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: