Journal of Network and Computer Applications ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Review

A survey of network anomaly detection techniques

Q1 Mohiuddin Ahmed, Abdun Naser Mahmood, Jiankun Hu

School of Engineering and Information Technology, UNSW Canberra, ACT 2600, Australia

ARTICLE INFO

Article history:
Received 10 June 2015
Received in revised form
29 October 2015
Accepted 19 November 2015

Keywords:
Intrusion detection
Computer security
Anomaly detection
Classification
Clustering
Information theory
Q2 Statistics

ABSTRACT

Information and Communication Technology (ICT) has a great impact on social wellbeing, economic growth and national security in todays world. Generally, ICT includes computers, mobile communication devices and networks. ICT is also embraced by a group of people with malicious intent, also known as network intruders, cyber criminals, etc. Confronting these detrimental cyber activities is one of the international priorities and important research area. Anomaly detection is an important data analysis task which is useful for identifying the network intrusions. This paper presents an in-depth analysis of four major categories of anomaly detection techniques which include classification, statistical, information theory and clustering. The paper also discusses research challenges with the datasets used for network intrusion detection.

Contents

http://dx.doi.org/10.1016/j.jnca.2015.11.016
1084-8045/

1. Introduction

Computer security has become a necessity due to proliferation of information technologies in everyday life. The mass usage of computerized systems has given rise to critical threats such as zero-day vulnerabilities, mobile threats, etc. Despite research in the security domain having increased significantly, are yet to be mitigated. The evolution of computer networks has greatly exacerbated computer security concerns, particularly internet security in today's networking environment and advanced computing facilities. Although Internet Protocols (IPs) were not designed to place a high priority on security issues, network administrators now have to handle a large variety of intrusion attempts by both individuals with malicious intent and large botnets (Papalexakis et al., 2012). According to Symantecs Internet Security Threat Report, there were more than three billion malware attacks reported in 2010 and the number of denial of service attacks increased dramatically by 2013 (Symantec internet security threat report, 2014). As stated in Verizon's Data Breach Investigation Report 2014, 63,437 security breaches carried out by hackers (Verizon's data breach investigation report, 2014). The Global State of Information Security Survey 2015 (The Global State of Information Security Survey, 2015) found an increase in great rise of incidents. Figure 1 shows the security incidents growth from 2009 to 2014. Therefore, the detection of network attacks has become the highest priority today. In addition, the expertise required to commit cyber crimes has decreased due to easily available tools (Hacking and cracking tools, 2014).

Anomaly detection is an important data analysis task that detects anomalous or abnormal data from a given dataset. It is an interesting area of data mining research as it involves discovering enthralling and rare patterns in data. It has been widely studied in statistics and machine learning (Ahmed et al., 2014), and also synonymously termed as outlier detection, novelty detection, deviation detection and exception mining. Although an anomaly is defined by researchers in various ways based on its application domain, one widely accepted definition is that of Hawkins (Hawkins, 1980): '*An anomaly is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*'. Anomalies are considered important because they indicate significant but rare events and can prompt critical actions to be taken in a wide range of application domains; for example, an unusual traffic pattern in a network could mean that a computer has been hacked and data is transmitted to unauthorized destinations; anomalous behavior in credit card transactions could indicate fraudulent activities, and an anomaly in a MRI image may indicate the presence of a malignant tumor (Ahmed et al., 2015a). Anomaly detection has been widely applied in countless application domains such as medical and public health, fraud detection, intrusion detection, industrial damage,

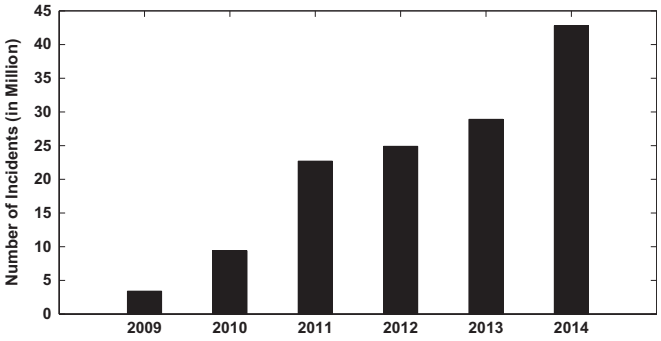

Fig. 1. Growth of information security incidents (The Global State of Information Security Survey, 2015).

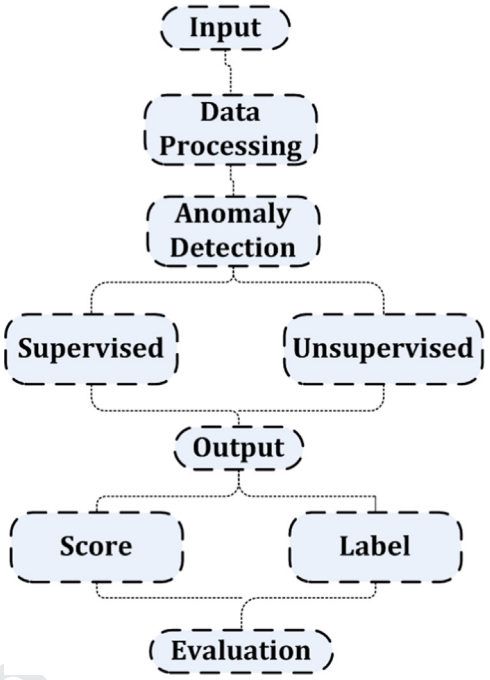

Fig. 2. Generic framework for network anomaly detection.

image processing, sensor networks, robots behavior and astronomical data (Mahmood et al., 2010; Ahmed et al., 2015b).

Figure 2 displays a generic framework for network anomaly detection. The input data requires processing because the data are of different types, for example, the IP addresses are hierarchical, whereas the protocols are categorical and port numbers are numerical in nature (Mahmood et al., 2008). Processing techniques are based on the individual anomaly detection techniques. Then, the anomaly detection techniques (broadly categorized in two: supervised and unsupervised) are applied on the data. For evaluation of the output, either scores or labels are used (discussed in Section 2.2).

Although network anomaly detection seems very straightforward, we need to find the data that do not follow normal behavioral patterns. Despite the many techniques available, following are the research challenges.

- A lack of universally applicable anomaly detection technique; for example, an intrusion detection technique in a wired network may be of little use in a wireless network.
- Data contains noise which tends to be an actual anomaly and, therefore, is difficult to segregate.
- A lack of publicly available labeled dataset to be used for network anomaly detection.
- As normal behaviors are continually evolving and may not be normal forever, current intrusion detection techniques may not be useful in the future (Qin et al., 2011). A need for newer and more sophisticated techniques because the intruders are already aware of the prevailing techniques.

Due to the aforementioned challenges, network anomaly detection has been more challenging than it was before. As most existing supervised techniques are based on knowledge provided by an external agent, they require labeled data and are unable to detect zero-day vulnerabilities. The research community has increased interest about proactive network security systems.

During the last decade several surveys of network intrusion detection have been conducted. One of the earliest was that of Towards a taxonomy of intrusion-detection systems (1999) who

Please cite this article as: Ahmed M, et al. A survey of network anomaly detection techniques. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.11.016

Table 1
Comparison of this and related surveys (∗ - this survey).

Survey	Network anomaly detection	Data issue
Ahmed[∗]	√	√
Debar-1 (Towards a taxonomy of intrusion-detection systems, 1999)	√	×
Chandola et al. (2009)	√	×
Patcha and Park (2007)	×	×
Debar-2 (Debar et al., 2000)	√	×
Phua et al.	×	×
Hodge and Austin (2004)	×	×
Axelsson (1998)	√	×
Markou and Singh (2003)	×	×
Furnell et al. (2007)	√	×
Beckman and Cook (1983)	×	×
Estevez-Tapiador et al. (2004)	√	×
Gogoi et al. (2011)	√	×
Ahmed et al. (2015a)	×	√
Xie et al. (2011)	√	×
Liao et al. (2013)	√	×

classified intrusion detection systems based on detection method, network behavior, audit source location and/or frequency of usage. The same authors extended their work to include the detection paradigm as state- or transition-based (Debar et al., 2000). Another popular survey was by Axelsson et al (Axelsson, 1998) which focused on the detection principle and operational aspects. In the taxonomy matrix of intrusion systems proposed by Furnell et al. (2007), the matrix is formed from the output type and data scale elements of the taxonomy. Another survey by Estevez-Tapiador et al. (2004) considered anomaly detection methods for network intrusion detection in wired networks. Most recently, Gogoi et al. (2011) classified outlier detection techniques as supervised and unsupervised but their taxonomy is slightly confusing because they consider proximity-based approaches under the supervised category.

