RESEARCH ARTICLE **SPECIAL ISSUE**

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S039

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Study of Anomalous Subgraph Detection in Social Networks

Anagha Ajoykumar [1], *Venkatesan M* [1]

[1]*Department of Computer Science and Engineering, NIT Puducherry,India.*

Emails: anagha.melepat@gmail.com, venkatesan.msundaram@nitpy.ac.in

Article History

Received: 28 February 2023
Accepted: 18 March 2023

Keywords:

Online social networks;
Social graphs;
Graph theory;
Anomaly detection;
Anomalous subgraphs;
Graph-based anomaly detection

Abstract

The reliance on the internet has made it possible for a number of internet networks to arise, each with a distinct user base. Intentionally or not, we are all members of a wide range of social networks. Online interpersonal and professional interactions are significantly influenced by social networking. It has a tremendous effect on a global scale and an individual one, affecting a wide range of industries including education, healthcare, entertainment, banking, and telecommunications. As their dependency on social media increases, users are publishing a lot of information about themselves online, leaving their data and themselves vulnerable to the outside world and making them ideal targets for criminals which not only jeopardizes the security of the social network's data but also make way to a slew of other potentially harmful situations, ranging from identity theft to major cybercrime such as hacking, cyber-bullying cyber threats, and even national security threats such as terrorism. This necessitated the development of methods and strategies to detect fraudulent users or abnormalities on social media. A graph framework is the most prominent form of mathematical modeling of a social network, hence deducing methods to identify abnormalities from a graph is critical. This paper gives a thorough review of graph-based anomaly detection methods, with a focus on identifying anomalous subgraphs. Since anomaly detection on subgraphs has received little attention from the researchers' community in contrast to other anomalous units, we examine the numerous research problems and outstanding questions in this domain.

1. Introduction

One cannot envision living without the internet in today's day and age. It has become a necessary and integral element of our very existence. It was a revolution that transformed the world's fundamental basic form of communication. From grocery shopping to socialization, it has now become the default mode of interaction in every aspect of modern life. It has evolved from its early days as a static network acting as a repository of knowledge maintained solely by professionals to being the world's biggest computer network, through many changes and becoming as a worldwide parallel society in its own capacity. Throughout the short history of the Internet, the advent of Web 2.0 in the initial decade of the 21st century marked a turning point by encouraging the creation of interactive, crowd-sourced communication platforms. Online social networks emerged as a result of this (Mislove et al.).

From a basic platform for information exchange to a sophisticated interdisciplinary tool, internet

is now a tool for content creation, interaction and even to relax and unwind. In the previous decade, social networks have demonstrated their effectiveness in a variety of disciplines, with a massive surge in usage and applications. It aids in the connection of people from all across the world and provides quick and easy communication, like in a friends' networks (Mislove et al.), co-authorship networks (Barabsi et al.), mobile call networks (Nanavati et al.), e-mail communication networks (Al-Mukhaini, Al-Qayoudhi, and Al-Badi), instant messenger networks (Nanavati et al.). Apart from that, it has a significant and significant role to play in other vital and serious aspects of society, such as academia (Curran and Hugh), healthcare (M. Lee, Yoon, and K. .-.-S. Lee), legislation (Bright, Brewer, and Morselli), law enforcement (Garside et al.), and even more crucial areas, such as military and intelligence services (Willis and Delbaere) or pharmaceutical services Putting all of the positives aside, it really is no surprise that social media has a negative side (). Apart from the negative consequences on users such as strenuous lifestyle, sleep disruption, inattentiveness, procrastination, increased sense of social isolation, and so on, there are some severe risks in social networks induced by the presence of malicious users or suspicious activities by these users over the internet, which overwhelm the remaining users and hence give way to illegal behaviour. Frauds and scams, along with breach of privacy, data theft, identity theft, misleading information, cyber bullying, cyber-attacks, hacking, and other issues, are serious concerns with billions of fraudulent members on the network with unknown motives like even terrorism (Keyvanpour, Moradi, and Hasanzadeh) (Liu and Chawla) (R. Yu, He, and Liu) (Bindu and Thilagam). Hence, it's necessary to detect the presence of these fake users on the network and alert the other legitimate users.

Finding anomalies in a social network is therefore crucial. They signify unusual or illegal behaviour that is not expected or visible during network operations normally. It's possible that this node, edge, or subgraph is abnormal. We must look at how the various users of the network interact with one another in order to find these (Wasserman and Faust). Considering graphs are most commonly used to represent graphs, these anomalous units could be found using graph mining techniques. There are different methods for spotting fraudulent units on each social network because each one has a unique structure and set of features. The majority of social network researchers have developed a number of tools and methodologies for identifying abnormalities in social networks using structural properties.

Detection of anomaly in social networks offers a plethora of real-life applications. Due to the increase in fraudulent activity on social networks, more people are suffering financial loss as well as harm to other users. Individuals within the network, as well as unauthorised users, poses a threat to organisations. Anomaly detection aids in the identification of significant users and rare activities in a network, such as uncommon connections between nodes, in addition to detecting fraudulent, untrustworthy, or dangerous behaviour. This highlights the significance of digital forensics in this setting. Social Network Forensics (Keyvanpour, Moradi, and Hasanzadeh) is a relatively new field of study that focuses on detecting, analysing, preventing, and predicting undesirable activities in social networks. With anomaly detection, this information may be used to maintain the network safe and optimise its impact.

In this paper, we provide a comprehensive and systematic review of the research works done in the area subgraph anomaly detection in social networks. The main contributions of this paper can be summarised as follows:

- Addressing the necessity for subgraph anomaly detection in social networks.
- Identifying the critical components associated with subgraph anomaly detection in social networks,
- Providing a comprehensive review of the state of the art in subgraph anomaly detection.
- Exploring the open problems and research challenges in the field of subgraph anomaly detection in social networks.

The rest of the paper is organized as follows. Section 2 presents the background topics related to mining social networks for anomalies. Section 3 discusses the existing works on subgraph anomaly detection in social networks. After presenting a discussion on the open problems and research challenges in Section 4, we conclude the review in Section 5.

