Received July 15, 2019, accepted August 10, 2019, date of publication August 13, 2019, date of current version August 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2935154

Group-Wise Principal Component Analysis for Exploratory Intrusion Detection

JOSÉ CAMACHO[1], ROBERTO THERÓN[2], JOSÉ M. GARCÍA-GIMÉNEZ[1], GABRIEL MACIÁ-FERNÁNDEZ[1], AND PEDRO GARCÍA-TEODORO[1]

[1]CITIC, Network Engineering and Security Group (NESG), University of Granada, 18071 Granada, Spain
[2]Visual Analytics and Information Visualization Group (VISUSAL), University of Salamanca, 37008 Salamanca, Spain

Corresponding author: José Camacho (josecamacho@ugr.es)

This work was supported in part by the Spanish Government-MINECO (Ministerio de Economía y Competitividad), using the Fondo Europeo de Desarrollo Regional (FEDER), under Projects TIN2014-60346-R and Project TIN2017-83494-R.

ABSTRACT Intrusion detection is a relevant layer of cybersecurity to prevent hacking and illegal activities from happening on the assets of corporations. Anomaly-based Intrusion Detection Systems perform an unsupervised analysis on data collected from the network and end systems, in order to identify singular events. While this approach may produce many false alarms, it is also capable of identifying new (zero-day) security threats. In this context, the use of multivariate approaches such as Principal Component Analysis (PCA) provided promising results in the past. PCA can be used in exploratory mode or in learning mode. Here, we propose an exploratory intrusion detection that replaces PCA with Group-wise PCA (GPCA), a recently proposed data analysis technique with additional exploratory characteristics. A main advantage of GPCA over PCA is that the former yields simple models, easy to understand by security professionals not trained in multivariate tools. Besides, the workflow in the intrusion detection with GPCA is more coherent with dominant strategies in intrusion detection. We illustrate the application of GPCA in two case studies.

INDEX TERMS Principal component analysis, group-wise principal component analysis, anomaly detection, intrusion detection.

I. INTRODUCTION

The number of cybersecurity incidents, where strategic assets of corporations get exposed to cybercrime organisations, has experienced a boost in the last five years [1]. As a result, corporations are devoting more economic and human resources for incident detection [2]. Due to the shortage of specialised professionals, there is a need for efficient tools and mechanisms to aid in the detection, triaging and analysis of incidents. As part of this set of tools, anomaly-based Intrusion Detection Systems (IDS) [3] are paramount to unveil new attack strategies.

The use of Principal Component Analysis (PCA) for intrusion detection was proposed more than a decade ago [4], [5]. PCA yields a data factorisation based on the criterion of maximising variance [6], [7]. This factorisation makes it possible to perform anomaly detection in a complex data set, with almost any number of features. This capability is of utmost importance for intrusion detection because a high number of features from multiple and variate data sources can be combined in the IDS [8]. An additional benefit in the use of PCA is that detected anomalies can be interpreted using the model [9], [10], reducing the time between detection and response [11], [12]. This is typically referred to as the diagnosis step.

PCA can be used either in exploratory mode or in learning mode. In the exploratory mode [13], PCA is applied to a data block in order to find anomalies in that block. In the learning mode [14], PCA is calibrated from a data block to build a normality model, and then applied to a different block with new, incoming data, to find the anomalous events. While both approaches present different characteristics, they share a main advantage over black box models (*e.g.* neural networks or kernel methods): PCA is an interpretable model, and besides detecting anomalies, it can be useful to visualise and understand the patterns in the data collected from a network [15].

Unfortunately, the PCA factorisation is often challenging to interpret in highly dimensional data. This difficulty may hamper the practical application of PCA in software for intrusion detection. To overcome this limitation, PCA can be

The associate editor coordinating the review of this article and approving it for publication was Leandros Maglaras.

modified to a so-called simple model structure, improving model interpretability. This improvement can be achieved by means of rotation [6] or sparse methods like sparse principal component analysis (SPCA) [16], [17]. These approaches have been extensively used to simplify the interpretation of PCA models in several areas of knowledge, in particular in biological sciences [18]. However, they are not easy to apply in practice, requiring a certain level of expertise in their use.

In a recent paper, Camacho et *al.* proposed the Group-wise PCA (GPCA) algorithm [19]. GPCA follows an alternative approach than sparse approaches and rotation techniques to yield a simple model structure. With GPCA, we can identify anomalies in the data stream following a straightforward workflow, which is simple to use and understand by security professionals not trained in multivariate tools. This has the following additional advantages over the use of PCA:

i. Domain knowledge can be included in the workflow [15], as an effective means to reduce false positives, a main problem of IDSs.
ii. Sustained security problems, difficult to find with PCA, can be unveiled with GPCA.

In this paper, we propose an exploratory intrusion detection approach based on GPCA. The paper is organised as follows. Section II presents the related work. Section III introduces the use of PCA in exploratory intrusion detection. Section IV presents GPCA. Section V and VI compare the performance of PCA and GPCA in two case studies. Section VII discusses the results and Section VIII brings conclusions.

II. RELATED WORK

IDS paradigms rely on data analysis to determine the occurrence of potentially harmful activities. For that, machine and network events are usually considered as inputs to extract behavioural patterns [20]. However, the amount and variety of data to be processed becomes almost unmanageable in current networked systems, due to their complexity and high speed. Authors in [21] and [22] present the overall problem from different technical perspectives: feature selection, data reduction, information fusion and processing techniques. Here, we will focus on the data reduction approaches [23].

Rehman et *al.* present a review of methods used for data reduction in [24]. Meng *et al.* [25] propose to reduce the data volume for wireless intrusion detection in IoT environments by sampling traffic, either systematically or at random. Authors in [26] propose a framework in which two feature reduction algorithms, Canonical Correlation Analysis (CCA) and Linear Discriminant Analysis (LDA), are used for reducing the less important features for fast, efficient and accurate detection of intrusions in netflow records using Spark.

