behavior.

- *Dynamic labeled anomalies*

In a dynamic network when anomalous behavior is observed by considering labels of the vertices and edges also; 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 the same way as for a static network.

2.4. Based on behavior

Another class of anomalies namely, "white crow anomalies" and "in-disguise anomalies" (see Fig. 4) is presented by Chen et al. [19].

2.4.1. White crow anomaly

It arises when one data object deviates significantly from other observations resembling the basic anomaly definition. For example, while examining the student record, if a record is found where height of a student is entered as 56 ft, which is impossible, then it is taken as a white crow anomaly. These anomalies are mostly detected as particular nodes, edges, or subgraphs representing the abnormal behavior.

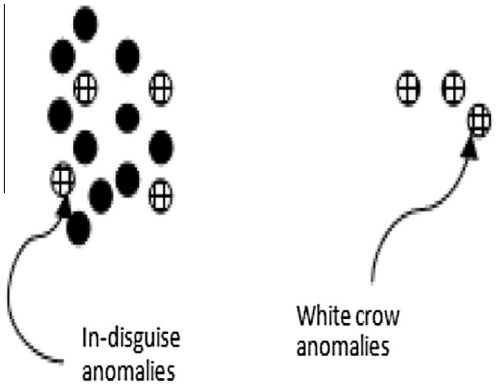

Figure 4 In-disguise and white crow anomalies.

2.4.2. In-disguise anomaly

It is considered as a small deviation from the normal pattern [20]. For example, anyone attempting to peep into someone's social network account would not want to get caught; therefore, he will try to behave in the same manner as a normal user. Such anomalies are recognized through strange patterns, which also include uncommon nodes or entity alterations. These are difficult to be detected as they are hidden inside the network.

2.5. Based on structural operations on network/graph structure

When dealing with the graphical structures like in social networks, anomalies can be classified according to the graphical properties as well. Eberle and Holder [20] classified anomalies according to the following three properties:

1. *Insertion*: Insertion deals with the existence of an unexpected vertex or an edge in the graph.
2. *Modification*: Modification deals with the presence of an unexpected label on a vertex or an edge.
3. *Deletion*: Deletion involves the absence of an expected vertex or an edge. Sometimes, it even incorporates the concept of dangling edges i.e. with the deletion of a particular vertex all the adjacent edges to it may also have been deleted.

2.6. Based on interaction pattern in network/graph structure

Types of interaction and links among nodes involve the study of anomalies in [6] as follows:

2.6.1. Near Stars/Cliques

Presence of completely disconnected (Near Stars) or all connected neighbors (Near Clique) is very rare and is considered anomalous

2.6.2. Heavy locality

Heavy weight around a particular area or a group is suspicious and hence determines the presence of an anomaly

2.6.3. Particular dominant links

An unexpected presence of heavy load at a particular node or link as compared to other nodes or links specifies an unusual activity

A data set or a network may contain more than one kind of anomaly. Some of these anomalies can be clubbed together to form a hybrid set. As an example, Savage et al. [13] studied the classification of anomalies as a combination of static/dynamic and labeled/unlabeled.

3. Data mining approaches to anomaly detection

Anomaly detection is defined as "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [21]. Many of the prevailing systems use signature based techniques to detect the strange behavior because of the fact that they produce very less false positives compared to the anomaly based techniques but still the latter is a better approach to use because of its benefit of detecting zero-day attacks (previously unknown attacks). From data mining perspective anomaly detection is broadly classified into the following three categories

- Supervised methods
- Semi-supervised methods
- Unsupervised methods

These methods usually work on the criteria whether a domain expert labeled sample of data of normal and anomalous objects is either available or not to build the anomaly detection model.

Supervised methods

These methods model both the normal and abnormal behaviors. It involves studying anomaly detection as a classification problem with the pre-labeled data, labeled either as normal or as anomalous. There are two applicable approaches for it.

- One, experts may pre-label the normal data and any such data which is not analogous to this model is considered anomalous.
- The other way is to do the opposite i.e. have the predefined set of anomalous data and any objects not corresponding to the set of anomalous data are considered as normal.

The major task involved in classification approach of supervised methods is to make the classifier learn. A classifier can be constructed in numerous ways. For example, it can be neural network based [22,23], support vector machine [15,16,24], Bayesian network based [25,26] etc. Supervised anomaly detection methods should keep in consideration the following two aspects:

- Anomalous objects are usually very less as compared to the normal data objects. Therefore, imbalanced class problem arises which needs to be resolved using methods such as oversampling (replicating) or other makeup "artificial anomaly" methods.

- Secondly, while choosing a classification method to detect anomalies major focus should be given on recall i.e. detecting accurately as many anomalies as possible rather than avoiding false positives.

Unsupervised methods

Unsupervised anomaly detection methods are used when labeled data objects are not available i.e. no predefined labels as "anomalous" or "normal" are attached to the data objects. Unsupervised methods are usually studied as a clustering problem.

These methods implicitly assume that the normal objects are a bit clustered forming one or more groups with distinct features. Hence, normal objects are expected to frequently follow a pattern whereas anomalies do not seem to behave in this manner as shown in Fig. 5.

But this assumption is not always true as sometimes it is the anomalies that form the similarity pattern or clusters such as collective anomalies as shown in Fig. 3 previously. So, here unsupervised methods work inefficiently by issuing a large number of false alarms especially when the normal objects are variedly scattered.

Generally, in unsupervised methods firstly the clusters based on a similarity measure of the objects are found and the objects not obeying any cluster are considered as anomalous. A prerequisite for supervised methods is the prior availability of complete data before any processing can occur. Therefore, they are mainly designed to handle only the static data.Two major challenges faced by unsupervised methods are as follows:

- Firstly, a data object not belonging to any cluster is considered as anomalous but many times this deliberation can be false because, such a data object can be a noise rather than an anomaly.
- Secondly, what is usually practiced is to firstly find the clusters and then the anomalies. But this methodology seems to be quite expensive as number of anomalies present in a data set is pretty less than normal data objects.

Semi-supervised methods

Semi-supervised methods work with two sets of data, labeled and unlabeled. So, these methods are used when out of the complete data set only few instances of data labeled as normal are available. Using the small amount of labeled data a classifier can be constructed which then tries to label the unlabeled data. Hence, a model for normal data objects is built which is used to detect the anomalies in a way that the objects not fitting the normal model are classified as anomalies. This is the simplest approach called **self-training** used under semi-supervised model. Another method called as **co-training** can be employed where two or more classifiers train each other. Self-training is more sensitive to errors than co-training. It is known as semi-supervised as it partially functions as supervised methods because only the normal class is taught and the algorithm learns to identify anomalies by itself.

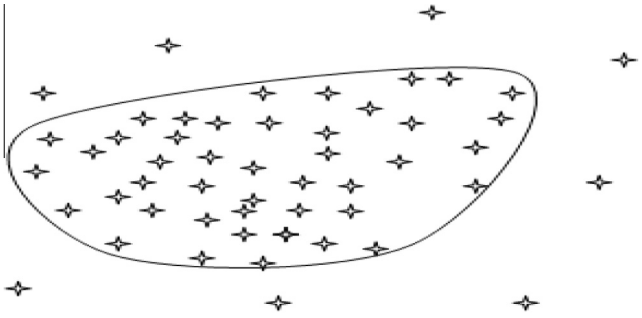

Figure 5 Unsupervised (clustering approach).

This approach is applicable to both the static and dynamic data and defines a boundary of normality. A new data object is considered as anomalous if it lies outside the boundary and normal otherwise.

The problem associated with the semi-supervised methods is that if instead of availability of labeled **normal** data only a small set of labeled **anomalous** data is present, and then it will be difficult to predict every possible anomaly by building a model for anomalies in the same manner as it is done for normal data.

Some of the prominent approaches under these three categories that are highly helpful in determining the anomalies and are covered in this paper are as follows:

- Proximity based.
- Clustering based.
- Classification based.

Graph Based Anomaly Detection (with special reference to social networks).

3.1. Proximity based (or nearest neighbor based) anomaly detection

Proximity and distance terms used to represent similarity and dissimilarity respectively are the key approaches used for detection of anomalies in any network. Proximity based anomaly detection approaches analyze each object with respect to its neighbors. It is assumed that normal data objects have a close proximity toward their neighbors i.e. they follow a dense neighborhood pattern whereas anomalous objects lie far away from their nearest neighbors. A number of k-nearest neighbor methods can be used which make use of various measures such as distance, density and other similarity measures to determine the proximity between the nodes. These proximity measures determine the efficiency of the methods. Proximity based methods can be mainly classified into the following two categories:

- Distance based (computes the anomaly score by using the distance of a data object to its k neighbors).
- Density based (computes the anomaly score by using relative density of each data object).

