Tackling the Big Data 4 Vs for Anomaly Detection

José Camacho, Gabriel Maciá-Fernández, Jesús Díaz-Verdejo and Pedro García-Teodoro
Dpt. of Signal Theory, Telematics and Communications - CITIC, University of Granada
Email:{josecamacho,gmacia,jedv,pgteodor}@ugr.es

***Abstract*—In this paper, a framework for anomaly detection and forensics in Big Data is introduced. The framework tackles the Big Data 4 Vs: Variety, Veracity, Volume and Velocity. The varied nature of the data sources is treated by transforming the typically unstructured data into a highly dimensional and structured data set. To overcome both the uncertainty (low veracity) and high dimension introduced, a latent variable method, in particular Principal Component Analysis (PCA), is applied. PCA is well known to present outstanding capabilities to extract information from highly dimensional data sets. However, PCA is limited to low size, thought highly multivariate, data sets. To handle this limitation, a kernel computation of PCA is employed. This avoids computational problems due to the size (number of observations) in the data sets and allows parallelism. Also, hierarchical models are proposed if dimensionality is extreme. Finally, to handle high velocity in analyzing time series data flows, the Exponentially Weighted Moving Average (EWMA) approach is employed. All these steps are discussed in the paper, and the VAST 2012 mini challenge 2 is used for illustration.**

I. Introduction

Network monitoring for security (NMS) shares a number of features with other Big Data problems, the so-called 4 Vs [1]:

(i) Variety: Data are varied in nature. Different sources, including unstructured and structured information, need to be properly combined in order to make the most of the analysis.

(ii) Veracity: The search for valuable information in large data sets is very much like the problem of finding the needle in a haystack. Big Data present low signal to noise ratio, and data mining techniques are needed to find patterns or trends of practical use, which are more reliable than punctual measures.

(iii) Volume: The amount of data that needs to be handled simultaneously makes parallelism a must. Exabytes, zettabytes, and even higher amounts of data are described in Big Data applications.

(iv) Velocity: In Big Data problems, a high rate of sampling is common. This further complicates the analysis and makes parallelism even more necessary.

Big Data treatment in NMS is gaining in importance for the R&D community (operators, service providers and system administrators in general). In NMS, typical data sources are network traffic (compressed or not), a variety of logs–Firewalls, Intrusion Detection Systems (IDSs), Operative Systems, Applications–which range from structured to unstructured data, Netflow data, statistics from management protocols, etc. The different sources of data need to be properly correlated in order to detect and correctly interpret anomalous events. This is further complicated by the volume and velocity of data: a simple network link at 1Gbps can generate more than a Terabyte of data in a daytime and a Petabyte in a year.

This paper discusses the extension of state-of-the-art techniques for anomaly detection and data visualization to Big Data. The proposed framework is tested in the VAST 2012 2nd mini challenge [2]. The rest of the paper is organized as follows. In Section II, the proposed system is introduced. Section III describes the VAST 2012 2nd mini challenge. The data of the challenge will be used throughout the paper for illustration purposes and without loss of generality, since the approach is broadly applicable. The following three sections are devoted to deal with the 4 Vs: Variety (Section IV), Veracity (Section V), Volume and Velocity (Section VI). In each of them, original contributions are highlighted and related works described. Also, the specific solutions to address the different Vs are illustrated using the data of the challenge. Finally, Section VII summarizes the conclusions of the work.

II. System Description

A diagram of the proposed anomaly detection system for NMS is depicted in Figure 1. The input to the system is a number of heterogeneous data sources, like network traffic, IDS and firewall logs, etc. Firstly, incoming data are preprocessed in two steps. In the first step, a number of features are computed from the data stream of each source. Also, the features corresponding to different sources are combined. This is what we call the parameterization step. Then, the new features are used to update a set of intermediate data structures which represent the current state of the network. The updating strategy follows an Exponentially Weighted Moving Average (EWMA) approach. The intermediate data structures are passed to the Analysis & Visualization system. There, they are used to compute a Principal Component Analysis (PCA) model. From the PCA model, a number of visualization techniques (MEDA, Time lines and oMEDA) are computed in order to detect and interpret anomalies and to unveil the common trends in the data. Since PCA is the core of the proposal, next section is devoted to introduce this technique.