2. Technical Preliminaries

In this section, we discuss the following background topics related to anomalous subgraph detection: Social Network, Graph Theory, Anomaly Detection in Graphs, Types of Anomalies in Social Network and Anomalous Subgraphs.

2.1. Social Network

A social network is a made up of a number of participants or agents, known as nodes, who are connected by various kinds of connections or links (R. J. Wilson). This is depicted in Figure 1. It represents interaction between social entities. These actors or participants could be individuals, organizations, communities, and so on. The relationship between these entities could be of any kind like friendship, common interest, beliefs among the others. Understanding and analysing a social structure, such as identifying local and global traits, influential actors, and network dynamics, is made easier when seen as a network. Some examples of social networks are friends' networks (Mislove et al.), telephone networks (Nanavati et al.), e-mail networks (Al-Mukhaini, Al-Qayoudhi, and Al-Badi), to name a few. The study of social networks' characteristics is known as social network analysis (R. J. Wilson). It enables us to look at the interactions between those connected via social networks and get understanding of the patterns present. Graphs are usually used to model social networks.

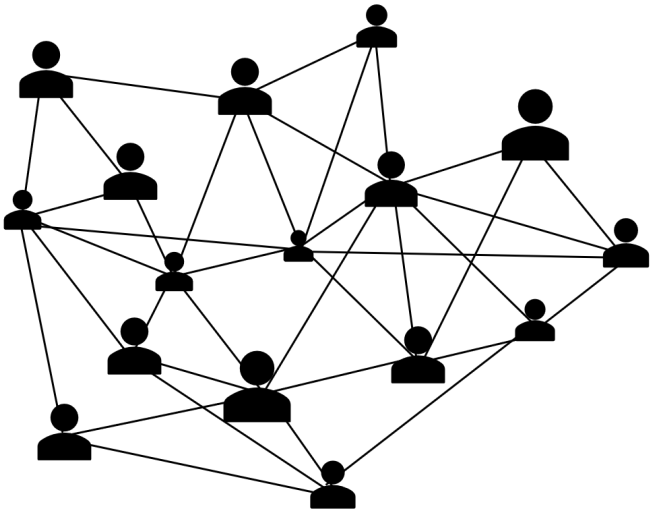

FIGURE 1. Representation of social network with various user interconnected with each other.

2.2. Graph Theory

In a very basic sense, graphs are simply set of points connected by a set of lines. According to (Biggs, Lloyd, and R. Wilson) (Bondy and Murty), a simple graph G consists of a non-empty finite set V(G) of elements called vertices (or nodes), and a finite set E(G) of distinct unordered pairs of distinct elements called edges. V(G) is called the vertex set and E(G) is called the edge set of the graph G. An edge {v, w} is said to join the vertices v and w, and is commonly written as vw. (Grubbs) defines graph as an ordered triple (V(G), E(G), ΨG) consisting of a nonempty set V(G) of vertices, a set E(G), disjoint from V(G), of edges, and an incidence function ΨG that associates with each edge of G an unordered pair of (not necessarily distinct) vertices of G. For example, Figure 2 represents a simple graph with six vertices and seven edges.

For the sake of this review, several additional notions relating to graph theory that are crucial are also covered as specified in (Biggs, Lloyd, and R. Wilson). If there is a one-to-one correspondence between the vertices of two graphs, then two graphs are said to be isomorphic if the number of edges connecting any two vertices in one graph equals the number of edges connecting the corresponding vertices in the other. An example is shown in Figure 3. Two vertices of a graph are said to be adjacent if there is an edge joining them, and those vertices are then incident with that edge. In the similar manner, two distinct edges are adjacent if they have a vertex in common. The degree of a vertex is the number of edges incident with it. An example is shown in Figure 4. A subgraph of a graph is the one whose edges are all members of the parent graph's edge set and whose vertices are all members of the parent graph. For example, Figure 5 denotes a subgraph of the graph in Figure 2 obtained by deleting the vertices E and F.

2.3. Anomaly Detection in Graphs

Anomaly is a term with lots of variations in definition as stated by different people and in different contexts and applications. (Barnett and Lewis) defines an outlier as an observation that stands out significantly from the other observations in the sample. It is an observation (or selection of observations), according to (John), that doesn't seem to match with the other data. According to (Aggarwal

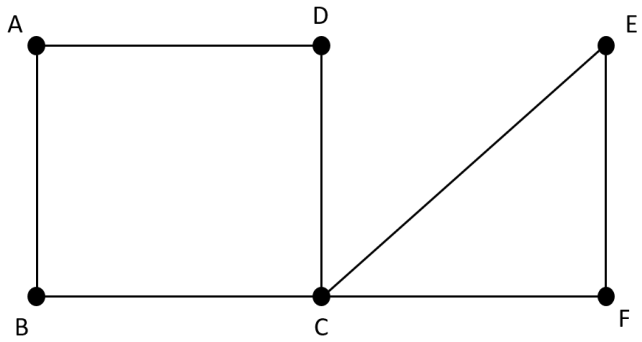

FIGURE 2. A simple graph with 6 vertices and 7 edges.

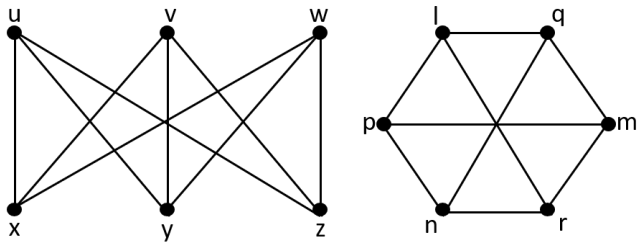

FIGURE 3. Isomorphic graphs under the correspondence u↔l, v↔m, w↔n, x↔q, y↔r, z↔p.

FIGURE 4. Adjacent vertices u and v; degree of each of u and v is 3; adjacent edges e and f.

