On the Use of Different Statistical Tests for Alert Correlation – Short Paper

Federico Maggi and Stefano Zanero

Politecnico di Milano, Dip. Elettronica e Informazione
via Ponzio 34/5, 20133 Milano Italy
{fmaggi,zanero}@elet.polimi.it

Abstract. In this paper we analyze the use of different types of statistical tests for the correlation of anomaly detection alerts. We show that the Granger Causality Test, one of the few proposals that can be extended to the anomaly detection domain, strongly depends on good choices of a parameter which proves to be both sensitive and difficult to estimate. We propose a different approach based on a set of simpler statistical tests, and we prove that our criteria work well on a simplified correlation task, without requiring complex configuration parameters.

1 Introduction

One of the most challenging tasks in intrusion detection is to create a unified vision of the events, fusing together alerts from heterogeneous monitoring devices. This *alert fusion* process can be defined as the *correlation* of *aggregated* streams of alerts. *Aggregation* is the grouping of alerts that both are close in time and have similar features; it fuses together different "views" of the same event. Alert *correlation* has to do with the recognition of logically linked alerts. "Correlation" does not necessarily imply "statistical correlation", but statistical correlation based methods are sometimes used to reveal these relationships.

Alert fusion is more complex when taking into account *anomaly detection* systems, because no information on the type or classification of the observed attack is available to the fusion algorithms. Most of the algorithms proposed in the current literature on correlation make use of such information, and are therefore inapplicable to purely anomaly based intrusion detection systems.

In this work, we explore the use of *statistical causality tests*, which have been proposed for the correlation of IDS alerts, and which could be applied to anomaly based IDS as well. We focus on the use of *Granger Causality Test* (GCT), and show that its performance strongly depends on a good choice of a parameter which proves to be sensitive and difficult to estimate. We redefine the causality problem in terms of a simpler statistical test, and experimentally validate it.

2 Problem Statement and State of the Art

The desired output of an *alert fusion* process is a compact, high-level view of what is happening on a (usually large and complex) network. In this work we use

C. Kruegel, R. Lippmann, and A. Clark (Eds.): RAID 2007, LNCS 4637, pp. 167–177, 2007.

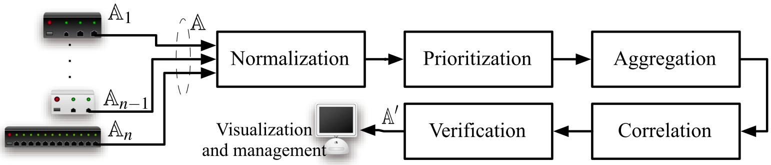

Fig. 1. A diagram illustrating alert fusion terminology as used in this work

a slightly modified version of the terminology proposed in [17]. Alerts streams are collected from different IDS sources, normalized and aggregated; alert correlation is the very final step of the process. In [17] the term "fusion" is used for the phase we name "aggregation", whereas we use the former to denote the whole process. Fig. 1 summarizes the terminology.

In [9] we propose a fuzzy time-based *aggregation* technique, showing that it yields good performance in terms of false positive reduction. Here, we focus on the more challenging *correlation* phase. Effective and generic correlation algorithms are difficult to design, especially if the objective is the reconstruction of complex attack scenarios.

A technique for alert correlation based on state-transition graphs is shown in [3]. The use of finite state automata enables for complex scenario descriptions, but it requires known scenarios signatures. It is also unsuitable for pure anomaly detectors which cannot differentiate among different types of events. Similar approaches, with similar strengths and shortcomings but different formalisms, have been tried with the specification of pre- and post-conditions of the attacks [15], sometimes along with time-distance criteria [12]. It is possible to mine scenario rules directly from data, either in a supervised [2] or unsupervised [5] fashion. Both approaches use alert classifications as part of their rules.

None of these techniques would work for anomaly detection systems, as they rely on alert names or classification to work. The best examples of algorithms that do not require such features are based on time-series analysis and modeling. For instance, [19] is based on the construction of time-series by counting the number of alerts occurring into sampling intervals; the exploitation of trend and periodicity removal algorithms allows to filter out predictable components, leaving *real* alerts only as the output. More than a correlation approach, this is a false-positive and noise-suppression approach, though.

