A database of anomalous traffic for assessing profile based IDS

Philippe Owezarski

A database of anomalous traffic for assessing profile based IDS

Philippe Owezarski*

CNRS; LAAS;
7 Avenue du colonel Roche, F-31077 Toulouse, France
Université de Toulouse;
UPS, INSA, INP, ISAE; LAAS;F-31077 Toulouse, France
owe@laas.fr

Résumé This paper aims at proposing a methodology and the required tools for evaluating current IDS (commercial ones, as well as prototypes resulting from advanced research projects) capabilities of detecting attacks targeting the networks and their services. This methodology tries to be as realistic as possible and reproducible, i.e. it works with real attacks and real traffic in controlled environments. It especially relies on a database containing attack traces specifically created for that evaluation purpose. By confronting IDS to these attack traces, it is possible to get a statistical evaluation of IDS, and to rank them according to their detection capabilities without false alarms. For illustration purposes, this paper shows the results obtained with 3 public IDS. It also shows how the attack traces database impacts the results got for the same IDS.

Keywords. Statistical evaluation of IDS, attack traces, ROC curves, KDD'99

I Motivation

I.1 Problematics

Internet is becoming the universal communication network, conveying all kinds of information, ranging from the simple transfer of binary computer data to the real time transmission of voice, video, or interactive information. Simultaneously, Internet is evolving from a single best effort service to a multiservice network, a major consequence being that it becomes highly exposed to attacks, especially to denial of services (DoS) and distributed DoS (DDoS) attacks. DoS attacks are responsible for large changes in traffic characteristics which may in turn significantly reduce the quality of service (QoS) level perceived by all users of the network. This may result in the breaking of SLA (*Service Level Agreement*) at the Internet Service Provider (ISP) fault, potentially inducing major financial losses for them.

* The author wants to thank all members of the MetroSec project. Thanks in particular to Patrice Abry, Julien Aussibal, Pierre Borgnat, Gustavo Comerlatto, Guillaume Dewaele, Silvia Farraposo, Laurent Gallon, Yann Labit, Nicolas Larrieu and Antoine Scherrer.

Detecting and reacting against DoS attacks is a difficult task and current intrusion detection systems (IDS), especially those based on anomaly detection from profile, often fail in detecting DDoS attacks efficiently. This can be explained via different lines of arguments. First, DDoS attacks can take a large variety of forms so that proposing a common definition is in itself a complex issue. Second, it is commonly observed that Internet traffic under normal conditions presents per se, or naturally, large fluctuations and variations in its throughput at all scales [PKC96], often described in terms of long memory [ENW96], self-similarity [PW00], multifractality [FGW98]. Such properties significantly impair anomaly detection procedures by decreasing their statistical performance. Third, Internet traffic may exhibit strong, possibly sudden, however legitimate, variations (flash crowds - FC) that may be hard to distinguish from illegitimate ones. That is why, IDS relying on anomaly detection by way of statistical profile often yield a significant number of false positives, and are not very popular.

These tools also lack efficiency when the increase of traffic due to the attack is small. This situation is frequent and extremely important because of the distributed nature of current denial of service attacks (DDoS). These attacks are launched from a large number of corrupted machines (called zombies) and under the control of a hacker. Each machine generates a tiny amount of attacking traffic in order to hide it in the large amount of cross Internet traffic. On the other hand, as soon as these multiple sources of attacking traffic aggregate on links or routers on their way to their target, they represent a massive amount of traffic which significantly decreases the performance level of their victim, and of the network it is connected to. The anomaly detection is easy close from the victim, but detecting it at this place is useless : targeted resources have been wasted, and the QoS degraded ; the attack is therefore successful. It is then essential for IDS to detect attacks close from their sources, when the anomaly just results from the aggregation of the attacking traffic of few zombies hidden in the massive amount of legitimate traffic.

I.2 The KDD'99 traditional evaluation method

It exists several approaches for statistically evaluating IDS. However, they all have in common the need of an attack database which IDS are confronted to. Up to now, the most used database is KDD'99.

The principle of a statistical evaluation deals with making the IDS to evaluate analyze the attack traces, and to count the number of true positives, true negatives, false positives, and false negatives. The relations between these values can then be exhibited by mean of a ROC curve (Receiver Operating Characteristic). The ROC technique has been used in 1998 for the first time for evaluating IDS in the framework of the DARPA project on the off-line analysis of intrusion detection systems, at the Lincoln laboratory at MIT [MIT]. At that time, it was the first intelligible evaluation test applied to multiple IDS, and using realistic configurations. A small network was set-up for this purpose, the aim being to emulate a US Air Force base connected to the Internet. The background traffic was generated by injecting attacks in well defined points of the network, and

collected with TCPDUMP. Traffic was grabbed and recorded for 7 weeks, and served for IDS calibration. Once calibration was performed, 2 weeks of traces containing attacks were used to evaluate the performance of IDS under study. Several papers describe this experience as Durst et al. [DCW+99], Lippmann et al. [LFG+00] and Lee et al. [LSM99]. This DARPA work in 1998 used a wide range of intrusion attempts, tried to simulate a realistic normal activity, and produced results that could be shared between all researchers interested in such topic. After an evaluation period, researchers involved in the DARPA project, as well as many others of the same research community, provided a full set of evaluation results with the attack database, which lead to some changes in the database, known nowadays under the name KDD'99. These changes mainly dealt with the use of more furtive attacks, the use of target machines running Windows NT, the definition of a security policy for the attacked network, and tests with more recent attacks.

Whereas the KDD'99 aimed at serving the needs of the research community on IDS, some important questions about its usability were raised. McHugh [McH01] published a strong criticism on the procedures used when creating the KDD'99 database, especially on the lack of verification of the network realism compared to an actual one. It was followed in 2003 by Mahoney et Chan [MC03] who decided to review into detail the database. Mahoney et Chan showed that the traces were far from simulating realistic conditions, and therefore, that even a very simple IDS can exhibit very high performance results, performances that it could never reach in a real environment. For example, they discovered that the trace database includes irregularities as differences on TTL between attacking and legitimate packets.

Unfortunately, despite all the disclaimers about this KDD'99 database for IDS evaluation, it is still massively used. This is mainly due to the lack of other choices and efforts for providing new attack databases. Such limitations of the KDD'99 database motivated ourselves for issuing a new one, as we needed to compare performances of existing profile based IDS with the ones of new anomaly detection tools we were designing in the framework of the MetroSec project [MET] (project from the French ACI program on Security & Computer science, 2004-2007)

I.3 Contribution

Despite these limitations related to the DARPA project contributions to IDS evaluation, the introduction of ROC techniques remains nevertheless a very simple and efficient solution, massively used since. It consists in combining detection results with the number of testing sessions to issue two values which summarize IDS performance : the detection ratio (number of detected intrusion divided by the number of intrusion attempts) and the false alarm rate (number of false alarms divided by the total number of network sessions). These summaries of detection results then represent one point on the ROC curve for a given IDS. The ROC space is defined by the false alarms and true positive rates on X and Y axis respectively, what in fact represents the balance between efficiency

and cost. The best possible IDS would then theoretically be represented by a single point curve of coordinates (0, 1) in the ROC space. Such a point means that all attacks were detected and no false alarm was raised. A random detection process would be represented in the ROC space by a strait line going from the bottom left corner (0, 0) to the upper right corner (1, 1) (the line with equation $y = x$). Points over this line mean that the detection performance is better than the one of a random process. Under the line, it is worse, and then of no real meaning.

Given the efficiency and simplicity of the ROC method, we propose to use it as a strong basis for our new profile based IDS evaluation methodology ; it therefore looks like the KDD'99 one. At the opposite, the KDD'99 trace database appears to us as completely unsuited to our needs for evaluating performances of IDS and anomaly detection systems (ADS). This is what our contribution is about. Let us recall here that in the framework of the MetroSec project, we were targeting attacks which can have an impact on the quality of service and performances of networks, i.e. on the quality of packets forwarding. It is not easy to find traffic traces containing this kind of anomalies. It would be required to have access to many traffic capture probes, close from zombies, and launch traffic captures when attacks arise. Of course, hackers do not advertise when they launch attacks and their zombies are unknown. We then haven't documented traffic traces at our disposal containing these kinds of anomalies, i.e. traces for which no obvious anomaly appears, but for which we would know that between two dates, an attack of that kind, and having a precise intensity was perpetrated. The lack of such traces is one of the big issues for researchers in anomaly detection.[1] In addition, it is not enough to validate detection methods and tools on one or two traces which anomalies would be detected in ; it would for instance forbid to quantity detection mistakes.

I.4 Paper structure

This paper presents a new evaluation and comparison method of profile based IDS and ADS performances, and that improves and adapts to new requirements the ancient KDD'99 method. The main contribution dealt with creating a new documented anomaly database - among which some are attacks. These traces contain, in addition of anomalies, a realistic background traffic having the variability, self-similarity, dependence, and correlation characteristics of real traffic, which massively distributed attacks can easily hide in. The creation process of this trace database, as well as the description of the main components used are described in section II. Then, section III shows for a given ADS - called NADA [FOM07] - the differences in evaluation results depending on the anomaly/attack trace database used. Section IV then shows by using our new evaluation method based on our new trace database - called MetroSec database - the statistical evaluation results got for 3 IDS or ADS publicly available. Last, section V

[1] This lack is even more important as for economical and strategic reasons of carriers, or users privacy, such data relating anomalies or attacks are not made public.

concludes this paper with a discussion on the strength and weaknesses of ours new evaluation method relying on our new MetroSec anomalies database.

II Generation of traffic traces with or without anomalies, by ways of reproducible experiments

II.1 The MetroSec experimental platform

One of the contributions of MetroSec was to produce controlled and documented traffic traces, with or without anomalies, for testing and validating intrusion detection methods.

For this purpose, we lead measurement and experimentation campaigns on a reliable operational network (for which we are sure that is does not contain any anomalies, or at least very few), and to generate ourselves attacks and other kinds of anomalies which are going to mix and interact with background regular traffic. It is then possible to define the kinds of attacks we want to launch, to control them (sources, targets, intensities, etc.), and to associate to the related captured traces a ticket indicating the very accurate characteristics of perpetrated attacks. In such a context, anomalies are reproducible (we can regenerate as often as wanted the same experimental conditions). This reproductivity makes possible the multiplication of such scenarios in order to improve the validation statistics of detection tools, or the comparison accuracy of our methods with others. The trace database produced in MetroSec is one of the significant achievements of the project, and guaranties the reliability of IDS evaluation.

The experimental platform which was used for creating the MetroSec trace database uses the Renater network, the French network for education and research. Renater is an operational network which is used by a significantly large community in its professional activity. Because of its design, Renater has the necessary characteristics for our experiments :

- it is largely over-provisionned related to the amount of traffic it is transporting. Its OC-48 links provide 2,4 Gbits/s of throughput, whereas a laboratory as LAAS, having at its disposal a link whose capacity is 100 Mbits/s, generates in average a traffic less than 10 Mbits/s [OBLG08]. As a consequence, Renater provides a service with a constant quality. Thus, even if we want to saturate the LAAS access link, the impact of Renater on this traffic, and the provided QoS would be transparent. Experimental conditions on Renater would be all the times the same, and therefore our experiments are reproducible ;
- Renater integrates two levels of security to avoid attacks coming from outside, but also from inside the network. Practically speaking, we effectively never observed any attack at the measurement and monitoring points we installed in Renater.

The laboratories involved in the generation of traces are ENS in Lyon, LIP6 in Paris, IUT of Mont-de-Marsan, ESSI in Nice, and LAAS in Toulouse. Traffic is captured at these different locations by workstations equipped with DAG cards

[CDG+00] and GPS for a very accurate temporal synchronization. In addition, if we want to perform massive attacks, the target is the LAASNETEXP network at LAAS [OBLG08], which is a network fully dedicated to risky experiments. We can completely saturate it in order to analyze extreme attacking situations.

II.2 Anomalies generation

Anomalies studied and generated in the framework of the METROSEC project consist of more or less significant increases of traffic in terms of volume. We can distinguish two kinds of anomalies :

- anomalies due to legitimate traffic. Let us for instance quote in this class flash crowds (FC). It is important to mention here that such experiments can hardly be fully controlled ;
- anomalies due to illegitimate traffic, as flooding attacks. This traffic, which we can have a full control on, is generated thanks to several classical attacking tools.

Details about anomalies generation are given in what follows.

• **Flash Crowd (FC).** For analyzing the impact on traffic characteristics of a flooding event due to legitimate traffic variations, we triggered flash crowds on a web server. For realism purpose, i.e. humanly random, we chose not to generate them using an automatic program, but to ask our academic colleagues to browse the LAAS web server (http ://www.laas.fr).

• **DDoS attack.** Attacks generated for validating our anomaly detection methods consist of DDoS attacks, launched by using flooding attacking tools (IPERF, HPING2, TRIN00 et TFN2K). We selected well known attacking tools in order to generate malicious traffic as realist as possible.

The IPERF tool [IPE] (under standard Linux environment) aims at generating UDP flows at variable rates, with variable packets rates and payloads. The HPING2 tool [HPI] aims at generating UDP, ICMP and TCP flows, with variable rates (same throughput control parameters as IPERF). Note that with this tool, it is also possible to set TCP flags, and then to generate specific signatures in TCP flows. These two tools were installed on each site of our national distributed platform. At the opposite of TRIN00 and TFN2K (cf. next paragraph), it is not possible to centralize the control on all IPERF and HPING2 entities running at the same time, and then to synchronize attacking sources. One engineer on each site is then in charge of launching attacks at a given predefined time, what induces at the target level a progressive increase of the global attack load.

TRINOO [Tri] and TFN2K [TFN] are two well known distributed attacking tools. They allow the installation on different machines of a program called zombie (or daemon, or bot). This program is in charge of generating the attack towards the target. It is remotely controlled by a master program which commands all the bots. It is possible to constitute an attacking army (or botnet) commanded by one or several masters.

TFN2K bots can launch several kinds of attacks. In addition of classical flooding attacks using UDP, ICMP and TCP protocols (sending of a large number of UDP, ICMP or TCP packet to the victim), many other attacks are possible.

The mixed flooding attack is a mix of UDP flooding, ICMP flooding and TCP SYN flooding. Smurf is an attacking technique based on the concept of amplification : bots use the broadcast address for artificially multiplying the number of attacking packets sent to the target, and then multiplying the power of this attack. TRIN00 bots, on their side, can only perform UDP flooding.

Tool	Attack type	Trace duration	Attack duration	Intensity
Campaign of November-December 2004				
HPING	TCP flooding	1h23mn 1h23mn 3h3mn	15mn 13mn 3mn	30.77% 27.76% 90.26%
		3h3mn 3h3mn 30mn	7mn 8mn 5mn	70.78% 45.62% 91.63%
	UDP flooding	16h20mn (MetroSec-Ref1)	5mn	99.46%
Campaign of June 2005				
IPERF	UDP flooding	1h30 1h30 1h30	30mn 30mn 30mn	17.06% (I) 14.83% 21.51% (III)
		1h30 1h30 1h30	41mn 30mn 30mn	33.29% 39.26% 34.94%
		1h30 1h30 1h30	30mn 30mn 30mn	40.39% 36.93% 56.40%
		1h30	30mn	58.02% (G)
Campaign of March 2006				
TRINOO	UDP flooding	2h 1h 1h	10mn 10mn 10mn	7.0% 22.9% 86.8%
Campaigns from April to July 2006				
TFN2K	UDP flooding	2h 1h 30mn	11mn 10mn 10mn	92% 4.0% 7.0%
	ICMP flooding	1h30 1h	20mn 10mn	13% 9.8%
	TCP SYN flooding	2h 1h	10mn 10mn	12% 33%
	Mixed flooding	1h	10mn	27.3%
	Smurf	1h	10mn	3.82%

Tab. I. Description of attacks in the trace database.

Attacks launched with the different attacking tools (IPERF, HPING2, TRINOO et TFN2K) have been performed by changing frequently the attack characteristics and parameters (duration, DoS flow intensity, size and rate of packets) in order to create different profiles for attacks to be detected afterwards. Main characteristics of generated attacks are summarized in table I. For each configuration, we captured the traffic before, during and after the attack, in order to mix the DoS period with two normal traffic periods. It is important to recall here that most of the times, we tried to generate very low intensity attacks, so that they do not have a significant impact on the global traffic (and therefore be not the cause of average traffic change). This emulates the case of a router receiving the packets from a small number of zombies, and then represents the most interesting problem of our problematic, i.e. detecting DDoS attacks close from their low intensity sources.

Our trace database contains nowadays around thirty captures of such kinds of experiments.

III Comparative evaluation with different anomalies databases

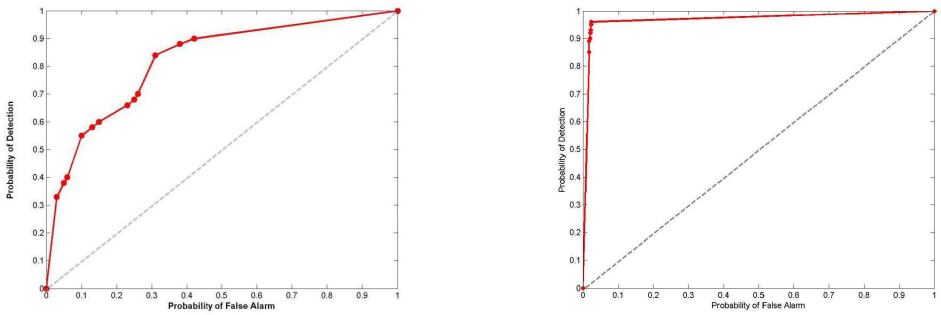
NADA (Network Anomaly Detection Algorithm) is an anomaly detection tool relying on the use of deltoids for detecting significantly anomalous variations