Table 1 shows the methods, application domains and data covered by us and related surveys. Chandola et al. (2009) provided an extensive survey encompassing various techniques and application domains however, not a stand-alone one for network anomaly detection. Phua et al. categorized, compared and summarized a great number of published technical and review articles on automated fraud detection, since the report was published newer attacks have emerged in the last couple of years. Patcha and Park (2007) and Hodge and Austin (2004) presented various anomaly detection techniques based on supervised, unsupervised and clustering methods with no priority on network anomaly detection. Markou and Singh (2003) and Beckman and Cook (1983) conducted surveys of anomaly detection but they considered only supervised methods. A comprehensive review focussing on signature-based, behavior-based and specification-based methods for intrusion detection system are presented in Liao et al. (2013). Xie et al. (2011) also provided a survey on anomaly detection in wireless sensor networks. One of the problems of the above study is that they do not include any discussion on the research challenges related to datasets. Our previous work (Ahmed et al., 2015a) addressed the data issue in the financial domain for fraud detection. Network anomaly detection techniques are generally tested using datasets (such as DARPA/KDD) developed at the end of last century (Creech and Hu, 2013), justified by the need for publicly available test data and the lack of any other alternative datasets. Widely accepted as benchmark, these datasets no longer represent relevant architecture or contemporary attack protocols, and are accused of data corruptions and inconsistencies. Hence, testing of network anomaly detection techniques using these datasets does not provide an effective performance metric, and contributes to erroneous efficacy claims. We consider this phenomenon as 'Data Issue' in the scope of this paper.

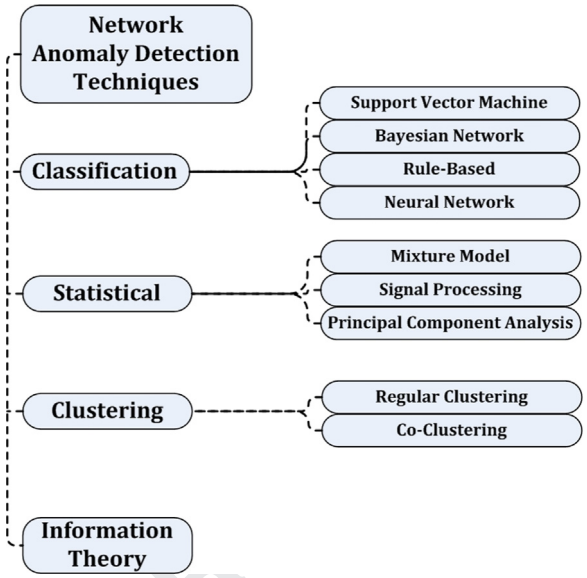

Fig. 3. Taxonomy of network anomaly detection techniques.

Based on the discussion on the existing surveys, this paper aims at the following:

- Mapping different types of anomaly with the major types of attacks based on the characteristics analysis.
- Providing a taxonomy of network anomaly detection based on classification, statistics, information theory and clustering (Fig. 3).
- Evaluating the effectiveness of different class of techniques using different criteria such as computational complexity, attack detection priority, output etc.
- Discussing the recent research related to overcome the issues of publicly available network intrusion evaluation datasets and a list of current repository.

1.1. Roadmap of the paper

In Section 2, we provide a brief introduction of anomaly detection, different types of attacks and the mapping of these attacks with different types of anomalies. Sections 3– analyse the four categories of techniques (Fig. 3). Section 7 discusses the dataset issues related to network traffic and Section 8 compares and contrasts different categories of network anomaly detection techniques. Section 9 concludes with future research directions in handling large volume of network traffic.

2. Preliminary discussion

2.1. Types of anomalies

Anomalies are referred to as patterns in data that do not conform to a well-defined characteristic of normal patterns. They are generated by a variety of abnormal activities, e.g., credit card fraud, mobile phone fraud, cyber attacks, etc., which are significant to data analysts. An important aspect of anomaly detection is the nature of the anomaly. An anomaly can be categorized in the following ways (Ahmed et al., 2014, 2015a).

Please cite this article as: Ahmed M, et al. A survey of network anomaly detection techniques. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.11.016

- *Point anomaly*: When a particular data instance deviates from the normal pattern of the dataset, it can be considered a point anomaly. For a realistic example, if a persons' normal car fuel usage is five litres per day but if it becomes fifty litres in any random day, then it is a point anomaly.
- *Contextual anomaly*: When a data instance behaves anomalously in a particular context, it is termed a contextual or conditional anomaly; for example, expenditure on a credit card during a festive period, e.g., Christmas or New Year, is usually higher than during the rest of the year. Although it can be high, it may not be anomalous as high expenses are contextually normal in nature. On the other hand, an equally high expenditure during a non-festive month could be deemed a contextual anomaly.
- *Collective anomaly*: When a collection of similar data instances behave anomalously with respect to the entire dataset, the group of data instances is termed a collective anomaly. For example, in a humans Electro Cardiogram (ECG) output, the existence of low values for a long period of time indicates an outlying phenomenon corresponding to an abnormal premature contraction (Lin et al., 2005) whereas one low value by itself is not considered anomalous.

2.2. Output of anomaly detection techniques

One important issue for anomaly detection is how anomalies are represented as output which, generally, is in one of the two following ways (Chandola et al., 2009).

- *Scores*: Scoring-based anomaly detection techniques assign an anomaly score to each data instance. Then, the scores are ranked and an analyst chooses the anomalies or uses a threshold to select them; for example, in Table 2, the data instances are represented as *a*, *b*, *c*, *d*, *e* and the corresponding anomaly scores within a range from 0 to 1.
- *Binary/label*: According to these techniques, outputs are considered in a binary fashion, i.e., either anomalous or normal. If we consider the data instances in Table 2, binary-based outputs will label each data instance as either normal or anomaly.

Techniques which provide binary labels are computationally efficient since each data instance does not need to provide or have an anomaly score.

2.3. Types of network attacks

The task of network security is to protect digital information by maintaining data confidentiality and integrity, and ensuring the availability of resources. In simple terms, a threat/attack refers to anything which has detrimental characteristics aimed at compromising a network. The poor design of a network, carelessness of its users and/or mis-configuration of its software or hardware can be vulnerable to attacks (Kendall, 1999).

1. *Denial of service*: (*DoS*) is a type of misuse of the rights to the resources of a network or host which is targeted at disrupting the normal computing environment and rendering the service unavailable. A simple example of a DoS attack is denying legitimate users access to a web service when the server is flooded with numerous connection requests. As performing a DoS attack requires no prior access to the target, it is considered a dreaded attack.
2. *Probe*: It is used to gather information about a targeted network or host and, more formally, for reconnaissance purposes. Reconnaissance attacks are quite common ways of gathering information about the types and numbers of machines connected to a network, and a host can be attacked to determine the types of software installed and/or applications used. A Probe attack is considered the first step in an actual attack to compromise a host or network. Although no specific damage is caused by these attacks, they are considered serious threats to corporations because they might obtain useful information for launching another dreadful attack.
3. *User to Root* (*U2R*): It is an attack launched, when an attacker aims to gain illegal access to an administrative account to manipulate or abuse important resources. Using a social engineering approach or sniffing password, the attacker can access a normal user account and then exploits a or some vulnerability to gain the privilege of a super user.
4. *Remote to User* (*R2U*): It is launched when an attacker wants to gain local access as a user of a targeted machine to have the privilege of sending packets over its network (also known as R2L). Most commonly, the attacker uses a trial and error method to guess the password by automated scripts, a brute force method, etc. There are also some sophisticated attacks whereby an attacker installs a sniffing tool to capture the password before penetrating the system.

Table 2
Outputs from anomaly detection techniques.

Data instance	Score	Binary/Label
a	0.3	Normal
b	0.4	Normal
c	0.2	Normal
d	0.8	Anomaly
e	0.1	Normal

2.4. Mapping of network attacks with anomalies

Based on the discussion on different types of anomalies and attacks in the previous section of this paper, in this section we identify the relationship among the attacks and anomalies. The DoS attack characteristics match with the collective anomaly (Ahmed and Mahmood, 2014a, 2015, 2014b). As stated in Section 2.1, when a collection of data instances behave anomalously, it is called collective anomaly but a single data instance from that group is not anomalous. In case of a DoS attack, numerous connection request to a web server is a collective anomaly but a single request is legitimate. So, we can consider the DoS attack as collective anomaly. Probe attacks are based on specific intention to attain information and reconnaissance. The authors of this paper map them with contextual anomaly. U2R and R2L attacks are condition specific and sophisticated. Launching suck attacks are not easy as compared to others. Therefore, these attacks are considered as point anomaly. Figure 4 illustrates the mapping of the attack classes.

3. Classification based network anomaly detection

Classification-based techniques rely on experts' extensive knowledge of the characteristics of network attacks. When a network expert provides details of the characteristics to the detection system, an attack with a known pattern can be detected as soon as it is launched. This is solely dependent on the attack's signature as a system, which is capable of detecting an attack only if its signature has been provided earlier by a network expert. This demonstrates a system which can detect only what it knows is vulnerable to new attacks, which are constantly appearing in different versions and more stealthily launched. Even if a new attack's signature is created and incorporated in the system, the

Please cite this article as: Ahmed M, et al. A survey of network anomaly detection techniques. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.11.016

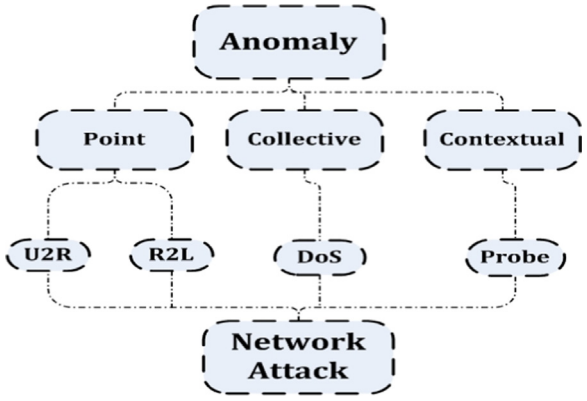

Fig. 4. Mapping of attacks with anomalies.

initial loss is irreplaceable and the repair procedure is extremely expensive.

