Security in Data Mining-A Comprehensive Survey

Niranjan A[1], Nitish A[2] and P Deepa Shenoy[3]

[1] UVCE

Received: 8 December 2015 Accepted: 3 January 2016 Published: 15 January 2016

Abstract

Data mining techniques, while allowing the individuals to extract hidden knowledge on one hand, introduce a number of privacy threats on the other hand. In this paper, we study some of these issues along with a detailed discussion on the applications of various data mining techniques for providing security. An efficient classification technique when used properly, would allow an user to differentiate between a phishing website and a normal website, to classify the users as normal users and criminals based on their activities on Social networks (Crime Profiling) and to prevent users from executing malicious codes by labelling them as malicious. The most important applications of Data mining is the detection of intrusions, where different Data mining techniques can be applied to effectively detect an intrusion and report in real time so that necessary actions are taken to thwart the attempts of the intruder. Privacy Preservation, Outlier Detection, Anomaly Detection and PhishingWebsite Classification are discussed in this paper.

Index terms— anamoly detection; classification; intrusion detection system; outlier detection; privacy preserving data minig.

1 I. Introduction

he term Security from the context of computers is the ability, a system must possess to protect data or information and its resources with respect to confidentiality, integrity and authenticity [1]. Confidentiality ensures that, a third party in no way would be able to read and understand the content while Integrity would not allow a third party to change or modify the content as a whole or even parts of it. Authenticity feature on the other hand would not allow a person to use, view or modify the content or the resource, if he is found to be unauthorised [2].

Those actions that compromise the availability, integrity or confidentiality of one or more resources of a computer could be termed as Intrusion. Preventing intrusions employing firewall and filtering router policies fail to stop these attacks. Inspite of all attempts to build secure systems, intrusions can still happen and hence they must be detected on their onset. An Intrusion detection system(IDS) [3] by employing data mining techniques can discover consistent patterns of features of a system that are useful can detect anomalies and known intrusions using a relevant set of classifiers. Using some of the basic data mining techniques such as Classification and Clustering, Intrusion can be detected easily. Classification techniques are helpful in analyzing and labelling the test data into known type of classes, while Clustering techniques are used to group objects into a set of clusters, such that all similar objects become the members of the same cluster and all other objects become members of other clusters [4]. Data mining, while allowing the extraction of hidden patterns or the underlying Figure 1: Privacy Preserving Data Mining Techniques knowledge from large volumes of data, might pose security challenges [5]. Privacy Preserving Data Mining(PPDM)aims at safeguarding sensitive information from an un-solicited or unsanctioned disclosure [6]. A number of PPDM approaches have been proposed so far. Some of them are listed as shown in Fig. 1, based on their enforcing privacy principle.

2 T a) Suppression

Any private or sensitive information pertaing to an individual such as name, age, salary, address and other information is suppressed before any computation takes place. Some of the techniques employed for this suppression are Rounding(Rs/-35462.33 may be rounded to 35,000), Generalization (Name Louis Philip may be replaced with the initials LP and Place Hamburg may be replaced with HMG and so forth). However when data mining requires full access to sensitive values, Suppression cannot be used. An alternate way of suppression is to limit the identity linkage of a record rather than suppressing thesensitive information present within a record. This technique is referred to as De-Identification. k-Anonymity is one such de-identification technique. It ensures that protection of the data released against Re-identification of the persons to which the data refer [7] [8]. Enforcing k-anonymity before all data are collected in one trusted place is difficult. A cryptographic solution based on Secret Sharing technique of Shamir could be used instead; this however incurs computation overhead.

3 b) Randomization

Assuming the presence of a central server of a company that accepts information present with many customers and performs data mining techniques for building an Aggregate Model; Randomization allows the customers to introduce controlled noise or randomly perturb the records and to take away true information present in it. Introduction of noise can be achieved in several ways by addition or multiplication of the values generated randomly. Perturbation helps Randomization technique to achieve preservation of the required privacy.

The individual records are generated by the addition of such randomly generated noise to the original data. The noise thus added to individual records cannot be recovered, resulting in the desired privacy. Randomization techniques typically involve the following steps: 1. Only after randomizing their data, the Data Providers transmit this data to the Data Receiver.

4 Data receiver computes the distribution by running a Distribution Reconstruction Algorithm. c) Data Aggregation

Data Aggregation Techniques, in order to facilitate data analysis: combine data together from various sources. This might allow an attacker to deduce private and invidual-level data and to identify the party. When the extracted data allows the data miner to identify specific individuals, his privacy is considered to be under a serious threat. To prevent data from being identified, it may be anonymized immediately after the aggregation process. However, the Anonymized data sets can still contain enough information that could be used for the identification of individuals [9].

5 d) Data Swapping

Data swapping process involves swapping of values across different records for the sake of privacypreservation. Without perturbing the lower order totals of the data, privacy of data can still be preserved allowing aggregate computations to be performed exactly as before. Since this technique does not follow randomization, it can be used in conjunction with other frameworks such as k-anonymity without violating the privacy definitions for that model.

6 e) Noise Addition/Perturbation

Differential privacy through the addition of controlled noise provides a mechanism that maximizes the accuracy of queries while minimizing the chances of identification of its records [10]. Some of the techniques used in this regard are: 1. Laplace Mechanism 2. Sequential Composition

7 Parallel Composition

The rest of this paper is structured as follows: Section-II covers a brief review of Classification and Detection of intrusions by employing various Data Mining Techniques, while Clustering techniques and their applications in Intrusion Detection are presented in Section-III. PPDM techniques and their necessity along with various types of PPDM are discussed in Section-IV. An overview of Intusion Detection System is discussed in Section-V. Phishing Website Classification using Data Mining Techniques are presented in Section-VI. Artificial Neural Networks(ANN) are presented in Section-VII. Section-VIII presents Anomaly Detection/Outlier Detection. Section-IX describes the various ways of Mitigating Code Injection Attacks.

8 II. Classification and Detection Using Data mining Techniques

Malware computer programs that replicate themselves in order to spread from one computer to another computer are called as worms. Malware includes worms, computer viruses, Trojan Horse, key loggers, adware, spyware Port scan worm, UDP worm, http worm, User to Root Worm and Remote to Local Worm and other malicious code [11]. Attackers write these programs for various reasons varying from interruption of a computer process, gathering sensitive information, or gaining entry to private systems. Detecting a worm on the internet is very important, because it creates vulnerable points and reduces the performance of the system. Hence it is essential to detect the worm on the onset and classify it using data mining classification algorithms much before it causes any damage. Some of the classification algorithms that can be used are Random Forest, Decision Tree, Bayesian and others [12]. A majority of worm detection techniques use Intrusion Detection System(IDS) as the underlying principle. Automatic detection is challenging because it is tough to predict what form the next worm will take. IDS can be classified into two types namely Network based IDS and Host based IDS. The Network based Intusion Detection System reflects network packets before they spread to an end-host, while the Host based Intusion Detection System reflects network packets that are already spread to the end-host. Moreover, the Host based detection studies encode network packets so

9 Global Journal of Computer Science and Technology

Volume XVI Issue V Version I 52 Year 2016 ( ) that the stroke of the internet worm may be struck. When we focus on the network packet without encoding, we must study the performances of traffic in the network. Several machine learning techniques have been used in the field of intrusion and worm detection systems. Thus, Data Mining and in particular Machine Learning Technique has an important role and is essential in worm detection systems. Using various Data Mining schemes several new techniques to build several Intrusion Detection models have been proposed. Decision Trees and Genetic Algorithms of Machine Learning can be emoloyed to learn anomalous and normal patterns from the training set and classifiers are then generated based on the test data to label them as Normal orAbnormal classes. The data that is labelled as Abnormal could be a pointer to the presence of an intrusion.

10 a) Decision Trees

Quinlan's decision tree technique, is one of most popular machine learning techniques. The tree is constructed using a number of decision and leaf nodes following divide-andconquer technique [12]. Each decision node tests a condition on one of the attributes of the input data and can essentially have a number of branches, to handle a separate outcome of the test. The result of decision may be represented as a leaf node. A training data set T is a set of n-classes {C1, C2 ,..., Cn}. T is treated as a leaf when it comprises of cases belonging to a single class. If T is empty with no cases, it is still treated a leaf and the major class of the parent node is given the related class. A test based on an attribute ai of the training data is performed when T consists of multiple classes, T is split into k subsets {T1, T2, ..., Tk}, where k gives the number of test outcomes. The process is recursed over each Tj, where 1 <= j<= n, until every subset belongs to a single class. Choosing the best attribute for each decision node while constructing the decision tree is very crucial. The C4.5-DT adopts Gain Ratio Criterion for the same. According to this criterion, an attribute that provides maximum information gain and that reduces the bias in favor of tests is chosen. The tree thus built can then be used to classify the test data, whose features are same as that of the training data. The test is carried out starting from the root node. Based on the outcome, one of the branches leading to a child is followed. As long as the child is not a leaf, the process is repeated recursively. The class and its corresponding leaf node is given to the test case being examined.

11 b) Genetic Algorithms(GA)

A machine learning approach of solving problems by employing biological evolution techniques are called Genetic Algorithms(GA). They can be effectively used to optimize a population of candidate solutions. GA makes use of data structures that are modelled on chromosomes and they are subjected to evolution using genetic operators namely: selection, crossover and mutation [13]. Random generation of a population of chromosomes is performed in the beginning. The population thus formed comprises of all possible solutions of a problem and are considered the candidate solutions. Different positions of a chromosome called 'genes' are encoded as bits, characters or numbers. A function called Fitness Function evaluates the goodness of each chromosome based on the desired solution. Crossover operator simulates natural reproduction while Mutation operator simulates mutation of the species. The Selection operator chooses the fittest chromosomes [14]. **??**ig 2. depicts the operations of Genetic Algorithms. Before using GA for solving various problems, following three factors have to be considered 1. Fitness function 2. Individuals representation and 3. Parameters of GA Figure **??**: Flowchart for a GA GA based approach can be incorporated for designing Artificial Immune Systems. Using this approach, Bin et al., [15] have proposed a method for smartphone malware detection where static and dynamic signatures of malwares are extracted and malicious scores of tested samples are obtained.