The correlation approach investigated in [14] and based on the GCT also does not require prior knowledge, and it drew our attention as one of the few viable proposal for anomaly detection alert correlation in earlier literature. We will describe and analyze this approach in detail in Section 4.

3 Problems in Evaluating Alert Correlation Systems

Evaluation techniques for alert fusion systems are still limited to a few proposals, and practically and theoretically challenging to develop [9]. Additionally, the common problem of the lack of reliable sources of data for benchmarking impacts heavily also on the evaluation of correlation systems. Ideally, we need both host and network datasets, fully labeled, with complex attack scenarios described in detail. These data should be freely available to the scientific community. These requirements rule out real-world dumps.

The only datasets of this kind effectively available are the ones by DARPA (IDEVAL datasets). Of course, since this data set was created to evaluate IDS sensors and not to assess correlation tools, it does not include sensor alerts. The alerts have to be generated by running various sensors on the data. The 1999 dataset [7], which we used for this work, has many known shortcomings. Firstly, it is evidently and hopelessly outdated. Moreover, a number of flaws have been detected and criticized in the network traces [10, 11]. More recently, we analyzed the host-based system call traces, and showed [8, 21] that they are ridden with problems as well.

For this work these basic flaws are not extremely dangerous, since the propagation of attack effects (from network to hosts) is not affected by any of the known flaws of IDEVAL, and in fact we observed it to be quite realistically present. What could be a problem is the fact that intrusion scenarios are too simple and extremely straightforward. Additionally, many attacks are not detectable in both network and host data (thus making the whole point of correlation disappear). Nowadays, networks and attackers are more sophisticated and attack scenarios are much more complex than in 1999, operating at various layers of the network and application stack.

The work we analyze closely in the following [14] uses both the DEFCON 9 CTF dumps and the DARPA Cyber Panel *Correlation Technology Validation* (CTV) [4] datasets for the evaluation of an alert correlation prototype. The former dataset is not labeled and does not contain any background traffic, so in fact (as the authors themselves recognize) it cannot be used for a proper evaluation, but just for qualitative analysis. On the contrary, the DARPA CTV effort, carried out in 2002, created a complex testbed network, along with background traffic and a set of attack scenarios. The alerts produced by various sensors during these attacks were collected and given as an input to the evaluated correlation tools. Unfortunately, this dataset is not available for further experimentation.

For all the previous reasons, in our testing we will use the IDEVAL dataset with the following simplification: we will just try to correlate the stream of alerts coming from a single *host-based IDS* (HIDS) sensor with the corresponding alerts from a single *network-based IDS* (NIDS), which is monitoring the whole network. To this end, we ran two anomaly-based IDS prototypes (both described in [8, 20, 21]) on the whole IDEVAL testing dataset. We ran the NIDS prototype on `tcpdump` data and collected 128 alerts for attacks against the host `pascal.eyrie.af.mil` [6]. The NIDS also generated 1009 alerts related to other hosts. Using the HIDS prototype we generated 1070 alerts from the dumps of the host `pascal.eyrie.af.mil`. With respect to these alerts, the NIDS was capable of detecting almost 66% of the attacks with less than 0.03% of false positives; the HIDS performs even better with a detection rate of 98% and 1.7% of false positives.

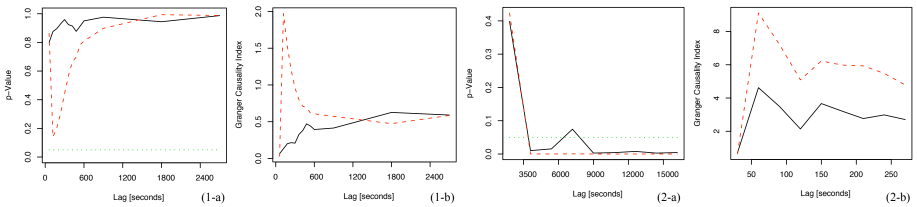

Fig. 2. p-value (-a) and GCI (-b) vs. p with $w = w_1 = 60s$ (1-) and $w = w_2 = 1800s$ (2-) “$NetP(k) \rightsquigarrow HostP(k)$” (dashed line), “$HostP(k) \rightsquigarrow NetP(k)$” (solid line)

In the following, we use this shorthand notation: Net is the substream of all the alerts generated by the NIDS. $HostP$ is the substream of all the alerts generated by the HIDS installed on `pascal.eyrie.af.mil`, while $NetP$ regards all the alerts (with `pascal` as a target) generated by the NIDS; finally, $NetO = Net \backslash NetP$ indicates all the alerts (with all but `pascal` as a target) generated by the NIDS.