on traffic characteristics. This tool also includes a classification mechanism of anomalies aiming at determining whether detected anomalies are legitimate and their characteristics. NADA has been issued in the framework of the METROSEC project. Thanks to the full control we have on its code, it was easier for us to run experiments with such a tool. Results were also easier to analyze with a full knowledge of the detection tool. For more details, interested readers can refer to [FOM07]. Anyway, as the evaluation methodology we are proposing is of the "black box" kind, it is not necessary to know how a tool is designed and developed for evaluating it. Just let us say that NADA uses a threshold k which aims at determining whether a deltoid corresponds to an anomalous variation. Setting the k threshold allows the configuration of the detection tool sensitivity, in particular related to the natural variability of normal traffic. The rest of this section aims at comparing NADA's performance evaluation results with our method depending on the anomalies database used, METROSEC or KDD'99.

III.1 Evaluation with the MetroSec anomalies database

The statistical evaluation of NADA was performed by using the traces with documented anomalies presented in section II.2. This means a total of 42 different traces, each of them containing at least one DDoS attack ; some contain up to four attacks of small intensity. Six traffic traces with flash crowds were also used for the NADA evaluation. In addition, the documented traces of the METROSEC database can be grouped according to the attacking tools used for generating the attacks/anomalies, and in each group differentiate them according to attack intensities, durations, etc. Such a differentiation is important as it makes possible to measure the sensitivity of the tool under evaluation, i.e. its capability of detecting small intensity anomalies (what of course cannot be done using the KDD'99 database, only able to provide binary results). The intensity and duration of anomalies are two characteristics which significantly have an impact on the capability of profile based IDS/ADS to detect them. Whereas the detection of strong intensity anomalies is well done by most of detection tools, it is in general not the case when small intensity attacks are considered. Therefore, a suited method for evaluating anomaly detection tools performance must be able to count how many times it succeeds or failed in detecting anomalies contained in the traces, and among which some have low intensity.

Figure 1.a shows the ROC curve got by evaluating NADA with the METROSEC anomaly database. It shows the detection probability (PD) according to the probability of false alarms (PF). Each point on the curve represents the average of all results obtained by NADA on all the anomalies of the METROSEC database for a given value of the k parameter, i.e. for a given sensitivity level. The curve analysis shows that NADA is significantly more efficient than a random tool, as all points are over the line $y = x$. Even when the detection probability increases, the NADA performance ROC curve exhibits a very weak false alarm rate. For example, with PD in [60%, 70%], the probability of false alarms is in [10%, 20%], which is a good result.



(a) (b)

Fig. 1. Statistical performances of NADA evaluated based on (a) the MetroSec database and (b) the 10% KDD database. Detection Probability (PD) vs. Probability of false alarms (PF), $PD = f(PF)$.

Base	DoS	Scan	U2R	R2L	Normal
10% KDD	391 458	4107	52	1126	97 277
Corrected KDD	229 853	4166	70	16 347	60 593
Full KDD	3 883 370	41 102	52	1 126	972 780

Tab. II. Characteristics of the KDD'99 database in terms of samples numbers.

III.2 KDD'99

KDD'99 database consists of a set of traces detailed in table II. During the *International Knowledge Discovery and Data Mining Tools* contest [Kd], only 10% of the KDD database were used for the learning phase [HB]. This part of the database contains 22 types of attacks and is a concise version of the full KDD database. This later contains a greater number of attack examples than normal connections, and the types of attacks do not appear in a similar way. Because of their nature, DoS attacks represent the huge majority of the database. On the other hand, the corrected KDD database provides a database with different statistical distributions compared to the databases "10% KDD" or "Full KDD". In addition, it contains 14 new types of attacks.