12 c) Random Forest

A classification algorithm that is made up of a collection of tree structured classifiers, and that chooses the winner class based on the votes casted by the individual trees present in the forest is called the Random Forest Algorithm. Each tree is constructed by picking up random data from a training dataset. The chosen dataset may be split up into training and testsets. The major chunk of the dataset goes into the training set while the minor chunk forms the test set. The tree construction involves the following steps:

13 d) Association Rule Mining (ARM)

Association-rule mining discovers interesting relations between a set of attributes in datasets [16]. The datasets and their inter-relationship can be represented as association rules. This information can be used for making strategic decisions about different activities such as, promotional pricing, shelf management and so on [17]. Traditional Association rule mining involves a data analyst being given datasets of different companies for the purpose of discovering patterns or asociation rules that exist between the datsets [18]. Although, we can achieve sophisticated analysis on these extremely large datasets in a cost-effective manner [19], it poses security risk [20] for the data owner whose sensitive information can be deduced by the dataminer [21]. Even today, association rule mining is one of the widely used pattern discovery methods in KDD.

Solving an ARM problem basically involves traversing the items in a database, which can be done using various algorithms based on the requirement [22]. ARM algorithms are primarily categorised into BFS (Breadth First Search) and DFS (Depth First Search) methods based on the strategy used to traverse the search space [23]. The BFS and DFS methods are further classified into Counting and Intersecting, based on how the support values for the itemsets are determined. The algorithms Apriori, Apriori-TID and Apriori-DIC are based on BFS with Counting strategies, while the Partition algorithm is based on BFS with Intersecting strategies. The FP-Growth algorithm on the otherhand, is based on DFS with Counting strategies while ECLAT is based on DFS with Intersecting [24] [25]. These algorithms can be optimized specifically for improving the speedup [26] [27]. BFS with Counting Occurences: The common algorithm in this category is the Apriori algorithm. It utilizes the downward closure property of an itemset, by pruning the candidates with infrequent subsets before counting their supports.The two metrics to be considered while evaluating the association rules are: support and confidence. BFS offers the desired optimization by knowing the support values of all subsets of the candidates in advance. The limitation of this approach is increased computational complexity in rule extraction from a large database. Fast Distributed Mining(FDM) algorithm is a modified, distributed and unsecured version of the Apriori algorithm [28]. The advancements in data mining techniques, have enabled organizations in using data more efficiently.

In Apriori, the candidates of a cardinality k are counted by a single scan of the entire database. Looking up for the candidates in each transaction forms the most crucial part of the Apriori Algorithm. For this purpose, a hashtree structure is used **??**29]. Apriori-TID an extension of Apriori, represents each transaction based on the current candidates it contains, unlike normal Apriori that relies on raw database. Apriori-Hybrid combines the benefits of both Apriori and Apriori-TID. Apriori-DIC another variation of Apriori, tries to soften the separation that exists between the processes, counting and candidate generation. This is done by using a prefix-tree. BFS with Intersections: A Partition Algorithm is similar to the Apriori algorithm that uses intersections rather than counting occurences for the determination of support values. The partitioning of itemsets could result in the exponential growth of intermediate results beyond the physical memory limitations. This problem can be overcome, by splitting the database up into a number of chunks that are smaller in size and each chunk is treated independently. The size of a chunk is determined such that all intermediate lists can fit into memory. An additional scan can optionally be performed to ensure that the itemsets are not only locally frequent but also are globally frequent. DFS with Counting Occurences: In Counting, a database scan for each reasonable sized candidate set is performed. Because of the involvement of computational overhead in database scanning, the simple combination of DFS and Counting Occurences is practically irrelevant. FP-Growth on the otherhand uses a highly compressed representation of transaction data called FP-Tree. An FP-Tree is generated by counting occurences and performing DFS. DFS with Intersections: The algorithm ECLAT combines DFS with the list intersections to select agreeable values. It makes use of an optimization technique called Fast Intersections. It does not involve the process of splitting up of the database since complete path of classes beginning from the root would be maintained in the memory. As this method eliminates most of the computational overhead the process of mining association rules becomes faster.

14 III. Clustering

Clustering is one of the widely used discovery methods in data mining. It allows to group a set of data in such a way that, Intra-Cluster similarity are maximized while minimizing the Inter-Cluster similarity are minimized. Clustering involves unsupervised learning of a number of classes that are not known in advance. The clustering algorithms can be broadly clasified into the following types and are listed in Fig. Unweighted Pair Group Method with Arithmetic Mean (UPGMA), or Average Linkage Clustering. Selecting appropriate clusters from the available hierarchy of clusters, could be achieved either using Agglomerative or Divisive Clustering.In Agglomerative Clustering, we begin with single objects and conglomerate them into clusters while in Divisive clustering, we start with the complete data set and isolate it into segments.

15 b) Centroid Based Clustering

Centroid-based clustering may have clusters that are represented by a vector, which necessarily is not a member of the data set or may have clusters strictly restricted to the members of the dataset. In kmeans Clustering algorithm, the number of clusters is limited to size k, it is required to determine k cluster centers and assigning objects to their nearest centers.

The algorithm is run multiple times with different k random initializations to choose the best of multiple runs [30]. In kmedoid clustering, the clusters are strictly restricted to the members of the dataset while in kmedians clustering, only the medians are chosen to form a cluster. The main disadvantage of these techniques is that the number of clusters k is selected beforehand. Furthermore, they result in incorrectly cut borders in between the clusters.

16 c) Distribution Based Clustering

Distribution-based clustering technique forms clusters by choosing objects that belong more likely to the same distribution. One of the most commonly preferred distribution techniques is the Gaussian Distribution. It suffers from the overfitting problem where a model cannot fit into set of training data.

17 d) Density Based Clustering

In this type of clustering, an area that is having higher density than the rest of the data set is considered as a cluster. Objects in the sparse areas are considered to be noise and border points. There are three commonly used Density-based Clustering techniques namely: DBSCAN, OPTICS and Mean-Shift. DBSCAN is based on connecting points that satisfy a density criterion within certain distance thresholds. The cluster thus formed may consist of all density-connected objects and objects that are within these objects range free to have an arbitrary shape.

18 e) Recent Clustering Techniques

All the standard clustering techniques fail for highdimensional data and so some of the new techniques are being explored. These techniques fall into two categories namely: Subspace Clustering and Correlation Clustering. In Subspace Clustering, the clustering model specifies a small list of attributes that should be considered for the formation of a cluster while in Correlaton Clustering,the model along with this list of attributes it also provides the correlation between the chosen attributes.

55 Year 2016

19 ( ) C f) Other Techniques

One of the most basic clustering techniques is the BSAS(Basic Sequential Algorithmic Scheme). Given the distance d(p, C) between a vector point p and a cluster C, the maximum number of clusters allowed q and threshold of dissimilarity 0, the BSAS constructs the clusters even when the number of clusters to be formed is not known in advance.

Every newly presented vector is either assigned to an already existing cluster or a new cluster is created, depending on the distance to the already present clusters.

20 g) Clustering applications in IDS

Clustering technique may be effectively used in the process of Intrusion Detection. The setup is depicted in Fig. **??**. Alerts generated by multiple IDSs belonging to both Network and Host types are logged into a centralized database. The alert messages arriving from diffrent IDSs will be in different formats. Before passing them into the server, a preprocessing step is needed to bring them all into some uniform format [31].

Best effort values are chosen for the missing attributes during the preprocessing stage. The timestamp information may have to be converted into seconds for the sake of comparison. Different IDSs may use different conventions for naming a single event and hence it is required to standardize

21 Global Journal of Computer Science and Technology

Volume XVI Issue V Version I the messages. Each alert may be added with an unique ID to keep track of the alerts. After preprocessing and normalizing alerts they are passed to the first phase to perform filtering and labeling functions. To minimise the number of Alerts, it is a good idea to employ Alert Fusion during which alerts with same attributes that differ by a small amount of time are fused together. Alert Fusion makes the generalization process fast. Generalization involves the addition of hierarchical background knowledge into each attribute. On every iteration of this process, the selected attribute is generalized to the next higher level of hierarchy and those alerts which have become similar by now are grouped together.

22 IV. Privacy Preserving Data Mining (ppdm)

Privacy Preserving Data Mining techniques aim at the extraction of relevant knowledge from large volumes of data while protecting any sensitive information present in it. It ensures the protection of sensitive data to conserve privacy and still allowing us to perform all data mining operations efficiently. The two types of privacy concerned data mining techniques are: 1. Data privacy 2. Information privacy Data privacy focuses on the modification of the database for the protection of sensitive data of the individuals while Information privacy focuses on the modification for the protection of sensitive knowledge that can be deduced from the database.

Alternatively we can say that Data privacy is corcerned about providing privacy to the input while Information privacy on the otherhand is about providing privacy to the output. Preserving personal information from revealation is the main focus of a PPDM algorithm [32]. The PPDM algorithms rely on analysing the mining algorithms for any side effects that are acquired during Data privacy. The objective of Privacy Preserving Data Mining is building algorithms that transform the original data in some mannner, so that both the private data and knowledge are not revealed even after a successful mining process. Only when some relevant adequate benefit is found resulting from the access, the privacy laws would allow the access.

Multiple parties may sometimes wish to share private data resulting after a successful aggregation [33] without disclosing any sensitive information from their end [34]. Consider for example, different Book stores with respective sales data that is in a way considered to be highly sensitive, may wish to exchange partial information among themselves to arrive at the aggregate trends without disclosing their individual store trends. This requires the use of secure protocols for sharing the information across multiple parties. Privacy in such cases should be achieved with high levels of accuracy [35].

57 Year 2016

23 ( ) C

The data mining technology by principle is neutral in terms of privacy [36]. The motive for which a data mining algorithm is used could either be good or malicious [37]. Data mining has expanded the investigation possibilities [38] to enable researchers to exploit immense datasets on one hand [39], while the malicious use of these techniques on the other hand has introduced threats of serious nature against protection of privacy [40].

