Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods

Mahbod Tavallaee, Natalia Stakhanova, and Ali Akbar Ghorbani, *Member, IEEE*

***Abstract*—Since the first introduction of anomaly-based intrusion detection to the research community in 1987, the field has grown tremendously. A variety of methods and techniques introducing new capabilities in detecting novel attacks were developed. Most of these techniques report a high detection rate of 98% at the low false alarm rate of 1%. In spite of the anomaly-based approach's appeal, the industry generally favors signature-based detection for mainstream implementation of intrusion-detection systems. While a variety of anomaly-detection techniques have been proposed, adequate comparison of these methods' strengths and limitations that can lead to potential commercial application is difficult. Since the validity of experimental research in academic computer science, in general, is questionable, it is plausible to assume that research in anomaly detection shares the above problem. The concerns about the validity of these methods may partially explain why anomaly-based intrusion-detection methods are not adopted by industry. To investigate this issue, we review the current state of the experimental practice in the area of anomaly-based intrusion detection and survey 276 studies in this area published during the period of 2000–2008. We summarize our observations and identify the common pitfalls among surveyed works.**

***Index Terms*—Anomaly detection, intrusion detection.**

I. Introduction

INTRUSION detection has been at the center of intense research in the past decade owing to the rapid increase of sophisticated attacks on computer systems. Typically, intrusion detection refers to a variety of techniques for detecting malicious and unauthorized activities commonly known as "attacks." There are three broad categories of detection approaches [1]: 1) signature-based technique that relies on pre-specified attack signatures; 2) anomaly-based approach, which typically depends on normal patterns classifying any deviation from normal as malicious; and 3) specification-based technique, which, although operates in a similar fashion to the anomaly-based approach, employs a model of valid program behavior in a form of specifications the development of which requires user guidance.

For years, research in the field of intrusion detection has been primarily focused on anomaly-based and misuse-based detection techniques. The latter method is traditionally favored in commercial products due to its predictability and high accuracy. In academic research, however, anomaly-detection approach is perceived as more powerful due to its higher potential to address novel attacks in comparison to misuse-based methods. While the academic community has proposed a wide spectrum of anomaly-based intrusion-detection techniques, adequate comparison of the strengths and limitations of these techniques that can lead to their potential adoption by industry is challenging. There are a variety of reasons for this shortcoming from simple underreporting of the experimental details to more serious issues that involve incorrect feature selection process and improper use of statistical analysis. To better understand the mismatch between the academic perception of anomaly-detection techniques and their potential for mainstream system implementation, we have carefully examined the evaluation practice of the anomaly-detection techniques.

In this paper, we focus on recent research in the area of anomaly detection, specifically, the work published during the period of 2000–2008. We analyze three major components in each study that, we believe, are critical for the evaluation and comparison of the intrusion-detection techniques. These components include the employed datasets, the characteristics of the performed experiments and the methods used for performance evaluation.

The remainder of this paper is organized as follows. Section II presents the overview of the related work. In Section III, we present our evaluation methodology. Sections IV–VI provide the results of our survey. Finally, Section VII concludes the survey and provides some future directions.

II. Related Work

The research on anomaly-based intrusion detection has been developing very rapidly since the first introduction of the anomaly-detection paradigm in 1987 by D. Denning. Since that time, the volume of research has grown tremendously; however, only in recent years has the question about the experimental practices of anomaly-detection research come up.

Gates and Taylor [2] questioned some of the core assumptions commonly accepted in network anomaly-detection studies (e.g., attacks are rare, simulated data are representative). They pointed out that many of these assumptions originate in Denning's early work, and thus, their applicability in the modern network domain should be revisited.

Ringberg *et al.* [3] also noted the inadequacy of some of the existing assumptions. Advocating for the use of simulation, they discussed four requirements necessary for a complete evaluation of anomaly-detection techniques, specifically, the existence of

Manuscript received May 31, 2009; revised December 16, 2009; accepted March 29, 2010. Date of publication May 17, 2010; date of current version August 18, 2010. The work of A. A. Ghorbani was supported by the Atlantic Canada Opportunity Agency (ACOA) through the Atlantic Innovation Fund (AIF) and Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2277441. This paper was recommended by Associate Editor E. R. Weippl.

The authors are with the Faculty of Computer Science, Information Security Center of Excellence, University of New Brunswick, Fredericton, NB E3B 5A3, Canada (e-mail: m.tavallaee@unb.ca; natalia@unb.ca; ghorbani@unb.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2010.2048428

"ground truth" in the anomaly-detection datasets, reproducibility of experiments, definition of anomaly and experimental control.

The validity, i.e., accuracy, and reliability, i.e., consistency, of the experiments in computer science have always been at the center of scientific criticism. Denning noted that the experimental practices of computer science do not adhere to the "traditional standard of science" [4]. The study performed by Tichy *et al.* [5] showed that relatively few papers in computer science had been published with experimentally validated results. The survey published by Wainer *et al.* [6] came to the same conclusion.

In between these surveys, several studies conducted more thorough analysis of the practices employed by the computer science community in conducting research. Zelkowitz and Wallace [7], [8] analyzed the use of experimentation in the software engineering community in the early 1990's. In the follow-up study in 2005, Zelkowitz noted the increasing use of experimental validation and the general improvement of the field research. Several other studies on the maturity of software engineering research [9], [10] generally confirmed Zelkowitz's observations.