3.1.1. Distance based anomaly detection method

Distance based anomalies are considered as "global anomalies". Generally, Euclidean or Mahalanobis distance is taken as the distance metrics.

General method. Neighborhood of an object determined by the distance threshold is analyzed for each object. If neighborhood of an object, o, misses out many objects from complete data set D, i.e. the defined neighborhood contains few elements, then 'o' is regarded as an anomaly [27,28]. Discussed methods make use of two global parameters d and β explained below.

As stated in [14], if d ($d \geqslant 0$) be a distance threshold, β ($0 < \beta \leqslant 1$) be a fraction threshold and dist(.,.) be a distance measure, then, object 'o' will be a **DB(d, β)-anomaly** if

$$\frac{\|\{o' | \{\text{dist}(o, o') \leq d\}\|}{\|D\|} \leq \beta \tag{1}$$

This computes the anomalies as follows:

- d determines the maximum distance (radius) allowed between the objects to belong to the neighborhood of the object.
- β specifies the fractional threshold that determines maximum number of objects that could be there in the neighborhood in order to behave as anomalous node. If this threshold is crossed then the behavior is assumed to be followed by the **normal** nodes and hence object 'o' is classified as a normal node.
- Summing up the both the factors it can be predicted that for every node o, the k-nearest neighbors, o_k, where $k \in [\beta||D||]$ are analyzed and all those nodes having less than k objects in their neighborhood and dist(o, o_k) > d are considered anomalous.

The above approach is defined in a similar way by Knorr et al. [28] as follows: "an object O in a data set T is a DB(p, D) outlier if at least fraction p of the objects in T lies greater than distance D from O."

The simplest approach in distance based anomaly detection is the use of nested loop method [28]. In this method, inner loop matches the β condition and concludes about the object whether it is anomalous or not based upon the number of elements present in the d-neighborhood of the object.

Even though being the simplest approach it takes $O(n^2)$ time and is encountered as too costly when viewed from following two aspects:

- It tests each object against whole of the data set. Rather, there must be a way to determine the anomalous behavior of an object from the nearby neighbors.
- Each object is checked one by one.

An alternative can be to make groups of the objects based on certain proximity measure and then check anomalies group by group.

2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2
2	2	1	1	1	2	2	2
2	2	1	**C**	1	2	2	2
2	2	1	1	1	2	2	2
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2

Figure 6 CELL grids.

To make such improvements in the behavior, a grid-based CELL approach [14,28] was formulated in which each cell consists of a number of objects that form a group as shown in Fig. 6 and each cell has a diagonal of length $d/2$, where d is the distance threshold value. A multidimensional grid may be constructed with length of each edge depending upon the number of dimensions i.e. with 'n'dimensions, length of each edge is $\frac{d}{2\sqrt{n}}$. For a 2-D set data (Fig. 6) the example can be summarized as follows:

Neighbors of each cell C is divided into 2 parts: Level-1 cells (immediate neighbors) and Level-2 cells (one or two cells away) with the following implicit properties:

- *Level-1 cell property:* For any point 'a' in cell C, and any point 'b' in level-1 cell, dist(a, b) $\leqslant d$ always hold true.
- *Level-2 cell property:* For any point 'a' in cell C, and any point 'b', if dist(a, b) $\geqslant d$, then 'b' belongs to a Level-2 cell.

Based upon the above two properties, for each level a pruning rule is defined to conclude about the presence or absence of anomaly in the whole group (cell). Let x, y_1, y_2 be the total number of data objects in cell C, level-1, level-2 respectively.

- *Level-1 cell pruning rule:* Assuming level-1 cell property holds true, if $x + y_1 > [\beta n]$, then all the objects in Cell C are not anomalous as they are satisfying the normal object behavior.
- *Level-2 cell pruning rule:* Using the Level-2 cell property, all the objects in cell C are distance based anomalies if $x + y_1 + y_2 < [\beta n] + 1$ as both the conditions for anomalous behavior are met.

For higher dimensions the approach can be improved using Hilbert space filling curve. The multi-dimensional space in grid based approach is extended by Angiulli and Pizzuti [29] to handle the high dimensional data more efficiently. Hilbert space filling curve is used along with HilOut algorithm, an algorithm defined to choose the anomalies based on their aggregate score with their neighbors rather than one absolute score. For each object o, weight w, is computed as [14]:

$$w(o) = \sum_{j=0}^{k} \text{dist}(o, nn_j(o)) \tag{2}$$

where $nn_1(o)$, $nn_2(o)$, ... $nn_k(o)$ are the k-nearest neighbors of node o. After the weight computation, all the objects are ranked in decreasing order and top-m are stated as anomalous. Use of space filling curve reduces the time and space complexity which otherwise increases with the increasing dimensionality. Similar approach is used by Ramaswamy et al. [30] in which the ranked retrieval of anomalies is there. Instead of applying the approaches to full dimensional space, high dimensional spaces can be reduced to low dimensional space using dimensionality reduction method. The best way to extract the lower dimensional space is by using principal component analysis (PCA) in which usually principal components having low variance are chosen since normal objects on these dimensions are expected to be closer to one another while anomalous nodes deviate from others. This application of PCA is classified into correlation based clustering methods. Other grid-based approaches include Clustering in QUEst (CLIQUE)

introduced by Chang and Jin [31], used for locating dense and spare clusters in subspaces.

As the dimensionality increases the question about why and up to what extent the data object is an anomaly is of more concern rather than just predicting out anomalies. One of the simplest approaches toward it is to compute sparsity coefficient. The more negative its value is, sparser a cell (hypercube) is and more likely the objects in C are anomalies.

3.1.2. Density based anomaly detection method

The major problem associated with distance based methods is its failure to detect local anomalies which can be easily overcome by density based methods. Density based approaches work by comparing the density of an object with density around its neighbors. For a normal object both densities are assumed to be same whereas for anomalous objects they are different. The concept of relative density is often used to measure the degree of anomalous behavior of an object.

The simplest approach developed in this domain is the Outlier Detection using In-degree Number (ODIN) proposed by Hautamäki et al. [32]. ODIN score of an object is the number of such k-neighbors of an object for whom this particular object is also a k-nearest neighbor. Inverse of such a node predicts the anomalous score.

The most popular density based anomaly detection approach is the Local Outlier Factor (LOF) computation given in [33]. LOF score is defined as the ratio of local reachability density of k-neighbors of object 'o' being examined with that of its own. This local reachability density used to compute LOF is a factor of both k-nearest neighborhood of object 'o' and the reachability distance measure. For an anomalous object LOF score is higher as relative density of an anomalous node is less than that of its neighbors. But for normal data objects both densities are approximately the same.

A small variation in the consideration of neighborhood set led to the formulation of a number of different approaches which otherwise worked same as the LOF. For example, in Connectivity-based Outlier Factor (COF) [34] neighborhood set of an object is constructed by adding the closest object to 'o' in the set and then incrementally updating the list with the objects having minimum distance to any other object present in the neighborhood set, until it reaches k. Similarly, another technique called Influential Outlier (INFLO) [35] also works on the different variant of a neighborhood set. It uses reverse k nearest neighbors set (RNNk) to get all those points which has object o in its neighborhood set. Computation of anomalous score in these methods is in the same manner as in LOF with certain added terminologies.

In order to make the approaches computationally less expensive some sort of statistical measures are added to them. For example, instead of using the densities as it is the computation of standard deviation of densities led toward an approach named as *Multi-granularity Deviation Factor* (MDEF) as suggested by Papadimitriou et al. [36]. Similarly, Local Outlier Probability (LoOP) method [37] also makes use of the statistical measures and estimates the probabilistic LOF as a factor of ratio of densities to finally compute the measure called as LoOP.

Merits of proximity based approaches

- Simplest data mining approach.
- Applicable to a number of domains.
- An easy and straight forward approach as the only major requirement for such methods is the identification of a distance or density measure.

Demerits of proximity based approaches

- Handling and detection of anomalies become difficult when we have several regions with widely differing densities.
- Also it becomes difficult to detect the group of anomalies if they are present close to each other.
- Proximity based methods are highly dependent on the proximity measures used for their efficient working which might not be available in certain situations.

3.2. Cluster based anomaly detection

As stated by Berkhin [38] clustering is considered as an unsupervised learning of a hidden data concept. Clusters of the data objects can be constructed using numerous methods such as, *K*-Means, *K*-Medoids for small data sets and CLARA [39], CLARANS [40] for large data sets and BIRCH [41], Chameleon [42] for performing macro clustering on micro clusters. Cluster based methods follow a simple assumption that usually anomalies either belong to a small sparse cluster or do not belong to any cluster whereas the normal objects are a part of large and dense clusters. So, cluster based anomaly detection approaches state the presence of anomaly in the following three cases:

- If the object does not belong to any cluster.
- If the distance between object and cluster to which it is closest is large.
- If the object is a part of small or sparse cluster, then not only the object but all the objects belonging to that cluster are considered as anomalous.

