Douglas Cunningham, Petra Hofstedt, Klaus Meer, Ingo Schmitt (Hrsg.): INFORMATIK 2015
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2015

A recommended framework for anomaly intrusion detection system (IDS)

Tho Le[1]

Abstract: Signature-based solutions for Intrusion Detection are dominant in practice despite of its incapability to detect zero-day attacks. Moreover, anomaly-based Intrusion Detection Systems (IDS), a promising approach against both known and unknown attacks, are not mature for a broad productive use. Therefore, the further development of anomaly based IDS is an imperative task to strengthen security in todays networked infrastructure. This motivates a detailed study to give a structured view of problems and challenges and of the current state in this field. For this purpose, a sound analysis of current limitations and a very comprehensive survey of research papers have been conducted. In this article, a short summary of the results of the survey is given. Furthermore, the survey led to important insights into future research efforts and a proposal for a promising future IDS architecture, which is presented in this work.

Keywords: Anomaly-based IDS, survey, problems and challenges, architecture.

1 Introduction

Intrusion Detection Systems (IDS) are an urgent requirement for network security. Most of today's IDS are based on signature detection (also known as *misuse detection*). They are easy to use and have a low false positive rate. However, there is a concerning disadvantage: they are not able to detect unknown attacks like zero-day exploits. Zero-day exploits are attacks that are not publicly known and because of that, no security patches are released and no security vendor has developed signatures to detect these security issues [SY14]. Therefore, zero-day exploits pose serious threats to organizations. In [SY14], the security company *Symantec* reports that in the year 2013, 23 new zero-day vulnerabilities (out of 6787 new vulnerabilities in total) are detected. It is an increase of more than 60% in comparison to 2012 (14 zero-days), which is the highest number of zero-days since 2006.

Besides signature-based IDS, anomaly-based systems are also developed. The basic idea is that attacks are detectable by abnormal behaviors. If the anomaly-based system observed the relevant system properties, it is able to detect unknown zero-day attacks. However, there are also disadvantages in these systems: anomaly-based systems have to learn the differences between normal and abnormal behaviors. Moreover, they have a high false positive rate. These drawbacks could be explanations of why anomaly-based detection systems are not widely applied.

[1] Heilbronn, Informatik, Max-Planck-Str. 39, 74081 Heilbronn, thole020287@gmail.com

In this paper, an overview about problems and challenges as well as trends of anomaly-based detection systems is given by analyzing more than 160 studies [Le14] published in the last 3 years. Based on these trends as well as the discussed problems and challenges of anomaly-based detection, the architecture for a novel anomaly detection system is proposed [Le14].

The remainder of this paper is organized as follows: Section 2 gives an overview of today's problems and challenges of anomaly detection systems. Section 3 presents statistic information about the state of the art systems. Finally, a novel architecture for anomaly detection systems is suggested in section 4 before the conclusion in section 5.

2 Problems and challenges

Beginning in 1987, anomaly-based detection has been capturing a number of research efforts in order to develop a more sophisticated detector in protecting computing systems from compromises, especially zero-day attacks [BBK14] [BBK11] [JPP11] [GT06] [Ga08] [TSG10]. However, after more than 27 years of research efforts, such kind of systems is still missing in practical deployment. Therefore, understanding the current drawbacks is a critical and imperative task for the development in this area. Although many authors have identified ahead problems [BBK14] [BBK11] [GT06] [Ga08] [PP07] [RRR08] [TSG10] [VTN13] [SP10] [Ko11], they individually gave a piece of the puzzle, hence the work is to collect, analyze and give a more advanced and broader picture of current shortcomings. From that, the direction for further development is shown. There are four identified domains of problems existing in anomaly-based detection technique, namely hypothesis, implementation, evaluation and operation.

2.1 Hypothesis

- **Definition of normality**: is one of the key factors influencing the performance of an anomaly-based detection system. This is proven to be a complicated task since normal and malicious activities are sometimes close to each other. Furthermore, expected behaviors can be various depending on a concrete situation, i.e. local site policies of target environments [GT06] [SP10]. For example, an event is considered as normal in one environment, but may be treated as attacks in others due to their security policies.
- **Autonomous activities and intrusions:** one of the principal pitfalls is fallen in its initial premise about the interrelationship between anomalous activities and intrusions. On February 1987, Dorothy Denning originally introduced a new intrusion detection model in her paper [De87]: "security violations could be detected from abnormal patterns of system usage". This statement implicitly indicates three assumptions [GT06]: Attacks are anomalous, attacks are rare and anomalous activities are malicious. However, Gate and Taylor [GT06] questioned

and critically examined those assumptions in context of network environment to prove inappropriateness.