4 The Granger Causality Test

In [14] Qin and Lee propose an interesting algorithm for alert correlation which seems suitable also for anomaly-based alerts. Alerts with the same feature set are grouped into collections of time-sorted items belonging to the same “type” (following the concept of type of [19]). Subsequently, frequency time series are built, using a fixed size sliding-window: the result is a time-series for each collection of alerts. The prototype then exploits the GCT [16], a statistical hypothesis test capable of discovering causality relationships between two time series when they are originated by linear, stationary processes. The GCT gives a stochastic measure, called *Granger Causality Index* (GCI), of how much of the history of one time series (the supposed cause) is needed to “explain” the evolution of the other one (the supposed consequence, or target). The GCT is based on the estimation of two models: the first is an *Auto Regressive* model (AR), in which future samples of the target are modeled as influenced only by past samples of the target itself; the second is an *Auto Regressive Moving Average eXogenous* (ARMAX) model, which also takes into account the supposed cause time series as an exogenous component. A statistical F-test built upon the model estimation errors selects the best-fitting model: if the ARMAX fits better, the cause effectively influences the target.

In [14] the unsupervised identification of “causally related” events is performed by repeating the above procedure for each couple of time-series. The advantage of the approach is that it does not require prior knowledge (even if it may use attack probability values, if available, for an optional prioritization phase). However, in a previous work [9] we showed that the GCT fails however in recognizing “meaningful” relationships between IDEVAL attacks.

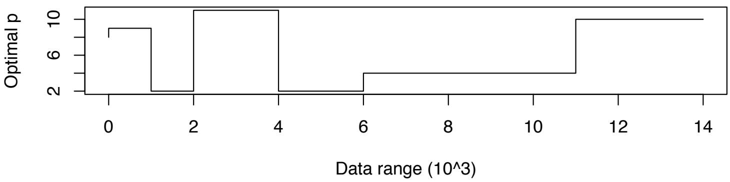

Fig. 3. The optimal time lag $\hat{p}$ given by the AIC criterion strongly varies over time

We tested the sensitivity of the GCT to the choice of two parameters: the sampling time, w, and the time lag p (that is, the order of the AR). In our simple experiment, the expected result is that $NetP \rightsquigarrow HostP$, and that $HostP \not\rightsquigarrow NetP$ (the $\rightsquigarrow$ indicates "causality" while $\not\rightsquigarrow$ is its negation). In [14] the sampling time was arbitrarily set to $w = 60s$, while the choice of p is not documented. However, our experiments show that the choice of these parameters can strongly influence the results of the test. In Fig. 2 (1-a/b) we plotted the p-value and the GCI of the test for different values of p ($w = 60s$). In particular, the dashed line corresponds to the test $NetP(k) \rightsquigarrow HostP(k)$, and the solid line to the test $HostP(k) \rightsquigarrow NetP(k)$. We recall that if the p-value is lower than the significance level, the null hypothesis is refused. Notice how different choices of p can lead to inconclusive or even opposite results. For instance, with $\alpha = 0.20$ and with $2 \leq p \leq 3$, the result is that $NetP(k) \rightsquigarrow HostP(k)$ and that $HostP(k) \not\rightsquigarrow NetP(k)$. As we detailed in [9] (Fig. 2 (2-a/b)), other values of p lead to awkward result that both $HostP(k) \rightsquigarrow NetP(k)$ and $NetP(k) \rightsquigarrow HostP(k)$.

A possible explanation is that the GCT is significant only if both the linear regression models are optimal, in order to calculate the correct residuals. If we use the *Akaike Information Criterion* (AIC) [1] to estimate the optimal time lag $\hat{p}$ over different windows of data, we find out that $\hat{p}$ wildly varies over time, as it is shown in Fig. 3. The fact that there is no stable optimal choice of p, combined with the fact that the test result significantly depends on it, makes us doubt that the Granger causality test is a viable option for general alert correlation. The choice of w seems equally important and even more difficult to perform, except by guessing.