A. Principal Component Analysis

PCA is applied to two-way data sets, where M variables or features are measured/computed for N observations or objects. The aim of PCA is to find the subspace of maximum variance in the M-dimensional feature space. The original features are linearly transformed into the Principal Components (PCs). These are obtained from the eigenvectors of $\mathbf{XX} := \mathbf{X}^T \cdot \mathbf{X}$, typically for mean centered $\mathbf{X}$. PCA follows the expression:

$$\mathbf{X} = \mathbf{T}_A \cdot \mathbf{P}_A^t + \mathbf{E}_A, \quad (1)$$

978-1-4799-3088-3/14/$31.00 ©2014 IEEE

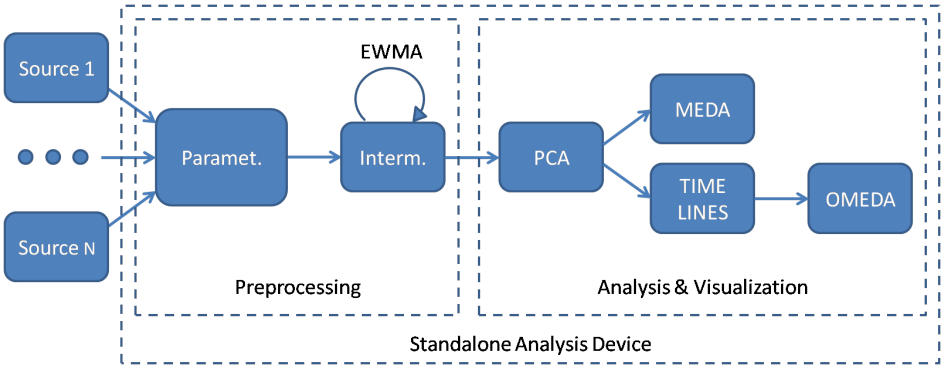

Fig. 1. Analysis strategy including tools and methodologies involved.

where $\mathbf{T}_A$ is the $N \times A$ score matrix, $\mathbf{P}_A$ is the $M \times A$ loading matrix and $\mathbf{E}_A$ is the $N \times M$ matrix of residuals.

PCA can handle very large dimensional data sets. For instance, PCA is a common analysis tool in genomic data [3], which can have up to a million of features. This capability is of utmost importance for NMS because a high number of features from multiple and variate data sources can be taken into account at the same time. Also, common and anomalous patterns can be interpreted from the joined contribution of the features involved. The more features in the model, the more information about these patterns can be extracted. In comparison, other network analysis tools are typically limited to univariate or low dimensional time series signals [4].

III. Case Study

The VAST 2012 2nd mini challenge [2] presents a corporate network where security incidents occur during two days. In particular, some staff members report unwanted messages and a non-legitime anti-virus program appearing on their monitors. Also, their systems seem to be running more slowly than normal. In summary, a forensics operation is required to discover the most relevant security events and their root causes.

The network infrastructure for the challenge is shown in Figure 1 of the supplemental material at [5]. Around 4,000 workstations and approximately 1,000 servers operate 24 hours a day. The data provided with the VAST 2012 mini challenge 2 consist of Cisco ASA firewall logs including a total of 23,711,341 data records, and IDS logs including 35,948 data records. The data set details are available at [6], from which the high complexity of the problem should be concluded.

IV. Addressing Variety

Variety in the information coming from heterogeneous data sources is addressed here by measuring/computing a high number of features from the incoming data. Although parameterization is a well-known issue in the literature, our approach presents some specific points worth to be mentioned. On the one hand, the parameterization has to take into account the analysis method employed, PCA. On the other hand, the nature of the data should also be considered, and therefore some level of problem-specific expertise is required. To the best of our knowledge, this is the first time that feature definition for PCA and NMS is discussed from a general perspective.