The analysis of the use of network simulation in 1999 led Pawlikowski *et al.* [11] to conclude that simulation-based performance evaluation studies of telecommunication networks are in a "deep crisis of credibility." The follow-up study by Kurkowski *et al.* [12] on the use of simulation in mobile ad hoc network (MANET) research found no significant improvement, thereby calling the results "discouraging."

III. Evaluation Methodology

We conducted a survey of research work in the area of anomaly-based intrusion detection published during the period of 2000–2008. To avoid selection bias, we collected all research papers indexed by the Google Scholar and the Digital Bibliography and Library Project (DBLP) databases for the reviewed time period. From this set, we excluded short papers, extended abstracts, nonpeer-reviewed research, and papers not available in the English language and those containing no evidence of experimental study. During this process, we encountered seven cases of self-plagiarism (nearly identical papers, i.e., more than 90% content overlap, published in different venues by the same authors) and one case of plagiarism (an original work was later republished by different authors). For these cases, we chose to retain the earlier copy of each work.

To narrow our focus, we further selected research work relevant to anomaly-based methods for intrusion detection. Thus, any methods specifically developed for fault, fraud detection, etc., were excluded. The final set of 276 papers, containing 61 journals and 215 conference/workshop papers, was reviewed manually without any means of automatic search techniques. Each of the selected papers went through at least two rounds of evaluation to reduce classification error.

To provide better perspective on the survey results, we divided the papers into four categories: journal papers published in ISI journals (40 papers) and non-ISI journals (11 papers), and conference/workshop papers that were further classified according to the venue impact factor (IF) [13] into two broad categories: frequently cited (FC) category including 88 papers, and rarely cited (RC) category including 138 papers.[1] The details of the surveyed papers and the top five venues for each category are given in Tables I and II.[2]

TABLE I
Details of the Surveyed Papers

Papers by intrusion detection types	
Host-based studies	93 papers out of 276
Network-based studies	163 papers out of 276
Application-based studies	20 papers out of 276
Applied intrusion detection methods	
Classification-based methods:	160 papers out of 276
NN	25 papers out of 160
HMM	36 papers out of 160
SVM	20 papers out of 160
Bayesian networks	14 papers out of 160
Other methods	65 papers out of 160
Statistics-based methods	62 papers out of 276
Clustering	36 papers out of 276
Misc. methods (control-flow graph, finite-state automata, etc.)	46 papers out of 276

A. Survey Focus

There are two factors commonly used in statistics to assess the scientific rigor of a study: *reliability*, i.e., the consistency of the measurements, and *validity*, which refers to the accuracy and quality of the conclusions of a study. Both factors are key components of a study's trustworthiness. The lack of reliability brings into question a method's robustness and its ability to withstand the uncertainties of the deployment environment. The consistency of the method performance is especially important in the anomaly-detection domain, where the occurrence of certain anomalous events might be exceedingly rare. On the other hand, the lack of validity questions a study's relevancy and usefulness. Even the most reliable method has little value in intrusion detection if it fails to provide accurate recognition of attacks.

We applied these two metrics, validity and reliability, to evaluate the scientific rigor of the studies in the area of anomaly-based intrusion detection. Each study was assessed along the set of factors that directly or indirectly address the validity and reliability measures.

To assess the reliability of a study, i.e, its repeatability, we analyzed: 1) the experimental setup, such as the details provided about the employed datasets, tools, environment, and initial configuration of the algorithms; 2) the experimental process, e.g., the number of experiments performed; and 3) overall provided documentation of the experiments.

Since the concept of validity encompasses several perspectives of the study such as internal validity, external validity, conclusion, and construct validity [14], its assessment was performed with regard to these aspects. Internal validity relates to

[1]The full list of venues with the corresponding IF can be found in http://iscx.ca/ADsurvey.

[2]In Table I, papers using more than one method are considered in several categories.

TABLE II
PUBLICATION VENUES OF SURVEYED PAPERS (TOP FIVE VENUES)

Frequently Cited (FC) category	
Information Assurance Workshop (IAW)	7
International Symposium on Recent Advances in Intrusion Detection (RAID)	5
ACM Symposium on Applied computing (SAC)	5
ACM Conference on Computer and Communications Security (CCS)	3
IEEE International Conference on Data Mining (ICDM)	3
Rarely Cited (RC) category	
International Conference on Machine Learning and Cybernetics (ICMLC)	13
International Symposium on Neural Networks (ISNN)	6
International Conference on Availability, Reliability and Security (ARES)	5
Fuzzy systems and knowledge discoveryconference (FSKD)	3
Security and Management conference (SAM)	1
ISI journals	
Computers & Security, Elsevier	10
Computer Communications, Elsevier	6
Computer Networks, Elsevier	3
ACM Transactions on Information and System Security (TISSEC)	2
IEEE Transactions on systems, man, and cybernetics	2
Non-ISI journals	
Journal of software	2
Journal of Information Assurance and Security	1
International Journal of Non-Standard Computing and Artificial Intelligence	1
International Journal of Computer Science and Network Security	1
Information Management & Computer Security	1

the experimental design of a study and allows one to conclude whether the produced outcome of the study can be attributed to the proposed approach rather than other factors not accounted for in the experimental study design. Internal validity was analyzed through the data-preparation process, e.g., the use of normalization techniques in order to avoid bias.