3.2.1. If the object does not belong to any cluster

For this case, simply the density based clustering approaches can be used simplest of which is the DBSCAN and its numerous variants. DBSCAN [43] checks the density around each object and the one being isolated or of lower density than others is considered as an anomaly. One of the striking features of this method is that it can detect the clusters of arbitrarily any shape. A number of improved variants of DBSCAN such as, FDBSCAN [44], L-DBSCAN [45], C-DBSCAN [46], P-DBSCAN [47], and TI-DBSCAN [48] have also been applied to detect the anomalies efficiently. Apart from the DBSCAN other applicable approaches are C^2P [49], CURE [50], SNN [51] etc. Out of all such measures the prominent one is the Shared Nearest Neighbor (SNN) method in which the similarity between the data points is identified based upon the number of nearest neighbors shared and hence the core points around which the clusters are to be built are identified. This approach helps to identify the dense as well as medium and sparse clusters.

3.2.2. If the distance between object and cluster to which it is closest is large

No doubt, the proposed methods help to identify the anomalies but they focus more toward finding the clusters and considering any point not related to any cluster as noise which in a way is assumed to be anomalous. Some of the cluster based methods also avoid finding the degree of anonymous behavior shown by each data object. In order to encounter such problems, numerous advanced approaches have been proposed. For example, Cluster based Local outlier factor (CBLOF) [52] and the corresponding algorithm FindCBLOF are used to mine the encountered anomalies. CBLOF is measured as a factor of both the *cluster size* to which object belongs and its *distance* from the cluster it is closest to. FindCBLOF uses the Squeezer algorithm which constructs the clusters out of which a set of large and small clusters are formed and CBLOF is calculated for every data point. In a similar fashion a number of other techniques using different distance measures have also been proposed like, Self-organizing maps (SOM) an unsupervised method proposed by Kohonen [53], *k*-means clustering [54,55], *k*-means + + [56,57]. As these techniques involve the computation of distance factor, therefore, they are a good way to handle the second case. Some of the semi-supervised methods proposed by Wu and Zhang [58], Vinueza and Grudic [59] can also be used.

3.2.3. If the object is a part of small or sparse cluster

This case is handled by defining a threshold value for the clusters and the objects belonging to low value clusters are considered as anomalous. FindCBLOF algorithm detects both the individual objects and points belonging to small clusters as anomalous by computing the similarity between the objects in the small cluster and the closest large cluster. CBLOF value for such points comes out to be very low. Apart from this, other applicable approaches are described in [52,60,61]. Distance or densities of the small clusters generated are compared with those of large clusters and anomalies are detected. Numerous efficient techniques such as *k*-d trees and CD-trees are used to partition the data into clusters.

Along with the above techniques concepts of local clustering, co-clustering, bi-clustering and subspace clustering have been used by Beutel et al. [62] to select the set of attackers in social network domain. In subspace clustering a cluster is defined as a subset of data objects which are similar to one another in terms of the above defined similarity measures such as distance, density or other such variants for a particular subspace. For example, one of the subspace clustering algorithms, CLustering In QUEst (CLIQUE) proposed by Chang and Jin [31] is used for locating dense and spare clusters in subspaces. Similarly, bi-clusters allow both the objects and attributes to be clustered at the same time allowing a particular object or attribute to be involved in multiple clusters, or not in any cluster at all.

Merits of clustering based approaches

- The major advantage of cluster based approaches is its unsupervised nature where no predefined set of labeled classes of data objects is required.
- These methods involve fast comparison process as once clusters are constructed it is faster to compare objects to clusters because a number of clusters available are comparatively less than number of objects.

Demerits of Clustering Based approaches

- Incur high computation cost when the clusters are to be found before detecting anomalies.
- A data object not belonging to any cluster may be a noise rather than an anomaly.
- Computational complexity for such methods is highest of all the data mining methods applied.
- Clustering approaches are a costly procedure for large data sets
- Sometimes clustering process involves anomalous objects depicting similar behavior and hence forming the clusters. As anomalies follow a presumption to be belonging to either no cluster or a small cluster, so, objects in the above encountered clusters might be considered as normal.

3.3. Classification based approaches

Classification is defined in [14] as a supervised method with two steps i.e. a learning step and a classification step. In the learning stage a trained set of labeled data instances are used to construct a classification model and in the classification step the constructed model is used to predict the class labels for the data. Both the steps are respectively stated as the training and the testing stages. For detection of anomalies, the training data objects are labeled as 'normal' and 'anomalous'. Numerous classifiers are available which can be used for the detection of anomalies. Classification based approaches can use either a one class model [63] or a multiclass model. A simple brute force approach may not be much effective as the number of normal data objects is much larger than number of anomalous data objects. Hence, there arises a need for such class models.

In one class model, only a single labeled class is defined i.e. classifier is constructed to only define the normal class and all those data objects that belong to that class are treated as normal whereas the ones that do not fit in the defined class are treated as anomalies. Some of the examples of one class models used for anomaly detection are one-class SVM [64], Gaussian model description (GAUSSD) [65], Principal component analysis description (PCAD) [66], Parzen window classifier (PWC) [67] etc. In each of them a decision boundary is set up. The data objects falling outside the decision boundary are treated as anomalous. One-class models help to detect new anomalous objects that are far from the other anomalous objects present in the given training set.

The other set of model called the multiclass model is used when the available data objects not only belong to a single class but to multiple classes. For example, the classification of a set of images of fruits into the probable classes of apples, oranges or mangoes. Every data object may be assigned only one label. Just like, a fruit may be classified as either one out of the three categories but not more than one at the same time.

On the contrary, in one-class model only class defined is that of the fruit. Irrespective of the type, every fruit object is put in this class and the remaining objects are classified as anomalous.

A number of classifiers can be used for the classification process. Some of them which are best suited for the detection of anomalies are discussed in the following subsections.

3.3.1. Bayesian classifier

These are the *statistical classifiers* making use of the prediction probabilities about any data object belonging to a particular class. Heckerman and David [68] gave a detail picture of Bayesian networks for data mining including its use in supervised and unsupervised learning.

The simplest Bayesian classification involves the use of Naïve Bayesian Classifiers. There are two properties that it exhibits. One, it makes use of the Bayes' theorem of posterior probability. Second, it obeys the class conditional independence property i.e. any effect of an attribute value on a certain class is independent of the value of other attributes.

Another variant of Bayesian classification is the Bayesian belief classifiers. They are the *probabilistic graphical models* having a directed acyclic graph with each node representing a discrete random variable of interest and a set of conditional probability tables representing casual relationships on which learning can be performed. They differ from naïve Bayesian classifiers in a manner that they allow dependencies between attribute values to be defined. It outperforms the latter's performance because of the explicit representation of causal structure and the presence of human expertise knowledge thereby increasing the learning rate.

Theoretically, Bayesian classifiers are more accurate than other classifiers with minimum error rate. But practically this might not always hold true. Also, even though the inferences drawn from it and threshold-based systems are found to be quite similar this approach requires much greater computational cost and effort [69].

Kruegel et al. [69] studied the detection of anomalous behavior using Bayesian classifiers. Using this approach, all the previously unknown attacks are also identified but with the generation of a number of false positives. A model of the normal behavior is available and deviations from these behaviors are identified as anomalies. There are a number of models which evaluate different set of features and return different probabilistic values as anomalous scores which are aggregated into a single value. But in every such model the final decision process is conducted using a Bayesian classifier.

3.3.2. Support vector machine [70]

It makes use of a hyperplane as a decision boundary to separate the tuples of different classes from one another. The major task involved in SVM is the selection of best separating hyperplane from among several of them. An approach toward this issue is the use of **maximum marginal hyperplane** (MMH) i.e. the one with largest margin is considered most accurate for classification. SVM as a classification measure can be used to detect the anomalies in various applications.

Otherwise being a two class model approach, it may sometimes be used as a one class algorithm with an addition of the fact that only a positive data set is taken as a class and the "anomalies" that are detected are treated as the other class. Using a one class support vector machine model, anomalous behavior has been detected by Cortes and Vapnik [70]. Similar behavior has been described by Manevitz and Yousef [71] for classification of various documents represented in different formats. In [16], time series novel data considered to be anomalous has been detected using one class support value regression technique and ranked by attaching a confidence level to each anomaly. Eskin et al. [72] defined a feature space to which data objects are mapped and are classified as anomalous depending upon their position in it. For example, in the feature space, objects in sparse region are classified as anomalies. Another SVM based approach that uses Multiple-Classifier Payload-based Anomaly Detector (McPAD) constituting one-class classifiers is used by Perdisci et al. [73]. Piciarelli et al. [74] detect the anomalies by the use of trajectory analysis used in traffic monitoring and video surveillance. The deployed method makes use of one-class SVM clustering to detect anomalous trajectories.