Principal Component Analysis (PCA) is a processing technique recurrently used in the literature to reduce dimensionality [27], [28]. The most referred work on PCA intrusion detection is that of Lakhina *et al.* [13], where the authors propose the use of PCA over link counts of traffic for detecting network-wide anomalies. For this, a PCA model is fitted from the complete traffic capture, following the exploratory mode, and anomalies are searched for in the residuals of this model, using the so-called Q-statistic or SPE. The underneath assumption is that the structural correlation captured by PCA represents the normal, free of anomalies, traffic behaviour. This assumption in fact leads to the main shortcoming of the approach: anomalies of large magnitude, and therefore of large variance, can pollute the normality model. This situation, in turn, makes the approach very sensitive to calibration settings [29].

In another work, Lakhina *et al.* [30] also explore the combination of counts of bytes, counts of packets and counts of IP flows as the input to the monitoring system. They state that for monitoring *more diverse data*, the model subspace should also be inspected for anomalies. For that, they suggest the use of the Hotellin's T^2 statistic, also referred to as the D-statistic when used with PCA. Thus, the detection is based on both the Q-statistic and the D-statistic, following standard practices in PCA anomaly detection in the process industry [31], [32]. Camacho *et al.* [8] follow this approach and extend the data parameterisation to combine traffic data with any source of security data, like traditional IDS logs or firewall logs.

Some contributions on multivariate analysis for security anomaly detection have opted for combining PCA with other detection schemes. Thus, Aiello *et al.* [33] combine PCA with mutual information for profiling DNS tunnelling attacks. Fernandes *et al.* [34] combine PCA with a modified version of Dynamic Time Warping for network anomaly detection. They also propose an alternative approach based on Ant Colony Optimization. Jiang *et al.* [35] apply PCA over a wavelet transform of the network traffic for network-wide anomaly detection. Chen *et al.* [36] use a similar approach with Multiscale PCA. Peng et *al.* propose in [37] a clustering method based on Mini Batch K-means with PCA (PMBKM). More recently, authors in [38] combine the approaches of information gain (IG) and PCA with an ensemble classifier based on a support vector machine (SVM), Instance-based learning algorithms (IBK), and a multilayer perceptron (MLP).

Authors in [14] introduce the Multivariate Statistical Network Monitoring (MSNM) approach, where the PCA model is used in learning mode, rather than in the exploratory mode originally proposed by Lakhina. PCA is first employed to estimate a normality model for both structural and residual sub-spaces in the calibration data, and this model is afterwards contrasted with future data for real-time anomaly detection. In the first step, calibration data needs to pass through a cleaning process where the D and Q statistics are employed to explore data for outliers. The identification and extraction of outliers are typically performed on an iterative basis, in which the data are visualised, outliers are isolated and the model re-calibrated. This is often a challenging process, which may lead to anomaly detection systems too sensitive or too insensitive to anomalies.

In a recent paper [15], we proposed a new tool for intrusion detection where we combined the PCA exploratory approach with visual analytics and GPCA. In it, GPCA takes a secondary role, and its potential contribution to the intrusion detection paradigm was not determined. In this paper, we extend that work by defining a clear workflow for the application of GPCA in intrusion detection, study how the analyst can use domain knowledge in the analysis, making it more efficient, and evaluate the performance of GPCA in two case studies.

III. PCA FOR EXPLORATORY INTRUSION DETECTION

PCA applies to data sets with M features corresponding to N observations, which can be arranged in a matrix $\mathbf{X}$ of M columns and N rows. For intrusion detection, features (columns) correspond to quantitative values obtained from any security-related source of data, including traffic and logs of applications and systems. Typically, the observations (rows) correspond to consecutive time intervals, which is suitable for real-time monitoring.

PCA aims to find the subspace of maximum variance in the M-dimensional feature space. The original features are linearly transformed into the Principal Components (PCs), using the eigenvectors of $\mathbf{X}^T \cdot \mathbf{X}$, typically for mean centred $\mathbf{X}$ and sometimes also after auto-scaling (normalising to unit variance). PCA follows the expression:

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

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

In their original publication, Lakhina *et al.* [13] propose to monitor only the residual subspace of PCA. For that, they compute the Q-statistic (Q-st) or SPE:

$$Q_c = e_c \, e_c^t \qquad (2)$$

where e_c is the residual vector in the c-th row of $\mathbf{E}$ in eq. (1). To identify anomalies, Lakhina et *al.* use the expression proposed by Jackson and Mudholkar [39] for the Upper Control Limit (UCL) at significance level α:

$$UCL(Q)_\alpha = \theta_1 \cdot \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \qquad (3)$$

where $\theta_n = \sum_{a=A+1}^{rank(\mathbf{X})} (\lambda_a)^n$, with $rank(\mathbf{X})$ the rank of the matrix of data $\mathbf{X}$ and λ_a the eigenvalues of matrix $\frac{1}{N-1} \cdot \mathbf{E}^T \cdot \mathbf{E}$; $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$; and z_α is the $100 \cdot (1-\alpha)\%$ standardised normal percentile. The most common approach is to set α to 0.01, in order to define a 99% control limit. All observations with a value of Q_c above the control limit are signalled as anomalies.

As already discussed, the same authors propose later [30] the combination of the Q-statistic with the D-statistic (D-st):

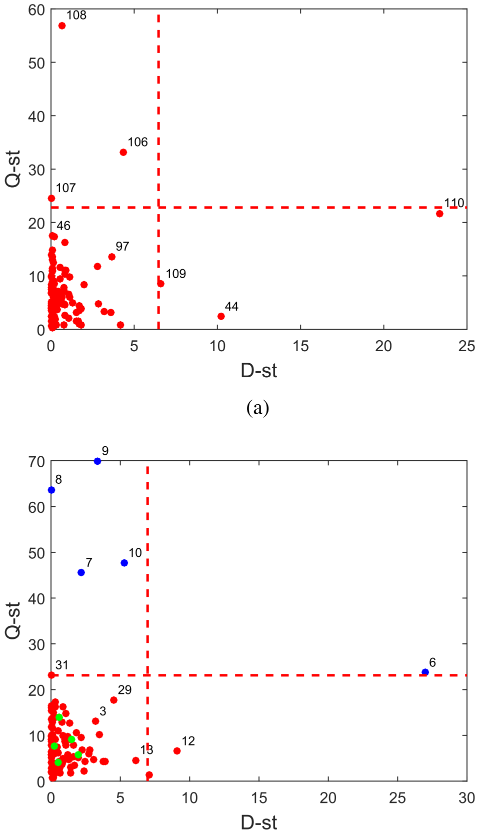
$$D_c = t_c \Lambda^{-1} t_c^t \qquad (4)$$

where t_c is the score vector in the c-th row of $\mathbf{T}$ in eq. (1) and $\Lambda = \frac{1}{N-1} \cdot \mathbf{T}^t \cdot \mathbf{T}$. They also define the corresponding UCL at significance level α following [7]:

$$UCL(D)_\alpha = \frac{A(N^2-1)}{N(N-A)} F_{(A,(N-A)),\alpha} \qquad (5)$$

with $F_{(A,(N-A)),\alpha}$ the F-distribution with A and $N-A$ degrees of freedom at significant level α.