Of course, our testing is not conclusive: the IDEVAL alert sets may simply not be adequate for showing causal relationships. Another, albeit more unlikely, explanation, is that the Granger causality test may not be suitable for anomaly detection alerts: in fact, in [14] it has been tested on misuse detection alerts. But in fact there are also theoretical reasons to doubt that the application of the Granger test can lead to stable, good results. First, the test is asymptotic w.r.t. p meaning that the results reliability decreases as p increases because of the loss of degrees of freedom. Second, it is based on the strong assumption of *linearity* in the auto-regressive model fitting step, which strongly depends on the observed phenomenon. In the same way, the stationarity assumption of the model does not always hold.

5 Modeling Alerts as Stochastic Processes

Instead of interpreting alert streams as time series (as proposed by the GCT-based approach), we propose to change point of view by using a stochastic model in which alerts are modeled as (random) events in time. This proposal can be seen as a formalized extension of the approach introduced in [17], which correlates alerts if they are fired by different IDS within a "negligible" time frame, where "negligible" is defined with a crisp, fixed threshold.

For simplicity, once again we describe our technique in the simple case of a single HIDS and a single NIDS which monitors the whole network. The concepts, however, can be easily generalized to take into account more than two alert streams, by evaluating them couple by couple. For each alert, we have three essential information: a timestamp, a "target" host (fixed, in the case of the HIDS, to the host itself), and the generating sensor (in our case, a binary value).

We reuse the scenario and data we already presented in Section 4 above. With a self-explaining notation, we also define the following random variables: T_{NetP} are the arrival times of network alerts in $NetP$ (T_{NetO}, T_{HostP} are similarly defined); ε_{NetP} (ε_{NetO}) are the delays (caused by transmission, processing and different granularity in detection) between a specific network-based alert regarding `pascal` (not `pascal`) and the corresponding host-based one. The actual values of each $T_{(\cdot)}$ is nothing but the set of timestamps extracted from the corresponding alert stream. We reasonably assume that ε_{NetP} and T_{NetP} are stochastically independent (the same is assumed for ε_{NetO} and T_{NetO}).

In an *ideal* correlation framework with two equally perfect IDS with a 100% DR and 0% FPR, if two alert streams are correlated (i.e., they represent independent detections of the same attack occurrences by different IDSes [17]), they also are "close" in time. $NetP$ and $HostP$ should evidently be an example of such a couple of streams. Obviously, in the real world, some alerts will be missing (because of false negatives, or simply because some of the attacks are detectable only by a specific type of detector), and the distances between related alerts will therefore have some higher variability. In order to account for this, we can "cut off" alerts that are too far away from a corresponding alert in the other time series, presuming them to be singletons. In our case, knowing that single attacks did not last more than $400s$ in the original dataset, we tentatively set a cutoff threshold at this point.

Under the given working assumptions and the proposed stochastic model, we can formalize the correlation problem as a set of two statistical hypothesis tests:

$$H_0 : T_{HostP} \neq T_{NetP} + \varepsilon_{NetP} \; vs. \; H_1 : T_{HostP} = T_{NetP} + \varepsilon_{NetP} \tag{1}$$

$$H_0 : T_{HostP} \neq T_{NetO} + \varepsilon_{NetO} \; vs. \; H_1 : T_{HostP} = T_{NetO} + \varepsilon_{NetO} \tag{2}$$

Let $\{t_{i,k}\}$ be the observed timestamps of T_i $\forall i \in \{HostP, NetP, NetO\}$, the meaning of the first test is straightforward: within a random amount of time, ε_{NetP}, the occurring of a host alert, $t_{HostP,k}$, is preceded by a network alert, $t_{NetP,k}$. If this does not happen for a statistically significant amount of events, the test result is that alert stream T_{NetP} is *uncorrelated* to T_{HostP}; in this case,