3.3.3. Neural network based

Neural network based classification methods, also called as backpropagation method or connectionist learning consist of weighted connected components in which at the initial phase of learning the associated weights are adjusted in order to make a correct prediction of the classes. Despite its difficulty for interpretation by human beings and long training times, these methods are highly used in classification and other prediction tasks due to their capability to classify the untrained data and high tolerance toward noisy data. These methods are also applicable both to the single class and multiclass environment. Multiclass problem in [75] is solved using neural network to not only detect the normal or the attack pattern but also the type of the attack. NNID, an off-line anomaly detection system is examined using Multi Layer Perceptron (MLP) neural network. A number of important commands are set to depict the user's behavior. The process is carried out by identifying each user's profile and detection of intrusions for every user based upon the evaluation of their commands. Another approach called Principal Component Analysis (PCA) used by Liu et al. [76], proved to be quite a beneficial classification approach to detect various threats.

In general, the classification process can be summed up as a sequence of following 4 steps

1. Start by discovering a set of class attributes and classes from the training data.
2. From those attributes, identify the ones suitable for classification.
3. A model using the training data is learnt.
4. The model is then used for the classification of unknown data objects.

Merits of classification based approaches

- Usually a fast process, especially the testing phase, because a classification model has already been learnt which just needs to be analyzed for testing process.
- Classification based techniques help in improving the efficiency especially when ensemble methods incorporating integration of a number of classifiers are used.

Demerits of classification based approaches

- One of the concerns for the classification method is the heavy dependency and reliability on training data which if not properly available may lead to the degradation of performance.
- Many of the times we may encounter a class imbalance problem in which only few objects represent the main class. Although a number of sampling strategies have been proposed still this issue needs to be addressed.

Combination of both classification and clustering approaches as shown in Fig. 7 helps to find the anomalies in semi-supervised manner by forming the clusters first and then applying class models to it. The points not classified in any of the cluster are considered as anomalies.

Other inter-related domains such as intrusion detection, misuse detection, and novelty detection are equally being explored to detect different novel and miscellaneous events. But somehow all such techniques are working with the similar objective in mind. Therefore, researchers nowadays are drifting toward integrating these independent domains to develop a hybrid approach. A number of researchers have inferred that the use of multiple detection techniques helps to achieve a better performance than the use of a single technique. For example, recently Kim et al. [77] proposed a novel hybrid approach for intrusion detecting by hierarchically clubbing anomaly detection and misuse detection methods. C4.5 decision tree algorithm was used for misuse detection and many one-class SVMs were used to handle different decomposed training subsets. The experiments were performed on renowned KDD data set. Similar approach has been used to detect malicious web pages also [78]. The implication used is that misuse detection method is an efficient way to identify the well known attacks but fails to detect new ones whereas the anomaly detection method is good in detecting the unknown events, no doubt most of the times with a high false positive rate. Therefore, to overcome the shortcomings of both the methods, an effort has been made to combine both the domains by implementing a two-phase process. First phase involves misuse detection (supervised learning) and the second phase undergoes anomaly detection (unsupervised learning). In a similar fashion, a novel approach named Cluster center and nearest neighbor approach (CANN) has been proposed in which two distance measures are computed and summed [79]. As the name suggests, first distance is computed between each subset of data and its cluster center whereas the other one is calculated between the data object and its nearest neighbor in the same cluster. Finally *k*-NN classifier is used to classify nodes in each subset of data.

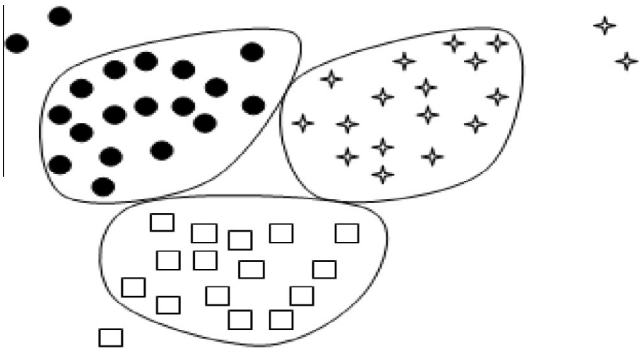

Figure 7 Combining classification and clustering approaches.

4. Anomaly detection in social networks

Online social networks being the center of attraction for number of applications are best viewed as a graphical structure with nodes and edges depicting the users and their interaction activities respectively. The nodes and edges in a network can be labeled or not depending upon the network structure being studied. Most of the cases involve considering only the binary and static social links in which mere presence of a link is considered sufficiently good without giving any importance to the actual communication activity of users. But going through the literature, it has been observed that earlier research analyzed the significance of users' actual interactions also. "No matter what resources are available within a structure, without communication activity those resources will remain dormant, and no benefits will be provided for individuals" [80]. Taking into consideration actual communication activities and interactions of users, the resulting graph, usually called an activity graph [81,82] are drawn. This activity graph can be categorized as a basic activity graph or a weighted activity graph. A graph containing similar kind of edges in every pair of nodes irrespective of strong or weak ties between them is called a basic activity graph but weighted activity graph represents a graph structure in which strength of the activity link is also taken into account.

The increasing trend of social networks attracted their misuse by number of malicious individuals also. Hence, the detection of anomalous activities becomes the need of the hour. Sometimes, it becomes difficult to analyze the social networks because of their large size and complex nature and it becomes necessary to prune the networks to include only the most relevant and significant relationships [83]. Usually, the presence of an anomaly is considered as a binary property in which anomaly is either present or not, but in some applications the extent to which anomaly is present is considered by giving degree of being an outlier to each object in the data set. As an example, Breunig et al. [33] referred this degree as Local Outlier factor (LOF).

OSN are often represented as graphs in which users are represented as nodes and interactions among users as edges which can be either labeled or not. In most of the cases, binary and static social links are considered in which only the mere presence of a link is considered sufficiently good but users' actual communication activity is given no importance. However, it has been found that earlier research focused on the importance of users' actual interactions also. "No matter what resources are available within a structure, without communication activity those resources will remain dormant, and no benefits will be provided for individuals" [80]. The graph resulting from such networks involving user interaction activities is called an activity graph [81,82] which can either be a basic activity graph or a weighted one. Basic activity graph represents the graph in which every pair of nodes has similar kind of edges irrespective of strong or weak ties in between them whereas weighted activity graph is the one in which strength of the activity link is also taken into account.

Sometimes, large size and complex nature of social networks make them difficult for analysis purpose. Hence, in such

cases, pruning of the networks to include only most relevant relationships is done [83]. In most of the cases, the presence of an anomaly is considered as a binary property in which anomaly is either present or not, but in some applications the extent to which anomaly is present is considered by giving degree of being an outlier to each object in the data set. For example, this degree has been referred to as Local Outlier factor (LOF) by Breunig et al. [33].

It has been seen that for any kind of social network, analysis of one or more of the three influence factors is targeted [84]:

- a subject node (node influencing others),
- a tie or a social link (communication link between nodes), or
- an object node (the node being influenced).

Out of these three factors first two have been the center of study by most of the researchers but the last one has got very little attention and is a current topic of research.

The graphical representation of the social networks leads to the applicability of different anomaly detection techniques. Proximity or similarity measures defined in data mining techniques do not seem much appropriate for social networks. Similarity in social networks may be defined on the following basis:

- *Structure context-based similarity*: It is a local cluster or neighborhood based similarity in a way that nodes having similar neighborhood are considered as similar. For example, in social networks, different users getting recommendation about a page or a community etc. from a number of mutual friends usually make similar decisions and help in determining how close they are.
- *Similarity based on random walks*: This type of similarity could be well described by this example. Suppose, an information or message needs to be forwarded to multiple users. But at an initial stage it is sent to only two users A and B who forward it to their friends. Now, the closeness or similarity could be measured by the simultaneous receipt of the message from both A and B to the nodes. So, here similarity is addressed as a random walk measure over the network.

In social networks, the nodes disobeying these similarity measures by following behavior different from the other nodes are detected as anomalous.

Broadly, anomaly detection techniques in social networks can be categorized as follows:

- Behavior based techniques
- Structure (Graph) based techniques
- Spectral Based techniques

4.1. Behavior based techniques

Behavior based techniques handle the behavioral properties of the users such as number and content of messages, content of the items shared, number of likes or comments on a post and duration of a conversation. Some of the popular behavior based techniques are discussed below.