The classification based approaches rely on the normal traffic activity profile that builds the knowledge base and consider activities deviate from baseline profile as anomalous. The advantage lies in their capability to detect attacks which are completely novel, assuming that they exhibit ample deviations from the normal profile. Additionally, as normal traffic not included in the knowledge base is considered an attack, there will be inadvertent false alarms. Therefore, training is required for anomaly detection techniques to build a normal activity profile which is time-consuming and also depends on the availability of completely normal traffic datasets. In practice, it is very rare and expensive to obtain attack-free traffic instances. Moreover, in today's dynamic and evolving network environments, it is extremely difficult to keep a normal profile up-to-date. Among a large pool of classification-based network anomaly detection techniques available, we discuss four major techniques as follows.

3.1. Support vector machine

The basic principle of the Support Vector Machine (SVM) is to derive a hyperplane that maximizes the separating margin between the positive and negative classes (Eskin et al., 2002). An interesting property of SVM is that it is an approximate implementation of the structure risk minimization principle, based on statistical learning theory. The standard SVM algorithm is a supervised learning technique, which requires labeled data to create a classification rule. However, it can also be adapted as an unsupervised learning algorithm whereby it tries to separate the entire set of training data from its origin whereas the regular supervised SVM attempts to separate two classes of data in a feature space by a hyperplane. In a paper by Eskin et al. (2002), the concept of the unsupervised SVM is used to detect anomalous events. The algorithm finds hyper planes which separate the data instances from their origins with the maximal margin and then an optimization problem is solved to determine the best hyperplane (for more details, Please see Cristianini and Shawe-Taylor, 2000).

The optimization problem is solved using a variant of the Sequential Minimal Optimization algorithm (Platt, 1999). Using a similar concept to that of the One-class SVM (OCSVM) but in a supervised manner, in the paper by Heller et al. (2003), a new approach called Registry Anomaly Detection (RAD) is developed to monitor Windows registry queries. It is usual that during normal computer activity, a certain set of registry keys is accessed by the Windows program. Based on the fact that users tend to frequently use certain programs and registry activities are normal, deviations from these activities will be considered anomalous. The OCSVM is applied to the RAD system to detect anomalous activities in the Windows registry. RAD maps the input data into a high-dimensional feature space via a kernel and iteratively finds the maximal margin hyperplane to separate two classes of data.

Interestingly, in a paper by Hu et al. (2003), an anomaly detection method which ignores noisy data is developed using the Robust SVM (RSVM) is presented. In practice, training data often contain noise which invalidates the main assumption of the SVM that all the sample data for training are independently and identically distributed. As a result, the standard SVM results in a highly non-linear decision boundary which leads to poor generalization. In this scenario, the RSVM incorporates the averaging technique in the form of a class centre to make the decision surface smoother and automatically control regularization. In addition, the number/quantity of support vectors in the RSVM is significantly less than the standard SVM which results in a reduced run time. More recently a patent is published which contains method and system for confident anomaly detection in computer network traffic where SVM is explored (Balabine and Velednitsky, 2015).

3.2. Bayesian network

A Bayesian network is an efficient approach for modeling a domain containing uncertainty. A discrete random variable is represented using a directed acyclic graph (DAG), where each node reflects the state of the random variable and contains a conditional probability table (CPT). The task of the CPT is to provide the probability of a node being in a specific state. In a Bayesian network, a parent–child relationship exists among the nodes which indicate that a variable represented by a child node is dependent on those represented by the parent nodes. As this network can be used for an event classification scheme, it is also applicable for network anomaly detection. In a paper by Kruegel et al. (2003), two major problems caused in high false positives in anomaly detection techniques are identified. It is assumed that anomaly detection systems contain a number of models for analyzing different features of an event. The first problem is that models which provide a score or probability for the normality/abnormality of an event require the anomaly detection system to aggregate their different outputs which result in high false positives. Secondly, anomaly detection systems cannot handle behaviors which are unusual but legitimate, for example, a sudden increase in CPU utilization, memory usage, etc. If this problem occurs, additional information can explain unusual behaviors that are not anomalous is ignored.

Based on the concept of the Bayesian network, the authors of Kruegel et al. (2003) proposed an approach for solving the aforementioned problems. For an ordered stream of input events ($S = e_1, e_2 \ldots$), the event classification system decides whether an event is normal or abnormal. This decision is based on the outputs ($o_i | i = 1, 2, \ldots, k$) from k models ($M = m_1, m_2, \ldots, m_k$) and possible additional information (I). The models analyze the features of a given input event and compare their results with those from previously established models. The result from an event classification system (EC) is defined as:

$$EC(o_1, o_2, \ldots, o_k, I) = e \text{ is normal} : \sum_{i=1}^{k} o_i \leq I \,|\, e \text{ is anomalous} : \sum_{i=1}^{k} o_i > I \quad (1)$$

A Bayesian network is applied to identify anomalous events by introducing the root node which represents a variable with two states. One child node is used to capture the model's outputs and the child node is connected to the root node, it is expected that the output events will be different when the input is either abnormal

Please cite this article as: Ahmed M, et al. A survey of network anomaly detection techniques. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.11.016

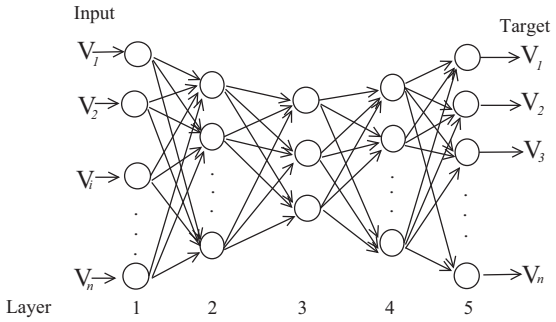

Fig. 5. Schematic of replicator neural network, adapted from Hawkins et al. (2002).

or normal. A more recent application of Bayesian network can be found in a telecommunication network (Deljac et al., 2015).

3.3. Neural network

The strength of a neural network for classifying data has also been used for network anomaly detection. Neural networks have been applied in various application domains, such as image and speech processing, but they have high computational requirements. For network anomaly detection, a neural network has been merged with other techniques, such as a statistical approach and variants of it. In Hawkins et al. (2002), a Replicator Neural Network (RNN) is used to provide an outlyingness factor for anomalous network traffic. It is a feed-forward multi-layer perception with three hidden layers placed between the input and output layers. Its objective is to reproduce the input data pattern at the output layer with a minimized error via training. Figure 5 presents a schematic view of a RNN. The $S_k(I_{ki})$ function produces the output from unit i for layer k as

$$\theta = I_{ki} = \sum_{j=0}^{L_{k-1}} w_{kij} Z_{(k-1)j} \tag{2}$$

where I_{ki} is the weighted sum of the inputs to the unit, Z_{kj} the output from the jth unit of the kth layer and L_k the number of units in the kth layer. The *outlier factor* is defined using the trained RNN as follows, where x_{ij} is the input value and o_{ij} the output value from the RNN.

$$OF_i = \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - o_{ij})^2 \tag{3}$$

In Zhang et al. (2001), a hierarchical intrusion detection system in which neural networks are combined with statistical models to detect network attacks is proposed. The output from the neural network classifier is represented as a continuous variable (t), where -1 means an intrusion with absolute certainty and 1 no attack. In addition, Self-organizing Maps (SOM) are used for network anomaly detection. Ramadas et al. (2003) suggested that, using SOM, network traffic can be classified in real time. SOM relies on the hypothesis that network attacks can be characterized by different sets of neurons that cover larger areas compared to others on an output neuron map. In Poojitha et al. (2010) a feed forward neural network trained by back propagation algorithm is developed to detect the anomalies using a given dataset with the information related to the computer network during normal and during anomalous behavior.

3.4. Rule-based

Rule-based anomaly detection techniques are widely used in supervised learning algorithms (Lee et al., 1999). The basic idea is to learn the normal behavior of a system and anything not encompassed within it considered anomalous. These techniques consider both single and multi-label learning algorithms.

From a machine learning point of view, single-label classification aims to learn from a set of instances each of which is associated with a unique class label from a set of disjoint class labels. However, multi-label classification allows one instance to be associated with more than one class which can be correlated with fuzzy clustering. For a given training set ($S = (x_i, y_i);\ 1 \le i \le n$) consisting of n training instances ($x_i \in x, y_i \in y$) which are independent and identically distributed, multi-label learning produces a multi-label classifier ($n : x \to y$) that optimizes the specific evaluation function. In Yang et al. (2013) a rule-based method for IEC 60870-5-104 driven SCADA networks using an in-depth protocol analysis and a deep packet inspection method is proposed.