Conclusion validity describes whether the relationship found between the data and the outcome is reasonable. For example, if the proposed algorithm employs certain features then how was determined that these features are significant for the study.

External validity relates to the generalization ability of the reported results, i.e., whether the produced outcome would hold for other data, or in a different deployment environment. Experimental validity is commonly analyzed through the use of sampling and feature-selection techniques and employed datasets.

Construct validity refers to the validity of the measurements and answers the question of whether we are measuring what we really intend to measure in a study. In other words, when we measure how well the proposed algorithm detects anomalies, is that what we really measure? We assessed construct validity through several factors: 1) definition of anomaly, i.e., what was being detected in the study; 2) evaluation measures, i.e., what measures were used to evaluate the approach; and 3) what type of evaluation was performed.

To summarize, the factors addressing the validity and reliability of a study can be broadly divided into three groups: factors related to *the employed data*, *the performed experiments*, and *the performance evaluation*. We review the experimental practice of the published research in the area of anomaly detection along these three dimensions.

IV. DATASETS

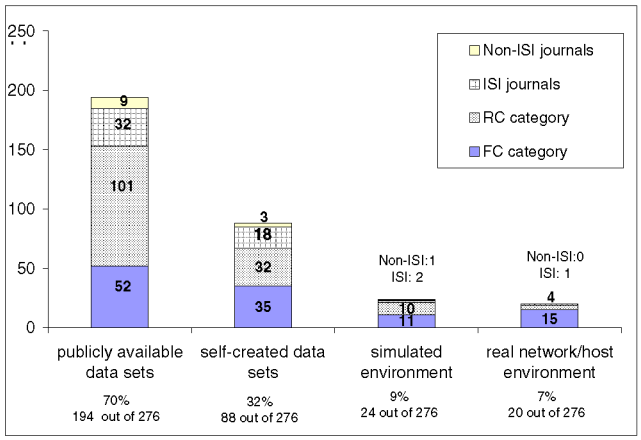
The evaluation data play an important role in the validation of any intrusion-detection method. The data quality not only allows us to judge the proposed method ability in detecting intrusive behavior, but also shows its potential effectiveness in the deployed operating environment.

This becomes a main challenge due to the criticism of existing datasets and the obstacles that prevent employing real traffic (e.g., privacy and reliability guarantee).

Among the surveyed works, the most prevalent approach to evaluation of the anomaly-based methods was based on fully or partially synthetic datasets, i.e., data fully or partially created in isolated environments. As Fig. 1 shows, 70% of works employed publicly available datasets, and 32% created a dataset for a particular study. Among the research studies preparing their own dataset, 85% gathered the data directly from the network or host environment and 15% generated synthetic data. A common approach to the generation of synthetic sets reported in the reviewed works is with the help of individuals (often graduate students) who are asked to use certain programs for a given period of time. While this approach is likely to produce the required amount of data, it hardly reflects a real environment.

Among the reviewed papers, 9% conducted the experiments with the use of simulation tools and only 7% attempted to test the robustness of the proposed approach in the "wild," i.e., through deployment in the real network.

There are several datasets publicly available for intrusion detection testing and evaluation. However, the most widely



Publicly available data sets	
24% (67 out of 276)	employ DARPA data
28% (77 out of 276)	employ KDD data set
36.5% (34 out of 93)	of host-based studies employ UNM data.
26% (24 out of 93)	of host-based studies employ DARPA data set.
6% (16 out of 276)	employ other sets
15% (41 out of 276)	inserted additional attack traces to the sets.

Self-created data sets for a given study	
0.7% (2 out of 276)	Released data
31% (86 out of 276)	Not released data 34 papers out of 86 are FC 32 papers out of 86 are RC 17 out of 86 are ISI journal papers 3 out of 86 are non-ISI journal papers
28% (24 out of 86)	Among studies that did not release data sets: not provided detailed data description 8 out of 24 are FC 10 out of 24 are RC 4 out of 24 are ISI journals 2 out of 24 are non-ISI journals
85% (75 out of 88)	collected real traffic/host events to generate sets.
24% (21 out of 88)	inserted synthetic attack traces into the data.
23% (20 out of 88)	inserted real attack traces into the data.

Fig. 1. Data usage in surveyed works.

used are the Defense Advanced Research Projects Agency (DARPA) intrusiondetection evaluation (DARPA) data (24%) and the knowledge discovery and data mining (KDD) set (28%), which together are used in more than 50% of studies. DARPA datasets [15] were generated in 1998, 1999, and 2000 in MIT Lincoln Laboratories, specifically for testing purposes. The sets consist of simulated host and network normal traffic and manually generated network-based attacks. The KDD set [16], known as KDD 99 intrusion data, is derived from the DARPA 98 dataset.