4.1.1. Content-based filtering

Content based filtering is one of the prominent and well-known behavior based approaches in which anomalous behavior is detected by looking at the internal content of the sent and received messages. A trained classification model that may be used in the analysis phase is built using the content of the messages. As an example, Vanetti et al. [85] found a Filtered Wall system in which certain set of filtering rules were used by the users to avoid the unwanted and irrelevant posts from their walls. A Blacklist may be created using these filtering rules in which a number of constraints are imposed like 'constraint on message creators', 'constraints on message contents', and 'action taken in the form of blocking, publishing or notification'. A user is regarded as anomalous or not based on the relative frequency of unknown activities that node performs by disobeying the set of filtering rules continuously. Depending upon the anomalous activities performed respective constraint may be imposed and the users may be put into the blacklist called as BL. However, some smart malicious users are intelligent enough to befool and deceive others by behaving similar to the legitimate users. For example, in social network scenario two of the famous attacks called Sybil attacks and cloning attacks are quite popular nowadays [86]. Though a number of techniques have been projected to handle such type of attacks yet most of them seem to fail because of one or the other reason. For example, some of the simple techniques such as clustering coefficient (CC), and voting scheme are botched by the spammers by behaving or creating a similar network structure to that of a normal user. In Clustering coefficient method, for normal users value of clustering coefficient is high whereas that for spammers is close to 0. But in order to present themselves as legitimate, the spammers increase their CC value by making the neighborhood structure similar to that of the genuine users. Similarly, in voting schemes the illegitimate users make a number of fake profiles to increase the votings in the form of likes, views etc. or to avoid being classified as spam during voting. Even the advanced techniques such as Honeypots [87] proposed to detect the spammers fail to attract anomalous users in most of the situations.

Recently, an unsupervised statistical anomaly detection technique known as Principle component analysis (PCA) was used by Viswanath et al. [88] to detect the anomalous behavior in individuals. Unlabeled Facebook data set was used and a number of fake and compromised users were identified. The criteria for normal and anomalous distributions were judged by observing the 'like' activities of the users, for example, by studying the pages 'liked' by a user, number of posts/pages liked at a particular time period. The motivation for implementation of this technique was the increase in purchase of fake Facebook likes, fake reviews for reviewing websites, followers on Twitter etc. Apart from these, a significant contribution made by them was the detection of click spams highly prevalent nowadays in ads. Either the users are unintentionally made to click on the spam links which seem to be genuine or some sort of malware hacks a person's account and clicks 'likes', posts comments or reviews without the knowledge of user. By experiments it was inferred that most of the clicks on such sites were done by anomalous users.

Xiao et al. used the profile information of a user to detect fake accounts in online social networks using certain supervised machine learning techniques for feature extraction and

cluster building [89]. The proposed technique is a faster and efficient way to identify fake accounts as it only uses the attributes entered by a user during registration i.e. profile creation. The employed technique is a first in its form to detect the clusters of fake accounts usually created by a single user on a particular social network, thereby superseding the existing techniques which only work and make deduction for a single account. The system was found to restrict around 2,50,000 fake accounts.

4.2. Structure based techniques

Structure based methods work on the basic principle of using structural properties to check the characteristics of normal and anomalous users. A particular graph metric is figured out for different nodes or structures and the nodes showing different values than other users are considered as anomalous. The properties or metrics used may range from the simple properties such as number of nodes, edges to highly complex centrality measures. Just like supervised learning, here also a predefined normal pattern is already known and any deviation from that known pattern depicts the anomalous behavior.

The structural properties have been used by most of the researchers working in social network domain to define a number of new approaches for identifying anomalies in online social networks. As an example, Link mining, used by Getoor and Diehl [90] studies the structural properties of the networks to predict different behaviors of individuals in social networks. For instance, a normal tendency shows that consumers, whose friends spend a lot, spend a lot themselves. The concept of link analysis is applicable for both homogeneous and heterogeneous networks, but in the concerned work the graphical structure of heterogeneous networks with different types of nodes or edges is given more focus. By analyzing the association between different nodes it is usually found that the linked objects often have a set of correlated attributes. In other words, connectivity of two users can be checked by examining the common properties as what is usually observed is that the objects sharing some sort of common features are often found to be linked with each other. Getoor and Diehl [90] covered eight link mining tasks with their respective algorithms and grouped the defined tasks under three categories, namely object-related, link-related and graph-related. Most of the structure based link prediction methods show poor performance because of the involvement of prediction of future relationships likely to occur [91]. Earlier also a number of advanced tasks such as anomalous link discovery (ALD) were proposed which involved only the prediction of anomalous relationships rather than all the involved relationships [92]. It was seen that almost every prediction model performed quite well for ALD.

In social networks, link prediction is highly useful for detecting friendship links between different users as such techniques are a good way to examine connected, missing and corrupted links [93]. Therefore, they easily help to analyze the dynamics and prediction of future link behaviors. These techniques help to identify dynamic unlabeled anomalies by predicting future events and analyzing previous network behavior which is a prerequisite for dynamic anomalies.

Shrivastava et al. [17] proposed a generic approach for detection of attacks, more specifically mentioned as Random Link Attacks (RLAs). The basic motive behind such an attack resembles that of the Sybil attacks as it also makes use of multiple fake profiles and their use to communicate with a number of random users. These attacks are quite prominent in email spams, viral marketing etc. with a fact that the victims are chosen randomly with each one having the equal probability to be a victim. This helps to analyze and detect the attacks efficiently as for an attacker, structure of a set of random nodes in its neighborhood will be quite different from that of a good node. A set of two properties namely, a clustering test and a neighborhood independent test are conducted on the suspicious nodes which after creating groups mark them as anomalous. Two heuristic algorithms GREEDY and TRWALK algorithm were proposed to detect the attackers.

Many already existing node-based and egonet-based features were studied recursively by Henderson et al., [94]. Some aggregate values were calculated on the already existing characteristics. Neighborhood information was retrieved using both node and egonet-based features and behavioral information was extracted using recursive features.

Akoglu et al. [6] utilized another structure based approach in which a number of pattern and law discoveries were used by to detect different types of anomalies in social network graph. To spot some strange nodes especially in weighted graphs an Oddball algorithm was proposed in which a number of new rules (power laws) were discovered to detect the deviation from the known normal behavior. A set of features were grouped into certain set of carefully chosen pairs and anomalous behavior was analyzed by examining the group structure. Groups were formed where the patterns of normal behavior (power laws) were observed and the points deviated from discovered patterns were flagged out to be considered as anomalous. A number of anomalous relationships were observed namely *Near Stars* or *Near Cliques*, *Heavy Vicinities* and *Dominant Edges*.

A number of researchers used this Oddball algorithm to propose various new approaches and examine the relationship between different graph metrics. For example, Hassanzadeh et al. [95] explored various new social network metrics and used the power laws defined in Oddball algorithm to analyze the relationship among them, thereby detecting the anomalous relationships between different users. Among the different metrics used it was seen that the relationship between number of edges and average betweenness centrality of a user's direct neighborhood helped to better predict the anomalous nodes.

Similarly, Rezaei et al. [96] used the same approach and analyzed the number of nodes and edges behavior to predict Near Star/Clique behavior in Twitter data set.

Both the works followed a five step process as follows:

- Determination of the anomaly detection graph metrics.
- Computation of fitting curve.
- Calculation of the anomaly score.
- Labeling of nodes for further evaluation.
- Computation of the final threshold score.

4.3. Spectral anomaly detection techniques

Spectral anomaly detection techniques help in detecting anomalies using some spectral characteristics in the spectral

space of a graph. Different complex measures such as eigenvalues or eigenvectors applicable to the adjacency matrix [97] or the different hypergraph algorithms used for Laplacian graphs [98] are focused in these methods. In most of the techniques a social network graph is partitioned into different groups or communities. Partitioning is done either by eliminating the links between different nodes or using certain clustering/classification algorithms and measures. Even some of the advanced techniques use the structural concept of centrality. For example, community structures were worked upon by Girvan and Newman [99]. As shown in Fig. 8, communities in the form of different friendship groups were created in which the strength of links between the nodes within a community or friendship group is dense whereas among different groups is sparse.

The concept of betweenness centrality formulated by Freeman [100] is modified to work for edges instead of vertices to find the number of shortest paths between a set of vertices that pass through the edge under consideration. The implication used is that the edges with high value of betweenness centrality state the points where a network is expected to break and hence are separated. Generally, in online social networks high betweenness centrality is found to be at the intersection of densely connected network groups. As a result, a number of significant groups could be determined by removing the set of links from a graph, a concept also used by Newman [101].