NADA evaluation was limited to the 10% KDD database. Several reasons motivated this choice : first, despite this database is the simplest, it is also the most used. Second, our intension is to show that the KDD database is not suited for evaluating current ADS (and we will show that the reduced database is enough to demonstrate it). Last, it is a first good test for observing NADA behavior with high intensity DoS attacks. Figure 1.b shows the NADA performance ROC curve obtained with the 10% KDD database. It especially shows the detection probability (PD) according to the probability of false alarms (PF),

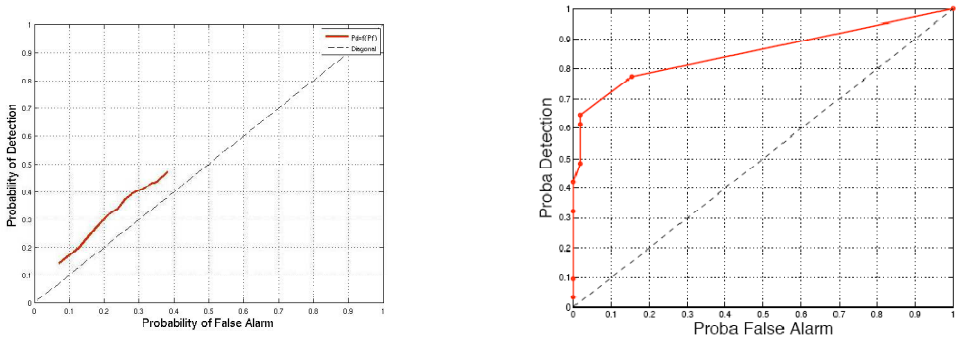
and its analysis shows that NADA got very good results. Applied to the KDD'99 database, NADA exhibits a detection probability close to 90%, and a probability of false alarms around 2%. These results are extremely good, but unfortunately unrealistic if we compare them with the results obtained with the MetroSec database ! DoS attacks are detected in a very reliable way, but certainly because the database is excessively simple : around 98% of attacks are DoS attacks of the same type, presenting in addition very strong intensities. The differences of NADA performances when applied to the two MetroSec and KDD'99 databases underlines the importance of the anomaly database for evaluating profile based IDS and ADS. It is obvious that the MetroSec database presents more complex situations for NADA than KDD'99. Therefore, the evaluation results got with the MetroSec database are certainly closer from the real performance level of NADA than the ones got with KDD'99.

IV Evaluation of 2 other IDS/ADS with the MetroSec database

For illustrating the real statistic evaluation efficiency of the MetroSec method proposed in this paper to distinguish between real capabilities of profile based IDS and ADS, this section shows the comparative results between NADA and two other tools or approaches : the Gamma-FARIMA based approach [SLO+07], and the PHAD tool (experimental Packet Header Anomaly Detection) [MC01]. The Gamma-FARIMA approach is also one of the achievements of the MetroSec projects. PHAD was selected as, in addition of being freely available, it aims at both detecting and classifying anomalies and attacks similarly to what the objectives of MetroSec were, and that lead to the design of NADA and Gamma-FARIMA approach. In addition, PHAD, funded by the American NSF, is said to be the ultimate intrusion and anomaly detection tool. The argumentation of its authors mainly relies on tests lead with KDD'99 traces : on these traces, PHAD gives perfect results.

• **PHAD.** The evaluation of PHAD using the MetroSec database was performed by making the K threshold vary ; K also represents the probability of generating a correct alarm. Figure 2.a shows the ROC curve obtained. Its analysis shows that PHAD behaves just a little bit better than a random detection process.

Figure 3 exhibits the probabilities of detection vs. false alarms with a given threshold K. It is shown that when K increases, PD and PF increase too. Such behavior suggests that PHAD is very inefficient : it detects many anomalies, but most of the times they are false alarms. With small K values, the PD curve gets close from the PF curve, and are not far from the X axis. Such behavior is not a good sign for PHAD performances as it exhibits a problem of too much false alarms... and then certainly a problem of the underlying model or of the learning process. It was also observed that when facing low intensity attacks, PHAD detects none of them.



(a) (b)

Fig. 2. Statistical performances of (a) PHAD and (b) Gamma-FARIMA approach evaluated thanks to the METROSEC database. Detection Probability (PD) vs. Probability of false alarms (PF), $PD = f(PF)$.

When comparing with the performances obtained with the KDD'99 database, there is a big gap. In fact, it seems that PHAD only detects anomalies of the KDD'99, and no other.