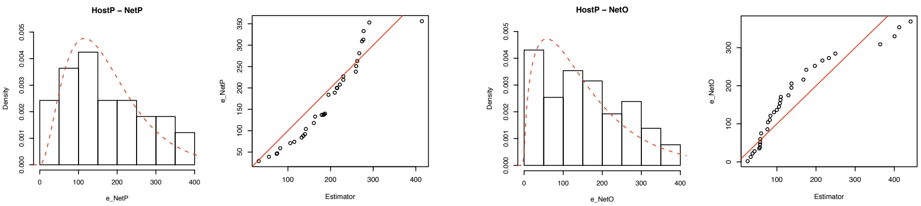

Fig. 4. Histograms vs. est. density (red dashes) and Q-Q plots, for both $\hat{f}_O$ and $\hat{f}_P$

we have *enough statistical evidence* for refusing H_1 and accepting the null one. Symmetrically, refusing the null hypothesis of the second test means that the $NetO$ alert stream (regarding to all hosts but `pascal`) is correlated to the alert stream regarding `pascal`.

Note that, the above two tests are strongly related to each other: in an ideal correlation framework, it cannot happen that both "$NetP$ is correlated to $HostP$" and "$NetO$ is correlated to $HostP$": this would imply that the network activity regarding to all hosts but `pascal` (which raises $NetO$) has to do with the host activity of `pascal` (which raises $HostP$) with the same order of magnitude of $NetP$, that is an intuitively contradictory conclusion. Therefore, the second test acts as a sort of "robustness" criterion.

From our alerts, we can compute a sample of ε_{NetP} by simply picking, for each value in $NetP$, the value in $HostP$ which is closest, but greater (applying a threshold as defined above). We can do the same for ε_{NetO}, using the alerts in $NetO$ and $HostP$.

The next step involves the *choice of the distributions* of the random variables we defined above. Typical distributions used for modeling random occurrences of timed events fall into the family of exponential *Probability Density Functions* (PDF)s [13]. In particular, we decided to fit them with Gamma PDFs, because our experiments show that such a distribution is a good choice for both the ε_{NetP} and ε_{NetO}.

The estimation of the PDF of ε_{NetP}, $f_P := f_{\varepsilon_{NetP}}$, and ε_{NetO}, $f_O := f_{\varepsilon_{NetO}}$, is performed using the well known *Maximum Likelihood* (ML) technique [18] as implemented in the `GNU R` software package: the results are summarized in Fig. 4. f_P and f_O are approximated by Gamma$[3.0606, 0.0178]$ and Gamma $[1.6301, 0.0105]$, respectively (standard errors on parameters are 0.7080, 0.0045 for f_P and 0.1288, 0.009 for f_O). From now on, the estimator of a given density f will be indicated as $\hat{f}$.

Fig. 4 shows histograms vs. estimated density (red, dashed line) and quantile-quantile plots (Q-Q plots), for both $\hat{f}_O$ and $\hat{f}_P$. We recall that Q-Q plots are an intuitive graphical "tool" for comparing data distributions by plotting the quantile of the first distribution against the quantile of the other one.

Considering that the samples sizes of $\varepsilon_{(\cdot)}$ are around 40, Q-Q plots empirically confirms our intuition: in fact, $\hat{f}_O$ and $\hat{f}_P$ are both able to explain real data well, within inevitable but negligible estimation errors. Even if $\hat{f}_P$ and $\hat{f}_O$ are both Gamma-shaped, it must be noticed that they significantly differ in their

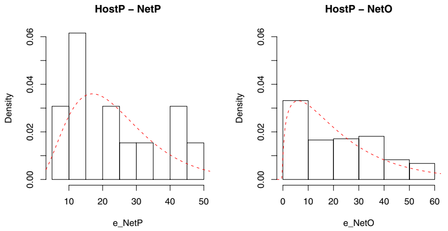

Fig. 5. Histograms vs. est. density (red dashes) for both $\hat{f_O}$ and $\hat{f_P}$ (IDEVAL 1998)

parametrization; this is a very important result since it allows to set up a proper criterion to decide whether or not ε_{NetP} and ε_{NetO} are generated by the same phenomenon.