Ying et al. [97] identified the malicious nodes by computing the spectral coordinates or the spectra i.e. the eigenvalues or eigenvectors for both the normal and anomalous user with a special reference to RLA's. The use of RLA's was stressed upon because of the absence of prior knowledge regarding which node is attacker and which one the victim node. The presence of fake links or nodes affect the value of the graph spectra. Spectral coordinates of a victim node are used to analyze the interdependency between the victim and the attacker nodes, thereby calculating the spectral coordinates for attacking nodes. It was observed that malicious users govern the attack set and each attacking node is linked to a number of victim nodes as shown in Fig. 9.

The aforementioned paper presented a novel method for the detection of collaborative attacks and a number of other related approaches. One of the simplest approaches although hardly ever required to be used to detect attackers is the concept of outdegree of nodes which is found to be high for attackers. In another approach, a number of non-

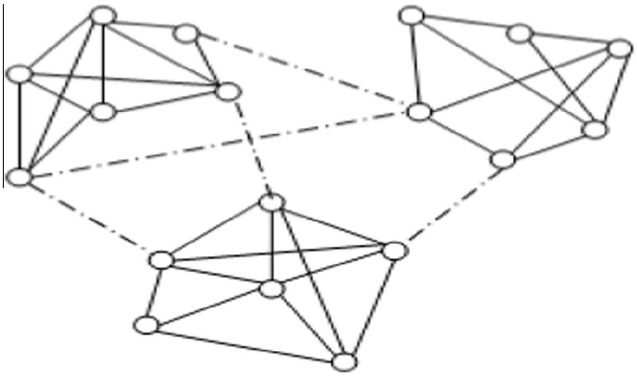

Figure 8 Friendship links depicting centrality also.

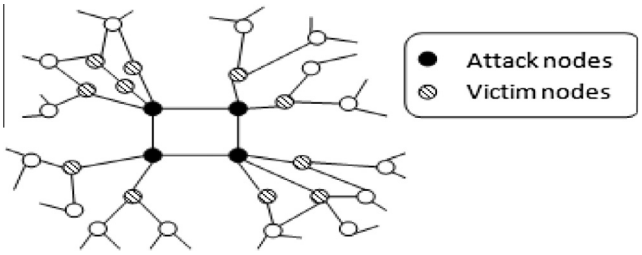

Figure 9 Describing relationship between attacking and victim nodes.

randomness tests are applied which involves the computation of different non-random measures specially the non-randomness for nodes. This non-randomness characteristic is used by a popular algorithm known as SPCTRA to identify the anomalous users. A number of different subgraphs are created where attackers or anomalous nodes are likely to have dense subgraphs. From these subgraphs, a set of nodes are chosen for RLA groups. Finally, all the dense subgraphs formed by regular nodes are removed and hence, the only ones left are the subgraphs of attacking nodes. The proposed approach supersedes the previous approaches because of the effectiveness of spectral characteristics.

4.4. Other graph based approaches

Based upon the nature and type of anomaly being studied a variety of other graph based techniques have been proposed and implemented in the social network domain.

For example, Savage et al. [13] surveyed on different techniques applicable for each of the static/dynamic unlabeled/labeled anomalies. Like the various techniques discussed in structure based approaches are used to identify **static unlabeled anomalies**. In a similar fashion, for detecting **dynamic unlabeled anomalies** apart from link prediction, other techniques such as Bayesian analysis and scan statistical approaches (mainly applicable to hypergraphs) are used with each approach having its own application and benefits. In case of labeled anomalies, a number of techniques have been proposed for static as well as dynamic networks. A number of approaches were discussed in the survey paper for **static labeled anomalies** such as contextual anomalies, heavy vicinities, and opinion spam. As an example, for the detection of opinion spam a belief propagation method has been applied which deals with a set of hidden labels. One more approach called Trust Rank was discussed that follows a link analysis perspective in which it is assumed that good nodes would never point to bad nodes. Two basic principles followed are as follows:

- If you point to a Bad node, you're Bad.
- If a Good node points to you, you're Good.

So, a fundamental principle followed in this trust rank method is that trustworthy pages are unlikely to be linked immediately or within a predefined range to spam pages [102]. One of the prominent methods employed for such static labeled anomalies is the use of Information theory, a quantitative measure exercising the measures such as entropy to detect the anomalies. Likewise, the approaches used for unlabeled networks can be used to handle the **dynamic labeled anomalies**

as well. The only constraint imposed is the addition of information regarding the attributes. In most of the approaches the network structure is considered as static for a fixed time period and in order to add the dynamic concept the behavior of different nodes/modules is compared at different time intervals. Signal processing works on such principles by using the probabilistic features.

Similarly, Akoglu et al. [103] gave a survey of different graph based anomaly detection methods covering both the static/dynamic and labeled/unlabeled constraints. In each network structure, different quantitative and qualitative techniques have been very well categorized into different sub modules such as structure based, window based, community based and feature based. Moreover, researchers have described a number of real-world applications where graph based anomaly detection methods could be fit, for example, opinion spams, auction networks, social networks, telecommunication networks, trading networks, cyber crimes, security networks to name a few.

Recently, there has been an inclination toward detecting anomalies in dynamic networks. Therefore, a number of researchers are adding dynamic concept into their research work. For example, a number of anomaly detection techniques specially related to dynamic networks are recently surveyed by Ranshous et al. [104]. For instance, a scoring function is used to identify various types of anomalies. Categorization of anomalous behavior is based upon the scoring function being used along with the application area under consideration.

Also, the most significant and pertinent subset of nodes is used by Vigliotti and Hankin [105] to detect anomalous patterns in huge dynamic networks. In their work the experiments were performed on the temporal networks. Temporal information from two data sets namely VAST data set (2008) and Twitter data set was taken. In VAST data set, the telephonic calls among different nodes are examined. Also, the already available techniques are used to predefine an anomalous pattern and the projected approach is just validated over the working data set. But for the Twitter network being used no prior knowledge pertaining to anomalous patterns is already known, anomalous patterns and nodes need to be assumed and it has to be tested whether the stated hypothesis regarding anomalous or non-anomalous behavior is true or not.

Lately, community outliers have gained much attention and a number of approaches have been proposed for them. Detection of communities in online social networks is itself a huge and tedious task. Harenberg et al. [106] studied various disjoint and overlapping community detection techniques used in large-scale networks. Disjoint communities involve participation of an individual node in at most one community whereas in overlapping communities a node can participate in multiple communities. For the detection of disjoint communities different clustering or graph partitioning algorithms are frequently used. Similarly, the detection of overlapping communities makes use of various blockmodelling, clustering or clique extraction methods. Gao et al. [107] worked extremely well in the field of detecting community anomalies differentiating them from local and global anomalies. A simple approach that comes to mind to detect community anomalies is to make use of the approaches used for both the local and global anomalies i.e. DNODA (for local anomalies) and GLODA (for global anomalies) and infer that in order to detect community anomalies it is necessary to use information present at both the current data object and its neighborhood. Such designed approach referred to as Community Neighbor Algorithm (CAN) follows a two step procedure using both network information and data object information:

- Network or link information helps in partitioning the network structure into different communities.
- Each individual's object information facilitates in the identification of anomalies.

But Gao et al. [107] proposed an advancement in the above approach by integrating both the network and data object information to detect the community anomalies. The proposed approach is called Community Outlier Detection algorithm (CODA) which makes use of a probabilistic mixture model designed for multivariate data objects (objects with multiple attributes). Statistical anomaly detection approaches were used to detect the community anomalies in which depending upon the type of data associated, different distributions were analyzed where normal data objects were assumed to follow the defined distribution whereas anomalous objects deviate from it or follow some other distributions. In the proposed technique, two types of data objects were used- continuous data and text data and for normal behavior they were found to follow Gaussian and multinomial distribution respectively. It was found that any encountered anomaly followed a uniform distribution. A set of hidden variables for data objects and Hidden Markov Random Field (HMRF) for the network links are worked upon by the defined ICM and EM based algorithms. In order to make it more effective a set of hyper graph parameters like, threshold (indicating few anomalies for its high value and more anomalies for the low value), link importance (for the prediction of confidence level), number of components (small determining global anomalies and large the local ones) were also defined and used.

5. Conclusion and future scope

The paper presented a wide variety of approaches applicable for anomaly detection in data mining and social network domain. As it would have been very difficult to cover each and every technique in the review paper, best efforts have been made to cover most of the important ones.

The paper is structured into five major sections. Section 1 described the importance and growing trend toward social networks along with the presence of anomalous activities in it. A set of widely accepted formal definitions of anomaly have been tabulated. Section 2 classified the anomalies into various categories based upon different parameters. Finally, Sections 3 and 4 described the most prominent applicable approaches for detecting anomalies in data mining and social networks respectively. Each approach places its importance and relevant application based upon the type of anomaly to be detected.

In spite of enormous work done in anomaly detection domain there remains a number of shortcomings that could be addressed and worked upon in future.