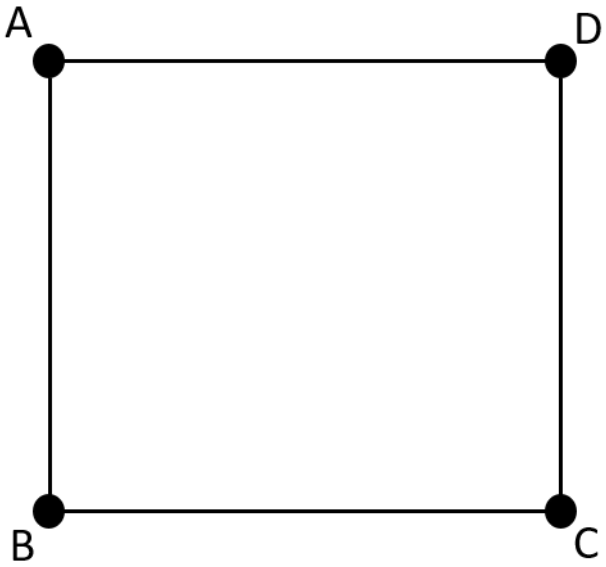

FIGURE 5. Subgraph of the graph in Figure 2 obtained by deleting the vertices E and F.

and P. S. Yu), an outlier can also be described as shocking real-world data, which is when a point is mistakenly classified as belonging to class B when it actually belongs to class A, shocking the observer. Like in Figure 6, outliers are noise points that are outside of a specified set of clusters or points that are outside of the specified set of clusters but still independent from the noise (Chandola, Banerjee, and Kumar). (Savage et al.) defines anomalies as data patterns that do not fit a recognised pattern of expected behaviour. (Vengertsev and Thakkar) defines it as portions of the network with a structure that differs from what you may expect from the network's typical structure. Anomaly detection refers to the problem of locating certain patterns or substructures that are unexpected, undesirable, and should be identified to safeguard the network and its users (Kaur and Singh).

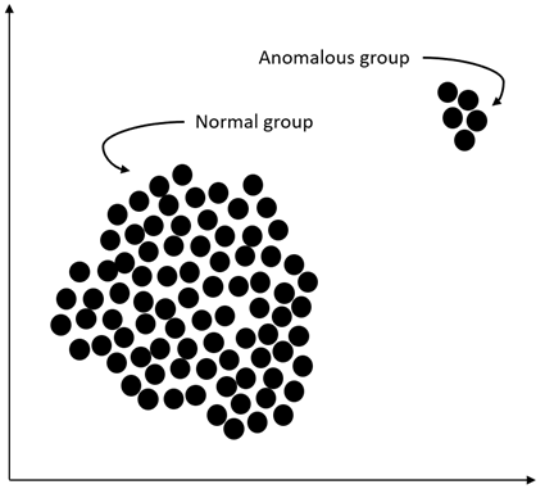

FIGURE 6. Example of anomalies in a simple two-dimensional graph.

2.4. Types of Anomalies in Social Network

Based on a variety of factors, anomalies can be divided into many categories (Gao et al.). The three primary categories of anomaly are point anomaly, contextual anomaly, and collective anomaly. This could be dependent on the nature and extent of anomalous (Savage et al.). A point anomaly is a single data point or user that behaves differently from the rest of the data. A data collection will contain contextual anomalies, which are conditional anomalies that show up when a data object deviates significantly from the context. When a group of data items behaves differently from other groups, even though

the individual data items itself might not be abnormal, this is called a collective anomaly. Anomalies, like in (Vengertsev and Thakkar), can be categorised as static or dynamic according on the network topology being used, as well as labelled or unlabelled depending on the type of information provided at a node or an edge. White crow anomalies and in-disguise anomalies, a different kind of anomalies, were introduced in (W. Eberle and L. Holder). In a situation where one data object significantly differs from other observations, a phenomenon known as the ”white crow anomaly” arises. That seemed almost impossible in this scenario. An in-disguise anomaly is an extremely subtle deviation from the norm in behaviour that is difficult to detect. The graphical qualities of anomalies or structural processes like insertion, deletion, and modification can also be used to categorise them (Akoglu, Mcglohon, and Faloutsos). Anomalies in (Ma et al.) such as Near Stars/Cliques, Heavy Locality, and Particular Dominant Links are based on types of communication and linkages among nodes. Near Stars have neighbours who are entirely linked, whereas Near Cliques have neighbours who are completely isolated. Particular dominant link implies significant load around a certain entity, whereas heavy locality implies abnormally heavy load around a specific group.Deciding which parameter must be considered to define the categories of anomalies depends on the application for which the abnormalities are discovered. The detection approaches are used to find anomaly units in networks, such as edges, nodes, subgraphs, and/or events (Wasserman and Faust). When identifying users whose behaviour considerably deviates from the norm, we examine a group of nodes as anomalies. If we need to identify unexpected or irregular interactions between users, a subset of edges may be considered anomalous. Anomalous subgraph identification seeks for subnetworks where the way the nodes interact is different from how it is throughout the rest of the network. Events in dynamic networks are the fixed time intervals where the social network diverges considerably from the previous and following networks in the sequence.

2.5. Anomalous Subgraphs

In reality, anomalies may collaborate and act in concert with others to get advantages. Fake users in a review sites network, for example, may publish fraudulent reviews to promote or disparage specific products. These data are shown as graphs, and the interactions among them typically yield suspected sub-graphs (Ranshous et al.). Because it is extremely difficult to enumerate every conceivable subgraph in even a single graph, discovering subgraphs with unexpected behaviour requires a different approach than detecting anomalous vertices or edges (Greene, Doyle, and Cunningham). As a result, the subgraphs that are analysed or identified are largely limited, such as those discovered with community detection methods. Matching algorithms, such as the community matching approach, are required in these cases to track the subgraphs through time steps (Cook and L. B. Holder). To detect subgraph anomalies in static social networks, various techniques such as Network Structure based approach and signal processing-based approach are used, whereas in dynamic social networks, community-based approach, matrix/tensor decomposition-based approach, probability-based approach, and so on are used. In the last several years, a lot of work has gone into employing deep learning approaches to solve this problem (Wasserman and Faust). Due to the versatility of heterogeneous graphs in depicting intricate interactions between various types of real objects, deep network representation approaches have been utilised in several recent articles to identify real-world abnormalities (Ranshous et al.).