Discovering the base of privacy preserving data mining algorithms and connected privacy techniues is the need of the hour [41]. We are required to answer few questions in this regard such as 1. Evaluation of these algorithms with respect to one another 2. Should privacy preserving techniques be applied to each of the data mining algorithms? Or for all applications? 3. Expanding the places of usage of these techniques.

24 Investigating their use in the fields of Defense and

Intelligence, Inspection and Geo-Spatial applications.

25 The techniques of combining confidentiality, privacy

and trust with high opinion to data mining.

To answer these questions, research progresses in both data mining and privacy are required. Proper planning towards developing flexible systems is essential [42]. Few applications may demand pure data mining techniques while few others may demand privacy-preserving data mining [43]. Hence we require flexible techniques in data mining that can cater to the the changing needs [44]. The research progress made so far in the area of PPDM is listed in Table 1.

26 Distributed Privacy Preserving Data Mining(DPPDM):

The tremendous growth of internet in the recent times is creating new opportunities for distributed data mining [52], in which, mining operations performed jointly using their private inputs [53]. Often occurence of mining operations between untrusted parties or competitors, result in privacy leakage [54]. Thus, Distributed Privacy Preserving Data Mining(DPPDM) [10][55] algorithms require a high level of collaboration between parties to deduce the results or to share mining results that are not sensitive. This could sometimes result in the disclosure of sensitive information.

Distributed data mining are classified as Horizontally Partitioned Data and Vertically Partitioned Data. In a Horizontally partitioned data framework, each site maintains complete information on an unique set of entities, and the integrated dataset consists of the union of all of these datasets. Vertically Partitioned Data framework on the otherhand involves each site, maintaining different types of information and each dataset and has only limited information about same set of entities.

Privacy feature can limit the information leakage caused by the distributed computation techniques [56].

Each non-trusting party can compute its own functions for unique set of inputs, revealing only the defined outputs of the functions. Apart from hiding sensitive information, the privacy service also controls the information and its uses by involving various number of negotiations and tradeoffs between hiding and sharing.

All efficient PPDM algorithms are based on the assumption that it is acceptable to release the intermediate results obtained during the data mining operations. Encryption techniques solve the data privacy problem and their use would make it easy to perform data mining tasks among mutual untrustworthy parties, or between competitors. Due to its privacy concern, Distributed Data Mining Algorithms employ encryption techniques.

Encryption is used in both approaches(horizontally and vertically partitioned data) of Distributed Data mining without much stress on the effiency of encryption technique used.

If the data are stored on different machines and partitioning is done row-wise, it is called horizontal partitioning and if the data are stored and partitioned column wise then it is called vertical partitioning. An overview of the same is depicted in Fig. 5.

The objective of data mining techniques is to generate high level rules or summaries and generalize across populations, rather than revealing information about individuals but they work by evaluating individual data that is subject to privacy concerns. Since much of

27 Global Journal of Computer Science and Technology

Volume XVI Issue V Version I

28 C

this information held by various organizations has already been collected, providing privacy is a big chalenge. To prevent any correlation of this information, control and individual safeguards must be separated to be able to provide acceptable privacy. Unfortunately, this separation makes it difficult to use the information for the identification of criminal activities and other purposes that would benefit the society. Proposals to share information across agencies to combat terrorism and other criminal activities, would also remove the safeguards imposed by separation.

Many of the complex socio-technical systems suffer from an inadequate risk model that focuses on the use of Fair Information Practice Principles(FIPPs). Anonymization suffers from the risk of failure, since the circumstances surrounding its selection are ignored. A Hybrid approach that combines privacy risk model with an integrated anonymization framework involving anonymization as the primary privacy risk control measure can be considered instead [57]. Public-Key Program Obfuscation: The process of making a program uncomprehensible without altering its functionality is called Program Obfuscation. A program that is obfuscated should be a virtual black box meaning, if it is possible for one to compute something from it, it should also be possible to compute the same even from the input-output behavior of the program. Secure Multi-party Computation: Distributed computing involves a number of distinct,and connected computing devices that wish to carry out a combined computation of some function. For example, servers holding a distributed database system, may wish to update their database. The objective of secure multiparty computation is to allow parties to carry out distributed computing tasks in a secure way [33]. It typically involves the parties carrying out a computation based on their private inputs and neither of them willing to disclose its own input to other parties. The problem is conducting such a computation by preserving the privacy of their inputs. This problem is called the Secure Multi-party Computation problem (SMC) [34]. Consider the problem of two-parties who wish to securely compute the median. The two parties have with them two separate input sets X and Y. The parties are required to jointly compute the median of the union of their sets X U Y, without revealing anything about each other's set. Association Rules can be computed in an environment where different information holders have different types of information about a common set of entities.

29 V. Intrusion Detection System(ids)

Intrusion detection systems aim at the detection of an intrusion on its onset [58]. A high level of human expertise and significant amount of time are required for the development of a comprehensive IDS [59]. However, IDSs that are based on the Data Mining techniques require less expertize and yet they perform better. An Intrusion Detection System detects network attacks against services that are vulnerable [60], attacks that are data driven on applications, privilege escalation [61], logins that are un-authorized and access to files that are 59 Year 2016

30 ( ) C

sensitive in nature [62]. The data mining process also efficiently detects malware from the code [63], which can be used as a tool for cyber security [64] [65]. An overview of an Intrusion Detection System is presented in Fig 6.

An IDS is basically composed of several components such as, sensors, a console monitor and a central engine [66]. Sensors generate security events while all events and alerts are monitored and controlled by the Console Monitor and the Central Engine records events in a database and generate alerts based on a set of rules [67]. An Intrusion detection system [68] can be classified depending on the location and the type of Sensors and based on the technique used by the Central engine for the generation of alerts. A majority of IDS implementations, involve all of the three components integrated into a single device.

Current virus scanner methodology makes use of two parts namely a Detector based on signatures and a Classifier based on the heuristic rules for the detection of new viruses. The signature-based detection algorithms rely on signatures that are unique strings of known malicious executables for the generation of detection models.The disadvantages of this approach are: it is more time-consuming and fails in detecting new malicious executables. Heuristic classifiers on the other hand are generated by a set of virus experts for the detection of new malicious executables.

31 i. Network Based IDS

Because of their increasingly vital roles in modern societies, computer networks have been targeted by enemies and criminals. For the protection of our systems, it is very essential to find the best possible solutions. Intrusion prevention techniques such as, authentication technqiues involving passwords or biometrics [69], programming errors avoidance, and protection of information using encryption techniques have been widely used as a first line of defense. Intrusion prevention techniques as the sole defense mechanism are not sufficient enough to combat attacks. Hence, it can therefore be used only as a second line of defense for the protection of computer systems [70].

An Intrusion Detection system must protect resources such as accounts of users [71], their file systems and the system kernels of a target system and must be able enough to characterize the legitimate or normal behavior of these resources by involving techniques that compare the ongoing system activities with already established models and to identify those activities that are intrusive [72] [73]. Network packets are the data source for Network-Based Intrusion Detection Systems. The NIDS makes use of a network adapter to listen to and analyse network traffic as the packets travel across the network. A Network based IDS generates alerts upon detecting an intrusion from outside the perimeter of its enterprise [74]. The network based IDSs are categorically placed at strategic points on LAN to observe both inbound and outbound packet [75]. Network based IDSs are placed next to the firewalls to alert about the inbound packets that may bypass the firewall [76]. Few Network-Based IDSs take custom signatures from the user security policy as input, permitting limited detection of security policy violations [77]. When packets that contain intrusion originated from authorized users, the IDS may not be able to detect [78] In a Host-based IDS, the monitoring sensors are placed on network resources nodes so as to monitor logs that are generated by the Host Operating System or application programs.

These Audit logs contain records of events or activities that are occuring at individual Network resources [81]. Since a Host-Based IDS is capable of detecting attacks that cannot be seen by a Network based IDS, an attacker can misuse one of trusted insiders [82]. A Host based system utilizes Signature Rule Base that is derived from security policy that is specific to a site. A Host Based IDS can overcome all the problems associated with a Network based IDS as it can alarm the security personnel with the location details of intrusion, he can take immediate action to thwart the intrusion. A Host based IDS can also monitor any unsuccessful attempts of an attacker. It can also maintain seperate records of user login and user logoff actions for the generation of audit records.

32 Advantages

Some of the advantages of a Host Based IDS are as follows: 1. Can detect attacks that are not detected by a Network Based IDS. 2. Operates on Operating System audit log trails, for the detection of attacks involving software integrity breaches.

33 Disadvantages

The disadvantages are: 1. Certain types of DoS(Denial of Service)attacks can disable them [83]. 2. Not suited for detecting attacks that target the network. 3. Difficult to configure and manage every individual system. iii. Hybrid IDS Since Network and Host-based IDSs have strengths and benefits that are unique over one another, it is a good idea to combine both of these strategies into the next generation IDSs [84]. Such a combination is often referred to as a Hybrid IDS. Addition of these two components would greatly enhance resistance to few more attacks. a. DM techniques for IDS Some of the techniques and applications of data mining required for IDS include the following 1. Pattern Matching 2. Classification and 3. Feature Selection Pattern Matching: Pattern Matching is a process of finding a particular sequence of a part of data (substring or a binary pattern), in the whole data or a packet to get a desired information [87]. Though it is fairly rigid, it is indeed simple to use. A Network Based IDS succeeds in detecting an intrusion only when the packet in question is associated with a particular service or, destined to or from a particular port. That is, only few fields of the packet such as Service, Source/Destination port address and few others have to be examined thereby reducing the amount of inspection to be done on each packet.

However, it makes it difficult for systems to deal with Trojans and their associated traffic that can be moved at will. The pattern matching can be classified into two categories based on the frequency of occurrence namely: a) Frequent Pattern Matching and b) Outlier Pattern Matching a) Frequent Pattern Matching

These are the type of patterns which occur frequently in an audit data, i.e., the frequency of occurrence of these patterns is more compared to other patterns in the same data [82].

34 Global Journal of Computer Science and Technology

Volume XVI Issue V Version I 60 Year 2016 ( ) Determining frequent patterns in a big data helps in analyzing and forecasting of a particular characteristic of the data. For example, by analyzing the sales information of an organization, frequent pattern matching might help to predict the possible sales outcome for the future. It also helps in decision making. The frequent pattern mining in ADAM project data is done by mining the repository for attack-free (train) data which is compared with the patterns of normal profile (train) data. A classifier is used to reduce the false positives.