Figure 1(a) illustrates the intrusion detection approach based on PCA in exploratory mode, using a scatter plot of the D-st versus the Q-st. There is only one block of observations, which are used to calibrate the PCA model, and to compute the statistics and the corresponding control limits. Then the same observations are contrasted to those limits in the monitoring chart in order to identify anomalies. In the figure, we can identify several anomalies, which exceed the UCL either in the D-st (observations 110, 44, and to a lesser extent 109) and/or the Q-s (observations 108, 106, and to a lesser extent 107).



(a)

(b)

FIGURE 1. Illustration of PCA detection in (a) exploratory and (b) learning mode.

In learning mode, a new block of observations is projected on the PCA model, and new Q-st and D-st are computed. This is illustrated in Figure 1(b), where we use one block of observations (red dots) for the calibration of model and control limits, and then the model is used to monitor new observations, which can be classified as normal (green dots) or anomalous (blue dots).

To combine the D-st and the Q-st into a single triaging score, [11] defines the *Tscore* according to the following equation:

$$T_c = \alpha \cdot D_c/UCL^D_{.99} + (1 - \alpha) \cdot Q_c/UCL^Q_{.99} \tag{6}$$

Once an anomaly is identified, either with the D-st and Q-st or the Tscore, the PCA model can also be used to provide a first diagnosis of the problem, by identifying those features related to the anomalous value of an observation. There are several approaches for that, see [10] for a review. The discussion on the performance of different diagnosis methods is out of the scope of this paper. The interested reader is referred to [40].

The diagnosis is illustrated in Figure 2 for one of the observations in blue color in Figure 1(b). The particular method used is named oMEDA [41], and noted as d^2_A for A the number of PCs. It is a bar plot of the features, built to compare two groups of observations. Each bar represents the contribution of the feature to the difference between both groups. A positive bar implies that the first group of observations presents a higher value in the corresponding feature than the second group. A negative bar reflects the opposite. A bar close to zero means that both groups of observations have a similar value in that feature. From the plot, we can conclude that the anomalous observation under analysis (first group) showed larger values than normal observations (second group) in the two first features.

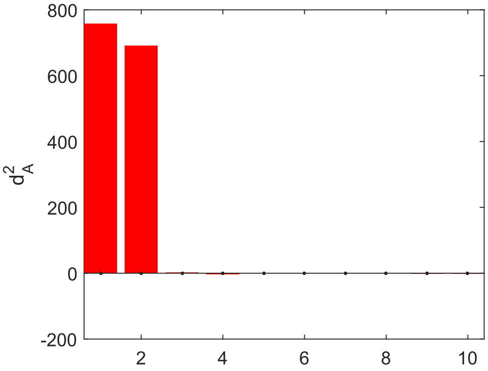


FIGURE 2. Illustration of PCA diagnosis.

IV. GPCA

GPCA [19] is a recent sparse PCA variant. Every component contains non-zero loadings for a single group of correlated features. GPCA starts with the identification of a set of K (possibly overlapping) groups of correlated features obtained from a map $\mathbf{M}$, with elements $m_{i,j} \in [-1, 1]$ containing the strength of the relationship between features i and j. An example of this map is the correlation matrix of $\mathbf{X}$. In the original formulation of GPCA, the MEDA approach (Missing-data for Exploratory Data analysis) [42] was implemented to define $\mathbf{M}$. MEDA uses a missing data strategy to estimate the correlation between any two variables. This approach has been found to be effective in filtering out noise when estimating correlations.

Once the groups have been defined, the GPCA algorithm first computes K candidate loading vectors, one per each of the groups of features. From these, only the loading vector with the largest variance is retained, and residuals are computed. The algorithm iterates until a set of sparse components is extracted.

Figure 3 illustrates the MEDA plot for a simulated data set. In the plot, the data features cluster in three, clear groups. Colours in the plot reflect the level and direction (positive, red, or negative, blue) of the correlation between features. The Group Identification Algorithm (GIA) was defined in [19] to automatically identify groups in a MEDA plot.

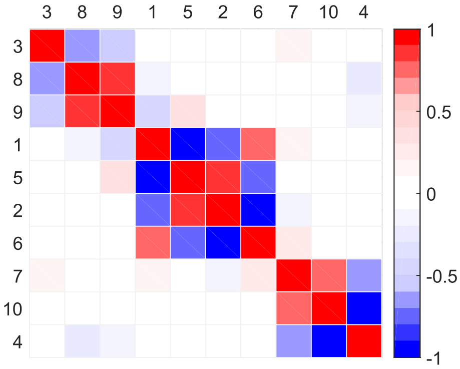


FIGURE 3. Illustration of the MEDA map.

Once the groups are identified, they can be visualised using colours with contextual (*e.g.*, security-related) information. For instance, we can show the level of security relevance of the group of features. See Figure 4 for an illustration of this approach. From this figure, the security analyst can focus on those feature groups with more security-relevant information. Using GPCA, we compute the temporal evolution of the observations corresponding to a feature-group of interest. See Figure 5 for an illustration, where we can identify anomalies in time (those spikes surpassing the control limits).