3. Existing methods of Subgraph Anomaly Detection

Almost all research papers on subgraph anomaly detection have been considered for this study. Since, subgraph anomaly detection is a slightly under-explored area, the number of works carried out is less and there are even lesser works that use deep learning techniques. Most non-deep learning technique-based works can be broadly classified in to methods which are applied on static graphs, or dynamic graphs, as well as attributed or non-attributed graphs. Table 1 summarises the techniques reviewed and their limitations.

In Subdue (Noble and Cook), frequent substructures are found via a greedy beam search, and they are then rated according to the Minimum Description Length (MDL) concept. Substructures that are

TABLE 1. Summary of reviewed anomalous subgraph detectiontechniques

Input	Technique	Limitations
	Non-deep Learning methods	
Static Unattributed Graph	Minimum description length (Noble and Cook)	Does not work for numeric values or continuous attributes
Static Unattributed Graph	Minimum description length (Rattigan and Jensen)	Cannot operate on unweightrd graph with discrete vertex and edge labels
Static Unattributed Graph	Randomized graph traversal (Thompson and Eliassi-Rad)	Cannot detect in time-evolving social networks
Dynamic Unattributed Graph	Product rule for the central limit theorem (Miller, Bliss, and Wolfe)	Does not study edge correlations
Static Unattributed Graph	Eigenvector l1 norms (Newman)	Cannot detect subgraphs than can be separated from background in space of small number of eigenvector
Static Attributed Graph	Extension of subdue (Mongiovì et al.)	Cannot be used for online detection of anomalies using dynamic graphs
Dynamic attributed Graph	Heaviest subgraph detection with fixed length moving window (Gupta et al.)	Not realistic in dynamic running conditions and system operations
Static Attributed Graph	Query based paradigm using egonets (Zhao and Han)	Cannot be used on temporal graphs or high dimensional data
Static Unattributed Graph	Signal processing on chung lu random networks (Hong)	Connections between anomalous nodes is not established
Static Attributed Graph	Tree approximation and dynamic programming (Berk and Jones)	Cannot be used on dynamic multi-attributed heterogenous networks
	Deep Learning methods	
Static Unattributed Graph	Dense block detection approach (Ester et al.)	Does not work on non-bipartite graphs
Static Unattributed Graph	Dense block detection approach (Akoglu, Tong, and Koutra)	Does not work on non-bipartite graphs
Dynamic Unattributed Graph	Residual matrix-based convolutional neural network (H. Wang et al.)	Cannot be used on attributed graphs

more frequent in the graph have lower Description Lengths (DL), which implies that substructures with a high DL are more anomalous. (W. bibinitperiod Eberle and L. Holder) presented an algorithm based this heuristic for anomaly detection. It preserves a parent list at the start, an ordered list of all detected substructures. All of the substructures are repeatedly removed from the parent list, their extensions are generated, evaluated, and then added to the list. A second list of the top substructures found so far is kept up to date when new substructures are produced. The substructure with the highest value is reported, and before the next iteration starts, each instance of the substructure is replaced with a new vertex representing it. It operates under the premise that because anomalous substructure typically contains few common patterns and is therefore more easily detectable than other subgraphs, it tends to experience less compression than other subgraphs. However, it does not work for numeric values or continuous attributes.

(Rattigan and Jensen) developed three approaches for graph-based data fraud prevention and detection. For the purpose of anomaly recognition, they classify graph changes into three categories: modification, vertex/edge deletion, and

vertex/edge insertions. One of these subtypes is the main focus of each algorithm. They primarily employ the theory of the minimum description length (Shrivastava, Majumder, and Rastogi) to find the normative pattern, and then they take a different route to find specific anomalous kinds. The first algorithm uses the normative patterns to look for patterns whose cost of transformation is below a given threshold. More anomalous patterns are those that have a lower value for a combination of cost and frequency. The second algorithm examines and assesses the likelihood of the presence of extensions of normative substructures. Less probable patterns are more anomalous. The third algorithm chooses patterns that are ancestors of the substructure and have the highest potential substructure of the normative pattern. More anomalous patterns have lower transformation costs. Each approach can detect anomalies on graphs with various sizes with high detection accuracies and low false positive rates, but it is unable to do so on unweighted graphs with discrete vertex and edge labels.

(Thompson and Eliassi-Rad) provides two effective methods to mine subgraphs satisfying the Random Link Attack (RLA) property. Using a random selection algorithm, the attacker node chooses a group of victim nodes to connect with in a wide variety of assaults on communication networks. The primary feature that distinguishes the assault group from a social subgraph is the existence of exterior triangles, which the attackers establish with the rest of the network and consist of one attack node and two non-attackers. The number of these triangles will be quite minimal for a malicious node. In order to create a potential attack cluster, the first technique, known as GREEDY, iteratively attaches nodes with a greater extent of connectedness to the attack cluster. An attack node will connect to numerous additional attack nodes located throughout the network in order to avoid being found. It is unlikely that many victims will have edges to the same good node in the neighbourhood of a subset of attackers and a few victims in a neighbourhood made up of victims, attackers, and a few good nodes. Nodes in the neighbourhood with powerful link to the subset will thus either be an attacker or a victim. If the node has more triangles than a certain threshold, it is probably an attacker. In the second method, referred to as triangle random walk (TRWALK), a randomised graph traversal is carried out, each time beginning at a questionable node. A triangle is randomly selected from its surroundings and then swapped to another triangle whose edge it shares and is repeated until a collection of nodes visited during the TRWALK is acquired. An attack set is likely to result from an iteration that does not cross any exterior triangles. The subset is examined for an RLA instance before moving on to the following suspect.