4. Statistical anomaly detection

Intrusion detection techniques have also been developed using statistical theories; for example, the established *chi-square* theory is used for anomaly detection in Ye and Chen (2001). According to this technique, a profile of normal events in an information system is created. The basic idea in this approach is to detect both a large departure of events from normal as anomalous and intrusions. A distance measure based on the *chi-square* test statistic is developed as

$$X^2 = \sum_{i=1}^{n} \frac{(X_i - E_i)^2}{E_i} \tag{4}$$

- X_i = the observed value of the ith variable,
- E_i = the expected value of the ith variable,
- n = the number of variables.

X^2 has a low value when an observation of the variables is near the expected. Following the $\mu \pm 3\sigma$ rule, when an observation, X^2 is greater than $\overline{X^2} + 3S_X^2$, is considered an anomaly.

Krügel et al. (2002) proposed a statistical processing unit for detecting anomalous network traffic, more specifically to detect the attacks which are rare such as R2L and U2R. A metric is developed which allows the system to automatically search identical characteristics of different service requests. The anomaly score of a request is calculated based on the following three main characteristics:

- the type of request;
- the length of the request; and
- the payload distribution.

The network administrator defines a threshold to raise alarms for anomalous requests. The anomaly score is derived as in Eq. (5) where the payload distribution is given more weight than the other properties.

$$AS = 0.3 \times AS_{type} + 0.3 \times AS_{length} + 0.4 \times AS_{payload} \tag{5}$$

Based on the principles of the statistical theory, different types of techniques have been developed to detect anomaly discussed next.

4.1. Mixture model

Based on the concept that an anomaly lies within a large number of normal elements, Eskin (2000) proposed a mixture model for detecting anomalies from noisy data. Generally, in mixture models, each element falls into one of the following two classes:

Please cite this article as: Ahmed M, et al. A survey of network anomaly detection techniques. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.11.016

- having a small probability of λ; or
- the majority of elements having the probability of $1-\lambda$.

The authors of Eskin (2000) assumed from an intrusion detection perspective that the set of system calls with the probability of $1-\lambda$ is a legitimate use of the system and the intrusions have the probability of λ. From a mixture model perspective, the two probability distributions which generate the data are called the *majority* (**M**) and *anomalous* (**A**) distributions, with an element (x_i) generated from either. When the generative distribution for the data is **D**, it can be represented as

$$D=(1-\lambda)M+\lambda A \tag{6}$$

The data elements generated from the **A** distribution are considered anomalous.

4.2. Signal processing technique

Although signal processing is an interesting research area, using such a technique for anomaly detection has hardly been explored. In Thottan and Ji (2003), a statistical signal processing technique based on an abrupt change detection is presented. The authors describe network anomalies in two ways:

- Anomalies correspond to network failures and performance problems;
- Encompasses security-related issues such as DoS attacks.

In Thottan and Ji (2003), management information bases are used to produce a network health function that can be used to raise alarms corresponding to anomalous networks. The unusual behaviors in these bases are determined by finding abrupt changes in their statistics. A hypothesis test based on the general likelihood ratio (GLR) is used to detect the changes to provide the degree of abnormality on a scale of between 0 and 1.

4.3. Principal component analysis (PCA)

Shyu et al. (2003) presented an easier way to analyze high dimensional network traffic dataset using PCA. PCAs are linear combinations of p random variables $(A_1, A_2, \ldots A_p)$ and can be characterized:

1. uncorrelated,
2. with their variances sorted in order from high to low or
3. their total variance equal to the variance of the original data.

A brief mathematical formulation of PCA is as follows. Let **A** be an $n*p$ data matrix of n observations on each of p variables $(A_1, A_2, \ldots A_p)$ and **S** a $p*p$ sample covariance matrix of $A_1, A_2, \ldots A_p$. If $(\lambda_1, e_1), \ldots (\lambda_p, e_p)$ are the p eigenvalue–eigenvector pairs of the matrix **S**, the ith principal component is as follows, where $i=1,2, \ldots, p$ and $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$.

$$y_i = e_i(x-\overline{x}) = e_{i1}(x_1-\overline{x_1})+e_{i2}(x_2-\overline{x_2})+\ldots\ldots+e_{ip}(x_p-\overline{x_p}) \tag{7}$$

An anomaly detection technique based on PCA (Shyu et al., 2003) has the benefits of:

- being free from any assumption of statistical distribution;
- being able to reduce the dimension of the data without losing any important information; and
- having minimal computational complexity which supports real-time anomaly detection.

According to the approach by Shyu et al. (2003), it is assumed that the number of normal instances is much higher than that of anomalies. The principal component classifier (PCC) contains two scores; one of each of the major and minor components, and a data instance (x) is classified as an anomaly if

$$\sum_{i=1}^{q}\frac{y_i^2}{\lambda_i} > c_1 \quad \text{or} \quad \sum_{i=p-r+1}^{p}\frac{y_i^2}{\lambda_i} > c_2 \tag{8}$$

and is normal when/if

$$\sum_{i=1}^{q}\frac{y_i^2}{\lambda_i} \leq c_1 \quad \text{or} \quad \sum_{i=p-r+1}^{p}\frac{y_i^2}{\lambda_i} \leq c_2 \tag{9}$$

where c_1 and c_2 are outlier thresholds for creating a specific false alarm. It is also assumed that the data distribution is multivariate normal and the false alarm rate of the classifier

$$\alpha = \alpha_1+\alpha_2-\alpha_1\alpha_2 \tag{10}$$

where

$$\alpha_1 = P\left(\sum_{i=1}^{q}\frac{y_i^2}{\lambda_i} > c_1 \,|\, x \text{ is a normal instance}\right) \tag{11}$$

and

$$\alpha_2 = P\left(\sum_{i=p-r+1}^{p}\frac{y_i^2}{\lambda_i} > c_2 \,|\, x \text{ is a normal instance}\right) \tag{12}$$

5. Information theory

Information-theoretic measures can be used to create an appropriate anomaly detection model. In a paper by Lee and Xiang (2001), several measures, such as entropy, conditional entropy, relative entropy, information gain and information cost, are used to explain the characteristics of a dataset. We provide the following definitions of these measures.

- *Entropy* is a basic concept of information theory which measures the uncertainty of a collection of data items. For a dataset, D in which each data item belongs to a class ($x \in C_D$), the entropy of D relative to the $|C_D|$−wise classification is defined as

$$H(D) = \sum_{x \in C_D} P(x)\log\frac{1}{P(x)} \tag{13}$$

  where $P(x)$ is the probability of x in D.
- *Conditional entropy* is the entropy of D given that Y is the entropy of the probability distribution ($P(x|y)$) as

$$H(D|Y) = \sum_{x,y \in C_D, C_Y} P(x,y)\log\frac{1}{P(x|y)} \tag{14}$$

  where $P(x, y)$ is the joint probability of x and y and $P(x|y)$ the conditional probability of x given y.
- *Relative entropy* is the entropy between two probability distributions $p(x)$ and $q(x)$ defined over the same $x \in C_D$ as

$$relEntropy(p|q) = \sum_{x \in C_D} P(x)\log\frac{p(x)}{q(x)} \tag{15}$$

- *Relative conditional entropy* is the entropy between two probability distributions ($p(x|y)$ and $q(x|y)$) defined over the same

$x \in C_D$ and $y \in C_Y$ as

$$relCondEntropy(p|q) = \sum_{x,y \in C_D, C_Y} p(x,y) \log \frac{p(x|y)}{q(x|y)} \quad (16)$$

- *Information gain* is a measure of the information gain of an attribute or feature A in a dataset D and is

$$Gain(D,A) = H(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} H(D_v) \quad (17)$$

where values A is the set of possible values of A and D_v the subset of D where A has the value v.

Based on this knowledge, appropriate anomaly detection models can be built. Supervised anomaly detection techniques require a training dataset followed by a test data to evaluate the performance of a model. In this situation, firstly, information-theoretic measures are used to determine whether a model is suitable for testing the new dataset. Noble and Cook (2003) conducted experiments on the benchmark DARPA and UNM audit datasets to demonstrate the utility of information-theoretic measures and concluded that they can be used to create efficient anomaly detection models and also to explain their performances.

5.1. Correlation analysis

In Ambusaidi et al. (2014) a nonlinear correlation coefficient-based (NCC) similarity measure is suggested to extract both linear and nonlinear correlations between network traffic. The extracted correlative information is used to detect malicious network behaviours. Pearson's correlation coefficient is a basic linear correlation method to find out dependence between two variables (Ahmed et al., 2015c), however, there are datasets where nonlinear correlation exists between different variables such as in network traffic. The NCC is defined by Wang et al. (2005) as in Eq. (18), where $H^r(X)$ and $H^r(Y)$ are the revised entropies of the variable X and Y.

$$NCC(X;Y) = H^r(X) + H^r(Y) - H^r(X,Y) \quad (18)$$

Given a set of m normal training data instances, NCC is calculated first. For any incoming instance the NCC between incoming one and the normal instances is recorded as $NCC^{m,m+1}$. For a user defined threshold σ which is ranged between 0 and 1, an incoming traffic instance is considered as anomalous if the difference in NCC is greater than the σ (19).

$$\left| NCC^m - NCC^{m,m+1} \right| > \sigma \quad (19)$$

In Tan et al. (2014a), for DoS attack detection a system is proposed that uses multivariate correlation analysis (MCA) for accurate network traffic characterization by extracting the geometrical correlations between network traffic features. The detection process contains three major steps as shown in Fig. 6. In Step 1, basic features are generated in a well-defined time interval. Step 2 includes the multivariate correlation analysis, where the 'triangle area map generation' module is applied to extract the correlations between two distinct features within each traffic instance coming from the first step. In Step 3 contains decision-making based on training and testing phase.