PCA is based on (co)variance, which is a quantitative measure. Therefore, quantitative features should be defined for the analysis. Another relevant issue is that PCA anomaly detection is based on the identification of very high or very low values in the features and combinations of them. The features should be selected taking this into account as well.

In NMS systems, most of the information comes in the form of logs. Unfortunately, almost each vendor defines its own log format. The key for the selection of features in NMS systems is to identify means of translating log information into quantitative features. We propose the feature-as-a-counter approach, so that features are basically counters for the number of associated events. Each feature is defined as the number of times a given event takes place, according to the logs information, during a time window w. This is a very general definition which we have found to be capable of representing most of the types of information of interest in NMS. The window size w may be defined so that scarce measurement matrices are avoided, that is, it should be big enough so that a given event takes places more than once in most intervals. Also, by properly selecting w, the combination of different and variate sources may be simple (for different w values in different sources) or even straightforward (for the same w).

A. VAST 2012

Data from the firewall and IDS logs in the VAST 2012 mini challenge 2 have been parsed into M-dimensional vectors representing time intervals of one minute, as the resolution of the IDS entries prevent us from using shorter intervals. A total of 2,350 observations, each one with the information for one minute, are obtained. Notice that the parameterization is also a means to reduce the volume of data. In this case, a very large data set is parsed into a reduced number of observations. Although this may not be considered to be a Big Data set, it will work for illustration purposes in the present paper.

For every sampling period of one minute, we have defined a set of 112 features that represent the information from the two data sources: 69 features for the firewall log and 43 for the IDS log. By using the same sampling period, both data sources are

seamlessly combined by appending the data matrices. Every feature is labelled as: `source_typelabel`, where `source` indicates if the feature is coming from the firewall (fw) or from the IDS (ids) logs, `type` indicates the type of the variable (*e.g.*, `ip` stands for a range of IP addresses and `p` for port) and `label` gives some specific information. For example, the variable `ids_pdns` collects the number of IDS logs related to incidents where the DNS port is present.

V. Addressing Veracity

Latent variable models such as PCA have two features which make them especially suited for addressing uncertain data: they can handle highly dimensional data sets, which improves the confidence on the analysis findings, and they can identify and patterns within the data. PCA has been mainly applied to anomaly detection, *e.g.,* [7]. However, another main application of this technique is data visualization. PCA is suited to visualize a wide variety of data sources. On the contrary, most visual techniques used in NMS are specific to the type of data, like NetFlow data [8] or traffic data [9].

In this section, we combine the most extended charts for PCA-based anomaly detection, the Hotelling T^2 and the Q-statistic [10], with some recently proposed visualization techniques, namely MEDA [11] and oMEDA [12]. We will show that this combination is extremely powerful to detect and interpret anomalies and worth to be extended to Big Data scenarios. This is the first time that this combination of tools is proposed in the scientific literature.

MEDA plots are color maps of size $M \times M$, for M the number of features, where dark colors identify strong positive (red) or negative (blue) relationships among features (see Figure 2). To improve the visualization of MEDA, in this work we propose the use of the seriation method in [13], where the features are reordered in the plot according to a similarity criterion. Thus, groups of features can be easily identified and close groups of features present some degree of relationship.

MEDA does not present any temporal information. Two very useful plots with temporal information are the time lines commonly employed in on-line industrial process monitoring [10]: the leverage or Hotelling T^2 plot and the residuals or Q-statistic plot (see Figure 3). In the network security context, the relevant events can be easily identified from these time line plots as those points that present either a high leverage or a high Q-statistic, or both.

A main step in the forensics part of the system is to connect the anomalous events with the related features. In this paper, the oMEDA plots are used for that purpose. An oMEDA plot is a bar plot of the features where very high or very low values detect the most relevant features of a group of observations (*e.g.*, an interval of consecutive outliers) in comparison with any reference (*e.g.*, the rest of the data).