(Miller, Bliss, and Wolfe) employs a scalable method based on the Product Rule for the Central Limit Theorem to assess the likelihood of occurrences and identify anomalous activity in volatile time-evolving networks. In order to recognise an unusual occurrence, the method initially develops a baseline for normal behaviour by finding persistent patterns among vertices, which is a group of vertices that form a linked component and communicate often. It then makes use of this data to identify unusual behaviour on both a local and a global level. It simulates a weighted ”cumulative” graph from the database of the time-evolving network, which is a dynamic graph made up of a fixed set of vertices and a set of time-stamped edges. It includes all previous edges but prioritises more recent ones, making it useful for estimating connection strength on average. By taking linked edges with weights that are higher than a certain threshold and regularly recurring edges, it extracts persistent patterns. We compare the present activity at a given period with the activity that is anticipated based on prior behavioural trends in order to identify anomalies. If the actual activity differs noticeably from the anticipated activity, we define the occurrence as anomalous. A specified anomality threshold is used in this comparison, and anomalous behaviour is marked for examination and analysis.

(Newman) offers a framework that presents a signal processing-based detection theory for anomalies in unweighted, undirected graphs applying the L1 properties of the eigenvectors of the modularity matrix of the graph (Davis et al.). This measure is shown to have a reasonably low variance for numerous kinds of randomly created graphs and to accurately detect the presence of an anomalous subgraph when it is not intimately related with stronger sections of the background graph. By projecting the large graph into the space of its two principal

eigenvectors, computing a Chi squared test statistic, and comparing the result to a threshold, the analysis of the principal eigenvectors of the modularity matrix can also reveal the presence of a small, tightly connected component embedded in the larger graph. The ”strength” with which a vertex is a member of the linked community is correlated with the size of the vertex's component in an eigenvector. As a result, if a small group of vertices form a community, with few of them belonging to other communities, an eigenvector that is well aligned with this group will exist. This implies that the norm of this eigenvector will be lower than the norm of an eigenvector with a similar eigenvalue when there isn't an abnormally dense subgraph. So, the subgraph detection algorithm calculates the modularity matrix's eigen decomposition for the graph, determine the L1 norm for each eigenvector, then take away the anticipated value, normalising the result by the L1 norm. The presence of an anomalous subgraph embedding is indicated if any of these modified L1 norms falls below a predetermined threshold.

(Mongiovì et al.) described an algorithm that analyses labelled graphs for structural and numerical anomalies. By giving anomaly scores, it expands the original Subdue method to encompass numerical outliers. By changing the graph so that all normal edges have a constant value while anomalous edges evaluate to a collection of values using K-Nearest Neighbours, it distinguishes between normal values and anomalous ones. A collection of feature vectors is created using the characteristics of the vertices or edges under scrutiny. To get the k-distance to a vertex's kth closest neighbour, the feature vector for each vertex is compared to the full set. If the k-distance is normal, the constant value is returned; if not, the outlierness index is calculated using this distance. Each vertex is given an anomaly score and then split up into sets so that similar types of vertex are grouped together. A vertex's type may depend on the label assigned to it, the kind of edges it is related to, or a variety of other criteria. For edges as well, the same procedure is done. On this, the Subdue technique is used to get common substructures, and the compressed graph's anomaly scores are computed. A weighted graph may be used to identify structural and numerical abnormalities by swapping out the numerical edge weights for anomaly scores. It only functions with static graphs, though.

(Gupta et al.) suggests a method to create a thorough list of all important anomalous locations in a dynamic network. It begins by outlining the regular behaviour of network edges, ranks edges over time according to how out of the ordinary their behaviour is, and then suggests a method for calculating extended areas of anomalous edges in order to locate anomalous locations. When given an edge and its weight at a specific time, the p-value is calculated as the percentage of timestamps where the same edge has an equal or greater weight recorded on it. The p-value of an observed score decreases as the observation becomes more abnormal. The approach is based on the NetAmoeba technique, which roughly approximates the Heaviest Dynamic Subgraph. The maximum score subsequence, which considers a given subgraph and determines the best subgraph for this interval by optimising the time interval that yields the highest score, computes the heaviest subgraph last. After receiving as inputs a score threshold and a parameter that sets the number of failures that must occur before stopping, the method outputs a collection of anomalous locations whose score exceeds the threshold. It runs NetAmoeba iteratively and begins the search with a seed generation process. The network is then cleared of the positive weights of edges that are located inside the recently found region. The algorithm stops when the final group of identified regions has a score below the threshold. It is unlikely that a region with a score higher than the requirement will be found in the future, hence the region is deemed anomalous if it is not found numerous times in a row.

(Zhao and Han) suggests a way for determining a subgraph's outlierness using a max-margin framework. This method compares the margin for linked to non-linked node pairings nearby a subgraph match in order to determine the outlier scores, which are used to rank such subgraph outliers. In order to allow the user, the freedom to find outliers complying to a particular architecture and conditionals stored in the form of a query, the study focuses on query-based outliers exploiting neighbourhood data. The first is to find every instance of the query that matches the given entity-relationship graph. The collection of all matches for the query is provided to us by an SPath-based solution (B. Wang

et al.). By simply listing all the graph edges that each match covers, it is simple to compute the set of all induced matches. The next step is to calculate the outlier score for each match. The outlier score of a match is computed using the margin for the max-margin hyperplane or the finest feature weight vector. After the matches are arranged according to their outlier scores in a non-ascending order, the top few matches might be returned as outliers. The suggested approach only applies to static networks and hence it is incompatible with temporal networks.

(Hong) provide a pre-processing method in which a local vertex set with a high likelihood of including the anomalous vertices is successfully obtained by subgraph search. The detection approach based on the local set may greatly improve detection performance due to the low noise power of the relational data corresponding to the local vertex set and the modest signal power loss. The sparse background graph, modelled by a Chung-Lu random graph, contains a dense abnormal graph fitted with the Erdos-Renyi model. A few priori adjacency matrices were also known. The anomalous vertices are described by the priori adjacency matrices. Using this a priori data, the subgraph search approach is developed to condense the global vertex set into a small set. The starting point of each vertex set is initially determined by the largest anomalous coefficient. Then, based on the biggest coefficient among the revised values for each of them, a vertex from the initial vertex's neighbouring matrix is chosen and the remaining vertices are added in the same way. The set with the highest coefficient is picked as the most anomalous among all sets. This is carried out for each graph snapshot to produce a number of local sets, which are then joined to create the final set. A detection statistic is applied to this final vertex set to determine if the graph is anomalous or not. This detection statistic is a random variable and has a Poisson binomial distribution (Shao et al.).