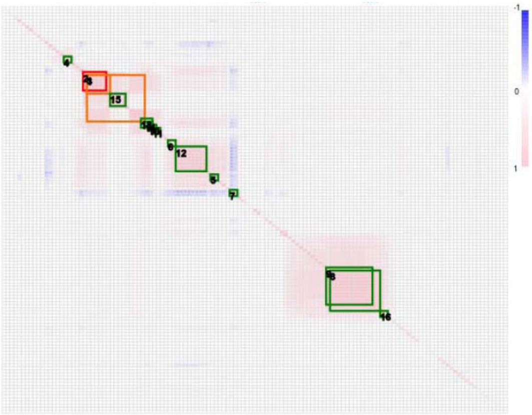


FIGURE 4. Illustration of GIA grouping with contextual information.

GPCA has two features that make it especially useful for forensic analysis and interpretation. On the first hand, GPCA

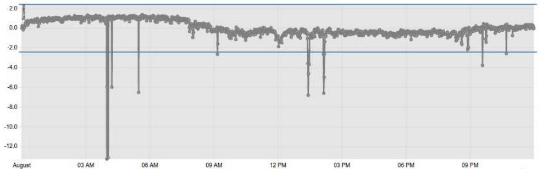


FIGURE 5. Illustration of scores of a group of features.

performs a separation of sources very similar to Independent Component Analysis (ICA) and other blind source separation techniques. Thus, the traffic is decoupled in different types of traffic, and the analyst can easily understand how these evolve in time. On the other hand, unlike ICA, GPCA is sparse, which means that this separation is linked to a reduced subset of features, simplifying interpretation.

When detecting intrusions with GPCA, we go from features to the observations, that is, we identify anomalies in the (groups of) features, and from them go to the time evolution of such anomalies. This workflow is complementary to that of PCA, and presents the advantages commented in the introduction: sustained security-related trends in the data can be unveiled, and domain knowledge can be used to reduce false positives.

V. CASE STUDY I: VAST CHALLENGE

This first case of study serves to illustrate the workflow for exploratory intrusion detection using GPCA and the main differences with the PCA approach.

A. EXPERIMENTAL FRAMEWORK

The data comes from the VAST 2012 2nd mini-challenge [43] and contains information captured in a corporate network during a timeframe of two days. The network infrastructure is comprised of approximately 1000 servers and 4000 workstations and is running 24 hours a day. Most of the company operations are carried out inside the network. However, some financial transactions have to go to data centres outside the network. During the capture, the users experienced several technical issues in their systems. Some staff members informed that their workstations were infected by spyware and that suspicious messages from a previously unseen antivirus software started popping up on their computers. From the official solution of the challenge, we know that a botnet compromised the network, causing the aforementioned performance problems and the emergence of the spyware.

The VAST2012 data set contains two semi-structured data sources: logs from an Intrusion Detection System (IDS) and logs from a Cisco ASA firewall (FW). A total of 23,711,341 data records from the FW and 35,948 records from the IDS are presented in CSV and in raw format. We parsed the raw data into a total of 265 features, 122 for the FW and 143 for the IDS. Tables 1 and 2 summarise the list of features, as well as the kind of information they contain. The features are computed for 1 minute intervals, yielding a 2345 × 265 matrix of parsed data. More details can be found in [12]. After auto-scaling, weights from 1 to 10 are assigned to each feature according to their security relevance.

TABLE 1. Firewall feature overview.

Number of features	Data referred in the features
6	Source and destination IPs
84	Source and destination ports
13	ASA Messages
8	Syslog priority
9	Options
2	Protocol

TABLE 2. IDS feature overview.

Number of features	Data referred in the features
6	Source and destination IPs
84	Source and destination ports
4	Snort priority
43	Snort Classification
6	IP headers

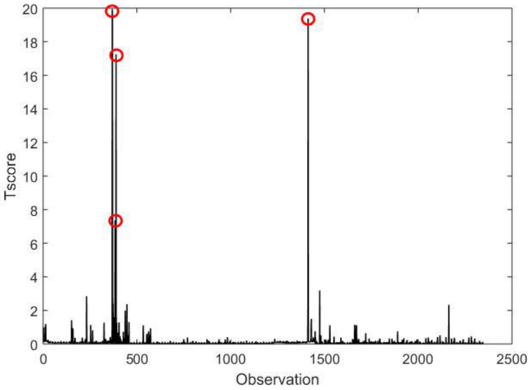


FIGURE 6. Tscore values for PCA anomaly detection.

Reproducibility of the results in this case study is possible by downloading the virtual machine at https://nesg.ugr.es/veritas/index.php/mbda

B. EXPLORATORY INTRUSION DETECTION

The results of applying the PCA methodology in this case study are summarised in Table 3. Figure 6 shows the time evolution of the Tscore, Eq. (6), according to which the anomalies were triaged. For more details on the derivation of the chart and table, please refer to [12]. The structure of the table corresponds to the workflow in PCA intrusion detection. First, a set of observations are triaged as the most relevant from the security perspective. Five observations are highlighted over the rest. In particular, an interval of 20 minutes around midnight of the second day includes 4 out of these 5 anomalies. To further investigate these anomalies, the PCA diagnosis points to the features in the fourth column of the table. Combining both the information in columns 3 and 4, the specific raw log entries corresponding to the anomalies can be identified and interpreted, as explained in [12].

The interpretation of the anomalies is listed in the last column of Table 3. First, several data exfiltration attempts

TABLE 3. VAST2102: Anomaly Report with PCA [12].

Index	Tscore	Timestamps	Features selected	Interpretation
369	19,94	06/04 00:04	fw_dport_telnet, ids_ssh_scan, ids_ssh_scan_outbound, ids_dport_ssh	Data exfiltration attempt by SSH and Telnet, the latter blocked by access lists.
370	19,52	06/04 00:05	fw_dport_snmp, fw_denyacl, ids_dport_snmp, ids_snmp_req, ids_successful-recon-limited, fw_error ids_brute_force, ids_attempted-recon	Attempted information leak using the SNMP protocol.
384	7,36	06/04 00:19	ids_scanbehav	Scan for windows RDP ports for vulnerabilities.
389	17,25	06/04 00:24	ids_attempted-recon, ids_successful-recon-limited, ids_prio2, ids_vnc_scan	Scan ports in range 5900-5920 looking for vulnerabilities.
1413	19,38	06/04 17:28	ids_policy-violation, ids_prio1, ids_dns_update	DNS server attack from multiple systems.