35 b) Outlier Pattern Matching

Patterns that are unusual and are different from the remaining patterns and that are not noise are referred to as Outlier Patterns. Preprocessing phase eliminates noise as it is not a part of the actual data while outliers on the other hand cannot be eliminated. Outliers exhibit deviating characteristics as compared to the majority of other instances. Outliers patterns are not usual and they occur less frequently and for this reason will have minimal support in the data. These patterns can quite often point out some sort of discrepancy in data such as transactions that are fraudulent, intrusion, abnormal behavior, economy recession etc.,. The outlier pattern mining algorithms can be of two types, one that looks for patterns only at fixed time intervals, and the other that calculates monitors patterns at all times. Outlier pappers make use of special data structures such as Suffix Tree and other String Matching Algorithms. Classification: Classification makes use of training examples for learning a model and to classify samples of data into known classes [88]. A wide range of classification techniques ranging from Neural Networks, Decision Trees, Bayesian classifier [89], Bayesian Belief Networks and others are used in applications that involve Data Mining techniques. Classification typically involves steps that are outlined below:

36 Feature Selection

Better classification Can consider NSL-KDD instances instead of their probabilities. Spam Mail classification and Text classification applications extensively use Naive Bayesian classifiers for they are less error prone. However, their disadvantage is that they require probabilities in advance. The probability information that is required by them is extremely huge which consist number of classes, their attributes and the maximum cardinality of attributes. The space and computational complexity of these classifiers increase exponentially.

37 Support Vector Machine(SVM):

Support Vector Machine is one of the learning methods extensively used for the Classification and Regression analysis of Linear and Non-linear data [90]. It maps input feature vectors into a higher dimensional space using non-linear mapping techniques. In SVM, the classifier is created by the linear separation of hyperpalnes and linear separation is achieved using a function called kernel.The Kernel transforms a linear problem by mapping it into feature spaces. Some of the commonly used kernel functions are Radial basis, sigmoid neural nets and polynomials. Users specify one of these functions while training the classifier and it selects support vectors along the surface of this function. The SVM implementation tries to achieve maximum separation between the classes [91]. Intrusion detection system involves two phases namely training and testing. SVMs are capable of learning a larger set of patterns and can provide better classification, because the categorizing complexity is independent of the feature space dimensionality [92]. SVMs can update the training patterns dynamically with the availability of new pattern during classification. For the efficient classification it is required to reduce the dimensionality of the dataset. To do this we have Feature Selection.

iii. Feature Selection(FS)

The process of reducing the dataset dimensionality by selecting a subset of the features from the given set of features is called Feature Selection [93]. FS involves discarding of redundant and irrelevant features. FS is considered to be an efficient machine learning technique that helps in building classification systems which are efficient. With the reduction in subset dimensionality, the time complexity is reduced with improved accuracy, of a classifier. Information Gain is a proposition of feature selection that can be used to compute entropy cost of each attribute. An entropy cost can be called as a rank. Rank of each feature represents its importance or association with an solution class that is used to recognize the data. So a feature with comparatively higher rank will be one of the most important features for classification. The three standard approaches that are commonly followed for feature selection are embedded technique, filter technique, and wrapper technique.

38 VI. Phishing Websites Classification

In the art of emulating a website of a trusted and creditable firm with the intention of grabbing users' private information (ussername, password) is called phishing. Fake websites are ususlly created by dishonest people to masquerade honest websites. Users unknowingly lose money due to phishing activities of attackers. Online trading therefore demands protection from these attacks and is considered a critical step. The prediction and classification accuracy of a website depends on the goodness of the extracted features. Most of the internetusers feel safe against phishing attacks by utilizing antiphishing tool, and hence the anti-phishing tools are required to be accurate in predicting phishing [94]. Phishing websites give us a set of clues within its content parts and through security indicators of the browsers [95]. A variety of solutions have been proposed to tackle the problem of phishing. Data mining techniques involving Rule based classification [96] serve as promising methods in the prediction of phishing attacks.

Phishing attack typically starts by, attacker sending an email to victims requesting personal information to be disclosed, by visiting a particular URL [97]. Phishers use a set of mutual features to create phishing websites to carry out proper deception [98]. We can exploit this information to successfully distinguish between phishy and non-phishy websites based on the extracted features of the website visited [94]. The two approaches that are commonly used in the identification of phishing sites are: black-list based, which involves comparison of the requested URL with those that are present in that list and Heuristic based method that involves the collection of certain features from the website to label it either as phishy or legitimate [99]. The disadvantage of Black-list based approach is that the black-list can not contain all phishing websites since, a new malicious website is launched every second [100]. In contrast, a Heuristic-based approach can recognize fraudulent websites that are new [101]. The success of Heuristic-based methods depend on the selection of features and the way they are processed. Data mining can be effectively used here to find patterns as well as relations among them [102]. Data mining is considered to be important for taking decisions, since decisions are made based on the patterns and rules derived using the data mining algorithms [103].

Although there is substantial progress made in the development of prevention techniques, phishing still remains a threat since most of the counter measures techniques in use are based still on reactive URL blacklisting [104]. Since Phishing Web sites will have shorter life time these methods are considered to be inefficient. Newer approaches such as Associative Classification (AC) are more suitable for these kinds of applications. Associative Classification technique is a new technique derived by combining Association rule and Classification techniques of data mining [105]. AC typically includes two phases; the training phase to induce hidden knowledge (rules) using Association rule and the Classification phase to construct a Classifier after pruning useless and redundant rules. Many research studies have revealed that AC usually shows better classifiers with reference to error rate than other standard classification approaches such as decision tree and rule induction.

39 VII. Artificial Neural Networks(ann)

An Artificial Neural Network is basically a connected set of processing units. Each connection has a specific weight that determines how one unit affects the other. Few of these units act as input nodes and few other as output nodes and remaining nodes consists of hidden layer. Neural network performs functionally, a mapping from input values to output values by activating each input node and allowing it to spread through the hidden layer nodes to the output nodes. The mapping is stored in terms of weight over connection. Fig. **??** shows the structure of HHNN [62].

ANN is one of the widely used techniques in the field of intrusion detection. **??**NN Feature selection is independent of the classifier used in case of Filter method, while in Wrapper method features are chosen specifically to the intended classifier. Filter method uses an arbitrary statistical way for the selection of features whereas wrapper method uses a learning algorithm to find the best subset of features. Wrapper approach is more expensive and requires more computational time than the filter approach but gives more accurate results compared to filter technique. **??**HHNN). Anomaly detection assumes that the intrusions always return as a number of deviations from the normal patterns. HHNN technique studies the relationship between the two sets of information, and generalizes it in getting new inputoutput pairs reasonably. Neural networks can be used hypothetically for the identification of attacks and look for these attacks in the audit stream. Since there is no reliable method at present to realize causes of association, it cannot clarify the reason behind the classification of the attack. The research progress made in HHNN is summarized in Table 3.

40 VIII. Anomaly Detection/Outlier Detection

Anomaly detection is a process that involves finding nonconforming patterns to the expected behavior. Such patterns are called anomalies. Different application domains term them differently as outliers or aberration or surprises or peculiarities or It is ineffective against new types of attacks which makes it susceptible to evasion methods.

Anomaly Based IDS on the other hand, records normal behavior and classifies the deviations from normal behavior as anomalies. It is considered to be robust and reliable to unknown attacks and prevent attacks from malicious users who improvise their attacking strategy. The widely used implementation of Anomaly Based IDS is by the extensive use of data through the same modules: Feature Extractor and Feature Selector, that is finally evaluated by the already trained Classifier. When the sample is found to be deviating from normal profiles, an alarm is raised. The profiles are required to be updated at regular intervals of time and Classifier training is also carried out periodically, so as to minimize the false alarm rate. For Feature selection, we can either employ the Ranking methods or the Filter methods. The Ranking methods output the feature set sorted in descending order according to a particular evaluation measure. The top variables in the feature set are considered to be the most discriminant features. It is therefore essential to determine a threshold to discard features that are considered to have little or no contribution to the classification process. Information Gain(IG) is one of the commonly used evaluation measures.

A variant of IG, with improvisation is the Gain Ratio (GR).

The GR overcomes the bias found in IG towards features resulting in a smaller set of features. For the purpose of Feature Selection we can employ a ranking method that is unsupervised called Principal components analysis(PCA).

The advantage of Filter methods for Feature Selection is that they automatically choose a set of selected features based on a particular evaluation measure. One of the widely employed Filtering methods for Feature Selection is the Best First Search(BFS). It makes use of Forward Selection and Backward Elimination to search through the feature space adopting a Greedy approach. When performance is found to be dropping, it backtracks to the previous feature subsets that have better performance and start all over again from there. BFS is computationally expensive for larger sets. Genetic Algorithms [109] is another type of Filtering technique that is considered to be very effective in practice [110].

41 IX. Mitigating Code Injection Attacks

A code injection attack typically involves writing of new machine code into the vulnerable programs memory [111], and after exploiting a bug in the program the control is redirected to the new code [112]. The protection technique [113], W+X mitigates this attack by allowing only either a Write or Execute operations on memory but never allows both [114].

The research progress made so far in this regard is summarized in Table 4.

42 a) Types of Code Injection

Some of the flavours of Code Injection attacks are: SQL Injection [121], HTML Script Injection [122], Object Injection [123], Remote File Injection [124] and Code Reuse Attacks(CRAs) [125].

43 i. SQL Injecton

A technique that uses SQL syntax to input commands that can alter read or modify a database is called SQL Injection. Consider for example a web page having a field on it to allow users to enter a password for authentication. The code behind the page usually a script code, will generate a SQL query to verify the matching password entered against the list of user names:

SELECT UsrList.Username FROM UsrList WHERE UsrList. Password = 'Password'

The access is granted when the password entered by the user matches the password specified in the query. If the malicious user can inject some valid code ('password' OR '1'='1') in the Password field. An attacker by leaving the password field empty makes the condition "'1'='1"' to become true and gains access to the database.