2.2 Implementation

- **Paradigm**: anomaly-based detection system was originally introduced to protect at single host and employed automatically to networking level, therefore, most of existing anomaly-based IDSs have not adapted properly to the modern networking paradigm [PP07] [BBK11] for both wired and wireless communication.
- **Challenges of using machine learning:** machine learning has been successfully deployed in various areas, such as voice recognition, email spam detection and it is one of the most favorable techniques in deploying anomaly-based detectors along with data mining. However, the similar success has not been found in the IDS application. Sommer and Paxson [SP10] analyzed specific characteristics in this application to find fundamental challenges in employing machine learning techniques for anomaly detection: novel attack detection and high cost of errors.

2.3 Evaluation

- **Lack of evaluation data**: lack of benchmark datasets, which can simulate realistic host and network environments, is a major issue that researchers are facing with when assessing their anomaly-based detectors [BBK14] [BBK11] [TSG10] [PP07] [SP10]. While the majority of studies were still evaluated on KDD99 dataset (published 1999) [KD14] as illustrated in section 3 and in [TSG10], it is out of date and no longer valid to represent current environments, hence such assessment results are often not reliable.
- **Lack of evaluation procedure:** many researchers [TSG10] [PP07] [SP10] believed that the inadequate evaluation process is one of the root causes that hinder anomaly-based IDS from coming into business. To such kind of critical infrastructure systems, before deploying into a real environment, it is essential to well understand and reliably evaluate its operation. Unfortunately, there is currently neither a standard framework of assessment process nor agreed evaluated metrics. Consequently, the validity of experimental results is relatively low and this may partially explain the limited success of anomaly detection in operational environments.

2.4 Operation

- **Training data:** obtaining labeled or “normal” data in real environment for anomaly training process is complicated and time consuming. Since supervised and semi-supervised algorithms require labeled and clean data respectively to learn and build behavior profiles of targeted environments, it is critical to prepare qualified training data for optimal performance. Unfortunately, there is currently no proven process for automatically extracting those required features in a real environment. Gate and Taylor [GT06] even said, “The assumption that there exists attack-free data for training a detector outside of simulated data is not a realistic assumption”.

- **High false alarm rate:** is the agreed problem shared by most of researchers for the limited widespread deployment of anomaly-based intrusion detection in real environments. Although some systems have recently achieved relatively low false alarm rates such as [MRR13] with 0.1% false alarm rate and 99.6% of detection rate, the small proportion of a large number is still a big one as mentioned in [GT06]. Therefore, reducing false positive rate is properly the most critical and imperative task toward the wide deployment of anomaly-based IDS in business.
- **Adaptability:** is the core of an effective anomaly-based detection system [GT06] [IX13]. Since intruders constantly evolve malicious behaviors to evade IDS systems, anomaly-based detectors must be able to frequently re-train their behavior profiles and adapt to new threats as well as new situations without performance influence.
- **Real time operation:** is another challenge to anomaly-based IDS [BBK14]. With the rapid increase of computer networks and heavy applications running on top of it, an anomaly-based detector is expected to process a large amount of data in a timely manner.

3 The current state of art

A comprehensive survey has been conducted to provide an up-to-date view of the current pace in anomaly-based IDS field. The survey collects 169 studies published in the last three years (2012 to 2014), from five popular digital libraries, namely: Institute of Electrical and Electronics Engineers (IEEE), Springer, Association for Computing Machinery (ACM), Google Scholar and Science Direct. From that, some interesting statistical summaries are presented with a notice that clustering, neural networks and support vector machine methods are the most employed techniques in this field. Some remarkable works can be found at [Le12], [QMG12], [MRR13], [Ab14]. More details of the survey can be referenced in the master thesis [Le14].

The survey observes a dramatic decline in the number of researches as it can be seen clearly from the chart 3.1. In 2012, 86 studies related to anomaly-based IDS were carried out, making this number the highest in the last 3 years. In 2013, the number has declined to 51 and continued to decrease to 26 reports published in 2014. This linear drop over the last three years may be suggested by many reasons, one of which could imply for the reduction of interest in this area, especially in the situation that this technique is still considered as immature after 27 years of continuous development.