A. VAST 2012

After the data preprocessing described in Section IV.A for the VAST data, a matrix of 2,350 time ordered observations on 112 features is obtained. From these features, 17 did not show any variation around its average and were discarded from the analysis. The dimension of the remaining data set is $2,350 \times 95$. It is necessary to select the number of PCs to be considered. In the present example, 6 PCs captured almost 60% of the total variance. This illustrates that although the data set is highly dimensional, a few components can considerably simplify its interpretation.

#	Int.	Relevant features
7	23-26	ids_ipfwhq ids_prio2 ids_leak ids_lvnc
9	41-44	ids_ldns ids_privacy ids_pdns ids_prio1
10	45-51	ids_pnstd ids_ipws ids_ipdc ids_psmb ids_command ids_lnetbios ids_prio3
15	78-80	ids_misc ids_ipweb ids_lirc
16	82-83	fw_opempty fw_empty

TABLE I
SELECTED GROUPS OF RELEVANT FEATURES IN THE SECOND VAST 2012 MINI CHALLENGE DATASET ACCORDING TO MEDA.

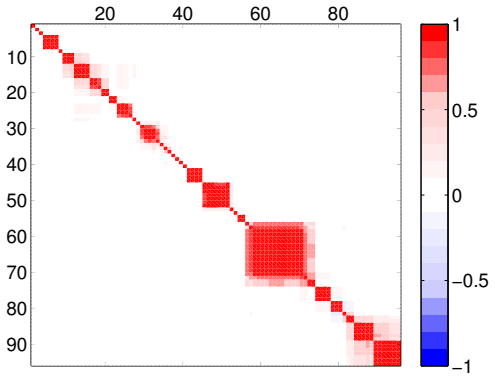

Fig. 2. MEDA plot for the dataset of the second VAST 2012 Mini Challenge.

A first visualization of the data can be obtained from the MEDA graph in Figure 2, where the correlation among the features is shown. As already commented, the features are seriated so that groups of related features can be easily identified. The groups found with MEDA are listed in Table II of the supplemental material at [5]. A selected number of groups is presented in Table I for the subsequent discussion. Each entry in the table has three columns: a group ID number, the interval of features in the MEDA plot belonging to the group and the labels of the features. The network manager should carefully inspect the complete table to understand the general trends in the data, which provide information about the state of the network. Since data sources only include security logs and not normal events (such as common traffic logs), the information will be mainly related to security events.

Let us focus on some specific groups in Table I in order to illustrate how the output of MEDA can be interpreted. The inspection of Groups 7 and 9 may alert the network manager. Group 7 represent IDS logs of medium priority (`ids_prio2`) reporting attempts of information leak (`ids_leak`) to the network firewall (`ids_ipfwhq`) using the VNC protocol (`ids_lvnc`). Group 9 represent IDS logs of high priority (`ids_prio1`) reporting potential corporate privacy viola-

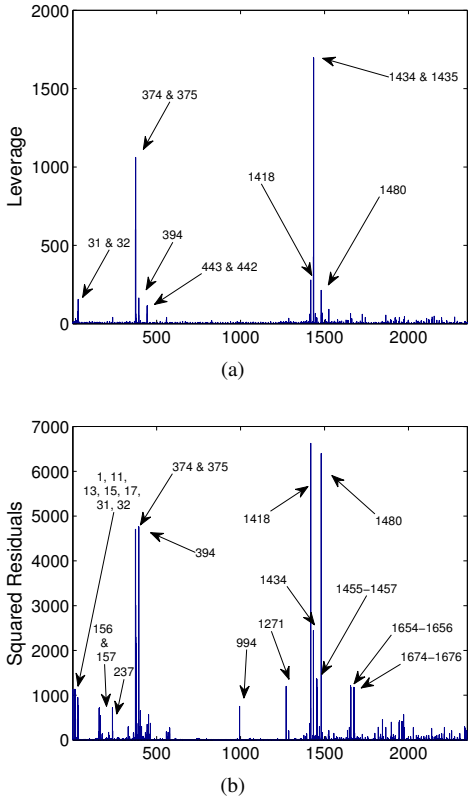

Fig. 3. Leverage (a) and residual (b) time lines.