Given the above estimators, a more precise and robust hypotheses test can be now designed. The Test 1 and 2 can be mapped into two-sided *Kolmogorov-Smirnov* (KS) tests [13], achieving the same result in terms of decisions:

$$H_0 : \varepsilon_{NetP} \sim f_P \ vs. \ H_1 : \varepsilon_{NetP} \not\sim f_P \qquad (3)$$

$$H_0 : \varepsilon_{NetO} \sim f_O \ vs. \ H_1 : \varepsilon_{NetO} \not\sim f_O \qquad (4)$$

where the symbol $\sim$ means "has the same distribution of". Since we do not know the real PDFs, estimators are used in their stead. We recall that the KS-test is a *non-parametric* test to compare a sample (or a PDF) against a PDF (or a sample) to check how much they differs from each other (or how much they fit). Such tests can be performed, for instance, with `ks.test()` (a `GNU R` native procedure): resulting p-values on IDEVAL 1999 are 0.83 and 0.03, respectively.

Noticeably, there is a significant statistical evidence to accept the null hypothesis of Test 3. It seems that the ML estimation is capable of correctly fitting a Gamma PDF for f_P (given ε_{NetP} samples), which double-checks our intuition about the distribution. The same does not hold for f_O: in fact, it cannot be correctly estimated, with a Gamma PDF, from ε_{NetO}. The low p-value for Test 4 confirms that the distribution of ε_{NetO} delays is completely different than the one of ε_{NetP}. Therefore, our criterion doest not only recognize noisy delay-based relationships among alerts stream *if they exists*; it is also capable of detecting if such a correlation does not hold.

We also tested our technique on alerts generated by our NIDS/HIDS running on IDEVAL 1998 (limiting our analysis to the first four days of the first week), in order to cross-validate the above results. We prepared and processed the data with the same procedures we described above for the 1999 dataset. Starting from almost the same proportion of host/net alerts against either `pascal` or other hosts, the ML-estimation has computed the two Gamma densities shown in Fig. 5: f_P and f_O are approximated by $\mathrm{Gamma}(3.5127, 0.1478)$ and $\mathrm{Gamma}(1.3747, 0.0618)$, respectively (standard errors on estimated parameters are 1.3173, 0.0596 for f_P and 0.1265, 0.0068 for f_O). These parameter are very similar to the ones we estimated for the IDEVAL 1999 dataset. Furthermore, with p-values of 0.51 and 0.09, the two KS tests confirm the same statistical discrepancies we observed on the 1999 dataset.

The above numerical results show that, by interpreting alert streams as random processes, there are several (stochastic) dissimilarities between net-to-host delays belonging to the same net-host attack session, and net-to-host delays belonging to different sessions. Exploiting these dissimilarities, we may find out the correlation among streams in an unsupervised manner, without the need to predefine any parameter.

6 Conclusions

In this paper we analyzed the use of of different types of statistical tests for the correlation of anomaly detection alerts, a problem which has little or no solutions available today. One of the few correlation proposals that can be applied to anomaly detection is the use of a *Granger Causality Test* (GCT). After discussing a possible testing methodology, we observed that the IDEVAL datasets traditionally used for evaluation have various shortcomings, that we partially addressed by using the data for a simpler scenario of correlation, investigating only the link between a stream of host-based alerts for a specific host, and the corresponding stream of alerts from a network based detector.

We examined the usage of a GCT as proposed in earlier works, showing that it relies on the choice of non-obvious configuration parameters which significantly affect the final result. We also showed that one of these parameters (the order of the models) is absolutely critical, but cannot be uniquely estimated for a given system. Instead of the GCT, we proposed a simpler statistical model of alert generation, describing alert streams and timestamps as stochastic variables, and showed that statistical tests can be used to create a reasonable criterion for distinguishing correlated and non correlated streams. We proved that our criteria work well on the simplified correlation task we used for testing, without requiring complex configuration parameters.

This is an exploratory work, and further investigations of this approach on real, longer sequences of data, as well as further refinements of the tests and the criteria we proposed, are surely needed. Another possible extension of this work is the investigation of how these criteria can be used to correlate anomaly and misuse-based alerts together, in order to bridge the gap between the existing paradigms of intrusion detection.

Acknowledgments

We need to thank prof. Ilenia Epifani and prof. Giuseppe Serazzi for their helpful comments, as well as the anonymous reviewers.