Despite the significant contribution of these datasets to the research on intrusion detection, their accuracy and ability to reflect real-life conditions have been extensively criticized [17], [18]. There were several efforts to provide alternative real traffic sets [19], which have been also questioned due to lack of proper documentation and reliability in assessment. In the host-based intrusion detection, the most commonly used synthetic dataset is the University of New Mexico (UNM) dataset that is employed in 36.5% of reviewed works. The DARPA dataset was employed in 26% of the host-based studies.

Overall, due to the lack of better datasets, the majority of the research in the field of network intrusion detection is still based on the synthetic sets.

In this context, data preprocessing gains special attention. Knowing the shortcomings of the data, it is important to prepare the data to ensure consistent and accurate evaluation. While this process and its requirements are well described in the data-mining literature, much of the published research in the intrusion detection field skips or overlooks this stage.

The following are the common pitfalls identified from our survey of published works (see Table III).

Definition of anomaly: The ability of an approach to detect intrusive, or rather abnormal behavior, is a cornerstone of the anomaly-based intrusion-detection methods. The primary problem in this context is the definition of the anomaly. The uniform understanding of anomaly as activity different from normal brings many challenges in the practical setting. Academic research broadly defines anomaly as abnormal behavior (e.g., outlier) [2], while system administrators narrow it down to only events potentially threatening their system/network.

The necessity for clear definition of anomaly in the context of given research has been also emphasized by Gates and Taylor [2] and Ringberg *et al.* [3]. As Ringberg *et al.* pointed out, it is quite challenging to accurately evaluate the anomaly-based intrusion detector capabilities without precise description what constitutes an anomaly.

Despite of this, the majority of current research on anomaly-based intrusion-detection methods, 88% of works, do not explicitly state what constitutes an anomaly in their study. While the general tendency among the surveyed works is to refer to anomalous behavior as behavior deviating from normal, the primary target in the evaluation of the proposed approach are the specific attacks.

Data normalization: The primary goal of data normalization is to ensure a common ground for the subsequent direct comparison of the data points. Depending on the nature of the data, the normalization might not be always necessary. However, omitting this step when the data units are not uniform introduces a bias toward data with larger units, consequently impacting the final outcome of the algorithm.

While the necessity of normalization cannot be judged in the datasets where little information is available, we can point out with certainty that publicly available datasets such as DARPA and KDD require such normalization. For example, the KDD set contains 41 features and the majority of features have a different scale, for example, destination host count is in the range of $0-255$, while source bytes range from 0 to 693 375 640. Thus, any type of anomaly analysis would require normalization of the employed features. Given this, only 21% of all studies on these datasets indicated the use of normalization techniques.

Features omission: The datasets often contain numerous features that can be not only redundant and unimportant, but also detrimental for the results' accuracy [20]. Thus, proper feature-selection practices can significantly affect the evaluation of a detection method's performance.

TABLE III
SURVEY RESULTS

Data sets preparation	
42% (69 out of 163)	used all features.
24% (40 out of 163)	employed the feature selection. 47% (19 out of 40) stated the feature selection method. 72.5% (29 out of 40) provided the selected features.
21% (35 out of 163)	indicated a use of normalization. 9 out of 163 are FC 22 out of 163 are RC 4 out of 163 are ISI journals
Experiments	
37% (72 out of 194)	of studies using public data sets, specified training and testing sets. 18 out of 194 are FC 39 out of 194 are RC 11 out of 194 are ISI journals 4 out of 194 are non-ISI journals
41% (36 out of 88)	of studies using self-created data sets, specified training and testing sets. 19 out of 88 are FC 9 out of 88 are RC 6 out of 88 are ISI journals 2 out of 194 are non-ISI journals
21% (34 out of 163)	of network-based studies specified the ratio of abnormal/normal activity in the testing set 23.5% (8 out of 34) used 1-2% abnormal/99-98% normal activity ratio in the testing set. 3% (1 out of 34) used 6-8% abnormal/94-92% normal activity ratio in the testing set. 3% (1 out of 34) used 19-20% abnormal/81-80% normal activity ratio in the testing set. 15% (5 out of 34) used 30-50% abnormal/70-50% normal activity ratio in the testing set. 56% (19 out of 34) used 80-82% abnormal/20-18% normal activity ratio in the testing set.
19% (52 out of 276)	conducted the performance study. 18 out of 52 are FC 21 out of 52 are RC 11 out of 52 are ISI journals 2 out of 52 are non-ISI journals
60% (31 out of 52)	of those that conducted the performance study stated the characteristics of the computer. 14 out of 31 are FC 10 out of 31 are RC 5 out of 31 are ISI journals 2 out of 31 are non-ISI journals
23% (64 out of 276)	did not specify the initial parameters of the algorithms. 15 out of 64 are FC 36 out of 64 are RC 12 out of 64 are ISI journals 1 out of 64 is non-ISI journal
12% (25 out of 212)	of those that specified the initial parameters justified the selected initial parameters of the algorithm.
20% (55 out of 276)	indicated methods to ensure validity and reliability of experiments. 6 out of 55 states the number of simulation runs 34 out of 55 indicated that the produced result is an average value 12 out of 55 reported the use of cross-validation 3 out of 55 reported a confidence interval for the obtained results
Evaluation	
35.5% (98 out of 276)	provided comparison of the proposed approach with earlier works. 29 out of 98 are FC 52 out of 98 are RC 14 out of 98 are ISI journals 3 out of 98 are non-ISI journals
26% (42 out of 163)	of network-based studies provided separate results for different attack types. 12 out of 42 are FC 24 out of 42 are RC 4 out of 42 are ISI journals 2 out of 42 are non-ISI journals
Publishing quality	
5% (13 out of 276)	produced unreadable plots/graphs. 5 out of 276 are FC 8 out of 276 are RC
1% (3 out of 276)	provided no explanation for the given plots/graphs. all 3 papers are in RC category