44 ii. HTML Script Injection

An attacker injects malicious code by making use of the <script>and </script>tags, within which he would change the location property of the document by setting it to an injected script.

iii. Object Injection PHP allows serialization and deserialization of objects. If an untrustworthy input is allowed into the deserialization function, it is possible to modify existing classes in the program and execute malicious attacks.

45 iv. Remote File Injection

Attackers might provide a Remote Infected file name as the path by modifying the path command of the script file to cause the intended destruction [126]. Attacks in which an attacker directs control flow through an already existing code with an erroneous result are called Code Reuse Attacks [127].

Attackers therfore have come out with codereuse attacks [128], in which a defect in the software is exploited to create a control flow through existing codebase to a malicious end [129]. The Return Into Lib C(RILC)is a type of code-reuse attack [130] where the stack is compromised and the control is transferred to the beginning of an existing library function such as mprotect() to create a memory region [131]that allows both write and execution operations on it to bypass W+X [132]. Such attacks can be effiently overcome using Data Mining techniques [133]. The source code is checked to find any such flaws and if so the instructions are classified as malicious [134]. Some of the classification Algorithms that can be used in this Regard are Bayesian [135], SVM [136] and Decision Tree [137].

46 vi. Return Oriented Programming

ROP attacks start when an attacker gains stack control [138] and redirects the control to a small snippet of code called gadget typically ending with a RET instruction [139]. Because attackers gain control over the return addresses [140], they can assign the RET of one gadget to the start of another gadget [141], achieving the desired functionality out of a large finite set of such small gadgets [142]. ROP Attacks inject no code and yet can induce arbitrary behavior in the targeted system [143]. A compiler-based approach has been suggested in [144] to combat any form of ROP. In [145], the authors present in-place code randomization that can be applied directly on third-party software, to mitigate ROP attacks. Buchanan et al., [146], have demonstrated that return-oriented exploits are practical to write, as the complexity of gadget combination is abstracted behind a programming language and compiler. Davi et al. [147] proposed runtime integrity monitoring techniques that use tracking instrumentation of program binaries based on taint analysis and dynamic tracing. In [148] a tool DROP, that detects ROP malicious code dynamically, is presented.

vii. Jump Oriented Programming In Jump Oriented Programming(JOP), an attacker links the gadgets using a finite set of indirect JMP instructions [149], instead of RET instructions. A special gadget called a dispatcher is used for flow control management among the gadgets [150].

1

Authors	Algorithm	Performance	Future enha ment
Boutet et al.(2015)[45]	kNN	Better than Randomization scheme	Can conside attacking r els
Tianqing et al.(2015)[46]	Correlated Differential Privacy (CDP)	Enhances the utility while answering a large group of queries on correlated datasets	Can experimente with Comple Applications
Bharath et al.(2015)[47]	PP k-NN classifier	Irrespective of the values of k, it is observed that SRkNNo is around 33% faster than SRkNN. E.g., when k=10, the computation costs of SRkNNo and SRkNN are 84.47 and 127.72 minutes, respectively (boosting the online running time of Stage 1 by 33.86%)	Parallelizatio not used
Nethravathi et al.(2015)[48]	PPDM	Reduced misplacement clustering error and removal of data that is sensitive and correlated	Works only numerical da
Mohammed et al.(2014)[49]	Differential Privacy	More secured under the Semi-Honest model	Overcoming vacy Attack
Vaidya et al.(2014)[50]	Distributed RDT	Lower Computation and Communication cost	Limited information that is still r vealed must checked
Lee(2014)[51]	Perturbation methods	Capable of performing RFM Analysis	Partial disclosure is still possibl

Figure 1: Table 1 :

2

1. Creation of a training dataset 2. Identification of classes and attributes 3. Identification of attributes that are useful for classification 4. Relevance analysis 5. Learning the Model using training examples 6. Training the set 7. Using the model for the classification of unknown data samples. Algorithm Performance M Vittapu et al.(2015)[85] SVM Classification TPR of 96% and FPR of 5% Mitchell et al.(2015)[61] Behavior Rule Analysis Better performance Jabez J et al.(2014)[98] Hyperboli Hopfiel Neural Network(HHNN) Detection rate of about 90% S Abadeh et al.(2014)[151] Genetic Fuzzy System Best tradeoff in terms of the mean F-measure,the average accuracy and the false alarm rate Soni et Bayesian Classifiers: Authors al.(2014)[86]	Future enhancement Can be experimented with other techniques Can be tested with other techniques Can be improved A Multi-objective Evolutionary Al-gorithm for maximizing performance metrics may be considered

Figure 2: Table 2 :

	Security in Data Mining-A Comprehensive Survey
Year 2016	
62	
Volume XVI Issue V Version I	FS runs as a part of data mining algorithms, in Embbedded technique.
)	
(C	
Global Journal of Computer Science and Technology	
	© 2016 Global Journals Inc. (US) 1

Figure 3:

Figure 4:

3

Authors	Algorithm	Performance	Future enhancement
C Cortes et al.(2016)[106]	Theoretical framework for analyzing and learning artificial neural networks	Optimizes generalization performance	Can be applied for different optimi tion tecniques and network architectures.
D T Bui et al.(2015)[107]	ROC and Kappa Index	MLP (90.2 %), SVM (88.7 %), KLR (87.9 %), RBF (87.1 %) and LMT (86.1 %).	Information Gain Ratio as feature selection can be tried.

Figure 7:

Figure 5: Table 3 :

4

Figure 6: Table 4 :

47 X. Conclusion

The purpose of this survey is to explore the importance of Data Mining techniques in achieving security [1] [2]

[1] Security in Data Mining-A Comprehensive Survey

[Checkoway et al.] , S Checkoway , L Davi , A Dmitrienko , A.-R .

[Stanley ()] , D M Stanley . *CERIAS Tech Report* 2013-19.

[Cortes et al. ()] , C Cortes , X Gonzalvo , V Kuznetsov , M Mohri , S Yang . arXiv:1607.01097v1. *AdaNet: Adaptive Structural Learning of Artificial Neural Networks* 2016. 1 (17) p. .

[N P Nethravathi et al. ()] 'A Brief Survey on Privacy Preserving Data Mining Techniques'. N P Nethravathi , J Vaibhav , Deepa Desai , M Shenoy , Indiramma , K R Venugopal . *Data Mining and Knowledge Engineering* 2016. 8 (9) p. .

[Liu et al. ()] 'A Collaborative Intrusion Detection Mechanism Against False Data Injection Attack in Advanced Metering Infrastructure'. X Liu , P Zhu , Y Zhang , K Chen . *IEEE Transactions on Smart Grid* 2015. 6 (5) p. .

[Lee et al. ()] 'A Data Mining Framework for Building Intrusion Detection Models'. W Lee , S J Stolfo , K W Mok . *Proceedings of the 1999 IEEE Symposium on Security and Privacy*, (the 1999 IEEE Symposium on Security and Privacy) 1999. p. .

[Darwish et al. ()] 'A Database Sanitizing Algorithm for Hiding Sensitive Multi-level Association Rule mining'. S M Darwish , M M Madbouly , M A El-Hakeem . *International Journal of Computer and Communication Engineering* 2014. 3 (4) p. .

[Sahu et al. ()] 'A Detail Analysis on Intrusion Detection Datasets'. S K Sahu , S Sarangi , S K Jena . *IEEE International on Advance Computing Conference(IACC)*, 2014. p. .

[Rauzy and Guilley ()] 'A Formal Proof of Countermeasures Against Fault Injection Attacks on CRT-RSA'. P Rauzy , S Guilley . *Journal of Cryptographic Engineering* 2014. 4 (3) p. .

[Gopalakrishna and Spafford ()] *A Framework for Distributed Intrusion Detection using Interest Driven Cooperating Agents*, R Gopalakrishna , E H Spafford . 2001. p. . (Technical Report)

[Cárdenas et al. ()] 'A Framework for Evaluating Intrusion Detection Architectures in Advanced Metering Infrastructures'. A A Cárdenas , R Berthier , R B Bobba , J H Huh , J G Jetcheva , D Grochocki , W H Sanders . *IEEE Transactions on Smart Grid* 2014. 5 (2) p. .

[Rasheed and Alhajj ()] 'A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences'. R Rasheed , Alhajj . *IEEE Transactions on Cybernetics* 2014. 44 (5) p. .

[Vaidya et al. ()] 'A Random Decision Tree Framework for Privacy-Preserving Data Mining'. J Vaidya , B Shafiq , W Fan , D Mehmood , D Lorenzi . *IEEE Transactions on Dependable and Secure Computing* 2014. 11 (5) p. .

[Lunt et al. ()] *A Real-Time Intrusion-Detection Expert System (IDES)*, T F Lunt , A Tamaru , F Gillham . 1992. p. . (Technical Report)

[Kumar and Spafford ()] 'A Software Architecture to Support Misuse Intrusion Detection'. S Kumar , E H Spafford . *Computer Science Technical Report* 1995. p. . Purdue University

[Tseng et al. ()] 'A Specification-Based Intrusion Detection System forAODV'. C.-Y Tseng , P Balasubramanyam , C Ko , R Limprasittiporn , J Rowe , K Levitt . *Proceedings of the 1st ACM Workshop on Security of Ad-hoc and Sensor Networks*, (the 1st ACM Workshop on Security of Ad-hoc and Sensor Networks) 2003. p. .

[Reddy et al. ()] 'A Study of Intrusion Detection in Data Mining'. E K Reddy , M Iaeng , V Reddy , Rajulu . *World Congress on Engineering* 2011. p. .

[Van Der Veen et al. ()] 'A Tough Call: Mitigating Advanced Code-Reuse Attacks at the Binary Level'. V Van Der Veen , E Göktas , M Contag , A Pawlowski , X Chen , S Rawat , H Bos , T Holz , E Athanasopoulos , C Giuffrida . *IEEE Symposium on Security and Privacy*, 2016. p. .

[Smola and Schölkopf ()] 'A Tutorial on Support Vector Regression'. A J Smola , B Schölkopf . *Statistics and Computing* 2004. 14 (3) p. .