The concept of multivariate correlation analysis approach in Tan et al. (2014a) is incorporated to characterize network traffic instances and to convert them into respective images. These images are used for DoS attack detection based on a widely used dissimilarity measure, namely Earth Mover's Distance (EMD) (Rubner et al., 1998). EMD considers cross-bin matching and provides a more accurate evaluation on the dissimilarity between distributions than some other well-known dissimilarity measures.

6. Clustering-based

Clustering refers to unsupervised learning algorithms which do not require pre-labeled data to extract rules for grouping similar data instances (Jain et al., 1999). Although there are different types of clustering techniques, we discuss the usefulness of regular clustering and co-clustering for network anomaly detection. The difference between regular clustering and co-clustering is the processing of rows and columns. Regular clustering techniques such as *k-means* (Ahmed and Naser, 2013) clusters the data considering the rows of the dataset whereas the co-clustering considers both rows and columns of the dataset simultaneously to produce clusters (Ahmed et al., 2015d).

The three key assumptions that are always made when using clustering to detect anomalies are briefly discussed below.

- *Assumption 1*: As we can create clusters of only normal data, any subsequent new data that do not fit well with existing clusters of normal data are considered anomalies; for example, as density-based clustering algorithms do not include noise inside clusters (Ester et al., 1996), noise is considered anomalous.
- *Assumption 2*: When a cluster contains both normal and anomalous data, it has been found that the normal data lie close to the nearest clusters centroid but anomalies are far away from centroids (Ahmed and Naser, 2013). Under this assumption, anomalous events are detected using a distance score.
- *Assumption 3*: In a clustering with clusters of various sizes, the smaller and sparser can be considered anomalous and the thicker normal. Instances belonging to clusters the sizes and/or densities below a threshold are considered anomalous.

6.1. Regular clustering

The approach used by Münz et al. (2007) to anomalous data is quite straightforward. They use *k-means* clustering to generate normal and anomalous clusters. Once clustering is achieved, it is analyzed using the following assumptions:

- An instance is classified as normal, if it is closer to the normal than anomalous clusters centroid and vice versa;

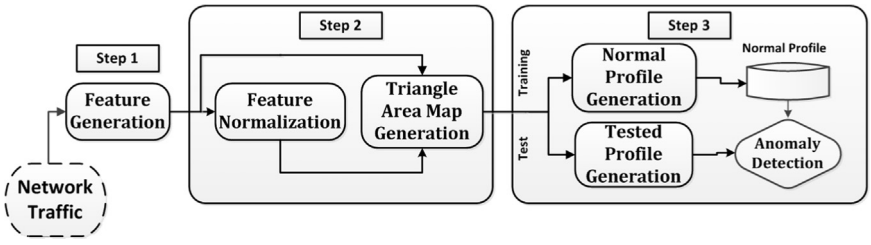

Fig. 6. MCA based framework for DoS attack detection, adapted from Tan et al. (2014a).

- If the distance between an instance and centroid is larger than a predefined threshold (d_{max}), the instance is treated as an anomaly; and
- An instance is treated as an anomaly, if it is closer to the anomalous than normal clusters centroid or if its distance to the normal clusters centroid is larger than the predefined threshold.

Petrovic et al. (2006) proposed a cluster-labeling strategy based on a combination of clustering evaluation techniques. The Davies–Bouldin clustering evaluation index and comparison of the centroid diameters of the clusters are combined in order to respond adequately to the properties of attack vectors. They consider the compactness of the corresponding clusters and the separation between them, and the principal parameters which distinguish between 'normal' and 'abnormal' behavior in the analyzed network. However, they do not explain the reason for use of $k=2$ for their *k-means* clustering. According to their approach, the attack vectors are often very similar, if not identical; for example, the corresponding cluster in the case of a massive attack is extremely compact and the *Davies–Bouldin* index of such a clustering is either 0 (when the non-attack cluster is empty) or very close to 0. Bearing in mind the expected similarity among attack vectors, as the diameter of the centroid of an attack cluster is expected to be smaller than that of a non-attack cluster they can distinguish between normal and anomalous clusters.

Portnoy et al. (2001) proposed clustering based on the width to classify data instances. The width is constant and remains the same for all clusters. Once clustering is performed, based on the assumption that normal instances constitute an overwhelmingly large proportion of the entire dataset, N percent of clusters are normal and the rest are anomalous. Using this assumption, Leung and Leckie (2005) proposed a density-and grid-based clustering algorithm which is suitable for unsupervised anomaly detection.

Syarif et al. (2012) investigated the performances of various clustering algorithms when applied for anomaly detection. They used five different approaches, the *k-means*, *improved k-means*, *k-medoids*, *Expectation Maximization* (*EM*) clustering, and distance-based anomaly detection algorithms. Table 3 demonstrates the performance evaluation of clustering algorithms used for network anomaly detection.

Arshad et al. (https://repository.lib.fit.edu/handle/11141/126) developed an approach to determine if a cluster is an outlier (CLAD). This technique relies on two properties of a cluster; its density and distance from other clusters. According to them the cluster density is dependent on the number of data instances. To determine the distance, they calculate the average inter-cluster distance (ICD) between one cluster and the others. According to Guan et al. (2003), if the population ratio of one cluster is above a given threshold, all the instances in that cluster are classified as normal; otherwise, they are labeled intrusive. Joshua et al. (Oldmeadow et al., 2004) proposed a solution for time-varying network intrusion detection and they also demonstrate how feature weighting can be improved in classification accuracy. Clustering techniques are also incorporated in other systems as hybrid techniques, for example in *k-NN* based classifiers for online anomaly detection (Su, 2011).

Table 3
Network anomaly detection: evaluation using NSL-KDD dataset (Syarif et al., 2012).

Algorithm	Accuracy (%)	False positive (%)
k-means	57.81	22.95
Improved k-means	65.40	21.52
k-medoids	76.71	21.83
EM clustering	78.06	20.74
Distance-based anomaly detection	80.15	21.14

More recently, In Ahmed and Mahmood (2014a), the authors also used *x-means* clustering (Pelleg and Moore, 2000) to detect collective anomaly such as DoS attacks. The performance of their technique was significantly better than other existing clustering based methods.

6.2. Co-clustering

Co-clustering can be simply considered a simultaneous clustering of both rows and columns (Govaert and Nadif, 2008; Banerjee et al., 2007). It can produce a set of **c** column clusters of the original columns (**C**) and a set of **r** row clusters of the original row instances (**R**). Unlike other clustering algorithms, co-clustering defines a clustering criterion and then optimizes it. In a nutshell, it simultaneously finds the subsets of rows and columns of a data matrix using a specified criterion. The benefits of co-clustering over the regular clustering are the following:

1. Simultaneous grouping of both rows and columns can provide a more compressed representation and it preserves information contained in the original data.
2. Co-clustering can be considered as a dimensionality reduction technique and it is suitable for creating new features.
3. Significant reduction in computational complexity. For example, traditional *k-means* algorithm has the $\mathcal{O}(mnk)$ as computational complexity where m=number of rows, n=number of columns and k is the number of clusters. But in co-clustering the computational complexity is $\mathcal{O}(mkl+nkl)$, here l is the number of column clusters. Obviously $\mathcal{O}(mnk) \gg \mathcal{O}(mkl+nkl)$.

According to Ahmed and Mahmood (2014b), co-clustering is beneficial for detecting DoS attacks and significant performance improvement is achieved while it is being used in the collective anomaly detection framework (Table 4 depicts the experimental results). However, in Papalexakis et al. (2012), the usage of co-clustering for detecting all types of network attacks is investigated. Table 5 shows the comparison of clustering purity between Ahmed and Mahmood (2014b) and Papalexakis et al. (2012) for identifying DoS attacks.

7. Intrusion detection datasets and issues

Due to privacy issues, the datasets used for network traffic analysis are not easily available. There are very few publicly available datasets and among them DARPA/KDD datasets are considered as benchmark. In this section, we discuss the limitations of publicly available DARPA/KDD datasets and will also provide a detailed introduction to a recent technique for building an intrusion detection dataset.

7.1. Limitations of DARPA/KDD datasets

Among the anomaly detection techniques discussed in the scope of this paper, more than 50% of them uses the DARPA/KDD datasets due to their availability. However, these datasets are criticized by Testing intrusion detection systems (2000) for the generation procedure and the analysis by Mahoney and Chan

Table 4
Evaluation results.