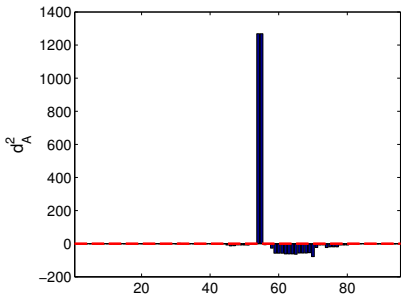

Fig. 4. oMEDA plots for interval {1, 11, 13, 15}.

tion (ids_privacy) in the DNS service (ids_pdns and ids_ldns). These groups may lead to the conclusion that attacks against the firewall and DNS server are potentially taking place. Group 15 shows logs of miscellaneous activity (ids_misc) between external websites (ids_ipweb) and internal nodes using IRC channels (ids_irc). Combining this information with that provided by the users of the network (see Section III), a command and control (C&C) activity may be suspected. MEDA may also show false alarms in IDS logs due to normal traffic. For instance, Group 10 captures the communication between internal workstations (ids_ipws) with client ports (ids_pntsd) to the Domain Control server (ids_ipdc) using Microsoft domains (ids_psmb). The group includes potential low level priority logs (ids_prio3) which may be caused by misconfiguration of the IDS (ids_command). Alternatively, the logs may be pointing out an attempt of denial of service attack or buffer overflow attack, which will give way to execute arbitrary code via a crafted SMB packet. To obtain more details, the network manager should inspect the actual IDS logs and also the logs of the Domain Control server. MEDA does not provide the information of which IDS logs to look at, and other visualization tools should be employed for that, as discussed afterwards. Finally, Group 16 shows that there is a certain amount of incomplete logs (ids_opempty and ids_empty). Again, logs should be accessed for more information. As we can see, the more the features considered in the analysis, the clearer the interpretation of the results. Therefore, over-parametrization is considered to be a useful approach.

The next step is to identify the anomalies in the period of study with the time lines in Figure 3. The anomalies can be detected as the highest bars in the plots. Relevant anomalies are listed in Table III of the supplemental material at [5].

Once a number of anomalies have been found, the analyst needs to investigate the associated causes. For this, oMEDA plots are very useful. Figure 4 illustrates one example of oMEDA plot associated to anomalies {1, 11, 13, 15}. This plot identifies the features that make this group of anomalies present such a high value in Figure 3(b). The oMEDA plot in Figure 4 highlights two consecutive features with very large values: fw_iplog (54) and fw_psyslog (55). With this information, the analyst may conclude that the anomalies are related to the syslog port in the log server. Using oMEDA, the features associated to the anomalies are included in Table III of the supplemental material at [5].

For illustration purposes, we will analyze the results of oMEDA for the highest intervals found in Figure 3. Intervals 374 (2012-04-06; 00:04), 375 (2012-04-06; 00:05) and 394 (2012-04-06; 00:24) report non-legitimate requests to the network firewall (ids_ipfwhq) of several services: telnet (fw_ptelnet), SSH (ids_pssh), SNMP (ids_psnmp), IMAP (ids_limap), POP3 (ids_lpop3) and VNC (ids_lvnc). These motivate the apparition of IDS logs of medium priority (ids_prio2) reporting attempts of information leak (ids_leak). These anomalies are related to the aforementioned Group 7 in Table I, and seem to be related to malicious behaviors aimed at gaining access or privileges to the system. Interval 1418 (2012-04-06; 17:28) is related to the aforementioned Group 9, reporting attacks to the DNS service. The analysis of the IDS logs of that interval, performed to gain more detail on the problem, reveals that 254 entries refer to DNS update attempts from different workstations (IPs 172.23.X.Y) to the DNS server (IP 172.23.0.10). This constitutes a well-known DNS poisoning attack, from which more severe and generalized attacks. Intervals 1434 and 1435 (2012-04-06; 17:43-17:44) are related to incomplete logs. Thanks to the time line plots, we have a specific location of these logs. Inspecting them, it was found that the information in the text logs and parsed logs in the csv files supplied by the organizers of the challenge were not coherent. Therefore, the