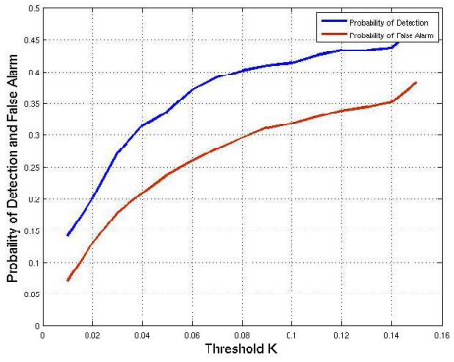


Fig. 3. PHAD : Probability of Detection (PD) and Probability False Alarm (PF) vs. threshold parameter k, $PD = f(k)$ and $PF = g(k)$.

• **Gamma-FARIMA.**

Figure 2.b shows the ROC curve obtained for the Gamma-FARIMA approach evaluated with the METROSEC database. More than 60% of anomalies are detected with a false alarm rate close to 0.

By observing all the ROC curves, it seems that the Gamma-FARIMA approach is the most efficient among the 3 when confronted to the MetroSec database. NADA also exhibits good performances, not too far from the ones of the Gamma-FARIMA approach. These two algorithms designed and developed in the framework of the MetroSec project aim at detecting and classifying known and unknown anomalies. Both algorithms reach this goal by using nevertheless different approaches. While NADA uses simple mathematical functions, the Gamma-FARIMA approach relies on more complex mathematical analysis which are the source of its advantage. However, the NADA simplicity coupled with its high performance level can be of interest. It is particularly true if NADA has to be combined with an identification process of malicious packets.

On the other side, PHAD presents a dramatically low performance level, when confronted to the MetroSec database (but is excellent when confronted to KDD'99 one). By going into a deep analysis of the PHAD algorithm, it seems that the use of 33 different parameters, and assigning a score to each of its anomalous occurrences introduce uncertainty without improving detection accuracy. In addition, these 33 parameters only play a minor role in the detection of most of the attacks [MC01].

V Conclusion

This paper presented a statistical evaluation method of profile based IDS and ADS that relies on a new traffic traces database containing both legitimate and illegitimate anomalies. This paper exhibited that the anomalies database has a major impact on the evaluation results. For example, it was shown that for NADA and PHAD the results obtained with KDD'99 and MetroSec traces are completely different. Such results seems to be well known in this research community, and is part of the common beliefs, but up to our knowledge, it was never published (at least Google does not know any paper demonstrating it).

This paper also shows that the KDD'99 does not permit satisfactory results when considering the evaluation accuracy of the different detection tools. It does not confront them to realistic enough conditions (and then complex enough), and in general the different evaluated tools pass the tests with good marks... marks that are not reproducible once installed in a real environment. Indeed, the KDD'99 evaluation is kind of binary : it only shows whether high intensity attacks can be detected. The MetroSec method plays with the intensities and durations of anomalies for determining levels at which an attack can be detected.

All these observations exhibit one of the big problems. If we assume that the MetroSec database is exhaustive for the current traffic anomalies phenomenon (we tried as much as possible to define generic anomalies on all dimensions of network traffic), how would it be possible to ensure its everlastingness ? Indeed, even if anomalies classes do not radically change, their shapes and intensities (especially related to the network capacities evolutions) will change. And the database, on a more or less long term, will lose some of its realism. Given the strategic aspect of traffic traces for operators, it is not sure at all that we could

continue having traffic traces in which anomalies could be injected. In addition, producing such traces containing anomalies is a very time consuming task which can hardly be supported by a single or few laboratories. It represents one of the main drawbacks against which it does not appear any solution.

The last problem exhibited by this paper is related to the role of experimental data we are exploiting. In fact, we use the same data for designing and validating/evaluating our tools. Thus, PHAD which works perfectly on KDD'99 database (it was designed for detecting anomalies and attacks by taking its inspiration in the KDD'99 database) shows incredibly low performances when tested with the MetroSec database. What would be the results of Gamma-FARIMA or NADA tools if they were evaluated with other anomalies databases than MetroSec or KDD'99 ones (we took our inspiration from these databases when designing our tools) ? Again, the solution would be to have a large number of anomalous traces database in order to separate, at least experimentally, design and evaluation. But we then fall back in the problem previously quoted of the lack of exploitable traffic traces.

Références

[CDG+00] (J.) CLEARY, (S.) DONNELLY, (I.) GRAHAM, (A.) MCGREGOR, and (M.) PEARSON. Design principles for accurate passive measurement. In *Passive and Active Measurements*, Hamilton, New Zealand, April 2000.

[DCW+99] (R.) DURST, (T.) CHAMPION, (B.) WITTEN, (E.) MILLER, and (L.) SPAGNUOLO. Testing and evaluating computer intrusion detection system. *Communications of the ACM*, 42(7), 1999.