In general, an anomaly-based detector can be trained to build a normal behavior profile via three ways, namely supervised, semi-supervised, unsupervised. Both supervised and semi-supervised systems require labeled and clean training data to build normal profiles respectively. However, unsupervised approaches can be trained with raw data. Those unsupervised systems require minimum effort from security experts since they are able to learn acceptable behaviors and adapt themselves to behavior drifts from targeted environments. Therefore, it is anticipated to be the most widely adopted and this is supported by the survey's result with 37% of total research works as illustrated in the figure 3.2. With less than 2%, supervised anomaly detectors also capture a considerable amount of interest, as they tend to produce less false alarms while gaining higher detection accuracy than unsupervised ones. Moreover, the semi-supervised mode is responsible for the smallest contribution in the anomaly field. This can be explained by the facts that: 1) exposing such IDS systems to only normal traffic is also a challenge since it requires intensive scanning and screening to remove various kind of hidden attacks, 2) semi-supervised systems are less efficient than supervised ones in terms of detection accuracy and false alarm generation.

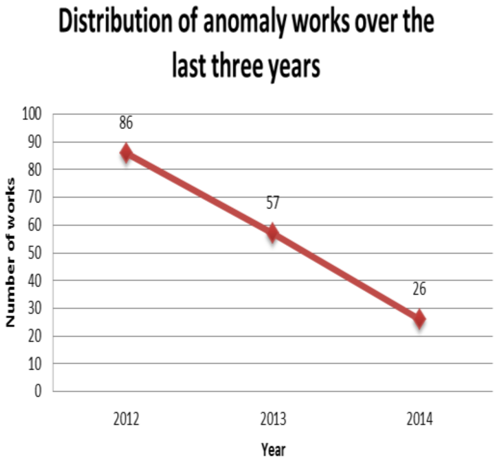

Fig. 3.1: Distribution of anomaly works [Le14]

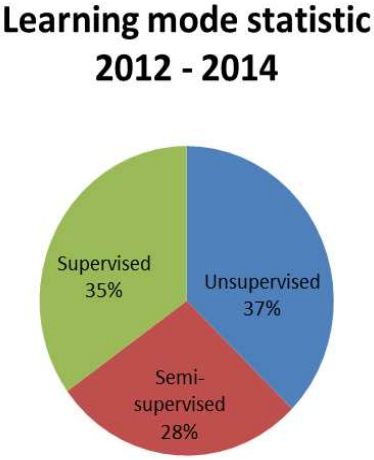

Fig. 3.2: Learning mode statistic [Le14]

Assessment is an integral part in developing an IDS tool to evaluate its reliability, efficiency and effectiveness before applying in real life scenarios. This phrase cannot be done without qualified evaluation datasets. Therefore, it would be necessary to survey datasets on which scientists are working to develop their detectors. As illustrated clearly in the figure 3.3, more than half of researches were experimented with DARPA families, including DARPA [LL14], KDD99 [KD14] and NSL-KDD [NS14], in which KDD99 is the most famous dataset, constituting 35% of total experiments. The other datasets are contributing much smaller proportions. While about a quarter of anomaly evaluations were performed with synthetic and real datasets individually, yet all other benchmarked datasets, such as Sendmail [CS14], Kyoto 2006+ [TA14] etc. are only accounting for 20% of total experiments conducted in the survey.

The pie chart 3.4 illustrates the proportion of under-lying algorithms contributing to the development of anomaly-based IDS during the three-year period. As expected, machine learning based and data mining based techniques are the main stream of favor, being responsible for more than 50% of total implementations. Slightly below those two, with 24%, the survey observes a new trend of combining multiple techniques within or across categories to improve detection performance called "combination". This new approach comes from the fact that each technique contains its own advantages and disadvantages. Therefore, a proper combination of those techniques could result in a better detection. Furthermore, statistical based algorithms are the least favorite technique used in during the last three years (21%). Noticeably there is no work applying knowledge-based techniques.

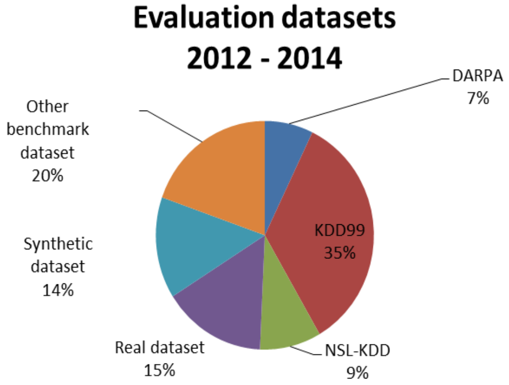

Fig. 3.3: Evaluation datasets [Le14]

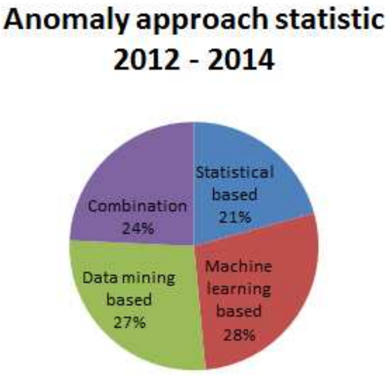

Fig. 3.4: Anomaly approach statistic [Le14]

4 The Recommended architecture

In order to provide meaningful recommendations, it is necessary to analyze carefully requirements on an anomaly IDS system as below:

- **Efficient detection performance**: detecting and alarming any intrusions are the primary functions of an IDS system, therefore, the first and most important requirement is to detect attacks as many as possible with a minimum number of false alarms. Furthermore, delay time for each decision must be small enough in the context of optimal computing resource usage.
- **Classification**: the conventional purpose of anomaly detection is to distinguish between normal and anomalous activities. However, in many cases, alerting anomalous events are not helpful enough to administrators, since they have to spend a considerable amount of effort and time on investigating types of attacks

before deciding appropriate responses. Therefore, it is more useful to alarm anomalous events along with their belonging categories, such as Probe, D.o.S etc.

- **Easy deployment**: this kind of systems should be straightforward to install and operate in targeted environments. Supervised and semi-supervised learning approaches require labeled and normal data for training process, which makes it complex to implement in business environments. Ideally, training phase in supervised and semi-supervised systems should be either bypassed or replaced by another process placing no stress on training data like in unsupervised approach.
- **Adaptability**: anomaly-based IDS must be able to adapt itself to changes of dynamic environments such as enterprise networks.
- **Real time operation**: It must be able to process a large volume of data in a timely manner.

Since signature and anomaly detections are two opposite directions that advantages of this are weaknesses of other, the idea of combining those two into a hybrid model seems to be very promising and that is also the cornerstone in the recommended architecture. Basically, there are three ways to combine signature and anomaly approaches together. Anomaly and signature models can run in parallel and two outcomes are merged or selected for a final decision, otherwise either anomaly or signature can be implemented on top of other. In this recommendation, the latter approach is chosen. The signature module is implemented in the first layer since it provides a better integration between the signature and anomaly modules. In this context, the signature module acts as a filter to remove all known attacks and leaving only normal and unknown attacks to the subsequent layer. Furthermore, the anomaly module inherits detection results from the upper module for the training process. Particularly, those detected known attacks along with the classification results from the unsupervised anomaly layer are used as labeled training data for a supervised algorithm. The general recommended architecture is clearly illustrated in the figure 4.1:

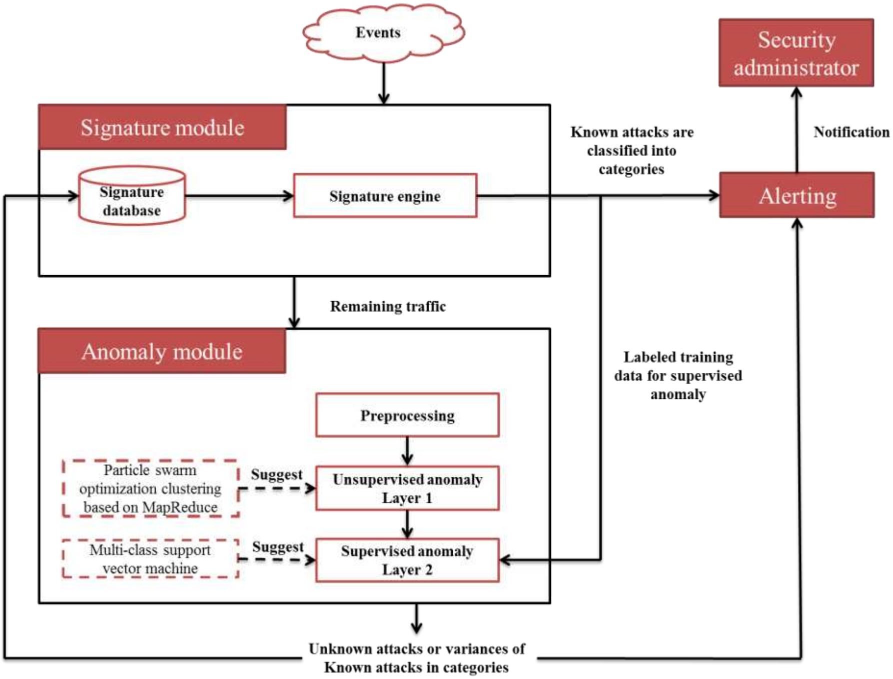
Events
Security administrator
Signature module
Signature database
Signature engine
Known attacks are classified into categories
Notification
Alerting
Remaining traffic
Anomaly module
Labeled training data for supervised anomaly
Preprocessing
Particle swarm optimization clustering based on MapReduce
Suggest
Unsupervised anomaly Layer 1
Multi-class support vector machine
Suggest
Supervised anomaly Layer 2
Unknown attacks or variances of Known attacks in categories

Fig. 4.1: The recommended architecture

- Events, which can be network traffic data or system call logs, are input into the signature module for the detection process.
- The signature database is utilized to detect known attacks in input data and classify them into predefined categories before sending to alerting module for notification. At the same time, those known attacks are removed from original data, leaving only normal and properly zero-day attack data to the anomaly module.
- The remaining traffic will be firstly preprocessed to select necessary features and normalize them for later detection process. The reduced data are then applied into the first layer, which employs an unsupervised algorithm to group instances into two categories: normal and anomalous. In order to process a large volume of data in a real time manner, it is proposed to employ parallel particle swarm optimization clustering algorithm based on MapReduce methodology [AL12] as inspired by the work [AL13]. Then results of each cluster in the first layer are forwarded to the supervised layer for final classification decisions. In this context, the multiple-class SVM technique [Le12] is suggested since it achieved the best performance in the conducted survey. The final detection of the anomaly module can be unknown attacks or variances of known attacks, which will be then sent to the alerting module for notification. At the same time, these will be updated to

signature database so that in next loops, those malicious activities are filtered in the signature layer.

Since the signature detection is not the focus of the thesis, its detailed operation is not covered in this section. Instead, the anomaly module is further elaborated with two working modes: training and operation.

- In the training mode as shown in figure 4.2, raw data, which can be normal or unknown attacks from the output of the upper signature module, are input into the unsupervised algorithm to build behavior profiles in forms of normal and anomalous clusters. For each cluster, a corresponding instance in the supervised layer is created as a second level of classification for sound detection decisions. Since the supervised algorithm requires labeled data for training, clustering results and known attacks detected in signature module are used to build classification models.