[Zhang et al.] *Academic Work Submitted to School of Computer Science at Research Showcase @ CMU*, Y Zhang , S Egelman , L Cranor , J Hong . (Phinding Phish: Evaluating Anti-Phishing Tools)

[Barbara et al. ()] 'ADAM: Detecting Intrusions by Data Mining'. D Barbara , J Couto , S Jajodia , L Popyack , N Wu . *Proceedings of the IEEE Workshop on Information Assurance and Security*, (the IEEE Workshop on Information Assurance and Security) 2001. p. .

[Mitchell and Chen ()] 'Adaptive Intrusion Detection of Malicious Unmanned Air Vehicles Using Behavior Rule Specifications'. R Mitchell , Chen . *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 2014. 44 (5) p. .

[Lee et al. ()] 'Adaptive Intrusion Detection: A Data Mining Approach'. W Lee , S J Stolfo , K W Mok . *Artificial Intelligence Review* 2000. 14 (6) p. .

[Bhatkar et al. ()] 'Address Obfuscation: An Efficient Approach to Combat a Broad Range of Memory Error Exploits'. S Bhatkar , D C Duvarney , R Sekar . *Usenix Security* 2003. 3 p. .

[Xu and Wang ()] ‘An Adaptive Network Intrusion Detection Method Based on PCA and Support Vector Machines’. X Xu , X Wang . *International Conference on Advanced Data Mining and Applications*, 2005. p. .

[Moore and Clayton ()] *An Empirical Analysis of the Current State of Phishing Attack and Defence*, T Moore , R Clayton . 2007. (Academic work)

[Jackson et al. ()] ‘An Evaluation of Extended Validation and Picturein-Picture Phishing attacks’. C Jackson , D R Simon , D S Tan , A Barth . *International Conference on Financial Cryptography and Data Security*, 2007. p. .

[Soni and Sharma ()] ‘An Intrusion Detection System Based on Data Using Data Mining Techniques and Feature Selection’. P Soni , P Sharma . *International Journal of Soft Computing and Engineering (IJSCE)* 2014. 4 p. .

[Zang and Bolot ()] ‘Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study’. H Zang , J Bolot . *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, (the 17th Annual International Conference on Mobile Computing and Networking) 2011. p. .

[Dwork et al. ()] ‘Calibrating Noise to Sensitivity in Private Data Analysis’. C Dwork , F Mcsherry , K Nissim , A Smith . *Theory of Cryptography Conference*, 2006. p. .

[N P Nethravathi et al. ()] ‘CBTS: Correlation Based Transformation Strategy for Privacy Preserving Data Mining’. N P Nethravathi , G Prashanth , Deepa Rao , Shenoy , K R Venugopal , M Indramma . *2015 IEEE International WIE Conference on Electrical and Computer Engineering*, (Dhaka, Bangladesh) 2015. p. . WIECON-ECE

[Musuvathi et al. ()] ‘CMC: A Pragmatic Approach to Model Checking Real Code’. M Musuvathi , D Y Park , A Chou , D R Engler , D L Dill . *ACM SIGOPS Operating Systems Review* 2002. 36 (5) p. .

[Farid et al. ()] *Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection*, D M Farid , N Harbi , M Z Rahman . arXiv:1005.4496. 2010. (arXiv preprint)

[Zhu et al. ()] ‘Correlated Differential Privacy: Hiding Information in Non-IID Data Set’. T Zhu , P Xiong , G Li , W Zhou . *IEEE Transactions on Information Forensics and Security* 2015. 10 (2) p. .

[Pinkas ()] ‘Cryptographic Techniques for Privacy-Preserving Data Mining’. B Pinkas . *ACM SIGKDD Explorations Newsletter* 2002. 4 (2) p. .

[Stinson ()] *Cryptography: Theory and Practice 3rd Edition*, D R Stinson . 2006. Text Book.

[Bloedorn et al. ()] *Data Mining for Network Intrusion Detection: How to Get Started*, E Bloedorn , A D Christiansen , W Hill , C Skorupka , L M Talbot , J Tivel . 2001. p. . (MITRE)

[Khan et al. ()] ‘Data Mining for Security Purpose & its Solitude Suggestions’. S Khan , A Sharma , A S Zamani , A Akhtar . *International Journal of Technology Enhancements and Emerging Engineering Research* 2012. 1 (7) p. .

[Schultz et al. ()] ‘Data Mining Methods for Detection of New Malicious Executables’. M G Schultz , E Eskin , F Zadok , S J Stolfo . *Proceedings of IEEE Symposium on Security and Privacy S&P*, (IEEE Symposium on Security and Privacy S&P) 2001. p. .

[Shetty and Shekokar ()] ‘Data Mining Techniques for Real Time Intrusion Detection Systems’. M Shetty , N Shekokar . *International Journal of Scientific & Engineering Research* 2012. 3 (4) p. .

[Friedman and Schuster ()] ‘Data Mining with Differential Privacy’. A Friedman , A Schuster . *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining) 2010. p. .

[Taylor ()] *Data Mining with Semantic Features Represented as Vectors of Semantic Clusters*, M Taylor . 2012. Springer-Verlag. p. .

[Bayardo ()] ‘Data Privacy Through Optimal k-Anonymization’. R J Bayardo , R . *21st International Conference on Data Engineering (ICDE’05)*, 2005. p. .

[Narayanan and Shmatikov ()] ‘De-Anonymizing Social Networks’. A Narayanan , V Shmatikov . *IEEE Symposium on Security and Privacy*, 2009. p. .

[Jacobson et al. ()] ‘Detecting Code Reuse Attacks with a Model of Conformant Program Execution’. E R Jacobson , A R Bernat , W R Williams , B P Miller . *International Symposium on Engineering Secure Software and Systems*, 2014. p. .

[Siddiqui et al. ()] ‘Detecting Internet Worms Using Data Mining Techniques’. M Siddiqui , M C Wang , J Lee . *Journal of Systemics, Cybernetics and Informatics* 2009. 6 (6) p. .

[Lu and Traore ()] ‘Detecting New Forms of Network Intrusion Using Genetic Programming’. W Lu , I Traore . *Computational Intelligence* 2004. 20 (3) p. .

[Barbara et al. ()] 'Detecting Novel Network Intrusions using Bayes Estimators'. D Barbara , N Wu , S Jajodia . *SDM*, 2001. p. .

[Fu et al. ()] 'Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (emd)'. A Y Fu , L Wenyin , X Deng . *IEEE Transactions on Dependable and Secure Computing* 2006. 3 (4) p. .

[Asha S Manek et al. ()] 'Detection of Fraudulent and Malicious Websites by Analysing User Reviews for Online Shopping Websites'. Deepa Asha S Manek , Chandra Shenoy , Mohan , K R Venugopal . *International Journal of Knowledge and Web Intelligence* 2016. 5 (3) p. .

[Tanachaiwiwat and Hwang ()] 'Differential Packet Filtering Against DDos Flood Attacks'. S Tanachaiwiwat , K Hwang . *ACM Conference on Computer and Communications Security (CCS)*, 2003. p. .

[Mcsherry and Mironov ()] 'Differentially Private Recommender Systems: Building Privacy Into the Net'. F Mcsherry , I Mironov . *Proceedings of the 15 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (the 15 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining) 2009. p. .

[Habibi et al. ()] 'DISARM: Mitigating Buffer Overflow Attacks on Embedded Devices'. J Habibi , A Panicker , A Gupta , E Bertino . *International Conference on Network and System Security*, 2015. p. .

[Zhang et al. ()] 'Distributed Data Mining with Differential Privacy'. N Zhang , M Li , W Lou . *IEEE International Conference on Communications*, 2011. p. .

[Chen et al. ()] 'Drop:Detecting Return-Oriented Programming Malicious Code'. P Chen , H Xiao , X Shen , X Yin , B Mao , L Xie . *International Conference on Information Systems Security*, 2009. p. .

[P Deepa Shenoy et al. ()] 'Dynamic Association Rule Mining using Genetic Algorithms'. P Deepa Shenoy , K G Srinivasa , K R Venugopal , L M Patnaik . *Intelligent Data Analysis* 2005. 9 (5) p. .

[Davi et al. ()] 'Dynamic Integrity Measurement and Attestation: Towards Defense Against Return-Oriented Programming Attacks'. L Davi , A.-R Sadeghi , M Winandy . *Proceedings of the 2009 ACM Workshop on Scalable Trusted Computing*, (the 2009 ACM Workshop on Scalable Trusted Computing) 2009. p. .

[Mitchell and Chen ()] 'Effect of Intrusion Detection and Response on Reliability of Cyber Physical Systems'. R Mitchell , R Chen . *IEEE Transactions on Reliability* 2013. 62 (1) p. .

[Qin and Hwang ()] 'Effectively Generating Frequent Episode Rules for Anomaly-based Intrusion Detection'. M Qin , K Hwang . *IEEE Symposium on Security and Privacy*, 2003.

[Diwate and Sahu ()] 'Efficient Data Mining in SAMS through Association Rule'. M R B Diwate , A Sahu . *International Journal of Electronics Communication and Computer Engineering* 2014. 5 (3) p. .

[Bhatkar et al. ()] 'Efficient Techniques for Comprehensive Protection from Memory Error Exploits'. S Bhatkar , D C Duvarney , R Sekar . *Proceedings of the 14th USENIX Security Symposium*, (the 14th USENIX Security Symposium) 2005.

[Lu et al. ()] 'EPPA: An Efficient and Privacy-Preserving Aggregation Scheme for Secure Smart Grid Communications'. R Lu , X Liang , X Li , X Lin , X Shen . *IEEE Transactions on Parallel and Distributed Systems* 2012. 23 (9) p. .

[P Deepa Shenoy et al. ()] 'Evolutionary Approach for Mining Association Rules on Dynamic Databases'. P Deepa Shenoy , K G Srinivasa , K R Venugopal , L M Patnaik . *7th Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD)*, (Seoul, South Korea) 2003. 2003. p. .