[ENW96] (A.) ERRAMILLI, (O.) NARAYAN, and (W.) WILLINGER. Experimental queueing analysis with long-range dependent packet traffic. *ACM/IEEE transactions on Networking*, 4(2) :209–223, 1996.

[FGW98] (A.) FELDMANN, (A.C.) GILBERT, and (W.) WILLINGER. Data networks as cascades : Investigating the multifractal nature of internet wan traffic. In *ACM/SIGCOMM conference on Applications, technologies, architectures, and protocols for computer communication*, pages 42–55, 1998.

[FOM07] (S.) FARRAPOSO, (P.) OWEZARSKI, and (E.) MONTEIRO. Nada - network anomaly detection algorithm. In *Proceedings of the 18th IFIP/IEEE Distributed Systems : Operations and Management (DSOM'2007)*, October 2007.

[HB] S. HETTICH and S. BAY. The uci kdd archive", irvine - university of california, department of information and computer science. http ://kdd.ics.uci.edu, 1999.

[HPI] HPING2. http ://sourceforge.net/projects/hping2.

[IPE] IPERF. The TCP/UDP bandwith Measurement Tool. http ://dast.nlanr .net/Projects/Iperf/.

[Kd] UCI KDD Archive KDD'99 datasets. http ://kdd.ics.uci.edu/databases/ kddcup99/kddcup99.html.

[LFG+00] (R.) LIPPMAN, (D.) FRIED, (I.) GRAF, (J.) HAINES, (K.) KENDALL, (D.) McCLUNG, (D.) WEBER, (S.) WEBSTER, (D.) WYSCHOGROD, (R.) CUNNINGHAM, and (Y.) ZISSMAN. Evaluating intrusion detection systems : The 1998 darpa off-line intrusion detection evaluation. In *DARPA Information Survivability Conference and Exposition*, pages 12–26, 2000.

[LSM99] (W.) LEE, (S.) STOLFO, and (K.) MOK. Mining in a data-flow environment : Experience in network intrusion detection. In *Proceedings of the ACM International Conference on Knowledge Discovery & Data Mining (KDD'99)*, pages 114–124, 1999.

[MC01] (M.) MAHONEY and (P.) CHAN. Phad : Packet header anomaly detection for identifying hostile network traffic. In *Technical Report CS-2001-04. Department of Computer Sciences - Florida Institute of Technology*, 2001.

[MC03] (M.) MAHONEY and (P.) CHAN. An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection. In *Recent Advances in Intrusion Detection (RAID 20003)*, pages 220–237, September 2003.

[McH01] (J.) McHUGH. Testing intrusion detection systems : A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Transactions on Information and System Security*, 3(4) :262–294, 2001.

[MET] METROSEC. http ://www.laas.fr/METROSEC.

[MIT] MIT. Lincoln Laboratory. http ://www.ll.mit.edu/mission/communications /ist/corpora/ideval, 2008.

[OBLG08] (P.) OWEZARSKI, (P.) BERTHOU, (Y.) LABIT, and (D.) GAUCHARD. Laasnetexp : a generic polymorphic platform for network emulation and experiments. In *Proceedings of the 4th International Conference on Testbeds and Research Infrastructure for the Development of Network & Communities (TRIDENTCOM'2008)*, March 2008.

[PKC96] (K.) PARK, (G.) KIM, and (M.) CROVELLA. On the relationship between file sizes, transport protocols, and self-similar network traffic. In *International Conference on Network Protocols*, pages 171–180, Washington, DC, USA, 1996. IEEE Computer Society.

[PW00] (K.) PARK and (W.) WILLINGER. Self-similar network traffic : An overview. In Kihong PARK and Walter WILLINGER, editors, *Self-Similar Network Traffic and Performance Evaluation*, pages 1–38. Wiley (Interscience Division), 2000.

[SLO+07] (A.) SCHERRER, (N.) LARRIEU, (P.) OWEZARSKI, (P.) BORGNAT, and (P.) ABRY. Non-gaussian and long memory statistical characterisations for internet traffic with anomalies. *IEEE Transaction on Dependable and Secure Computing*, 4(1), January 2007.

[TFN] TFN2K. An analysis. http ://packetstormsecurity.org/distributed/TFN2k Analysis-1.3.txt.

[Tri] Trinoo. The DoS Project's "trinoo" distributed denial of service attack tool http ://staff.washington.edu/dittrich/misc/trinoo.analysis.