Accuracy	Precision	Recall	F-measure	Attack cluster purity	Normal cluster purity
92.82%	**0.9236**	**0.9923**	**0.96**	**92.36%**	**95.6%**

Please cite this article as: Ahmed M, et al. A survey of network anomaly detection techniques. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.11.016

Table 5
Cluster purity comparison.

Purity	Ahmed and Mahmood (2014b)	Papalexakis et al. (2012)
Normal (%)	95.6	75.84
Attack (%)	92.36	92.44

Table 6
Attack structure, adapted from Creech and Hu (2013).

Payload/effect	Vector
Password bruteforce	FTP, SSH by Hydra (Thc-hydra, 2014)
Adding new super-user	Client side poisoned executable
Java-based meterpreter	Tiki Wiki vulnerability exploit (Tikiwiki, 2014)
Linux Meterpreter Payload	Client side poisoned executable
C100 webshell	PHP remote file inclusion vulnerability

(2003) found evidence of simulation artifacts that could result in over-estimations of anomaly detection performances. The KDD datasets were developed using a Solaris-based operating system to collect a wide range of data due to its easy deployment. However, we can see significant differences in today's operating systems which barely resemble Solaris. In this age of Ubuntu, Windows and MAC, Solaris has almost no market share. The traffic collector used in KDD datasets, TCPdump, is very likely to become overloaded and drop packets from a heavy traffic load. More importantly, there is some confusion about these datasets attack distributions. According to an attack analysis, Probe is not an attack unless the number of iterations exceeds a specific threshold while label inconsistency has been reported. A description of the KDD datasets states that there are 24 training and 14 test attacks. However, it is reported by Shafi and Abbass (2013) that the training data contain 22 attacks and the test data 17. This inconsistency has a significant impact on the class distribution of attacks. In this scenario, it is important to create intrusion detection datasets in modern-day computing to address the issues of DARPA/KDD. The next section discusses one such contemporary dataset for network traffic analysis.

7.2. Contemporary network attacks evaluation dataset: ADFA-LD12

Creech and Hu (2013) introduced a publicly available dataset ADFA-LD12, which is a representative of the modern attack structure and methodology. The dataset is developed using Ubuntu Linux version 11.04 (a modern Linux distribution widely used Ubuntu Linux, 2014) as the host operating system (Adfa intrusion detection datasets, 2014). To allow web-based attacks, Apache Version 2.2.17 (The apache software foundation, 2014) running on PHP Version 5.3.5 (PHP: Hypertext processor, 2014) was installed and enabled. TikiWiki Version 8.1 (Tikiwiki: Cms groupware, 2014) was installed as a web-based collaborative tool because it has a known vulnerability, as detailed in Tikiwiki (2014). The developers of the dataset selected the attacks carefully and focused on the methods of contemporary penetration testers and hackers. Also, there was a trade-off between the vulnerability of the targeted system and the realism required, with the intended target server for the ADFA-LD12 fully patched to create a realistic target. The vulnerability used in this scenario, such as TikiWiki remote code execution vulnerability (Tikiwiki, 2014) is considered to emulate a realistic and small flaw in a well-configured server, is an acceptable simulation of the real world. Table 6 shows the breakdown of payloads and vectors used to attack the Ubuntu OS. The ADFA-LD12 is a possible successor to the DARPA/KDD datasets as it uses a modern Linux operating system and up-to-date exploits. Interestingly, the performances of intrusion detection techniques on the KDD datasets vary from the ADFA-LD12 as KDD dataset fails to represent contemporary attacks.

7.3. Current network data repositories

Some publicly available network traffic datasets are based on the current operating systems and hardware (their sources and comments about them are listed below). However, several projects dedicated to the development of benchmark intrusion detection evaluation datasets are currently being undertaken.

- PREDICT (Predict, 2014): Stands for Protected Repository for the Defense of Infrastructure Against cyber threats. It is a US-based community of producers of security-relevant network operations data and consists of researchers of networking and information security. This dataset supports developers and evaluators by providing regularly updated network data relevant to cyber security research.
- CAIDA (Caida, 2014): Provides basic captured network traces but it is not labeled and lacks multiple-attack scenarios.
- Internet Traffic Archive (Internet traffic archive, 2014): It is a repository for supporting widespread access to traces of internet network traffic and is sponsored by ACM SIGCOMM. However, it suffers from heavy anonymization, lacks the necessary packet information, is not labeled and has no multiple-attack scenarios.
- DEFCON (Defcon, 2014): It is different from real network traffic and consists mainly of intrusive traffic and normally used for the alert correlation technique.
- ADFA Intrusion Detection Datasets (Adfa intrusion detection datasets, 2014): It covers both Linux and Windows, and are designed for the evaluation by a system call-based HIDS.
- NSL-KDD (NSL-KDD, 2014): It is a dataset suggested as a means of solving some of the inherent problems of the KDD dataset mentioned in Testing intrusion detection systems (2000).
- KYOTO (Kyoto Dataset, 2014): It contains traffic data from Kyoto University's 'Honeypots'.
- ISCX 2012 (Shiravi et al., 2012): It is developed by Information Security Centre of Excellence at University of New Brunswick. It contains seven days captured traffic with overall 2450324 flows including DoS attacks.
- ICS Attack (ICS Attack Dataset, 2014): Oak Ridge National Laboratories (ORNL) have created three datasets which include measurements related to electric transmission system normal, disturbance, control and cyber attack behaviors.

8. Evaluation of network anomaly detection techniques

This paper contains discussion on four major categories of network anomaly detection (Sections 3–6). In this section, we compare and contrast (Table 7) these techniques based on the following criteria:

Table 7
Evaluation of network anomaly detection techniques.

Technique	Output	Attack priority	Complexity
Classification	Label, score	DoS	Quadratic
Statistical	Label, score	R2L, U2R	Linear
Clustering	Label	DoS	Quadratic
Information theory	Label	Neutral	Exponential

- Computational Complexity (Papadimitriou, 1994): Linear, Quadratic, Exponential
- Preference of attack detection: DoS, Probe, R2L, U2R
- Output: Label, Score

As discussed earlier in Section 2.2, the anomaly detection techniques which only have the labelled output are more efficient than the score based outputs. In this scenario, clustering and information theory based techniques are better than the classification and statistical techniques. When the priority of attack detection is concerned, the classification and clustering-based techniques are more interested to identify DoS attacks. Moreover, the DoS attacks are the most easily launched attacks and yet they have detrimental impact on any network. Therefore, the techniques dealing with identifying DoS attacks are more demanding than others. Based on the computational complexity, statistical techniques are better than other techniques due to their linear complexity nature. It is usual to have linear complexity for fitting statistical distribution such as Gaussian, mixture models. However, in case of principal component analysis, the complexity is not linear because of the underlying computations. The clustering and classification-based techniques have quadratic complexity for the following reasons:

- Clustering techniques require pairwise distance computation.
- Classification techniques require quadratic optimization to separate two or more classes (e.g. SVM).

The information theory based techniques suffer from exponential time complexity because of the calculation of the measurements such as entropy, relative uncertainty, etc. These techniques also require dual optimization for minimizing the subset size and simultaneously reducing the complexity in the dataset.

From the above discussion, we can arrive at the following conclusions:

- As far as the output style is concerned, clustering and information theory based techniques are better than others. Clustering techniques are computationally efficient and has specific target for DoS attack detection, while the information theory based techniques have no specific attack target.
- Based on the attack preference, both clustering and classification targets DoS attack, while the statistical techniques have the preference for R2L and U2R attacks, which are very rare.
- Though the statistical techniques have the lowest complexity, they are not suitable for DoS attack detection.
- Although both classification and clustering techniques have similar complexity and the same target, the classification techniques are based on supervised learning which requires tracing of pre-labeled data. But clustering is completely unsupervised. Clustering cannot outdo the statistical techniques in complexity, yet it outperforms in all other criteria.

9. Conclusions and future research directions

The survey of literature reported in this paper has categorized the network anomaly detection methods on four major categories. For each category, we described the assumptions for segregating normal data instances from anomalous. These assumptions will provide a guideline to assess the efficiency of the techniques when applied in a particular domain. Compared to other surveys, this paper provided a discussion on network traffic dataset issues which are of significant concern to the research community in the area of network traffic analysis.

Data mining and machine learning techniques constantly attempt to improve the knowledge discovery process. The ubiquitous data streams generated from various applications are greater in volume (Ahmed and Mahmood, 2014; Ahmed et al., 2015e, 2015f, 2015c). Given the fact that internet traffic doubles each year and computer network traffic is increasing at a fast rate making it a challenging task to monitor a network in real time. Applications such as email, ftp, http and p2p generate a large amount of data, even for small networks, which cannot be analyzed in real time (Zhu, 2011). Consequently, many existing data mining techniques cannot be applied to data streams which are evolving and need to be mined in a single pass. As argued by Hoplaros et al. (2014), summarization is a potential solution to this issue. However, existing summarization processes are complex and struggles to find emerging patterns in huge volumes of data also known as 'Big Data'. This poses a challenge for the next-generation data mining community. From a network traffic perspectives, finding both normal and anomalous traffic patterns is important and could be an interesting area for future research.