As our survey showed, among network-based approaches, 24% of the works applied the automated feature-selection method. One of the concerns is that the selected features are not always specified: 27.5% of works did not disclose the employed features. This practice cannot be considered appropriate as it makes it nearly impossible to reproduce the experiment or produce a fair comparison in the future research.

Data reduction: Generally, datasets used for evaluation of intrusion detection systems (IDSs) are very large, which not only complicates the analysis, but also slows down the processing. Sampling is generally a popular data reduction technique used in many areas. However, in the intrusion-detection domain, numerous studies have pointed out that sampling brings significant distortion to the traffic statistics, essentially degrading the accuracy of the analysis [21], [22]. While potential improvements for estimating the accuracy of sampling have been also studied [21], [23], the question of the impact of sampling in intrusion detection remains open.

Mai *et al.* [22] performed experiments to show the impact of four sampling techniques: random packet, random flow, smart sampling and sample-and-hold, on the detection accuracy of port-scan and volume anomalies (e.g., Denial-of-service attacks). Their results show that all four methods introduce a significant bias that degrades the performance of the detection algorithms. Despite this research, our survey shows that 13% of reviewed works used random sampling and 9% selected instances using some predefined rules. However, the practice of the use of sampling techniques is inconsistent and varies depending on the needs of a particular study. We found the following sampling strategies used more commonly than others: random sampling only from training set or only from the test set to generate both training and testing sets for evaluation, sampling of attack instances of the dominant type or specified types, sampling of the particular sessions.

V. Experiments

Poor experimental practice is another factor in IDS methods evaluation that can bias its final outcome. One of the measures of a study's validity and reliability is the ability to reproduce the experiment. Such repeatability is ensured through proper experimentation procedures employed by the researchers, on the one hand, and full documentation of the experimental environment on the other hand.

Experiments setup: This phase often does not receive proper attention in the publications. However, the absence of proper documentation of experiments affects first of all the credibility of the study, making it hard to reproduce the experiments and make comparison with other methods, which essentially makes a study less believable. The experimental setup assumes *the proper description of the employed datasets*, including the indication of the testing and training sets. This is especially important in case of publicly available datasets, such as DARPA, KDD, that are already prepared and broken down into testing and training parts. Despite this, the selection of training and testing sets is not uniform among researchers. Often these sets are rearranged to suit the needs of a particular study.

Unfortunately, out of 194 papers using the publicly available datasets, more than half (63%) did not properly specify which sets were used for training and testing of the approach. Among experimental datasets that were not public, these numbers were slightly higher: out of 88 papers, 59% did not describe the training and testing sets.

Another aspect related to the employed testing set is the ratio of anomalous and normal records in the testing data. An assumption of rareness of anomalies, i.e., the existence of a small portion of anomalous records compared to the volume of normal activity, is common in the intrusion-detection domain [2]. Recently, there have been several studies showing that this picture is changing and nowadays abnormal traffic on the Internet (including scanning activity) cannot be quantified as rare [24], [25].

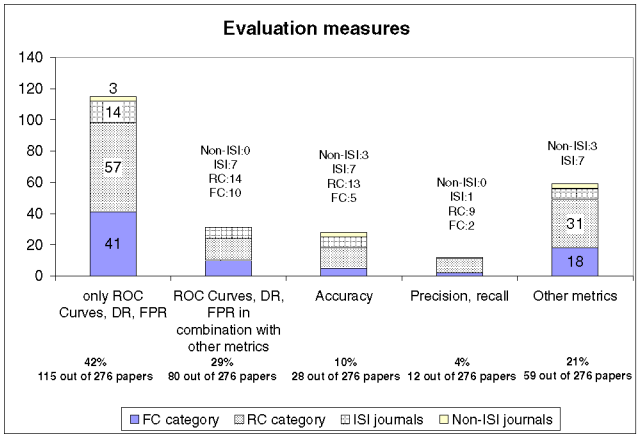
Reviewing the network-based studies on the DAPRA and KDD datasets, we noted a great variability in the employed ratio of abnormal to normal activity. Among 34 papers that specified this ratio for the testing set, the majority of the studies (24 out of 34) experimented with a high percentage of abnormal activity (30%−82%) in the data. It should be also noted that 19 of these studies worked with the KDD dataset that has 81% of abnormal activity. As some of these studies employed random sampling, a final percentage of abnormal activity ranged from 80% to 82%. Two papers experimented with 6%−20% of abnormal activity in the set, and only 8 papers out of 34 (23.5%) assumed a low probability of intrusive activity, using a 1%−2% abnormal to 99%−98% normal activity ratio.