In order to handle the issue of anomaly identification in multi-attributed networks, (Berk and Jones) suggests a generic framework called multi-attributed anomalous subgraphs and attributes scanning (MASA). It recognises the associated subset of abnormal properties as well as an anomalously linked subgraph. The framework optimises an anomalous scoring function using a set of sophisticated nonlinear nonparametric scan statistic functions. The Berk-Jones statistic (Neil et al.) is used as a case study in this study to show how anomalous an attribute is determined by its statistical p-value, which is calculated as the proportion of historical observations that have a greater or equal observation on this attribute (Luan et al.). The nonparametric scan statistics, created for computing the joint anomalousness of the p-values, are used to formulate the functions used to estimate the anomaly characteristic of the subgraphs and the corresponding subset of attributes. The graph is approximated as the tree from a randomly chosen root node using the tree approximation priors. The Steiner tree is used as a case study in this work. Then, for the aforementioned functions, finding the most anomalously linked subgraph and the qualities related to it may be roughly compared to finding the best subtree in the tree and the attributes associated to it. It seems sense that an attribute would have a greater anomalous value if its p-value were less.

The first effort at using deep learning technique for subgraph anomaly detection was made by (H. Wang et al.). The anomalous subgraph detection issue was formalised as a binary hypothesis test, where the null hypothesis represents a normal observed graph and the alternative hypothesis represents a background graph that contains an anomalous subgraph. They presented a framework for detecting subgraphs using Deep Neural Networks (DNN) that includes both an offline training phase and an online detection phase. In the offline phase, samples are delivered to the hidden layer to generate feature maps for capturing the state of the graph, and a training set is built based on the specific form of the neural network. The optimal detection statistic for the task of identifying anomalous subgraphs is determined using the Neyman-Pearson theorem. DNN uses back propagation to determine the optimum parameter. The trained DNN is fed the observed sample during the online phase to construct the feature vector, and the detection statistic is determined. This statistic, when compared to the threshold, determines whether or not the observed graph has an anomalous subgraph. Based on this framework, an algorithm known as the residual matrix-based convolutional neural network (RM-CNN) was created, which locates the graph's aberrant behaviour with the maximum likelihood of identification for a given false alarm probability..

The purpose of (Ester et al.) was to learn unusual occurrence representations of users such that benign users are dispersed over the vector space while suspicious users belonging to a single group will be close to one another. The suggested model, DeepFD, evaluates behavioural similarities between two users as the proportion of shared characteristics across all the things they have examined. This is in response to the discovery that user nodes associated with a particular fraudulent group are much more likely to have connections with identical item nodes. An autoencoder that follows the encoding-decoding procedure and was trained using three losses is then used to create user representations. The learned item representations and user representations can be used to correctly reconstruct the bipartite graph structure thanks to the first loss, the reconstruction loss. The second term keeps track of user resemblance data in the learned user representations. In other words, if two users engage in similar behaviours, their representations should do the same. The third loss Lreg is used to regularise all trainable parameters. The suspected dense blocks that are projected to produce dense areas in the feature space are then found using DBSCAN (Zheng et al.).

(Akoglu, Tong, and Koutra) uses the dense block detection approach to further detect both malicious users and associated modified products in online review networks that are modelled as bipartite graphs. FraudNE seeks to cluster suspicious users and objects from the same dense block together while distributing other items at random as opposed to encoding both nodes of various kinds into a common latent space like DeepFD does. FraudNE employs a source node and a sink node autoencoder—to understand user and item characterizations, respectively. Both autoencoders undergo lengthy training in order to effectively reduce their particular reconstruction losses and a shared loss function. Reconstruction losses assess the mismatch between the decoded characteristics of the inputs' extracted features from the graph structure. The shared loss function is proposed to restrict the learning of representations and guarantee that each interconnected pair of users and items receives comparable representations. FraudNE employs the DBSCAN (Zheng et al.) method, which is practical to utilise for dense area identification, to discriminate between dense sub-graphs created by suspect users and things.

4. Open Problems and Research Challenges

Taking into account the numerous methodologies in the literature, it can be observed that the process of finding anomalies in social networks is made up of two very distinct subprocesses, namely the suitable feature space and the detecting of anomaly in the space (Vengertsev and Thakkar). However, it wasn't clear why a specific set of attributes was being taken into consideration. Choosing an appropriate feature space may be very challenging in reality because there aren't many papers that explicitly explain the justifications for looking at a specific collection of characteristics. It is unclear which feature combinations should be utilised to capture essential concepts because many social network properties are more or less interconnected. In the absence of explicit justification for why a particular behaviour should be readily identifiable by a particular anomalous network feature, there is no reason to assume that a particular anomalous network characteristic will act in a specific way across different data sets, representing various social networks. Many of the approaches looked at were only tested on one or a few data sets, and as a result, they might have been skewed towards the particular anomalies seen in those data sets.

The enormous search area and combinatorial nature of the enumeration of potential graph substructures that are connected with more complicated anomalies present still another difficulty in the task of identifying the anomalies [58]. When the graphs are attributed, the possibilities expand even more because they cover both the attribute space and the graph structure. However, not all algorithms are applicable everywhere, and many contemporary techniques were developed with specific problem areas and data types in mind. While comparing existing and novel approaches, it is critical to consider how the size of the research, the volume of anomalies, and the magnitude of the gap between normal and unusual data affect the performance of the algorithms. There are currently very few publicly accessible data sets with established ground truths that may be used for such comparisons. So, a variety of approaches are assessed on a limited amount of data, and verifying findings is done for the highest anomalies using inspection, which is

extremely time-consuming and highly dependent on the level of disclosure by the target system's owners. Even if some data sets are reasonably well defined, anomalies discovered in these data sets may be easily compared to known sequences of occurrences, these data sets are typically small and may only be acceptable for a certain portion of problem domains. Therefore, creating synthesized large datasets might be a sensible course of action to overcome this difficulty.

There are still many problems that might be handled in the future despite the substantial amount of work done in this domain, especially with regard to handling anomalies in dynamic networks because comparably little progress has been made in this area (Gao et al.). Despite the fact that some approaches make use of temporal information, social networks have not given much attention to the time dimensions. For each of the social network techniques, such as behaviour-based, structure-based, or spectral-based, there is still potential for the exploration of a number of additional graph metrics that might be used to discover the new sorts of anomalies existing in distinct social networks. Relatively little research has been focused on it, but the focus of researchers right now is on looking for anomalies in massive data from social networks. Current solutions either focus on a predefined set of labelled data or examine the activity of randomly chosen nodes rather than studying the irregular behaviour of data in social networks. Although node and edge oddities have received considerable attention, subgraph anomalies were previously given less attention but have recently gained ground. As can be seen, deep learning has a lot of potential applications in this field and must be considered in the years ahead.

The fact that most modern techniques use deep learning technologies and social networks are frequently represented as graphs creates a great deal of complexity (Ranshous et al.). There is little to no prior knowledge regarding the features or patterns of anomalies in real applications due to the fact that tagged ground-truth anomalies are often inaccessible for research across a broad variety of industries. Graph anomalies will display various out-of-the-ordinary patterns in various types of graphs. The fact that there are several types of graph anomalies necessitates the need for detection systems to have precise definitions of anomalies as well as the ability to recognise audible cues about the abnormal patterns' atypical behaviour. These strategies must be able to handle the high dimensional and large data that real-world networks most frequently produce and be able to uncover abnormal patterns while adhering to practical resource and computing time constraints. Since real-world networks tend to be dynamic in nature, it is important to evaluate the various links between items that have been restored in traditional graphs or hypergraphs in order to account for the varying patterns of anomalies. Also, it ought to be resistive to concealed abnormalities and adaptable to newly discovered anomalies.

The requirements for anomaly detection in social media platforms will change quickly in the near future as ever-growing data volumes and more complex behaviours are taken into account. This might inspire the concept of special places with more complex design components. Thus, it will be helpful to develop any guidelines or tactics for translating actual behaviour into appropriate feature spaces. The fine line separating typical users and abnormal users would make it a lot harder to forecast the latter, necessitating the development of more potent and innovative strategies. Anomalies must not only be detected but also prevented because some domains or apps cannot allow the compromising of their sensitive data. As a result, they must be vigilant to any abnormal or malicious users long before they are actually discovered. However, it has been clear from the start that much more effort has gone into anomaly detection than towards its avoidance. Therefore, research must indeed concentrate strongly on these aspects in the coming years.

5. Conclusion

The study provides a thorough analysis of the various methods that have been suggested for identifying subgraph abnormalities in social networks that are represented as graphs. Due to the large size of the network and its dynamic nature, mining social networks for anomalies is a complex and computationally intensive task. In the last decade, a wide range of algorithms for detecting social network anomalies in various problem circumstances have been introduced. The state-of-the-art approaches are organized in this work, and the associated methodologies are briefly discussed. Starting with the

fundamental technical facts needed to comprehend the work done in this domain, it moves on to conventional anomaly detection approaches, which were eventually supplanted by graph-based anomaly detection due to its enhanced applicability and efficiency. In addition, following a whole slew of statistical methodologies, deep learning has gradually made its way into the domain and is now being used to detect anomalies in graph-based networks. However, there hasn't been much research on deep learning. The paper also goes through the many research problems and open issues for future study in this area, as well as how deep learning can be utilized to detect anomalies in social networks in the future. Choosing an algorithm is tough given the several techniques described. Many application-specific considerations must be taken into account when selecting an algorithm, including the nature of the network being analyzed and the kinds of abnormalities to be found. This comprehensive overview lists the numerous methods for searching social networks for anomalies that have been developed as well as suggestions for how to make the current methods more effective. Even though a lot of work has been done, there is indeed a lot more to be done in terms of refinement and attention.

Authors' Note

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

References

Aggarwal, Charu C and Philip S Yu. "Outlier detection for high dimensional data". *ACM SIGMOD Record* 30.2 (2001): 37–46.

Akoglu, L., H. Tong, and D. Koutra. "Graph based anomaly detection and description: a survey". *Data Mining and Knowledge Discovery* 29 (2014).

Akoglu, Leman, Mary Mcglohon, and Christos Faloutsos. "oddball: Spotting Anomalies in Weighted Graphs". *Advances in Knowledge Discovery and Data Mining*. Ed. Zaki, et al. Springer Berlin Heidelberg, 2010. 410–421.

Barabsi, A L, et al. "Evolution of the social network of scientific collaborations". *Phys. A: Stat. Mech. Appl* 311.3 (2002): 736–743.

Barnett, V and T Lewis. "" *Outliers in statistical data* 3 (1994).

Berk, Robert H and Douglas H Jones. "Goodness-of-fit test statistics that dominate the Kolmogorov statistics". *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 47.1 (1979): 47–59.

Bindu, P V and P Santhi Thilagam. "Mining social networks for anomalies: Methods and challenges". *Journal of Network and Computer Applications* 68 (2016): 213–229.

Bondy, J A and U S R Murty. "Graph Theory with Applications". 1976.

Bright, David, Russell Brewer, and Carlo Morselli. "Using social network analysis to study crime: Navigating the challenges of criminal justice records". *Social Networks* 66 (2021): 50–64.

Chandola, V, A Banerjee, and V Kumar. "Anomaly detection: a survey". *ACM Comput. Surv* 41.3 (2009): 15–15.

Cook, D J and L B Holder. "Graph-based data mining". *IEEE Intelligent Systems* 15.2 (2000): 32–41.

Curran, Kevin & and Michael Mc Hugh. "Social Networking and Health". *International Journal of Innovation in the Digital Economy* 4.2 (2013): 40–49.