[Parmar et al. ()] 'Exception-Oriented Programming: Retrofitting Code-Reuse Attacks to Construct Kernel Malware'. G Parmar , Dr. Kirti Mathur , ; L Deng , Q Zeng . *The Institution of Engineering and Technology* 2015. 2016. 5 (5) p. . (Indian Journal of Applied Research)

[Ben-David et al. ()] 'FairplayMP: A System for Secure Multi-Party Computation'. A Ben-David , N Nisan , B Pinkas . *Proceedings of the 15th ACM Conference on Computer and Communications Security*, (the 15th ACM Conference on Computer and Communications Security) 2008. p. .

[Hussein et al. ()] 'Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database'. M Hussein , A El-Sisi , N Ismail . *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 2008. p. .

[Dubiner et al. ()] 'Faster Tree Pattern Matching'. M Dubiner , Z Galil , E Magen . *Journal of the ACM (JACM)* 1994. 41 (2) p. .

[Onarlioglu et al. ()] 'G-Free: Defeating Return-Oriented Programming Through Gadget-Less Binaries'. K Onarlioglu , L Bilge , A Lanzi , D Balzarotti , E Kirda . *Proceedings of the 26th Annual Computer Security Applications Conference*, (the 26th Annual Computer Security Applications Conference) 2010. p. .

[Polychronakis ()] *Generic Detection of Code Injection Attacks using Network-Level Emulation*, M Polychronakis . 2009. (Ph.D. Thesis)

[Desale and Ade ()] 'Genetic Algorithm Based Feature Selection Approach for Effective Intrusion Detection System'. K S Desale , R Ade . *International Conference on Computer Communication and Informatics (ICCCI)*, 2015. p. .

[Mohan et al. ()] 'GUPT: Privacy Preserving Data Analysis made Easy'. P Mohan , A Thakurta , E Shi , D Song , D Culler . *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, (the 2012 ACM SIGMOD International Conference on Management of Data) 2012. p. .

[Frey et al. ()] 'Hide and Share: Landmark-Based Similarity for Private KNN Computation'. D Frey , R Guerraoui , A Kermarrec , A Rault , Franc¸ois Ta¨?ani , J Wang . *Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, (the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks) 2015. p. .

[Yeung and Ding ()] 'Host-Based Intrusion Detection Using Dynamic and Static Behavioral Models'. D.-Y Yeung , Y Ding . *Pattern Recognition* 2003. 36 (1) p. .

[Mitropoulos et al. ()] 'How to Train Your Browser: Preventing XSS Attacks Using Contextual Script Fingerprints'. D Mitropoulos , K Stroggylos , D Spinellis . *ACM Transactions on Privacy and Security* 2016. 19 (1) p. .

[Patel and Bhatt ()] 'Implementnig Data Mining for Detection of Malware from Code'. D K B Patel , S H Bhatt . *An International Journal of Advanced Computer Technology:Compusoft* 2014. 3 (4) p. .

[Improved Kernel Security through Code Validation, Diversification, and Minimization ()] *Improved Kernel Security through Code Validation, Diversification, and Minimization*, 2013. (Ph.D. Thesis)

[Mohammad et al. ()] 'Intelligent Rule-Based Phishing Websites Classification'. R M Mohammad , F Thabtah , L Mccluskey . *IET Information Security* 2014. 8 (3) p. .

[Hu et al. ()] *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, K Hu , Y Lu , L Zhou , C Shi . 1999. p. . (Integrating Classification and Association Rule Mining: A Concept Lattice Framework)

[Tsai et al. ()] 'Intrusion Detection by Machine Learning: A Review'. C.-F Tsai , Y.-F Hsu , C.-Y. Lin , W.-Y Lin . *Expert Systems with Applications* 2009. 36 (10) p. 0.

[Jabez and Muthukumar ()] 'Intrusion Detection System: Time Probability Method and Hyperbolic Hopfield Neural Network'. J Jabez , B Muthukumar . *Journal of Theoretical & Applied Information Technology* 2014. 67 (1) p. .

[Abadeh et al. ()] 'Intrusion Detection using a Fuzzy Genetics-Based Learning Algorithm'. M S Abadeh , J Habibi , C Lucas . *Journal of Network and Computer Applications* 2007. 30 (1) p. .

[Bletsch et al. ()] 'Jump-Oriented Programming: A New Class of Code-Reuse Attack'. T Bletsch , X Jiang , V W Freeh , Z Liang . *Proceedings ACM Symposium on Information, Computer and Communications Security*, (ACM Symposium on Information, Computer and Communications Security) 2011. p. .

[Kevin et al. ()] 'Just-In-Time Code Reuse: On the Effectiveness of Fine-Grained Address Space L]ayout Randomization'. S Kevin , Z , F Monrose , D Fabian , D Lucas , L Alexandra , S Christopher , R Ahmad . *2013 IEEE Symposium on Security and Privacy*, 2013. p. .

[Kumar et al. ()] 'K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset'. V Kumar , H Chauhan , D Panwar . *International Journal of Soft Computing and Engineering (IJSCE)* 2013. 3 (4) p. .

[Bk et al. ()] 'K-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data'. Bk , Y Samanthula , W Elmehdwi , Jiang . *IEEE Transactions on Knowledge and Data Engineering* 2015. 27 (5) p. .

[Rizvi and Haritsa ()] 'Maintaining Data Privacy in Association Rule Mining'. S J Rizvi , J R Haritsa . *Proceedings of the 28th International Conference on Very Large Data Bases*, (the 28th International Conference on Very Large Data Bases) 2002. p. .

[Gupta et al. ()] 'MARLIN: A Fine Grained Randomization Approach to Defend Against ROP Attacks'. A Gupta , S Kerr , M S Kirkpatrick , E Bertino . *International Conference on Network and System Security*, 2013. p. .

[Venugopal and Buyya ()] 'Mastering C++'. K R Venugopal , Rajkumar Buyya . *Tata McGraw-Hill Education* 2013.

[Thabtah et al. ()] 'MCAR: Multiclass Classification based on Association Rule'. F Thabtah , P Cowling , Y Peng . *The 3rd ACS/IEEE International Conference on Computer Systems and Applications*, 2005. p. .

[Lee et al. ()] 'Mining Audit Data to Build Intrusion Detection Models'. W Lee , S J Stolfo , K W Mok . *KDD-98 Proceedings*, 1998. p. .

[Al-Nashif et al. ()] 'Multi-Level Intrusion Detection System'. Y Al-Nashif , A A Kumar , S Hariri , Y Luo , F Szidarovsky , G Qu . *International Conference on Autonomic Computing ICAC'08*, 2008. p. .

[Hwang et al. ()] 'Netshield: Protocol Anomaly Detection with Datamining against DDOS Attacks'. K Hwang , P Dave , S Tanachaiwiwat . *Proceedings of the 6th International Symposium on Recent Advances in Intrusion Detection*, (the 6th International Symposium on Recent Advances in Intrusion DetectionPittsburgh, PA) 2003. p. .

[Mukherjee et al. ()] 'Network Intrusion Detection'. B Mukherjee , L T Heberlein , K N Levitt . *IEEE Network* 1994. 8 (3) p. .

[Abadeh et al. ()] 'On The Combination of Genetic Fuzzy Systems and Pairwise Learning for Improving Detection Rates on Intrusion Detection Systems'. S Abadeh , A Fernandez , A Bawakid , S Alshomrani , F Herrera . *Journal of Expert Systems with Applications* 2015. 42 (1) p. .

[Chen and Guo ()] 'Online Detection and Prevention of Phishing Attacks'. J Chen , C Guo . *First International Conference on Communications and Networking in China*, 2006. p. .

[Vasanthakumar et al. ()] 'PFU: Profiling Forum Users in Online Social Networks, A Knowledge Driven Data Mining Approach'. G U Vasanthakumar , Deepa Shenoy , K R Venugopal , L M Patnaik . *2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECONECE)*, 2015. p. .

[Abdelhamid et al. ()] 'Phishing Detection Based Associative Classification Data Mining'. N Abdelhamid , A Ayesh , F Thabtah . *Expert Systems with Applications* 2014. 41 (13) p. .

[Marchal et al. ()] 'PhishStorm:Detecting Phishing with Streaming Analytics'. S Marchal , J Franc¸ois , R State , T Engel . *IEEE Transactions on Network and Service Management* 2014. 11 (4) p. .

[Vasanthakumar et al. ()] 'PIB: Profiling Influential Blogger in Online Social Networks, A Knowledge Driven Data Mining Approach'. G U Vasanthakumar , Bagul Prajakta , Deepa Shenoy , K R Venugopal , L M Patnaik . *11 th International Multi-Conference on Information Processing (IMCIP)*, 2015. 54 p. .

[Aburrous et al. ()] 'Predicting Phishing Websites using Classification Mining Techniques with experimental Case Studies'. M Aburrous , M A Hossain , K Dahal , F Thabtah . *Seventh International Conference on Information Technology: New Generations (ITNG)*, 2010. p. .

[Mohanappriya ()] *Prediction and Pan Code Reuse Attack by Code Randomization Mechanism and Data Corruption*, N Mohanappriya , R . 2016. p. . (Techniques and Algorithms in Emerging Technologies)

[Vaidya and Clifton ()] 'Privacy Preserving Association Rule Mining in Vertically Partitioned Data'. J Vaidya , C Clifton . *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining) 2002. p. .

[Kiayias et al. ()] 'Privacy preserving Data Mining within Anonymous Credential Systems'. A Kiayias , S Xu , M Yung . *International Conference on Security and Cryptography for Networks*, 2008. p. .

[Dunning and Kresman ()] 'Privacy Preserving Data Sharing with Anonymous ID Assignment'. L A Dunning , R Kresman . *IEEE Transactions on Information Forensics and Security* 2013. 8 (2) p. .

[Zhan et al. ()] 'Privacy-Preserving Collaborative Association Rule Mining'. J Zhan , S Matwin , L Chang . *IFIP Annual Conference on Data and Applications Security and Privacy*, 2005. p. .

[Zhan ()] 'Privacy-Preserving Collaborative Data Mining'. J Zhan . *IEEE Computational Intelligence Magazine* 2008. 3 (2) p. .

[Yj and Lee ()] 'Privacy-preserving Data Mining for Personalized Marketing'. Yj , Lee . *International Journal of Computer Communications and Networks (IJCCN)* 2014. 4 (1) p. .