TABLE 4. VAST2102: Anomaly Report with GPCA.

GPC	Variance	Features group	Relevance	Indices	Timestamps	Interpretation
1	5.1422e+05	ids_vnc_scan, ids_prio2, ids_sport_private, ids_ttl_low, ids_attempted-recon, ids_successful-recon-limited	8	370 388 389 458 573	06/04 00:05 06/04 00:23 06/04 00:24 06/04 01:33 06/04 03:28	Data exfiltration attempted by SNMP. Scan ports in range 5900-5920 looking for vulnerabilities.
2	5.0630e+05	ids_policy-violation, ids_dns_update, ids_prio1, ids_dport_dns	10	14 1412 1413	05/04 18:07 06/04 17:27 06/04 17:28	DNS server attack from multiple systems.
3	4.5282e+05	fw_dport_telnet, ids_dport_ssh, ids_ssh_scan_outbound, ids_ssh_scan	7	369 438 553	06/04 00:04 06/04 01:13 06/04 03:08	Data exfiltration attempt by SSH and Telnet the latter blocked by access lists.
4	2.7724e+05	ids_defrag_DF, ids_sport_registered, ids_netbios, ids_dport_mds, ids_prio3, ids_protocol-command-decode, ids_dport_reserved, ids_dst_ip_private, ids_ttl_medium, ids_src_ip_private	5	-	Trend	Suspicious traffic in port 445, used by SMB. IDS detect attempted integer overflow
5	2.6675e+05	fw_critical, fw_asa_106001, fw_inbound, fw_sport_irc	7	-	Trend	IRC traffic blocked by FW
6	1.5433e+05	ids_sport_irc, ids_irc_auth, ids_misc-activity, ids_dport_register	5	-	Trend	IRC traffic detected by IRC

by Telnet and SSH were made. Subsequently, an information leakage carried out using SNMP was detected by the IDS. Multiple vulnerability scans targeting remote desktop services like VNC and RDP followed. Some 17 hours later, a coordinated attack from multiple infected systems targeted the DNS server. All these observations correspond to landmarks in the killchain of an attack, and were effectively highlighted by the PCA anomaly detector. However, part of useful information for diagnosing the problem is missing in this security report.

In comparison, the security report obtained with GPCA is shown in Table 4. We extracted 6 Group-wise Principal Components (GPCs), each of them of decreasing variance with respect to the previous one, using the correlation map computed by MEDA and shown in Figure 7. Each GPC models a group of features identified in the MEDA plot. These are shown in the third column of the table. By taking the maximum of the weight of the selected features, with values between 1 and 10, we can rate the GPCs according to their security relevance. This is shown in column 4. This relevance can be made visual as illustrated in Figure 7, so that a relevance of 10 is shown as a square in red colour, and a relevance below 3 is shown as a square in green colour. This is useful to guide the analyst, who would focus her attention in GPCs of high relevance.

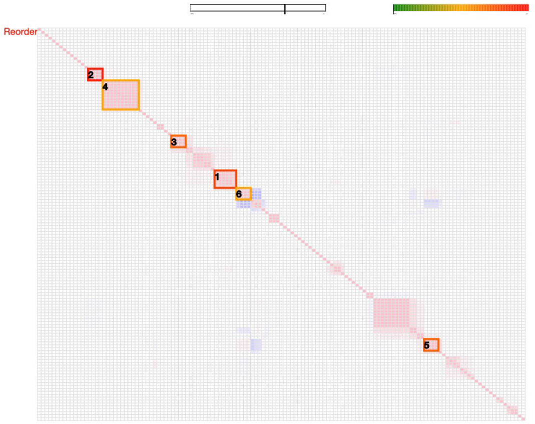


FIGURE 7. MEDA plot.

Following this idea, the most relevant GPC is GPC2, followed by GPC1, GPC3 and GPC5. These are shown in red and orange colours in Figure 7. The temporal evolution