Davis, Michael & et al. "Detecting anomalies in graphs with numeric labels". *Proceedings of the 20th ACM international conference on Information and knowledge management* (2011).

Eberle, William and Lawrence Holder. "Anomaly detection in data represented as graphs". *Intelligent Data Analysis* 11.6 (2007): 663–689.

Ester, M, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". *Proc. ACM Int. Conf. Knowl. Discov. Data Mining* (1996): 226–231.

Garside, Debbie & et al. “Secure Military Social Networking and Rapid Sensemaking in Domain Specific Concept Systems: Research Issues and Future Solutions”. *Future Internet* 4.1 (2012): 253–264.

Greene, Derek, Dónal Doyle, and Pádraig Cunningham. “Tracking the Evolution of Communities in Dynamic Social Networks”. *2010 International Conference on Advances in Social Networks Analysis and Mining* (2010): 176–183.

Grubbs, F E. “Procedures for Detecting Outlying Observations in Samples”. *Technometrics* 11.1 (1969): 1–21.

Gupta, Manish & et al. “Local Learning for Mining Outlier Subgraphs from Network Datasets”. *Proceedings of the 2014 SIAM International Conference on Data Mining* (2014).

Hong, Y. “On computing the distribution function for the Poisson binomial distribution”. *Computational Statistics & Data Analysis* 59 (2013): 41–51.

John, G H. “Robust decision trees: removing outliers from databases”. *Proc of KDD* (1995): 174–183.

Kaur, Ravneet and Sarbjeet Singh. “A survey of data mining and social network analysis based anomaly detection techniques”. *Egyptian Informatics Journal* 17.2 (2016): 199–216.

Keyvanpour, M, M Moradi, and F Hasanzadeh. “Digital forensics 2.0”. *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications*. Springer International Publishing, 2014. 17–46.

Liu, Y and S Chawla. “Social media anomaly detection: challenges and solutions”. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. 2317–2318.

Luan, Mingan, et al. “Anomalous Subgraph Detection in Given Expected Degree Networks With Deep Learning”. *IEEE Access* 9 (2021): 60052–60062.

Ma, Xiaoxiao & et al. “A Comprehensive Survey on Graph Anomaly Detection with Deep Learning”. *IEEE Transactions on Knowledge and Data Engineering* (2021): 1–1.

Miller, B A, N T Bliss, and P J Wolfe. “Subgraph detection using eigenvector L1 norms”. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010 (Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS* (2010).

Mislove, A, et al. “Measurement and analysis of online social networks”. *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. ACM, 2007. 29–42.

Mongiovì, Misael, et al. “NetSpot: Spotting Significant Anomalous Regions on Dynamic Networks”. *Proceedings of the 2013 SIAM International Conference on Data Mining* (2013).

Al-Mukhaini, Elham M, Wafa S Al-Qayoudhi, and Ali H Al-Badi. “Adoption Of Social Networking In Education: A Study Of The Use Of Social Networks By Higher Education Students In Oman”. *Journal of International Education Research (JIER)* 10.2 (2014): 143–154.

Nanavati, Amit A, et al. “On the structural properties of massive telecom call graphs”. *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM ’06*. ACM Press, 2006. 435–444.

Neil, Joshua, et al. “Scan Statistics for the Online Detection of Locally Anomalous Subgraphs”. *Technometrics* 55.4 (2013): 403–414.

Newman, M E J. “Finding community structure in networks using the eigenvectors of matrices”. *Physical Review E* 74.3 (2006).

Noble, Caleb C and Diane J Cook. “Graph-based anomaly detection”. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (2003).

Ranshous, Stephen & et al. “Anomaly detection in dynamic networks: a survey”. *Wiley Interdisciplinary Reviews: Computational Statistics* 7.3 (2015): 223–247.

Rattigan, Matthew J and David Jensen. “The case for anomalous link discovery”. *ACM SIGKDD Explorations Newsletter* 7.2 (2005): 41–47.

Savage, David, et al. "Anomaly detection in online social networks". *Social Networks* 39 (2014): 62–70.

Shao, Minglai, et al. "MASA: An efficient framework for anomaly detection in multi-attributed networks". *Computers & Security* 102 (2021): 102085–102085.

Shrivastava, Nisheeth & Anirban Majumder, and Rajeev Rastogi. "Mining (Social) Network Graphs to Detect Random Link Attacks". *2008 IEEE 24th International Conference on Data Engineering* (2008).

Wang, Bo, et al. "Anomaly Detection With Subgraph Search and Vertex Classification Preprocessing in Chung-Lu Random Networks". *IEEE Transactions on Signal Processing* 66.20 (2018): 5255–5268.

Wang, Haibo, et al. "Deep Structure Learning for Fraud Detection". *2018 IEEE International Conference on Data Mining (ICDM)* (2018): 567–576.

Willis, Erin and Marjorie Delbaere. "Patient Influencers: The Next Frontier in Direct-to-Consumer Pharmaceutical Marketing". *Journal of Medical Internet Research* 24.3 (2022): e29422–e29422.

Wilson, R J. *Introduction to Graph Theory*. New York: Prentice Hall/Pearson, 2010.

Yu, R, X He, and Y Liu. "Glad: group anomaly detection in social media analysis". *ACM Trans. Knowl. Discov. Data (TKDD)* 10.2 (2015): 18–18.

Zhao, Peixiang and Jiawei Han. "On graph query optimization in large networks". *Proceedings of the VLDB Endowment* 3.1-2 (2010): 340–351.

Zheng, Mengyu, et al. "FraudNE: a Joint Embedding Approach for Fraud Detection". *2018 International Joint Conference on Neural Networks (IJCNN)* (2018): 1–8.

Embargo period: The article has no embargo period.

To cite this Article: , Anagha Ajoykumar, and Venkatesan M . "**Study of Anomalous Subgraph Detection in Social Networks** ." International Research Journal on Advanced Science Hub 05.05S May (2023): 287–300. http://dx.doi.org/10.47392/irjash.2023.S039