One, dynamic labeled anomalies need to be focused more as comparatively less work has been done in this domain.

Secondly, temporal constraints need to be added to the approaches in order to add dynamicity. Although some approaches make use of the temporal information, for

example, use of a Markov chain model by Ye [108] to use the previous information for learning the defined model still it has been found that social networks have not been focused much with respect to the time dimensions.

Thirdly, the thin line of difference between the normal and anomalous users makes it difficult for the prediction of latter and hence more effective and novel techniques need to be framed.

Fourthly, not only the detection but prevention of anomalies is necessary because certain domains or applications cannot compromise with their sensitive information and hence need to be alert toward the presence of any anomalous or malicious individual far before its actual detection. But right from the beginning it has been seen that huge amount of work has been done toward anomaly detection rather than working on its prevention.

Fifthly, for each of the social network techniques namely behavior based, structure based or spectral based, there remains a scope for the exploration of a number of other graph metrics that could be used to detect the new kinds of anomalies present in different social networks.

Finally, analysis of big data in social networks for the presence of anomalies is the current focus of the researchers and very less work has been centered on it. Current techniques either focus on a predefined set of labeled data or observe the behavior of randomly chosen nodes rather than the unstructured behavior of data in social networks.

References

[1] Schneider F, Feldmann A, Krishnamurthy B, Willinger W. Understanding online social network usage from a network perspective. In: Proc of the ninth ACM SIGCOMM conference on Internet measurement conference; 2009. p. 35–48.

[2] Adamic L, Adar E. How to search a social network. Soc. Networks 2005;27(3):187–203.

[3] Ellison NB et al. Social network sites: definition, history, and scholarship. J Comput-Mediated Commun 2007;13(1):210–30.

[4] Wen C, Tan BCY, Chang KT-T. Advertising effectiveness on social network sites: an investigation of tie strength, endorser expertise and product type on consumer purchase intention. In: Proc of ICIS; 2009. p. 151.

[5] Chu Z, Widjaja I, Wang H. Detecting social spam campaigns on twitter. In: Applied cryptography and network security; 2012. p. 455–72.

[6] Akoglu L, McGlohon M, Faloutsos C. Oddball: spotting anomalies in weighted graphs. Adv Knowl Discov Data Min 2010:410–21.

[7] Bianca S. Stalkers, nude photos and beheadings: top facebook scams and malware attacks in 2014. Available: <http://www.hotforsecurity.com/blog/stalkers-nude-photos-and-beheadings-top-facebook-scams-and-malware-attacks-in-2014-11080.html>; 2014 [accessed: 23-Dec-2014].

[8] Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics 1969;11(1):1–21.

[9] Barnett V, Lewis T. Outliers in statistical data, vol. 3. New York: Wiley; 1994.

[10] John GH. Robust decision trees: removing outliers from databases. In: Proc of KDD; 1995. p. 174–9.

[11] Aggarwal CC, Yu PS. Outlier detection for high dimensional data. ACM Sigmod Rec 2001;30(2):37–46.

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

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

[14] Han J, Kamber M, Pei J. Data mining concepts and techniques. 3rd ed. Elsevier; 2012.

[15] Ma J, Perkins S. Online novelty detection on temporal sequences. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining; 2003. p. 613–8.

[16] Ma J, Perkins S. Time-series novelty detection using one-class support vector machines. In: Proc of the international joint conference on neural network, vol. 3; 2003. p. 1741–5.

[17] Shrivastava N, Majumder A, Rastogi R. Mining (social) network graphs to detect random link attacks. In: IEEE 24th international conference on data engineering (ICDE); 2008. p. 486–95.

[18] 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.

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

[20] Eberle W, Holder L. Anomaly detection in data represented as graphs. Intell Data Anal 2007;11(6):663–89.

[21] Hawkins DM. Identification of outliers, vol. 11. Springer; 1980.

[22] Brotherton T, Johnson T, Chadderdon G. Classification and novelty detection using linear models and a class dependent-elliptical basis function neural network. In: Proc. of IEEE world congress on computational intelligence on neural networks, vol. 2; 1998. p. 876–9.

[23] Augusteijn MF, Folkert BA. Neural network classification and novelty detection. Int J Remote Sens 2002;23(14):2891–902.

[24] Ratsch G, Mika S, Scholkopf B, Muller K. Constructing boosting algorithms from SVMs: an application to one-class classification. Pattern Anal Mach Intell IEEE Trans 2002;24(9):1184–99.

[25] Box GEP, Tiao GC. A Bayesian approach to some outlier problems. Biometrika 1968;55(1):119–29.

[26] Abraham B, Box GEP. Bayesian analysis of some outlier problems in time series. Biometrika 1979;66(2):229–36.

[27] Knox EM, Ng RT. Algorithms for mining distance based outliers in large datasets. In: Proceedings of the international conference on very large data bases; 1998. p. 392–403.

[28] Knorr EM, Ng RT, Tucakov V. Distance-based outliers: algorithms and applications; 2000. p. 237–53.

[29] Angiulli F, Pizzuti C. Fast outlier detection in high dimensional spaces. In: Principles of data mining and knowledge discovery. Springer; 2002. p. 15–27.

[30] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. ACM Sigmod Rec 2000;29(2):427–38.

[31] Chang J-W, Jin D-S. A new cell-based clustering method for large, high-dimensional data in data mining applications. In: Proceedings of the 2002 ACM symposium on applied computing; 2002. p. 503–7.

[32] Hautamäki V, Kärkkäinen I, Fränti P. Outlier detection using k-nearest neighbour graph. ICPR; 2004, p. 430–433.

[33] Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: identifying density-based local outliers. ACM Sigmod Rec 2000;29(2):93–104.

[34] Tang J, Chen Z, Fu AW-C, Cheung DW. Enhancing effectiveness of outlier detections for low density patterns. Adv Knowl Discov Data Min 2002:535–48.

[35] Jin W, Tung AKH, Han J, Wang W. Ranking outliers using symmetric neighborhood relationship. Adv Knowl Discov Data Min 2006:577–93.

[36] Papadimitriou S, Kitagawa H, Gibbons PB, Faloutsos C. Loci: fast outlier detection using the local correlation integral. In: Proc