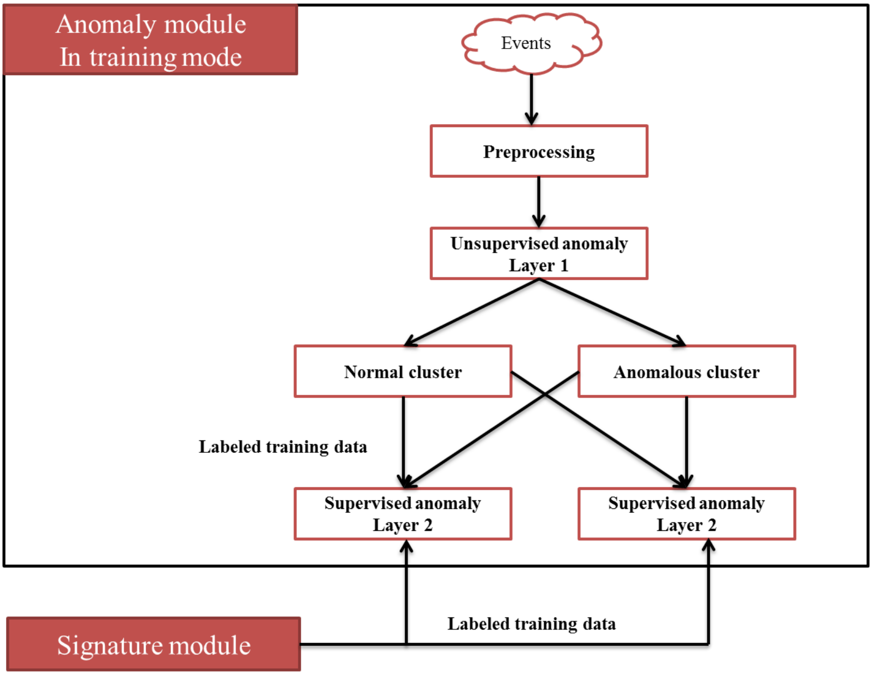

Fig. 4.2: Anomaly module in training mode

- In the operation mode, classification results of each cluster in layer 1 are forwarded to each corresponding instance in the supervised layer for final classification decisions as presented in figure 4.3.

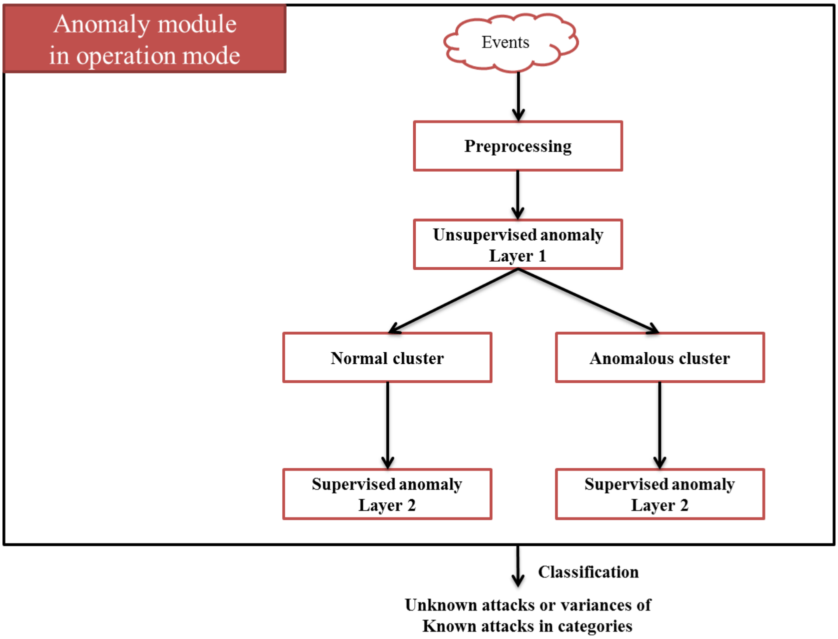
Anomaly module in operation mode
Events
Preprocessing
Unsupervised anomaly Layer 1
Normal cluster
Anomalous cluster
Supervised anomaly Layer 2
Supervised anomaly Layer 2
Classification
Unknown attacks or variances of Known attacks in categories

Fig. 4.3: Anomaly module in operation mode

5 Conclusion

In this paper, three main contributions have been presented:

- An advanced view on the current problems and challenges of the anomaly-based IDS.
- An overview about current trends and technologies that is based on the survey of more than 160 published studies.
- A recommended architecture is proposed for a more efficient IDS.

However, there are some limitations of the work: the suggested architecture is actually only proposed in literature without having a prototype to check its performance and quality of detection. Furthermore, because of the variety of publications, the survey reviewed only a short period of time (2012-2014); therefore, it may not reflect the entire current trends and tendencies. It is also important to note that the analyzed studies present only the state of art and not the state of practice.

For future works, I consider further researches to improve the understanding of the limitations of current technologies employed as very important. Existing problems, which limit the potential of the technique, should be analyzed in more detail. In addition, an implementation of the suggested architecture is undertaken to evaluate its efficiency in practice. On the other hand, the current state of art of anomaly-based IDS should be reviewed in a longer period for a better overview of research efforts and technology trends.

Acknowledgements

I would like to thank my supervisors, Prof. Dr. Friedbert Kaspar and Mr Fabian Berner, without their instructions and dedicated support through each step of the process, this study would not ever be accomplished.

References

[Ab14] Abuadlla.Y. et.al. : Flow-based Anomaly Intrusion Detection System using Two Neural Network Stages: Computer Science and Information Systems 2014, pp. 601 – 622, 2014.

[AL12] Aljarah, I; Ludwig, S: Parallel Particle Swarm Optimization Clustering Algorithm based on MapReduce Methodology: Nature and Biologically Inspired Computing, Mexico City, pp. 104 - 11, 2012.