References

1. Akaike, H.: A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19(6), 716–723 (1974)
2. Dain, O., Cunningham, R.: Fusing heterogeneous alert streams into scenarios. In: Proc. of the ACM Workshop on Data Mining for Security Applications, November 2001, pp. 1–13. ACM Press, New York (2001)
3. Eckmann, S., Vigna, G., Kemmerer, R.: STATL: An attack language for state-based intrusion detection. In: Proceedings of the ACM Workshop on Intrusion Detection, Atene, November 2000, ACM Press, New York (2000)
4. Haines, J., Ryder, D.K., Tinnel, L., Taylor, S.: Validation of sensor alert correlators. IEEE Security and Privacy 01(1), 46–56 (2003)
5. Julisch, K., Dacier, M.: Mining intrusion detection alarms for actionable knowledge. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, New York, NY, USA, pp. 366–375. ACM Press, New York (2002)
6. Lippmann, R., Haines, J.W., Fried, D.J., Korba, J., Das, K.: The 1999 darpa off-line intrusion detection evaluation. Comput. Networks 34(4), 579–595 (2000)
7. Lippmann, R., Haines, J.W., Fried, D.J., Korba, J., Das, K.: Analysis and results of the 1999 DARPA off-line intrusion detection evaluation. In: Debar, H., Mé, L., Wu, S.F. (eds.) RAID 2000. LNCS, vol. 1907, Springer, Heidelberg (2000)
8. Maggi, F., Matteucci, M., Zanero, S.: Detecting intrusions through system call sequence and argument analysis (submitted for publication, 2006)
9. Maggi, F., Matteucci, M., Zanero, S.: Reducing false positives in anomaly detectors through fuzzy alert aggregation (submitted for publication, 2006)
10. Mahoney, M.V., Chan, P.K.: An analysis of the 1999 DARPA / Lincoln laboratory evaluation data for network anomaly detection. In: Vigna, G., Krügel, C., Jonsson, E. (eds.) RAID 2003. LNCS, vol. 2820, pp. 220–237. Springer, Heidelberg (2003)
11. McHugh, J.: Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by lincoln laboratory. ACM Trans. on Information and System Security 3(4), 262–294 (2000)
12. Ning, P., Cui, Y., Reeves, D.S., Xu, D.: Techniques and tools for analyzing intrusion alerts. ACM Trans. Inf. Syst. Secur. 7(2), 274–318 (2004)
13. Pestman, W.R.: Mathematical Statistics: An Introduction. Walter de Gruyter, Berlin (1998)
14. Qin, X., Lee, W.: Statistical causality analysis of INFOSEC alert data. In: Vigna, G., Krügel, C., Jonsson, E. (eds.) RAID 2003. LNCS, vol. 2820, pp. 73–93. Springer, Heidelberg (2003)
15. Templeton, S.J., Levitt, K.: A requires/provides model for computer attacks. In: NSPW '00: Proceedings of the 2000 workshop on New security paradigms, New York, NY, USA, pp. 31–38. ACM Press, New York (2000)
16. Thurman, W.N., Fisher, M.E.: Chickens, eggs, and causality, or which came first? American Journal of Agricultural Economics (1998)
17. Valeur, F., Vigna, G., Kruegel, C., Kemmerer, R.A.: A comprehensive approach to intrusion detection alert correlation. IEEE Trans. Dependable Secur. Comput. 1(3), 146–169 (2004)
18. Venables, W., Ripley, B.: Modern Applied Statistics with S. Springer, Heidelberg (2002)
19. Viinikka, J., Debar, H., Mé, L., Séguier, R.: Time series modeling for IDS alert management. In: Proc. of the 2006 ACM Symp. on Information, computer and communications security, New York, NY, USA, pp. 102–113. ACM Press, New York (2006)

20. Zanero, S.: Analyzing TCP traffic patterns using self organizing maps. In: Roli, F., Vitulano, S. (eds.) ICIAP 2005. LNCS, vol. 3617, pp. 83–90. Springer, Heidelberg (2005)
21. Zanero, S.: Unsupervised Learning Algorithms for Intrusion Detection. PhD thesis, Politecnico di Milano T.U., Milano, Italy (May 2006)