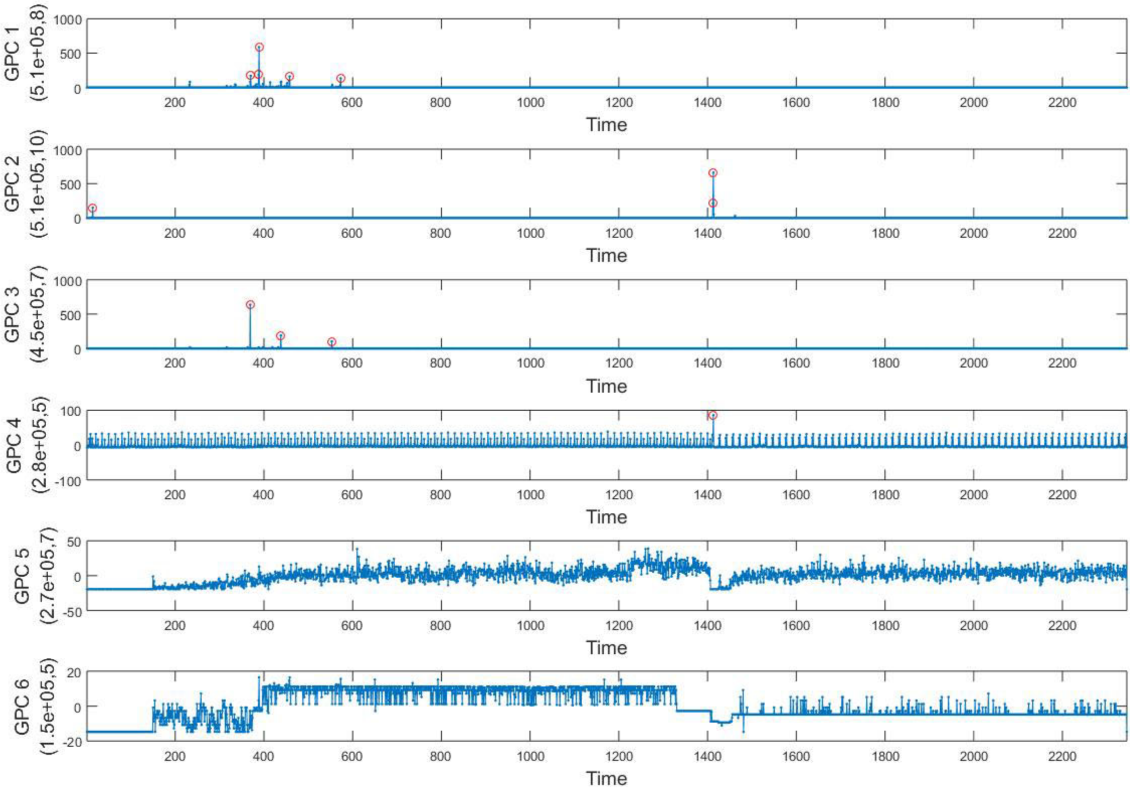


FIGURE 8. Group-wise Principal Components: variance and relevance between parenthesis.

of the GPCs, obtained with GPCA, is shown in Fig. 8. Starting with GPC2, in both the plot and the table we can see that there was an attack to the DNS server right at the beginning of the capture, at 05/04 18:07, much before than when it was noticed with PCA. GPC1 and GPC3 show the data exfiltration and scanning attempts also highlighted by PCA. However, GPC4, GPC5 and GPC6 show a sustained behaviour, not anomalous but present during the entire capture, that went unnoticed by the PCA anomaly detector. In particular, GPC5 and GPC6 reflect the use of IRC traffic, something which is not permitted according to the security policies of a bank network, and that shows the command and control communication of the botnet.

VI. CASE STUDY II: ISP NETWORK

In this second test case, we are interested in the analysis of data coming from a real network scenario, so that we can test the usefulness of GPCA for security analysts when applied to real traffic.

A. EXPERIMENTAL FRAMEWORK

For our purpose, we select the UGR'16 dataset [44]. This is a dataset obtained from a Tier-3 ISP where traffic is generated by hosted of companies, web and email servers, recursive DNS servers, virtualisation environments, etc. In this network, a set of sensors were deployed in the border routers so that ingress and egress network traffic flows were monitored. The dataset consists of Netflow traces corresponding to more than 16,000M connections for more than four months in 2016. It is divided into two sets: a *calibration set* for building models from ‘‘normal’’ traffic and a *test set*, where attacks using real hacking tools were generated from a set of 25 virtual machines. Details about the dataset are shown in Table 5.

TABLE 5. Features of the calibration and the test sets in the UGR'16 dataset.

Feature	Calibration	Test
Capture start	10:47h 03/18/2016	13:38h 07/27/2016
Capture end	18:27h 06/26/2016	09:27h 08/29/2016
Attacks start	N/A	00:00h 07/28/2016
Attacks end	N/A	12:00h 08/09/2016
Number of files	17	6
Size (compressed)	181GB	55GB
# Connections	≈ 13,000M	≈ 3,900M

We are interested in evaluating the differences in the workflows of GPCA and PCA when an exploratory analysis is applied to a subset of the dataset. For this purpose, we choose a 3 hours trace from t_0 = [08/06/2016 18:00h] to t_f = [08/06/2016 20:59h]. In the first hour of the trace, DoS and scan attacks are executed in the following manner (see details in [44]):

- At t_0: *DoS11*. A low rate DoS attack from one machine to another (one-to-one) is performed for 3 minutes.

The attack consists in sending SYN packets directed to port 80 (HTTP) in the victim machine.

- At $t_0 + 10m$: *DoS53s*. The same low rate DoS attack as in DoS11 is performed for 3 minutes, but now it is launched from 5 machines and directed to 3 machines. The attacks are all synchronized in time. This attack generates five times more traffic than the DoS11 attack.
- At $t_0 + 20m$: *DoS53a*. This attack is similar to DoS53s, but now the attack is spread out during 10 minutes so that during the first 3 minutes two machines attack a given victim. Then, after one minute with no attacks, a new attack burst is struck during 3 minutes (two machines against one). Finally, after another minute with no attacks, the last burst is started from a single machine to a victim. Note that these three bursts last 10 minutes. In the first two bursts, the traffic volume is higher than in DoS11 but lower than DoS53s. For the third burst, the amount of traffic is similar to DoS11.
- At $t_0 + 40m$: *Scan11*. A port scan attack is performed for 3 minutes from an attacker machine to a victim.
- At $t_0 + 50m$: *Scan44*. During 3 minutes, four machines are scanning four different victims.

During the second hour in this trace, a Neris botnet [45] is communicating with 20 infected machines in the network. Finally, the third hour is free of attacks (background traffic only).

In order to analyse the trace with PCA and GPCA, we first pre-process it and obtain 134 numeric features for every minute of traffic. These features are calculated following the feature-as-a-counter approach [14]. For example, the feature `sport_http` accounts for the number of flows in a minute that have the source port equal to 80. A summary of the features collected is shown in Table 6. Thus, for the three hours trace, our dataset is a matrix of 180 rows (observations, in minutes) for 134 features.

TABLE 6. Variable values considered as features in our detection system.

Variable	#features → values
Source IP	2 → *public, private*
Destination IP	2 → *public, private*
Source port	50 → *specific services, Other*
Destination port	50 → *specific services, Other*
Protocol	5 → *TCP, UDP, ICMP, IGMP, Other*
Flags	6 → *A, S, F, R, P, U*
ToS	3 → *0, 192, Other*
# Packets in	5 → *very low, low, medium, high, very high*
# Packets out	5 → *very low, low, medium, high, very high*
# Bytes in	5 → *very low, low, medium, high, very high*
# Bytes out	5 → *very low, low, medium, high, very high*

B. EXPLORATORY INTRUSION DETECTION

Beginning with the PCA analysis, a security analyst would first obtain the Tscore values for the different observations (minutes) as shown in Fig. 9. Then, for the main anomalies pointed out by the Tscore (signalled with red circles in Fig. 9), he/she would proceed with a detailed analysis *i)* identifying the features responsible for the anomaly (oMEDA analysis), *ii)* selecting those raw traces (connections) involved in the anomaly, and *iii)* interpreting them.