system is signaling a parsing error, which is evidenced from the number of empty logs. This was a relevant lesson learned from the case study: the inclusion of features which may point out to potential parsing errors may be very useful to avoid incorrect conclusions. This parsing error has not been reported elsewhere to the best of our knowledge. Finally, Interval 1480 (2012-04-06; 18:30) is related to two features: `fw_syswarn`, and `fw_pftp`. The combination of both features leads to identify FTP access attempts that generate warnings in the firewall. This FTP behavior is typical of infected bots which are trying to download C&C (command&control) messages from botnet handlers, which may be located in the websites.

As it has been illustrated, the information obtained from the proposed system is of special value for the network manager in order to derive hypothesis about the events taking place in the network and their associated causes. Once hypothetical causes are determined, the analyst may confirm the findings in selective anomalies by inspecting the information in the original logs pointed out by the system.

VI. Addressing velocity and volume

Addressing volume and velocity in Big Data are two similar problems whose goal is essentially the same: how to deal with large amounts of data. The proposed solution is based on the iterative computation of intermediate parameters, from which models and visualization techniques can be obtained. Using this approach, only the intermediate parameters need to be maintained in memory instead of the whole set of observations. Also, this is the most suitable approach for continuous model update and parallelization.

Traditional algorithms for PCA take the whole calibration data set $\mathbf{X}$, with N observations, as input. Due to limited computer resources, in particular computer memory, this approach is unfeasible when N grows beyond a certain number, as it is the case for Big Data sets. For a large number of observations, the loading vectors of PCA, $\mathbf{P}_A$, can be identified using the eigendecomposition (ED) of the cross-product matrix $\mathbf{XX}$.

In order to account for the non-stationarity in the data, an EWMA approach [14] is introduced. For a new batch of observations of the variables at time interval t, $\mathbf{X}_{(t)}$, with $B_{(t)}$ observations, the EWMA update of $\mathbf{XX}$ follows:

$$\mathbf{XX}_{(t)} = \lambda \cdot \mathbf{XX}_{(t-1)} + \tilde{\mathbf{X}}_{(t)}^T \cdot \tilde{\mathbf{X}}_{(t)}, \qquad (2)$$

where $\tilde{\mathbf{X}}_{(t)}$ stands for the preprocessed (mean centered and scaled) version of $\mathbf{X}_{(t)}$ and λ is a forgetting factor which takes values from 0 to 1. Preprocessing parameters are also updated following the EWMA approach [14]. The value of λ depends on the data dynamics and may be set to consider a given number $N_{(t)}$ of past intervals in the model, taking into account that for $B_{(t)} = B$ constant in time $N_{(t)}$ converges to $B/(1-\lambda)$. Also, a variable λ [14] may prevent from loosing relevant information in the model.

Eq. (2) is coherent with the computational efficient definition of the singular value decomposition (SVD) in [15], which firstly obtains matrix $\mathbf{XX}$ and then performs the ED.

Anomaly	Proposed	[18]	[19]	[20]
Attacks to the DNS/DC	X			
Firewall access attempts	X	X	X	
FTP attempts to outer nodes	X			X
Background IRC activity	X	X	X	
Parsing errors	X			

TABLE II
FINDINGS IN THE PROPOSED APPROACH AND PREVIOUS ANALYSIS.

The SVD is one of the main algorithms to compute PCA. The complexity of computing $\mathbf{XX}$ is $O(NM^2)$, and the SVD computation $O(M^3)$. In Big Data we have $M << N$. As commented, the computation of $\mathbf{XX}$ from $\mathbf{X}$ would be unfeasible since $\mathbf{X}$ cannot be stored in memory. In any case, this computation would be $O(NM^2)$ for a huge N. On the contrary, the complexity of the EWMA update in eq. (2) is $O(B_{(t)}M^2)$. For practical reasons, $B_{(t)} << N$.