Another concern of the experimental setup is *the use of the simulation tools and well-known algorithms*. The initial configurations of these tools and algorithms can significantly influence the final result. Our review shows that 24% of publications left out a definition for the initial parameters of the input variables. For example, the widely used libSVM tool [26] that implements support-vector machines defines 13 variables, each of which can affect the outcome of the algorithm. However, out of 19 studies using SVM algorithm, 8 papers (42%) did not state the initial configuration.

Among the 212 surveyed works that defined the initial configuration, the selected parameters were rarely justified. Only 12% of the studies revealed the motivation for their selection.

Results validation: Since the evaluation of the detection method in the majority of the surveyed studies is based on the simulation, one of the important aspects of the credibility of the obtained results is their statistical reliability. It is important to ensure that the evaluation results are consistent and not attributed to a coincident error.

In machine learning, a commonly used method of checking reliability is cross-validation. Cross-validation involves partitioning the data into training and testing portions and evaluating the method using the generated testing set. Cross-validation is generally performed in multiple rounds with the final result averaged over the result of each run. The application of cross-validation in network anomaly-detection evaluation has to be carefully guided. To provide a credible result, on the one hand, the researcher has to ensure that the attacks represented in the test stage do not fully repeat the attacks used in the training run



Evaluation metrics	
6.5% (18 out of 276)	employ ROC curve only 5 out 18 are FC 9 out 18 are RC 3 out 18 are ISI journal papers 1 out 18 is non-ISI journal paper
26% (71 out of 276)	state only DR and FP rates 28 out 71 are FC 34 out 71 are RC 8 out 71 are ISI journal papers 1 out 71 is non-ISI journal paper
2% (6 out of 276)	employ Area Under Curve (AUC) metric
2.5% (7 out of 276)	present results using visual aids (graphs, pictures) without any link to known evaluation metrics
4% (12 out of 276)	use their own evaluation metrics

Fig. 2. Evaluation measures in surveyed works.

to avoid memorization of training data by the detector. On the other hand, a sensible percentage of anomalies (attacks) has to be provided in all partitions used for validation.

Such caution is mainly dictated by the nature of the data employed in the anomaly detector evaluation. In both the synthetic data such as DARPA and KDD and the real network traffic, the representation of attack types is generally unequal, with some, usually easily detectable types clearly dominating the rest of the traffic. For example, such dominating types are the denial-of-service attacks in DARPA 1998 data or scanning activity in the Internet backbone traffic [25].

Although the statistical reliability of results shows the consistency of the method's performance, it does not allow to assess its validity, i.e., accuracy of the obtained results. There are a number of techniques used to ensure statistical validity of simulation result [27]. One of the methods employed in statistics is based on the central limit theorem, which is used to determine the sufficient number of simulation runs needed to guarantee certain confidence in the results of a given study.

Unfortunately, in spite of their significance, the issues of reliability and validity of the experiments were often ignored in the surveyed works. 80% of the papers did not discuss any methods used to ensure validity and reliability of the experiments. Among the rest, 20% of the studies (6 out of 55) directly stated the number of simulation runs and 34 studies indicated that the produced result was an average of the performed evaluation runs. At the same time, a single execution of the simulation will rarely produce a credible result. It was not uncommon to see the justification of the reported result as the best result obtained in the experiments. A total of 12 papers reported the use of cross-validation. Three of the surveyed works stated the confidence interval.

VI. Performance Evaluation

The effectiveness of anomaly-detection techniques is generally evaluated from two perspectives:

1) the ability of the approach to distinguish normal vs. intrusive/abnormal;
2) the efficiency of the method according to the time required for training the model and the time taken during the detection process.

A. Normal Versus Intrusive

An approach's ability to correctly classify behavior is essentially interpreted in terms of four possibilities: *true-positive events*, known attacks instances detected as abnormal, *false-positive* (FP) events that are incorrectly classified as abnormal, *true-negative events*, which are correctly identified normal behavior and *false-negative* (FN) events that present abnormal behavior incorrectly classified as normal. intrusion-detection assessment metrics are derived on the basis of these four indicators. The most commonly used metrics are *detection rate* (DR), a ratio between the number of correctly detected attacks and the total number of attacks and *FP rate* (FPR), computed as the ratio between the number of normal connections that are incorrectly misclassified as attacks and the total number of normal connections. Often these metrics are displayed as a receiver operating characteristic (ROC) curve to emphasize the tradeoff between them. However, the ROC curve alone, or DR and FPR might be misleading or simply incomplete for understanding the strengths and weaknesses of the proposed approach [17], [28]–[30].

As such, Lazarevic *et al.* [31] showed that the standard metrics: DR and FPR are subjective toward bursty attacks (e.g., denial-of-service and probing) due to a high number of connections in a short period of time, and show better performance in comparison with other types of attacks often characterized by a single connection (such as user-to-root or remote-to-local). Ulvila and Gaffney [29] pointed out the flaws of the ROC curve metric from the perspective of multiple IDS comparison.

Generally, all these studies lead to two points to address potential inaccuracies and allow comprehensive evaluation of the detection method performance assessment: 1) a comprehensive set of evaluation metrics should be considered, and 2) the attacks of different types should be evaluated separately.