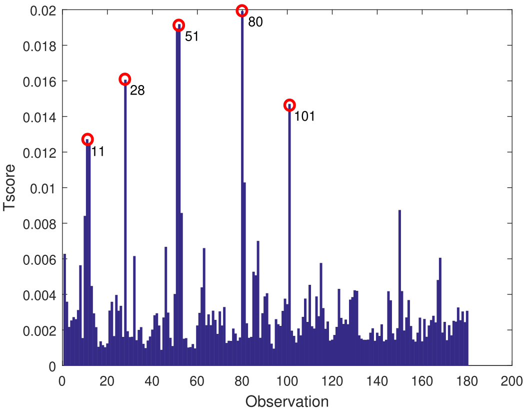


FIGURE 9. T-score values for PCA anomaly detection UGR'16 trace scenario.

Now we explain the analysis we have done for these anomalies. A summary is given in Table 7.

- Observation 11: The involved features indicate that the anomaly is triggered by HTTP traffic (`dport_http`, `sport_http`), with connections that transport a low number of bytes in the range $[150, 1000)$ (`nbytes_low`), that use ports which are in the range [0, 1024] (`sport_reserved` (in this case only port 80 – HTTP is used), and are failed connections (`tcpflags_RST`). Our interpretation of this traffic is that the anomaly is generated by a DoS attack struck with HTTP traffic. This attack pattern corresponds to the DoS53s attack in the UGR'16 trace.
- Observation 28: The involved features point out to ICMP traffic (`protocol_icmp`, `sport_zero`) and Telnet traffic (`sport_telnet`). After observing the raw traces for Telnet and ICMP, we conclude that, while we do not appreciate any odd behaviour in Telnet traffic, an anomaly in ICMP traffic is actually present. It is an ICMP Scan from the IP 224.231.46.145 to the whole range of addresses in the ISP.
- Observation 51: The number of features involved in this anomaly is large, and all of them are related to different ports, both as source and destination. This is a clear indication of a Port Scanning attack. It actually corresponds to the UGR'16 Scan44 attack.
- Observation 80: Now the features are clearly pointing out to an anomaly in the DNS traffic (`sport_dns`, `dport_dns`, `protocol_udp`). After exploring the raw traces in this minute we find that infected bots (Neris botnet) are the responsible nodes for this anomaly.
- Observation 101: The features are indicating that the anomaly is caused by HTTPS traffic (`dport_https`) with the URGENT flag activated (`tcpflags_URG`). We explore the values of these two features (see Fig. 10), and find out that the amount of packets with URG flag

TABLE 7. UGR'16: Anomaly Report with PCA using the Multivariate Big Data Analysis (MBDA) approach [12] for the UGR'16 trace.

Index	Tscore	Timestamps	Features selected by oMEDA	Interpretation
11	0.0127	08/06 18:10	dport_http, sport_http, nbytes_low, tcpflags_RST, sport_reserved	DoS53s
28	0.0160	08/06 18:27	sport_telnet, protocol_icmp, sport_zero	ICMP scan
51	0.0189	08/06 18:50	sport_kpasswd, sport_cups, sport_finger, sport_nntp, sport_quote, sport_echo, sport_daytime, sport_discard, sport_ldaps, dport_citrix, dport_msnmessenger, sport_gopher, sport_ldap, dport_kpasswd, sport_chwhereen, dport_cups, dport_discard, dport_mgc, sport_kerberos, dport_finger, dport_quote, dport_daytime, dport_echo, dport_ldap, dport_kerberos, dport_ldaps, dport_emule, dport_chwhereen, dport_syslog, dport_nntp, dport_multiplex, dport_gopher	Scan44 (Port scan)
80	0.0199	08/06 19:19	tcpflags_ACK, protocol_udp, dport_dns, sport_dns	DNS anomaly (botnet)
101	0.0146	08/06 19:40	tcpflags_URG, dport_https	Not an attack (see explanation)

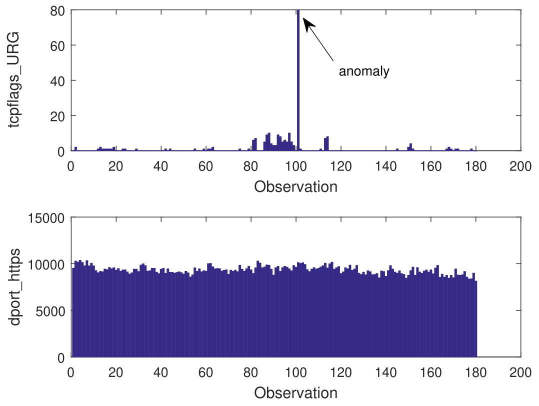


FIGURE 10. Time evolution of features `tcpflags_URG` and `dport_https` in the UGR'16 trace.

is very reduced (around 80) and the IP addresses from where this traffic is generated do not follow a clear pattern. Thus, we conclude that this anomaly is not an actual attack.

Now we are interested in showing how the GPCA workflow would simplify the security analyst goal of interpreting the different anomalies. Following the GPCA methodology, an analyst would first obtain the MEDA plot shown in Fig. 11, revealing the groups of variables that exhibit a minimum correlation level. Then, from this set of groups, the analyst can prioritise the most important GPCs, mainly according to the amount of variance captured and the relevance of features included in the groups according to her expert knowledge. Like in the previous example, we have introduced the expert knowledge by establishing a weighted score (in the interval [1, 10]) to every possible feature as shown in Table 8. In this example, we are specially concerned about IRC (relevance 8), Emule (relevance 8) and Metasploit traffic (relevance 10), as these types of traffic should not normally exist in this network. We also want to prioritise somehow anomalies in which the amount of traffic is very high or very low (relevance 5).

In Table 9, we show the analysis for the six most relevant GPCs. First, the analyst would give an interpretation of every group of features. Our interpretation (last column of Table 9) is obtained as follows:

TABLE 8. Weights assigned to the features in the UGR'16 dataset as expert knowledge from the security analyst.