Reference [16] introduces the iterative computation of MEDA from the cross-product matrix $\mathbf{XX}$ and oMEDA for weighted sum of elements. For more detail, please refer to the cited references. On the other hand, time lines are computed following the same procedure in an iterative setup and in the normal case, since they are directly computed from incoming data as they arrive to the system.

Finally, for sufficiently large data flows in terms of velocity and/or volume, parallelization is a must. Data should be processed in independent units and then put together. Two forms of parallelization are possible: in the observations dimension and in the variables dimension. Parallelization of the observations is straightforward given the iterative computation described. Parallelization in the variables can be achieved with Hierarchical PCA [17]. We propose the following application of Hierarchical PCA to this problem: one PCA model is computed and maintained in each data source unit (*e.g.,* the firewall, the IDS, etc.) and only scores are sent to a central unit to study inter-source correlation with a high level PCA. This reduces the amount of data to be passed among processing units. The modification of the proposed system incorporating parallelization in the variables is depicted in Figure 2 of the supplemental material at [5].

A. VAST 2012

We derive an experiment to illustrate parallelization in the variables. A hierarchical model is built as follows: a PCA model is obtained from the variables corresponding to the firewall and a PCA model from the IDS. The scores of each model are joined in the top level PCA model. Combining the residuals of the two low level models and the top level model, we obtain the residual time line of Figure 5(a). Although not exactly identical, the plot signals the same anomalies as Figure 3(b). The leverage time line (not shown) is obtained from the top level alone. MEDA and oMEDA plots can also be obtained for each of the models.

VII. Discussion and Comparison

This section is devoted to summarize the findings of the system proposed in the VAST challenge. We also compare our

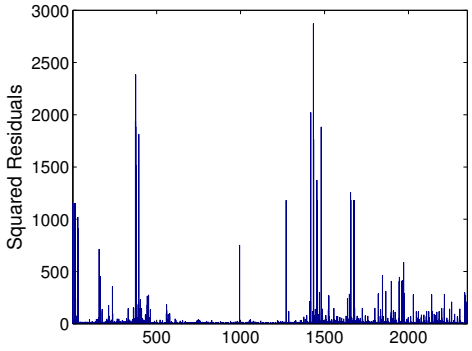

Fig. 5. Residual time line for the hierarchical model.

proposal to the three award-winning submissions to the VAST 2012 Challenge [18], [19], [20]. The main findings reported in the paper are summarized in Table II. These identify a "growing botnet infection" in the network, as reported by the organizers [2]. Table II includes which of these findings are also reported in the award-winning papers, showing the superiority of the proposed system.

It should be noted that one of the findings reported, the background IRC activity, was detected by MEDA, while the rest were clearly found in both MEDA and the time lines/oMEDA. Also, the times lines provide the specific location in time of the anomalous events, which is necessary to identify the logs that should be inspected to gain more detailed information on the problem. Therefore, the combination of tools proposed has proven to be of great value for anomaly detection and forensics. While the application of the time lines is extended in the literature, their combination with MEDA and oMEDA has been used here for the first time. This combination is of especial interest when analyzing an already infected network, since common trends pointed out in MEDA give an insight into the problem.

VIII. Conclusions

We have presented an anomaly detection and forensics system to detect and interpret anomalous events in a network environment generating Big Data. Three main advantages are obtained. First, we can deal with different and heterogeneous sources of information in a straightforward manner, what makes it suitable for security monitoring environments. Second, it is possible to deal with a huge amount of different features to describe the information. Indeed, as the number of features increases, the semantic information provided by our tool is more complete. Third, the number of events to be analyzed is dramatically reduced in Big Data scenarios.

The VAST 2012 mini challenge 2 dataset has been used as an experimental framework to validate our approach. The results obtained are consistent and coherent with the information provided by the challenge organizers, and they outperform the results of the three award-winning proposals in the challenge.

Acknowledgments

Research in this paper is supported by the Spanish Ministry of Science and Technology through grant TEC2011-22579.