In spite of this, the majority of research in the field of anomaly detection still follows the beaten path. As Fig. 2 shows,[3] the ROC curve and DR and FPR are still the most commonly used

[3]In Fig. 2, papers using several methods are considered in corresponding categories.

metrics. 42% of the surveyed works employed only these measures and 29% of the papers complemented these metrics with other evaluation measures. Other traditional metrics commonly employed for evaluation are accuracy (10% of the works) and precision and recall (4%). Among less-frequently used metrics are the area under curve (AUC), response/detection time, cost measure, etc. Although the AUC metric is widely used in many data-mining studies, it was employed in 2% of the surveyed papers, usually in combination with accuracy and ROC curve. 2.5% of the papers lacked any identifiable evaluation metrics, presenting results using various visual aids such as pictures, graphs.

Unfortunately, the second point, i.e., presentation of separate results for each type of attack, is also far from being a standard practice. Among 163 network-based studies only 26% of studies presented separate results for the types of attacks present in the dataset.

B. Evaluating Method Efficiency

Another facet of the IDSs' evaluation is the analysis of performance requirements that generally includes the processing memory and time overhead. In our survey, 19% of the papers conducted a performance study. However, among them, 40% did not indicate the characteristics of the computer used to obtain these results (see Table III). Such oversight essentially produces results that cannot be reproduced and cannot be used for comparison with other techniques.

VII. Conclusion

Summarizing our observations about a scientific rigor of the anomaly-based intrusion-detection studies, we find that the majority of surveyed works do not satisfy these requirements. Overall, our findings confirm a common trend in the experimental computer science field that shows a lack of a scientific rigor in academic research.

While one of the common perceptions of the highly ranked publication venues is the better quality of their published papers, our survey results do not support that. The focus of our survey was primarily on the experimental part of the research; thus, we have not attempted to analyze the quality of the proposed methods. Nevertheless, the review of the published research along three analyzed components of the experimental study: datasets, performed experiments, and the evaluation, show that studies from all categories fail to follow basic principles of scientific experimentation. Since our survey is based on an analysis of published documents, it is possible that many of the identified pitfalls were avoided in the conducted research, but not reported. Unfortunately, even the best research work can lose its value behind an ambiguous, unclear and unsound presentation.

We hope that these results will help academic community to overcome common pitfalls discovered in this survey in their future research. In the future, we plan to repeat this study to see how research in the area of anomaly-based intrusion detection is changing and whether any significant trends can be noted.

References

[1] R. Sekar, A. Gupta, J. Frullo, T. Shanbhag, A. Tiwari, H. Yang, and S. Zhou, "Specification-based anomaly detection: A new approach for detecting network intrusions," in *Proc. 9th ACM Conf. Comput. Commun. Security (CCS)*, 2002, pp. 265–274.

[2] C. Gates and C. Taylor, "Challenging the anomaly detection paradigm: A provocative discussion," in *Proc. 2006 Workshop N. Security Paradigms (NSPW)*, pp. 21–29.

[3] H. Ringberg, M. Roughan, and J. Rexford, "The need for simulation in evaluating anomaly detectors," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 1, pp. 55–59, 2008.

[4] P. J. Denning, "ACM President's Letter: What is experimental computer science?" *Commun. ACM,* vol. 23, no. 10, pp. 543–544, 1980.

[5] W. F. Tichy, P. Lukowicz, L. Prechelt, and E. A. Heinz, "Experimental evaluation in computer science: a quantitative study," *J. Syst. Softw.*, vol. 28, no. 1, pp. 9–18, 1995.

[6] J. Wainer, C. N. Barsottini, D. Lacerda, and L. R. M. de Marco, "Empirical evaluation in computer science research published by ACM," *Inf. Softw. Technol.*, vol. 51, no. 6, pp. 1081–1085, 2009.

[7] M. V. Zelkowitz and D. R. Wallace, "Experimental models for validating technology," *Computer*, vol. 31, no. 5, pp. 23–31, 1998.

[8] M. V. Zelkowitz, "An update to experimental models for validating computer technology," *J. Syst. Softw.*, vol. 82, no. 3, pp. 373–376, 2009.

[9] J. E. Hannay, O. Hansen, V. By Kampenes, A. Karahasanovic, N.-K. Liborg, and A. C. Rekdal, "A survey of controlled experiments in software engineering," *IEEE Trans. Softw. Eng.*, vol. 31, no. 9, pp. 733–753, Sep. 2005.

[10] C. Zannier, G. Melnik, and F. Maurer, "On the success of empirical studies in the international conference on software engineering," in *Proc. 28th Int. Conf. Softw. Eng. (ICSE)*, New York, NY: ACM, 2006, pp. 341–350.

[11] K. Pawlikowski, H.-D. Jeong, and J.-S. R., "On credibility of simulation studies of telecommunication networks," *IEEE Commun. Mag.*, vol. 40, no. 1, pp. 132–139, Jan. 2001.

[12] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation studies: the incredibles," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 9, no. 4, pp. 50–61, 2005.

[13] Estimated Venue Impact Factors. (2009, Jan.). [Online]. Available: http://citeseerx.ist.psu.edu/stats/venues

[14] W. Trochim and J. P. Donnelly, *The Research Methods Knowledge Base*. Mason, OH: Atomic Dog, 2006.