Existing anomaly detection techniques are mostly for monitoring a single system or a single network by carrying out local analysis for attacks. Hence, between instances of such a standalone anomaly detection techniques, no communication and interaction exists. Certainly, such a solution will not be able to detect sophisticated and highly distributed attacks (Vasilomanolakis et al., 2015; Tan et al., 2014b). Thus, for the security of large networks and large IT ecosystems (i.e. cloud services), collaborative techniques are extremely efficient which consist of several monitors that act as sensors and collect data. Due to the unavailability of implementations of collaborative techniques such as CIDSs (Collaborative Intrusion Detection Systems), future research efforts are necessary for extensive quantitative evaluation with state-of-the-art network infrastructure.

References

Adfa intrusion detection datasets, accessed: 2014-12-29. URL ⟨http://seit.unsw.adfa.edu.au/staff/sites/hu/⟩.

Ahmed M, Mahmood A. Clustering based semantic data summarization technique: a new approach. In: 2014 IEEE 9th conference on industrial electronics and applications (ICIEA), 2014, p. 1780–5.

Ahmed M, Mahmood A. Network traffic analysis based on collective anomaly detection. In: 2014 IEEE 9th conference on industrial electronics and applications (ICIEA), 2014. p. 1141–46.

Ahmed M, Mahmood AN. Network traffic pattern analysis using improved information-theoretic co-clustering based collective anomaly detection. In: Security and privacy in communication networks, Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, vol. 153. Springer International Publishing, 2014. p. 1–16.

Ahmed M, Mahmood A. Novel approach for network traffic pattern analysis using clustering-based collective anomaly detection. Ann. Data Sci. 2015;2(1):111–30.

Ahmed M, Naser A. A novel approach for outlier detection and clustering improvement. In: 2013 8th IEEE conference on industrial electronics and applications (ICIEA), 2013, p. 577–82.

Ahmed M, Mahmood AN, Hu J. Outlier Detection, CRC Press, New York, USA, 2014. p. 3–21, Chapter 1 (in book: The State of the Art in Intrusion Prevention and Detection).

Ahmed M, Mahmood AN, Islam MR. A survey of anomaly detection techniques in financial domain, Future Generation Computer Systems, http://dx.doi.org/10.1016/j.future.2015.01.001.

Ahmed M, Anwar A, Mahmood AN, Shah Z, Maher MJ. An investigation of performance analysis of anomaly detection techniques for big data in scada systems. EAI Endorsed Trans Ind Netw Intell Syst 2015b;15(3):1–16.

Ahmed M, Mahmood AN, Maher MJ. An efficient technique for network traffic summarization using multiview clustering and statistical sampling. EAI Endorsed Trans Scalable Inf Syst 2015c;15(5):1–9.

Ahmed M, Mahmood A, Maher M. Heart disease diagnosis using co-clustering. In: Jung JJ, Badica C, Kiss A, editors. Scalable information systems, Lecture notes of the institute for computer sciences, Social informatics and telecommunications engineering, vol. 139. Springer International Publishing; 2015d. p. 61–70.

Ahmed M, Mahmood A, Maher M. A novel approach for network traffic summarization. In: Jung JJ, Badica C, Kiss A, editors. Scalable information systems, Lecture notes of the institute for computer sciences, Social informatics and

telecommunications engineering, vol. 139. Springer International Publishing; 2015e. p. 51–60.
Ahmed M, Mahmood A, Maher M. An efficient approach for complex data summarization using multiview clustering. In: Jung JJ, Badica C, Kiss A, editors. Scalable information systems, lecture notes of the institute for computer sciences, Social informatics and telecommunications engineering, vol. 139. Springer International Publishing; 2015f. p. 38–47.
Ambusaidi MA, Tan Z, He X, Nanda P, Lu LF, Jamdagni A. Intrusion detection method based on nonlinear correlation measure. Int J Internet Protoc Technol 2014;8(2/3): 77–86.
Axelsson S. Technical report: Research in intrusion-detection systems: A survey, no. 98–17, SE–412 96, Göteborg, Sweden, 1998.
Balabine I, Velednitsky A. Method and system for confident anomaly detection in computer network traffic. Google Patents, 2015.
Banerjee A, Dhillon I, Ghosh J, Merugu S, Modha DS. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. J Mach Learn Res 2007;8:1919–86.
Beckman RJ, Cook RD. Outlier .s. Technometrics 1983;25(2):119–49.
Caida, accessed: 2014-12-29. URL ⟨www.caida.org⟩.
Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv 2009;41(3):15:1–58.
Creech G, Hu J. Generation of a new ids test dataset: time to retire the kdd collection, In: Wireless communications and networking conference (WCNC), 2013 IEEE, 2013. p. 4487–92.
Cristianini N, Shawe-Taylor J. An introduction to support vector machines: and other kernel-based learning methods. New York, NY, USA: Cambridge University Press; 2000.
Debar H, Dacier M, Wespi A. A revised taxonomy for intrusion-detection systems. Ann Des Télécommun 2000;55(7–8):361–78.
Defcon, accessed: 2014-12-29. URL ⟨www.defcon.org⟩.
Deljac Željko, Randic M, Krcelic G. Early detection of network element outages based on customer trouble calls. Decis Support Syst 2015;73:57–73.
Eskin E. Anomaly detection over noisy data using learned probability distributions. In: Proceedings of the seventeenth international conference on machine learning, ICML '00, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 2000. p. 255–62.
Eskin E, Arnold A, Prerau M, Portnoy L, Stolfo S. A geometric framework for unsupervised anomaly detection. in: Barbará D, Jajodia S (editors). Applications of data mining in computer security, Adv Inf Secur, vol. 6. Springer US, 2002. p. 77–101.
Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD'96, 1996. p. 226–31.
Estevez-Tapiador JM, Garcia-Teodoro P, Diaz-Verdejo JE. Anomaly detection methods in wired networks: a survey and taxonomy. Comput Commun 2004; 27(16):1569–84.
Furnell S, Tucker C, Furnell S, Ghita B, Brooke P. A new taxonomy for comparing intrusion detection systems. Internet Res 2007;17(1):88–98.
Gogoi P, Bhattacharyya D, Borah B, Kalita JK. A survey of outlier detection methods in network anomaly identification. Comput J 2011;54(4):570–88.
Govaert G, Nadif M. Block clustering with Bernoulli mixture models: comparison of different approaches. Comput Stat Data Anal 2008;52(6):3233–45.
Guan Y, Ghorbani A, Belacel N. Y-means: a clustering method for intrusion detection. In: IEEE CCECE 2003 Canadian conference on electrical and computer engineering, 2003, vol. 2; 2003. p. 1083–6.
Hacking and cracking tools, accessed: 2014-12-29. URL ⟨http://hackingncrackingtools.blogspot.com.au/⟩.
Hawkins D. Identification of outliers (monographs on statistics and applied probability). 1st ed. Springer; 1980.
Hawkins S, He H, Williams G, Baxter R. Outlier detection using replicator neural networks. In: Kambayashi Y, Winiwarter W, Arikawa M, editors. Data warehousing and knowledge discovery, lecture notes in computer science, vol. 2454. Berlin, Heidelberg: Springer; 2002. p. 170–80.
Heller KA, Svore KM, Keromytis AD, Stolfo SJ. One class support vector machines for detecting anomalous windows registry accesses. In: Proceedings of the workshop on data mining for computer security, 2003.
Hodge V, Austin J. A survey of outlier detection methodologies. Artif Intell Rev 2004;22(2):85–126.
Hoplaros D, Tari Z, Khalil I. Data summarization for network traffic monitoring. J Netw Comput Appl 2014;37(0):194–205.
Hu W, Liao Y, Vemuri VR. Robust anomaly detection using support vector machines. In: Proceedings of the international conference on machine learning; 2003.
ICS Attack Dataset, 2014, accessed: 2015-02-27. URL ⟨http://www.ece.msstate.edu/wiki/⟩.
Identifying outliers via clustering for anomaly detection, accessed: 2014-12-29. URL ⟨https://repository.lib.fit.edu/handle/11141/126⟩.
Internet traffic archive, accessed: 2014-12-29. URL ⟨http://ita.ee.lbl.gov/⟩.
Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv 1999; 31(3):264–323.
Kendall K. A database of computer attacks for the evaluation of intrusion detection systems. In: Proceedings of DARPA information survivability conference and exposition (DISCEX); 1999. p. 12–26.
Kruegel C, Mutz D, Robertson W, Valeur F. Bayesian event classification for intrusion detection. In: Proceedings of 19th annual computer security applications conference; 2003. p. 14–23.
Krügel C, Toth T, Kirda E. Service specific anomaly detection for network intrusion detection. In: Proceedings of the 2002 ACM symposium on applied computing, SAC '02, ACM, New York, NY, USA; 2002. p. 201–8.
Kyoto Dataset, accessed: 2014-12-29. URL ⟨www.takakura.com⟩.
Lee W, Xiang D. Information-theoretic measures for anomaly detection. In: Proceedings of 2001 IEEE symposium on security and privacy, 2001 S P 2001; 2001. p. 130–43.
Lee W, Stolfo S, Mok K. A data mining framework for building intrusion detection models. In: Proceedings of the 1999 IEEE Symposium on Security and Privacy; 1999. p. 120–32.
Leung K, Leckie C. Unsupervised anomaly detection in network intrusion detection using clusters. In: Proceedings of the twenty-eighth Australasian conference on computer science – vol. 38, ACSC '05, Australian Computer Society, Inc., Darlinghurst, Australia, Australia; 2005. p. 333–42.
Liao H-J, Lin C-HR, Lin Y-C, Tung K-Y. Intrusion detection system: a comprehensive review. J Netw Comput Appl 2013;36(1):16–24.
Lin J, Keogh E, Fu A, Van Herle H. Approximations to magic: finding unusual medical time series. In: Proceedings of the 18th IEEE symposium on computer-based medical systems, CBMS '05, IEEE Computer Society, Washington, DC, USA; 2005. p. 329–334.
Mahmood A, Leckie C, Udaya P. An efficient clustering scheme to exploit hierarchical data in network traffic analysis. IEEE Trans Knowl Data Eng 2008; 20(6):752–67.
Mahmood AN, Hu J, Tari Z, Leckie C. Critical infrastructure protection: resource efficient sampling to improve detection of less frequent patterns in network traffic. J Netw Comput Appl 2010;33(4):491–502.
Mahoney M, Chan P. An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection. In: Vigna G, Kruegel C, Jonsson E, editors. Recent advances in intrusion detection, Lecture notes in computer science, vol. 2820. Berlin, Heidelberg: Springer; 2003. p. 220–37.
Markou M, Singh S. Novelty detection: a review; part 2: neural network based approaches. Signal Process 2003;83(12):2499–521.
Münz G, Li S, Carle G. Traffic anomaly detection using kmeans clustering. In: In GI/ITG Workshop MMBnet, 2007.
Noble CC, Cook DJ. Graph-based anomaly detection. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '03, ACM, New York, NY, USA; 2003. p. 631–6.
NSL-KDD, accessed: 2014-12-29. URL ⟨http://nsl.cs.unb.ca/NSL-KDD/⟩.
Oldmeadow J, Ravinutala S, Leckie C. Adaptive clustering for network intrusion detection. In: Dai H, Srikant R, Zhang C, editors. Advances in knowledge discovery and data mining, Lecture notes in computer science, vol. 3056. Berlin, Heidelberg: Springer; 2004. p. 255–9.
Papadimitriou CM. Computational complexity. Reading, MA: Addison-Wesley; 1994.
Papalexakis EE, Beutel A, Steenkiste P. Network anomaly detection using co-clustering. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), ASONAM '12, IEEE Computer Society, Washington, DC, USA; 2012. p. 403–10.
Patcha A, Park J-M. An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput Netw 2007;51(12):3448–70.
Pelleg D, Moore AW. X-means: extending k-means with efficient estimation of the number of clusters. In: Proceedings of the seventeenth international conference on machine learning, ICML '00, San Francisco, CA, USA; Morgan Kaufmann Publishers Inc., 2000. p. 727–34.
Petrovic S, Alvarez G, Orfila A, Carbo J. Labelling clusters in an intrusion detection system using a combination of clustering evaluation techniques. In: Proceedings of the 39th annual Hawaii international conference on System Sciences, 2006. HICSS '06, vol. 6; 2006. p. 129b–129b.
PHP: Hypertext processor, accessed: 2014-12-29. URL ⟨http://www.php.net⟩.
Phua C, Lee VCS, Smith-Miles K, Gayler RW. A comprehensive survey of data mining-based fraud detection research, CoRR abs/1009.6119. URL arxiv.org/abs/1009.6119.
Platt JC. Advances in kernel methods. In: Fast training of support vector machines using sequential minimal optimization. MIT Press, Cambridge, MA, USA; 1999. p. 185–208.
Poojitha G, Kumar K, Reddy P. Intrusion detection using artificial neural network, In: 2010 International conference on computing communication and networking technologies (ICCCNT); 2010. p. 1–7.
Portnoy L, Eskin E, Stolfo S. Intrusion detection with unlabeled data using clustering. In: Proceedings of ACM CSS workshop on data mining applied to security (DMSA); 2001.
Predict, accessed: 2014-12-29. URL ⟨www.predict.org⟩.
Qin T, Guan X, Li W, Wang P, Huang Q. Monitoring abnormal network traffic based on blind source separation approach. J. Netw. Comput. Appl. 2011;34(5):1732–42 Dependable multimedia communications: systems, services, and applications.
Ramadas M, Ostermann S, Tjaden B. Detecting anomalous network traffic with self-organizing maps. In: Vigna G, Kruegel C, Jonsson E, editors. Recent advances in intrusion detection, Lecture notes in computer science, vol. 2820. Berlin, Heidelberg: Springer; 2003. p. 36–54.
Rubner Y, Tomasi C, Guibas L. A metric for distributions with applications to image databases. In: Sixth international conference on computer vision, 1998; 1998. p. 59–66.
Shafi K, Abbass H. Evaluation of an adaptive genetic-based signature extraction system for network intrusion detection. Pattern Anal Appl 2013;16(4):549–66.