References

[1] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: The real-world use of big data," IBM Institute for Business Value, IBM Institute for Business Value - Executive Report, 2012.

[2] K. Cook, G. Grinstein, M. Whiting, M. Cooper, P. Havig, K. Liggett, B. Nebesh, and C. L. Paul, "Vast challenge 2012: Visual analytics for big data," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology 2*. IEEE, 2012, pp. 151–155.

[3] H. Milting, A. Kassner, C. Oezpeker, M. Morhuis, B. Bohms, J. Boergermann, and J. Gummert, "Genomics of myocardial recovery in patients with mechanical circulatory support," *The Journal of Heart and Lung Transplantation*, vol. 32, no. 4, Supplement, p. 229, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053249813005950

[4] R. Marty, *Applied Security Visualization*. USA: Pearson Education, 2008.

[5] "http://nesg.ugr.es/index.php/es/descargas/viewdownload/3-recursos/9-supplemental-material-for-tackling-the-big-data-4-vs-for-anomaly-detection."

[6] "Vast challenge 2012, available at http://www.vacommunity.org/vast+challenge+2012."

[7] D. Brauckhoff, K. Salamatian, and M. May, "Applying pca for traffic anomaly detection: Problems and solutions," in *INFOCOM 2009. 28th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 19-25 April 2009, Rio de Janeiro, Brazil*. IEEE, 2009, pp. 2866–2870.

[8] P. Minarik and T. Dymacek, "Netflow data visualization based on graphs," in *Proceedings of the 5th international workshop on Visualization for Computer Security*, ser. VizSec '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 144–151.

[9] D. Rolls, G. Michailidis, and F. Hernández-Campos, "Queueing analysis of network traffic: methodology and visualization tools," *Computer Networks*, vol. 48, no. 3, pp. 447 – 473, 2005.

[10] J. Camacho and J. Picó, "Online monitoring of batch processes using multi-phase principal component analysis," *Journal of Process Control*, vol. 10, no. 16, pp. 1021–1035, 2006.

[11] J. Camacho, "Missing-data theory in the context of exploratory data analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 103, pp. 8–18, 2010.

[12] ——, "Observation-based missing data methods for exploratory data analysis to unveil the connection between observations and variables in latent subspace models," *Journal of Chemometrics*, vol. 25, no. 11, pp. 592–600, 2011.

[13] G. Caraux and S. Pinloche, "Permutmatrix: a graphical environment to arrange gene expression profiles in optimal linear order," 2005.

[14] B. S. Dayal and J. F. Macgregor, "Recursive exponentially weighted PLS and its applications to adaptive control and prediction," *Journal of Process Control*, no. 3, pp. 169–179.

[15] G. H. Golub and C. F. Van Loan, *Matrix computations (3rd ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[16] J. Camacho, *Exploratory Data Analysis using latent subspace models*. INTECH, 2012.

[17] N. Kettaneh, A. Berglund, and S. Wold, "PCA and PLS with very large data sets," *Computational Statistics & Data Analysis*, vol. 48, pp. 69–85, 2005.

[18] F. Fischer, J. Fuchs, F. Mansmann, and D. A. Keim, "Banksafe: A visual situational awareness tool for large-scale computer networks: Vast 2012 challenge award: Outstanding comprehensive submission, including multiple vizes." in *IEEE VAST*. IEEE Computer Society, 2012, pp. 257–258.

[19] Y. Cao, R. Moore, P. Mi, A. Endert, C. North, and R. C. Marchany, "Dynamic analysis of large datasets with animated and correlated views: Vast 2012 mini challenge 2 award: Honorable mention for good use of coordinated displays." in *IEEE VAST*. IEEE Computer Society, 2012, pp. 283–284.

[20] L. Shi, Q. Liao, and C. Yang, "Investigating network traffic through compressed graph visualization: Vast 2012 mini challenge 2 award: "good adaptation of graph analysis techniques"." in *IEEE VAST*. IEEE Computer Society, 2012, pp. 279–280.