[15] MIT Lincoln Labs. (2008, Feb.). *DARPA intrusion detection evaluation* [Online]. Available: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html

[16] KDD Cup 1999. (2008, Oct.) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[17] J. McHugh, "The 1998 Lincoln Laboratory IDS evaluation," in *Proc. 3rd Int. Workshop Recent Adv. Intrusion Detection (RAID)*. London, U.K.: Springer-Verlag, 2000, pp. 145–161.

[18] M. V. Mahoney and P. K. Chan, "An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection," in *Proc. 6th Int. Symp. Recent Adv. Intrusion Detection*. Berlin, Germany: Springer-Verlag, 2003, pp. 220–237.

[19] The Internet Traffic Archive. (2009, Apr.). [Online]. Available: http://ita.ee.lbl.gov/

[20] A. H. Sung and S. Mukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks," in *Proc. Symp. Appl. Internet (SAINT)*. Washington, DC: IEEE Comput. Soc., 2003, pp. 209–216.

[21] N. Hohn and D. Veitch, "Inverting sampled traffic," *IEEE/ACM Trans. Netw.*, vol. 14, no. 1, pp. 68–80, Jan. 2006.

[22] J. Mai, C.-N. Chuah, A. Sridharan, T. Ye, and H. Zang, "Is sampled data sufficient for anomaly detection?" in *Proc. 6th ACM SIGCOMM Conf. Internet Meas. (IMC)*. New York, NY: ACM, 2006, pp. 165–176.

[23] P. Tune and D. Veitch, "Towards optimal sampling for flow size estimation," in *Proc. 8th ACM SIGCOMM Conf. Internet Meas. (IMC)*. New York, NY: ACM, 2008, pp. 243–256.

[24] V. Yegneswaran, P. Barford, and J. Ullrich, "Internet intrusions: global characteristics and prevalence," *SIGMETRICS Perform. Eval. Rev.*, vol. 31, no. 1, pp. 138–147, 2003.

[25] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Profiling internet backbone traffic: behavior models and applications," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 169–180, 2005.

[26] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A library for support vector machines* [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm
[27] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation,and Modeling*. New York, NY: Wiley–Interscience, 1991.
[28] A. A. Cárdenas, J. S. Baras, and K. Seamon, "A framework for the evaluation of intrusion detection systems," in *Proc. 2006 IEEE Symp. Security Privacy (SP)*. Washington, DC: IEEE Comput. Soc., 2006, pp. 63–77.
[29] J. W. Ulvila and J. E. Gaffney, "Evaluation of intrusion detection systems," *J. Res. Nat. Inst. Standards Technol.*, vol. 108, no. 6, pp. 453–471, 2003.
[30] S. Axelsson, "The base-rate fallacy and its implications for the difficulty of intrusion detection," in *Proc. 6th ACM Conf. Comput. Commun. Security (CCS)*, New York, NY: ACM, 1999, pp. 1–7.
[31] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," presented at the 3rd SIAM Int. Conf. Data Mining, San Francisco, CA, 2003.

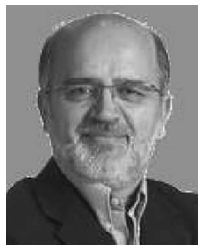
Mahbod Tavallaee received the Bachelor's degree in software engineering from Tarbiat Moallem University, Tehran, Iran, in 2004, and the Master's degree in information technology from the University of Tehran, Iran, in 2007. He is currently working toward the Ph.D. degree at the Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada.

He was a Research Associate in Iran Telecommunication Research Center for about six months. He is currently an active member of the Information Security Center of Excellence (ISCX), University of New Brunswick. His current research interests include intrusion detection systems.

Natalia Stakhanova received the Ph.D. degree in computer science from Iowa State University, Ames.

She is currently a Research Scientist at the Faculty of Computer Science, Information Security Center of Excellence (www.ISCX.ca), University of New Brunswick, Fredericton, NB, Canada. She has authored or coauthored more than ten journal and conference papers. She has two pending patents in the field of computer security. Her current research interests include intrusion detection and general information security field.

Dr. Stakhanova was the recipient of the "Nokia Best Student Paper Award" at the IEEE International Conference on Advanced Information Networking and Applications (AINA) in 2007.



Ali Akbar Ghorbani (M'94) received the B.Sc. degree from the University of Tehran, Iran, in 1976, the M.Sc. degree from George Washington University, Washington, DC, in 1979, and the Ph.D. degree from the University of New Brunswick (UNB), Fredericton, NB, Canada, in 1995.

He is currently a Professor and Dean with the UNB, where he is the Director of Information Security Center of Excellence, and is also the Coordinator of the Privacy, Security and Trust network annual conference. He holds UNB Research Scholar position and is the Coeditor-in-Chief of the *Computational Intelligence: An International Journal*, and an Associate Editor of the *International Journal of Information Technology and Web Engineering*. His current research interests include web intelligence, network security, complex adaptive systems, critical infrastructure protection, and trust and security assurance.

Dr. Ghorbani is a member of the Association for Computing Machinery, the IEEE Computer Society, and the Canadian Society for Computational Studies of Intelligence.