Please cite this article as: Ahmed M, et al. A survey of network anomaly detection techniques. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.11.016

Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Comput Secur 2012;31(3):357–74.

Shyu M-L, Chen S-C, Sarinnapakorn K, Chang L. A novel anomaly detection scheme based on principal component classifier. In: IEEE foundations and new directions of data mining workshop, in conjunction with ICDM'03, 2003. p. 171–9.

Su M-Y. Using clustering to improve the knn-based classifiers for online anomaly network traffic identification. J Netw Comput Appl 2011;34(2):722–30 (efficient and robust security and services of wireless mesh networks).

Syarif I, Prugel-Bennett A, Wills G. Unsupervised clustering approach for network anomaly detection. In: Benlamri R, editor. Networked digital technologies communications in computer and information science, vol. 293. Berlin Heidelberg: Springer; 2012. p. 135–45.

Symantec internet security threat report, accessed: 2014-12-29. URL ⟨http://www.symantec.com/⟩.

Tan Z, Jamdagni A, He X, Nanda P, Liu RP. A system for denial-of-service attack detection based on multivariate correlation analysis. IEEE Trans Parallel Distrib Syst 2014a;25(2):447–56.

Tan Z, Nagar UT, He X, Nanda P, Liu RP, Wang S, Hu J. Enhancing big data security with collaborative intrusion detection. IEEE Cloud Comput 2014:27–33.

Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Trans Inf Syst Secur 2000:3(4):262–94.

Thc-hydra, accessed: 2014-12-29. URL ⟨http://www.thc.org/thc-hydra/⟩.

The apache software foundation, accessed: 2014-12-29. URL ⟨http://apache.org⟩.

The Global State of Information Security Survey 2015, accessed: 2015-01-19. URL ⟨http://www.pwc.com⟩.

Thottan M, Ji C. Anomaly detection in ip networks. IEEE Trans Signal Process 2003;51(8):2191–204.

Tikiwiki cms groupware remote php code injection, accessed: 2014-12-29. URL ⟨http://www.exploit-db.com/exploits/18265/⟩.

Tikiwiki: Cms groupware, accessed: 2014-12-29. URL ⟨http://info.tiki.org/Tiki+Wiki+CMS+Groupware⟩.

Towards a taxonomy of intrusion-detection systems, Comput. Netw. 31 (9) (1999) 805–822.

Ubuntu Linux, accessed: 2014-12-29. URL ⟨http://www.ubuntu.com⟩.

Vasilomanolakis E, Karuppayah S, Mühlhäuser M, Fischer M. Taxonomy and survey of collaborative intrusion detection. ACM Comput Surv 2015;47(4):551–533.

Verizon's data breach investigation report 2014, accessed: 2014-12-29. URL ⟨http://www.verizonenterprise.com/DBIR/2014/⟩.

Wang Q, Shen Y, Zhang JQ. A nonlinear correlation measure for multivariable data set. Phys D: Nonlinear Phenom 2005;200(3–4):287–95.

Xie M, Han S, Tian B, Parvin S. Anomaly detection in wireless sensor networks: a survey. J Netw Comput Appl 2011;34(4):1302–25 (advanced topics in cloud computing).

Yang Y, McLaughlin K, Littler T, Sezer S, Wang H. Rule-based intrusion detection system for scada networks. In: Renewable power generation conference (RPG 2013), 2nd IET; 2013. p. 1–4.

Ye N, Chen Q. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. Qual Relaib Eng Int 2001;17:105–12.

Zhang Z, Li J, Manikopoulos CN, Jorgenson J, Ucles J. Hide: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification. In: Proceedings of IEEE workshop on information assurance and security; 2001. p. 85–90.

Zhu R. Intelligent rate control for supporting real-time traffic in wlan mesh networks. J Netw Comput Appl 2011;34(5):1449–58 (dependable multimedia communications: systems, services, and applications).

Please cite this article as: Ahmed M, et al. A survey of network anomaly detection techniques. Journal of Network and Computer Applications (2015), http://dx.doi.org/10.1016/j.jnca.2015.11.016