of 19th international conference on data engineering; 2003. p. 315–26.
[37] Kriegel H-P, Kröger P, Schubert E, Zimek A. LoOP: local outlier probabilities. In: Proceedings of the 18th ACM conference on information and knowledge management; 2009. pp. 1649–52.
[38] Berkhin P. A survey of clustering data mining techniques. Group Multidimens Data 2006:25–71.
[39] Kaufman L, Rousseeuw PJ. Clustering Large Applications (Program CLARA). Find Groups Data: Introduct Cluster Anal 2008:126–63.
[40] Ng RT, Han J. CLARANS: a method for clustering objects for spatial data mining. IEEE Trans Knowl Data Eng 2002;14(5):1003–16.
[41] HaiZhou D, YongBin L. An improved BIRCH clustering algorithm and application in thermal power. In: international conference on web information systems and mining (WISM), vol. 1; 2010. p. 53–6.
[42] Karypis G, Han E-H, Kumar V. Chameleon: hierarchical clustering using dynamic modeling. Comput (Long Beach Calif) 1999;32(8):68–75.
[43] Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd 1996;96:226–31.
[44] Zhou S, Zhou A, Jin W, Fan Y, Qian W. FDBSCAN: a fast DBSCAN algorithm. RUAN JIAN XUE BAO 2000;11(6):735–44.
[45] Viswanath P, Pinkesh R. l-dbscan: a fast hybrid density based clustering method. In: 18th international conference on pattern recognition (ICPR), vol. 1; 2006. p. 912–5.
[46] Ruiz C, Spiliopoulou M, Menasalvas E. C-dbscan: density-based clustering with constraints. Rough Sets, Fuzzy Sets Data Min Granul Comput 2007:216–23.
[47] Kisilevich S, Mansmann F, Keim D. P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In: Proceedings of the first international conference and exhibition on computing for geospatial research & application; 2010. p. 38.
[48] Kryszkiewicz M, Lasek P. TI-DBSCAN: clustering with DBSCAN by means of the triangle inequality. In: Rough sets and current trends in computing; 2010. p. 60–9.
[49] Nanopoulos A, Theodoridis Y, Manolopoulos Y. C2P: clustering based on closest pairs. In: Proceedings of the international conference on very large data bases; 2001. p. 331–40.
[50] Guha S, Rastogi R, Shim K. Cure: an efficient clustering algorithm for large databases. Inf Syst 2001;26(1):35–58.
[51] Ertöz L, Steinbach M, Kumar V. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: SDM; 2003. p. 47–58.
[52] He Z, Xu X, Deng S. Discovering cluster-based local outliers. Pattern Recognit Lett 2003;24(9):1641–50.
[53] Kohonen T. The self-organizing map. Neurocomputing 1998;21(1):1–6.
[54] Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. Appl Stat 1979:100–8.
[55] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 2002;24(7):881–92.
[56] Arthur D, Vassilvitskii S. k-means + +: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms; 2007. p. 1027–35.
[57] Bahmani B, Moseley B, Vattani A, Kumar R, Vassilvitskii S. Scalable k-means + +. Proc VLDB Endow 2012;5(7):622–33.
[58] Wu N, Zhang J. Factor analysis based anomaly detection. Information assurance workshop, IEEE Systems, Man and Cybernetics Society; 2003. p. 108–15.
[59] Vinueza A, Grudic G. Unsupervised outlier detection and semi-supervised learning. Technical report CU-CS-976-04. University of Colorado at Boulder; 2004.
[60] Pires A, Santos-Pereira C. Using clustering and robust estimators to detect outliers in multivariate data. In: Proceedings of the international conference on robust statistics; 2005.
[61] Portnoy Leonid, Eskin Eleazar, Stolfo Sal. Intrusion detection with unlabeled data using clustering. In: Proceedings of ACM CSS workshop on data mining applied to security (DMSA); 2001.
[62] Beutel A, Xu W, Guruswami V, Palow C, Faloutsos C. CopyCatch: stopping group attacks by spotting lockstep behavior in social networks. In Proceedings of the 22nd international conference on World Wide Web; 2013. p. 119–30.
[63] Moya MM, Koch MW, Hostetler LD. One-class classifier networks for target recognition applications. NASA STI/Recon Tech. Rep. N, vol. 93; 1993. p. 24043.
[64] Li K-L, Huang H-K, Tian S-F, Xu W. Improving one-class SVM for anomaly detection. Int Conf Mach Learn Cybernetics 2003;5:3077–81.
[65] Lumini A, Nanni L. Ensemble of on-line signature matchers based on OverComplete feature generation. Expert Syst Appl 2009;36(3):5291–6.
[66] David MJ. Tax. One-class classification; concept-learning in the absence of counter-examples. ASCI dissertation series, vol. 65; 2001.
[67] Raudys Š. On the effectiveness of Parzen window classifier. Informatica 1991;2(3):434–54.
[68] Heckerman D. Bayesian networks for data mining. Data Min Knowl Discov 1997;1(1):79–119.
[69] Kruegel C, Mutz D, Robertson W, Valeur F. Bayesian event classification for intrusion detection. In: Proc of 19th annual computer security applications conference; 2003. p. 14–23.
[70] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[71] Manevitz LM, Yousef M. One-class SVMs for document classification. J Mach Learn Res 2002;2:139–54.
[72] Eskin E, Arnold A, Prerau M, Portnoy L, Stolfo S. A geometric framework for unsupervised anomaly detection. Appl Data Min Comput Secur 2002:77–101.
[73] Perdisci R, Ariu D, Fogla P, Giacinto G, Lee W. McPAD: a multiple classifier system for accurate payload-based anomaly detection. Comput Networks 2009;53(6):864–81.
[74] Piciarelli C, Micheloni C, Foresti GL. Trajectory-based anomalous event detection. IEEE Trans Circuits Syst Video Technol 2008;18(11):1544–54.
[75] Moradi M, Zulkernine M. A neural network based system for intrusion detection and classification of attacks. In: Proceedings of the 2004 IEEE international conference on advances in intelligent systems-theory and applications; 2004.
[76] Liu G, Yi Z, Yang S. A hierarchical intrusion detection model based on the PCA neural networks. Neurocomputing 2007;70(7):1561–8.
[77] Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Syst Appl 2014;41(4):1690–700.
[78] 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).
[79] Lin W-C, Ke S-W, Tsai C-F. CANN: an intrusion detection system based on combining cluster centers and nearest neighbors. Knowl-Based Syst 2015;78:13–21.
[80] Butler BS. Membership size, communication activity, and sustainability: a resource-based model of online social structures. Inf Syst Res 2001;12(4):346–62.
[81] Heidemann J, Klier M, Probst F. Identifying key users in online social networks: a PageRank based approach; 2010.

[82] Nazir A, Raza S, Chuah C-N. Unveiling Facebook: a measurement study of social network based applications. In: Proceedings of the eighth ACM SIGCOMM conference on Internet measurement; 2008. p. 43–56.
[83] Singh L, Getoor L, Licamele L. Pruning social networks using structural properties and descriptive attributes. In: Proc of international conference on data mining; 2005.
[84] Li H, Cui J-T, Ma J-F. Social influence study in online networks: a three-level review. J Comput Sci Technol 2015;30(1):184–99.
[85] Vanetti M, Binaghi E, Carminati B, Carullo M, Ferrari E. Content-based filtering in on-line social networks. Privacy and security issues in data mining and machine learning. Springer; 2011. p. 127–40.
[86] Bhat SY, Abulaish M. Using communities against deception in online social networks. Comput Fraud Secur 2014;2014(2):8–16.
[87] Dagon D, Qin X, Gu G, Lee W, Grizzard J, Levine J, Owen H. Honeystat: local worm detection using honeypots. In: Recent advances in intrusion detection; 2004. p. 39–58.
[88] Viswanath B, Bashir MA, Crovella M, Guha S, Gummadi KP, Krishnamurthy B, Mislove A. Towards detecting anomalous user behavior in online social networks. In: Proceedings of the 23rd USENIX security symposium (USENIX Security); 2014.
[89] Xiao C, Freeman DM, Hwa T. Detecting clusters of fake accounts in online social networks. In: Proceedings of the eighth ACM workshop on artificial intelligence and security; 2015. pp. 91–101.
[90] Getoor L, Diehl CP. Link mining: a survey. ACM SIGKDD Explor Newslett 2005;7(2):3–12.
[91] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. J Am Soc Inf Sci Technol 2007;58(7):1019–31.
[92] Rattigan MJ, Jensen D. The case for anomalous link discovery. ACM SIGKDD Explor Newslett 2004;7(2):41–7.
[93] Zheleva E, Getoor L, Golbeck J, Kuter U. Using friendship ties and family circles for link prediction. Advances in social network mining and analysis. Springer; 2010. p. 97–113.
[94] Henderson K, Gallagher B, Li L, Akoglu L, Eliassi-Rad T, Tong H, Faloutsos C. 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.
[95] Hassanzadeh R, Nayak R, Stebila D. Analyzing the effectiveness of graph metrics for anomaly detection in online social networks. Web Inf Syst Eng 2012:624–30.
[96] Rezaei A, Kasirun ZM, Rohani VA, Khodadadi T. Anomaly detection in online social networks using structure-based technique. In: Eighth international conference on internet technology and secured transactions (ICITST); 2013. p. 619–22.
[97] Ying X, Wu X, Barbará D. Spectrum based fraud detection in social networks. In: IEEE 27th international conference on data engineering (ICDE); 2011. p. 912–23.
[98] Agarwal S, Branson K, Belongie S. Higher order learning with graphs. In: Proceedings of the 23rd international conference on machine learning; 2006. p. 17–24.
[99] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci 2002;99(12):7821–6.
[100] Freeman LC. A set of measures of centrality based on betweenness. Sociometry. JSTOR; 1977.
[101] Newman MEJ. Detecting community structure in networks. Eur Phys J B – Condens Matter Complex Syst 2004;38(2):321–30.
[102] Gyöngyi Z, Garcia-Molina H, Pedersen J. Combating web spam with trustrank. In: Proceedings of the thirtieth international conference on very large data bases, vol. 30; 2004. p. 576–87.
[103] Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: a survey. Data Min Knowl Discov 2014:1–63.
[104] Ranshous S, Shen S, Koutra D, Harenberg S, Faloutsos C, Samatova NF. Anomaly detection in dynamic networks: a survey. Wiley Interdiscip Rev Comput Stat 2015;7(3):223–47.
[105] Vigliotti MG, Hankin C. Discovery of anomalous behaviour in temporal networks. Soc Networks 2015;41:18–25.
[106] Harenberg S, Bello G, Gjeltema L, Ranshous S, Harlalka J, Seay R, Padmanabhan K, Samatova N. Community detection in large-scale networks: a survey and empirical evaluation. Wiley Interdiscip Rev Comput Stat 2014;6(6):426–39.
[107] Gao J, Liang F, Fan W, Wang C, Sun Y, Han J. On community outliers and their efficient detection in information networks. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining; 2010. p. 813–22.
[108] Ye N. A markov chain model of temporal behavior for anomaly detection. In: Proceedings of the 2000 IEEE Systems, Man, and Cybernetics information assurance and security workshop, vol. 166; 2000. p. 169.