[Ouda et al. ()] 'Privacy-Preserving Data Mining in Homogeneous Collaborative Clustering'. M Ouda , S Salem , I Ali , E.-S Saad . *International Arab Journal of Information Technology (IAJIT)* 2015. 12 (6) p. .

[Vaidya and Clifton ()] 'Privacy-Preserving Data Mining: Why, How, and When'. J Vaidya , C Clifton . *IEEE Security & Privacy* 2004. 2 (6) p. .

[Kantarcioglu and Clifton ()] 'Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data'. M Kantarcioglu , C Clifton . *IEEE Transactions on Knowledge and Data Engineering* 2004. 16 (9) p. .

[Giannotti et al. ()] 'Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases'. F Giannotti , L V Lakshmanan , A Monreale , D Pedreschi , H Wang . *IEEE Systems Journal* 2013. 7 (3) p. .

[Roughan and Zhang ()] 'Privacy-Preserving Performance Measurements'. M Roughan , Y Zhang . *Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data*, (the 2006 SIGCOMM Workshop on Mining Network Data) 2006. p. .

[Yao ()] 'Protocols for Secure Computations'. A C Yao . *23rd Annual Symposium on Foundations of Computer Science, 1982. SFCS'08*, 1982. p. .

[Friedman et al. ()] 'Providing k-Anonymity in Data Mining'. A Friedman , R Wolff , A Schuster . *The VLDB Journal* 2008. 17 (4) p. .

[Barrantes et al. ()] 'Randomized Instruction Set Emulation to Disrupt Binary Code Injection Attacks'. E G Barrantes , D H Ackley , T S Palmer , D Stefanovic , D D Zovi . *Proceedings of the 10th ACM Conference on Computer and Communications Security*, (the 10th ACM Conference on Computer and Communications Security) 2003. p. .

[Sadeghi et al. ()] 'Return-Oriented Programming Without Returns'. H Sadeghi , M Shacham , Winandy . *Proceedings of the 17th ACM Conference on Computer and Communications Security*, (the 17th ACM Conference on Computer and Communications Security) 2010. p. .

[Davi et al. ()] *Return-Oriented Programming Without Returns on ARM*, L Davi , A Dmitrienko , A.-R Sadeghi , M Winandy . HGI-TR-2010- 002. 2010. (Technical Report)

[Roemer et al. ()] 'Return-Oriented Programming: Systems, languages, and applications'. R Roemer , E Buchanan , H Shacham , S Savage . *ACM Transactions on Information and System Security (TISSEC)* 2012. 15 (1) p. .

[Yeh et al. ()] 'Robust Laser Speckle Authentication System through Data Mining Techniques'. C.-H Yeh , G Lee , C.-Y. Lin . *IEEE Transactions on Industrial Informatics* 2015. 11 (2) p. .

[Davi et al. ()] 'ROP Defender: A Detection Tool to Defend Against Return-Oriented Programming Attacks'. L Davi , A.-R Sadeghi , M Winandy . *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, (the 6th ACM Symposium on Information, Computer and Communications Security) 2011. p. .

[Graziano et al. ()] 'ROPMEMU: A Framework for the Analysis of Complex Code-Reuse Attacks'. Mariano Graziano , Davide Balzarotti , Alain Zidouemba . *11th ACM Asia Conference on Computer and Communications Security*, 2016. p. .

[Follner and Bodden ()] 'ROPocop -Dynamic Mitigation of Code-Reuse Attacks'. E Follner , Bodden . *Secure Software Engineering Group* 2015. 29 (3) p. .

[Zhuang et al. ()] 'Runtime Code Reuse Attacks: A Dynamic Framework Bypassing Fine-Grained Address Space Layout Randomization'. Y Zhuang , T Zheng , Z Lin . *SEKE* 2014. p. .

[Roughan and Zhang ()] 'Secure Distributed Data Mining and Its Application to Large Scale Network Measurements'. M Roughan , Y Zhang . *ACM SIGCOMM Computer Communication Review* 2006. 36 (1) p. .

[Vaidya and Clifton ()] 'Secure Set Intersection Cardinality with Application to Association Rule Mining'. J Vaidya , C Clifton . *Journal of Computer Security* 2005. 13 (4) p. .

[Mohammed et al. ()] 'Secure Two-Party Differentially Private Data Release for Vertically Partitioned Data'. N Mohammed , D Alhadidi , B Fung , M Debbabi . *IEEE Transactions on Dependable and Secure Computing* 2014. 11 (1) p. .

[Kayaalp et al. ()] 'Signature-Based Protection from Code Reuse Attacks'. M Kayaalp , T Schmitt , J Nomani , D Ponomarev , N A Ghazaleh . *IEEE Transactions on Computers* 2015. 64 (2) p. .

[Shapiro ()] 'Situating Anonymization within a Privacy Risk Model'. S S Shapiro . *2012 IEEE International Systems Conference(SysCon)*, 2012. p. .

[G¨oktas¸ et al. ()] 'Size Does Matter:Why Using Gadget-Chain Length to Prevent Code-Reuse Attacks is Hard'. E G¨oktas¸ , E Athanasopoulos , M Polychronakis , H Bos , G Portokalidis . *USENIX Security Symposium*, 2014. p. .

[Wu Bin et al. ()] 'Smartphone Malware Detection Model Based on Artificial Immune System'. Wu Bin , Lu Tianliang , Zheng Zheng Kangfeng , Lin Dongmei , Xing . *China Communications* 2014. 11 (13) p. .

[Wu et al. ()] 'Smartphone Malware Detection Model Based on Artificial Immune System'. B Wu , T Lu , K Zheng , D Zhang , X Lin . *China Communications* 2015. 11 (13) p. .

[Pappas et al. ()] 'Smashing the Gadgets: Hindering Return-Oriented programming using In-Place Code Randomization'. V Pappas , M Polychronakis , A D Keromytis . *2012 IEEE Symposium on Security and Privacy*, 2012. p. .

[Roesch ()] 'SNORT: Lightweight Intrusion Detection for Networks'. M Roesch . *Proceedings of LISA '99: 13th Systems Administration Conference*, (LISA '99: 13th Systems Administration Conference) 1999. p. .

[Venugopal et al. ()] *Soft Computing for Data Mining Applications*, K R Venugopal , K Srinivasa , L M Patnaik . 2009. Springer.

[Bui et al. ()] *Spatial Prediction Models for Shallow Landslide Hazards: A Comparative Assessment of the Efficacy of Support Vector Machines, Artificial Neural Networks, Kernel Logistic Regression, and Logistic Model Tree*, D Bui , T Tuan , H Klempe , B Pradhan , I Revhaug . 2015. Berlin Heidelberg: springer-Verlag. 13 p. .

[Allen et al. ()] *State of the Practice of Intrusion Detection Technologies*, J Allen , A Christie , W Fithen , J Mchugh , J Pickel . 2000. p. . (Technical Report)

[Ilgun et al. ()] 'State Transition Analysis: A Rule-Based Intrusion Detection Approach'. K Ilgun , R A Kemmerer , P A Porras . *IEEE Transactions on Software Engineering* 1995. 21 (3) p. .

[Roglia et al. ()] 'Surgically Returning to Randomized Lib (C)'. G F Roglia , L Martignoni , R Paleari , D Bruschi . *Computer Security Applications Conference*, 2009. p. .

[Chourse and Richhariya ()] 'Survey Paper on Intrusion Detection Using Data Mining Techniques'. S Chourse , V Richhariya . *International Journal of Emerging Technology and Advanced Engineering, ISO* 2008. 4 (8) p. .

[Joachims ()] 'Text Categorization with Support Vector Machines: Learning With Many Relevant Features'. T Joachims . *European Conference on Machine Learning*, 1998. p. .

[Heady et al.] *The Architecture of a Network Level Intrusion Detection System*, R Heady , G F Luger , A Maccabe , M Servilla . University of New Mexico (Academic Work submitted to the)

[Vittapu et al. ()] 'The Practical Data Mining Model for Efficient IDS Through Relational Databases'. M S Vittapu , V Sunkari , A Y Abate . *International Journal of Research in Engineering and Science* 2015. 3 (1) p. .

[Beaver et al. ()] 'The Round Complexity of Secure Protocols'. D Beaver , S Micali , P Rogaway . *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, (the 22nd Annual ACM Symposium on Theory of Computing) 1990. p. .

[Hong ()] 'The State of Phishing Attacks'. J Hong . *Communications of the ACM* 2012. 55 (1) p. .

[Wu et al. ()] 'Top 10 Algorithms in Data Mining'. V Wu , J R Kumar , J Quinlan , Q Ghosh , H Yang , G J Motoda , A Mclachlan , B Ng , S Y Liu , Philip . *Knowledge and Information Systems* 2008. 14 (1) p. .

[Herzberg and Gbara ()] *Trustbar: Protecting (even naive) Web Users from Spoofing and Phishing Attacks*, A Herzberg , A Gbara . 2004/155. `http://eprint.iacr.org/` 2004/155, 2004. (Cryptology ePrint Archive Report)

[Du and Zhan ()] 'Using Randomized Response Techniques for Privacy-Preserving Data Mining'. W Du , Z Zhan . *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining) 2003. p. .

[Buchanan et al. ()] 'When Good Instructions Go Bad: Generalizing Return-Oriented Programming to RISC'. E Buchanan , R Roemer , H Shacham , S Savage . *Proceedings of the 15th ACM Conference on Computer and Communi-cations Security*, (the 15th ACM Conference on Computer and Communi-cations Security) 2008. p. .

[Buchanan et al. ()] 'When Good Instructions Go Bad: Generalizing Return-Oriented Programming to RISC'. E Buchanan , R Roemer , H Shacham , S Savage . *Proceedings of the 15th ACM Conference on Computer and Communications Security*, (the 15th ACM Conference on Computer and Communications Security) 2008. p. .

[Gupta and Gupta ()] *XSS-SAFE:A Server-Side Approach to Detect and Mitigate Cross-Site Scripting (XSS) Attacks in JavaScript Code*, S Gupta , B B Gupta . 2015. Springer. 4 p. .

[Yao et al. ()] F Yao , J Chen , G Venkataramani . *International Conference on Computer Design*, 2013. p. .