[AL13] Aljarah, I; Ludwig, S: MapReduce Intrusion Detection System based on A Particle Swarm Optimization Clustering Algorithm: Evolutionary Computation, Cancun, pp. 955 - 962, 2013.

[BBK11] Bhuyan, M; Bhattacharyya, D; Kalita, J: Survey on Incremental Approaches for Network Anomaly Detection. International Journal of Communication Networks and Information Security 11/11, pp. 226 - 239, 2011.

[BBK14] Bhuyan, M; Bhattacharyya, D; Kalita, J: Network Anomaly Detection: Methods, Systems and Tools. Communications Surveys & Tutorials 02/14, pp. 303 - 336, 2014.

[CS14] UNM Synthethic Sendmail Data, http://www.cs.unm.edu/~immsec/data/synth-sm.html, as of 10.12.2014.

[De87] Denning, D: An Intrusion Detection Model: IEEE Transactions on Software Engineering, pp. 222 - 232, 1987.

[Ga08] García-Teodoro, P. et.al. : Anomaly-based Network Intrusion Detection: Techniques, Systems and Challenges. Computers & Security 08/08, pp. 18 - 28, 2008.

[GT06] Gates, C; Taylor, C: Challenging the Anomaly Detection Paradigm: A Provocative Discussion: 06 Proceedings of the 2006 Workshop on New Security Paradigms, New York, pp. 21 - 29, 2006.

[IX13] Ippoliti, D; Xiaobo, Z: A Self-Tuning, Self-Optimizing Approach for Automated Network Anomaly Detection Systems: Proceedings of The 9th International Conference on Autonomic Computing, pp. 85 - 90, 2013.

[JPP11] Jyothsna, V; Prasad, V; Prasad, K: A Review of Anomaly based Intrusion Detection Systems. International Journal of Computer Applications 08/11, pp. 26 - 35, 2011.

[KD14] KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, as of 10.12.2014.

[Ko11] Koch, R: Towards Next-Generation Intrusion Detection: Cyber Conflict (ICCC), 2011 3rd International Conference,Tallinnn, pp.1 - 18, 2011.

[Le12] Le, V. et.al. : Network Intrusion Detection based on Multi-class Support Vector Machine: ICCCI'12 Proceedings of the 4th International Conference on Computational Collective Intelligence: Technologies and Applications, pp. 536-543, 2012.

[Le14] Le.T: Anomalie-based Security Analysis, Master of Science Thesis, 2014.

[LL14] DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/ideval/, as of 10.12.2014.

[MRR13] Muniyandi, A; Rajeswari, R; Rajaram, R: Network Anomaly Detection by Cascading K-means Clustering and C4.5 Decision Tree Algorithm: International Conference on Communication Technology and System Design 2011, pp. 174 - 182, 2013.

[NS14] The NSL-KDD Dataset, http://nsl.cs.unb.ca/NSL-KDD/, as of 10.12.2014.

[PP07] Patcha, A; Park, J: An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends. Computer Networks 08/07, pp. 3448 - 3470, 2007.

[QMG12] Qazanfari, K; Mirpouryan, M.S; Gharaee, H: A Novel Hybrid Anomaly Based Intrusion Detection Method: Telecommunications (IST), Tehran, pp. 942 – 947, 2012.

[RRR08] Ringberg, H; Roughan, M; Rexford, J: The Need for Simulation in Evaluating Anomaly Detectors: ACM SIGCOMM Computer Communication, New York, pp. 55 - 59, 2008.

[SP10] Sommer, R; Paxson, V: Outside the Closed World: On Using Machine Learning for Network Intrusion Detection: IEEE Symposium on Security and Privacy, Oakland, pp. 305 - 316, 2010.

[SY14] Internet Security Threat Report 2014, http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf, as of 10.12.2014.

[TA14] Traffic Data from Kyoto University's Honeypots, http://www.takakura.com/Kyoto_data/, as of 10.12.2014.

[TSG10] Tavallaee, M; Stakhanova, N; Ghorbani, A: Toward Credible Evaluation of Anomaly-based Intrusion-Detection Methods. Systems, Man, and Cybernetics 09/10, pp. 516 - 524, 2010.

[VTN13] Viswanathan, A; Tan, K; Neuman, C: Deconstructing the Assessment of Anomaly-based Intrusion Detectors, Springer Berlin Heidelberg, 2013.