Weight	Features
1	Default value for all features
5	npackets_verylow, nbytes_verylow,
5	npackets_veryhigh, nbytes_veryhigh
8	sport_irc, dport_irc, sport_emule, dport_emule
10	sport_metasploit, dport_metasploit

- GPC1. This group is formed by a large number of features related to different ports (source or destination). Anomalies within this group will have the characteristic of being traffic that uses many different ports. Thus, we interpret that this group represents *port scan anomalies*. The relevance of group 1 is given by the most relevant feature included in this group, being in this case equal to 8 (`dport_emule`).
- GPC2. Here, the features are indicating that the abnormality is given by an unusual number of connections with very low number of packets (< 4) (`npackets_verylow`), which are TCP flows with SYN flag activated from public IP addresses (`ip_public`, `protocol_tcp`, `tcpflags_SYN`). Our interpretation is that this is generated when abnormal *bursts of traffic* occur. The relevance of this group is determined by the `npackets_verylow` feature, thus adopting a value of 5.
- GPC3. Observing the features included by GPCA in this group, we deduce that the anomalies detected are related to connections with a reduced number of bytes (`nbytes_low`), using HTTP port and RST flag. Our interpretation, in this case, is that this group represents *HTTP DoS attacks*. The relevance of this group is 1.
- GPC4. This group represents anomalies in DNS traffic (`port_dns`), using UDP (`protocol_udp`) traffic.
- GPC5. In this case, the features selected by the algorithm are not pointing us to an intuitive interpretation

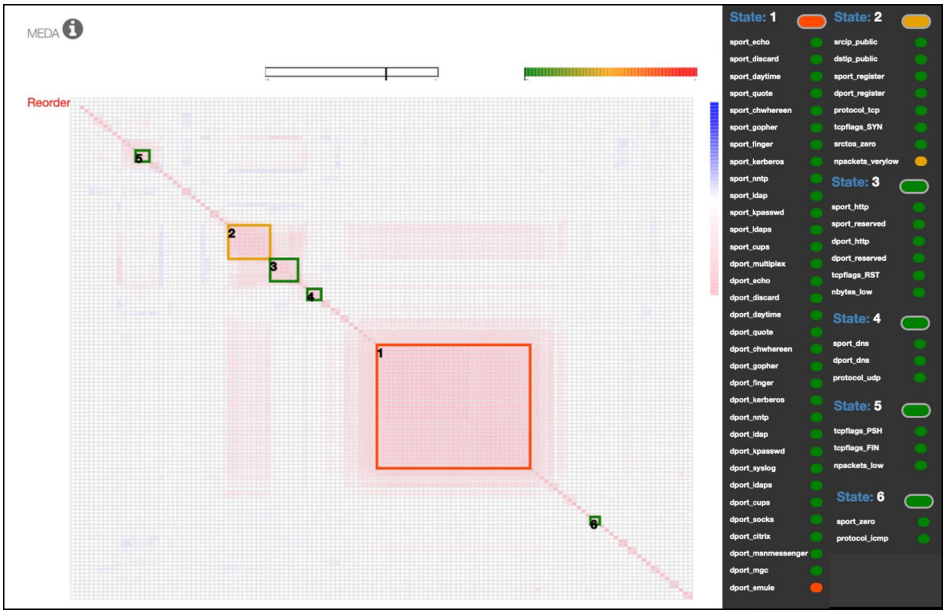


FIGURE 11. MEDA plot for the UGR'16 trace in the GPCA analysis.

TABLE 9. UGR'16: Anomaly Report with GPCA for the UGR'16 trace.

GPC	Variance	Features in the group	Relevance	Index (Time)	Attack	Interpretation
1	5653.33	sport_echo, sport_discard, sport_daytime, sport_quote, sport_chwhereen, sport_gopher, sport_finger, sport_kerberos, sport_nntp, sport_ldap, sport_kpasswd, sport_ldaps, sport_cups, dport_multiplex, dport_echo, dport_discard, dport_daytime, dport_quote, dport_chwhereen, dport_gopher, dport_finger, dport_kerberos, dport_nntp, dport_ldap, dport_kpasswd, dport_syslog, dport_ldaps, dport_cups, dport_socks, dport_citrix, dport_msnmessenger, dport_mgc, dport_emule	8	41 (18:40) 51 (18:50)	Scan11 Scan44	Port Scan
2	1359.67	srcip_public, dstip_public, sport_register, dport_register, protocol_tcp, tcpflags_SYN, srctos_zero, npackets_verylow	5	1 (18:00) 11 (18:10) 21-30 (18:20-18:30) 51 (18:50) 80 (19:19)	Dos11 Dos53s Dos53a Scan44 botnet	Bursts of traffic
3	995.12	sport_http, sport_reserved, dport_http, dport_reserved, tcpflags_RST, nbytes_low	1	1 (18:00) 11 (18:10) 21-30 (18:20-18:30)	DoS11 DoS53s DoS53a	HTTP Dos Attacks
4	531.02	sport_dns, dport_dns, protocol_udp	1	80 (18:19)	botnet	DNS anomalies
5	516.29	tcpflags_PSH, tcpflags_FIN, npackets_low	1			No interpretation
6	357.82	sport_zero, protocol_icmp	1	28 (18:27)	None	ICMP anomalies

of any type of anomaly. Thus, we leave this group with no interpretation for further inspection of the anomalies signalled by the group of features. It is important to recognise that even using GPCA, which eases the interpretability of the information, it is possible to find groups that are not meaninful in terms of security.

- GPC6. This group is directly related to *ICMP anomalies* (`sport_zero`, `protocol_icmp`).

The final step in the workflow followed by the analyst is the evaluation of the evolution of the scores associated to every GPC. This is represented in Fig. 12, where we see the anomalies identified by labels in every GPCA group. Let us analyse these results following a per attack type classification:

- Scan attacks. These attacks are directly revealed by GPC1 (Port Scan Anomalies), but also GPC2 (burst of traffic anomalies) is able to reveal the amount of traffic generated in Scan44. Yet, GPC2 is not able to signal Scan11 as an anomaly, while GPC1 is.
- DoS attacks. We can check how DoS attacks are detected by GPC2 (bursts of traffic anomalies) and GPC3 (HTTP DoS), and they are struck with HTTP traffic. We are also

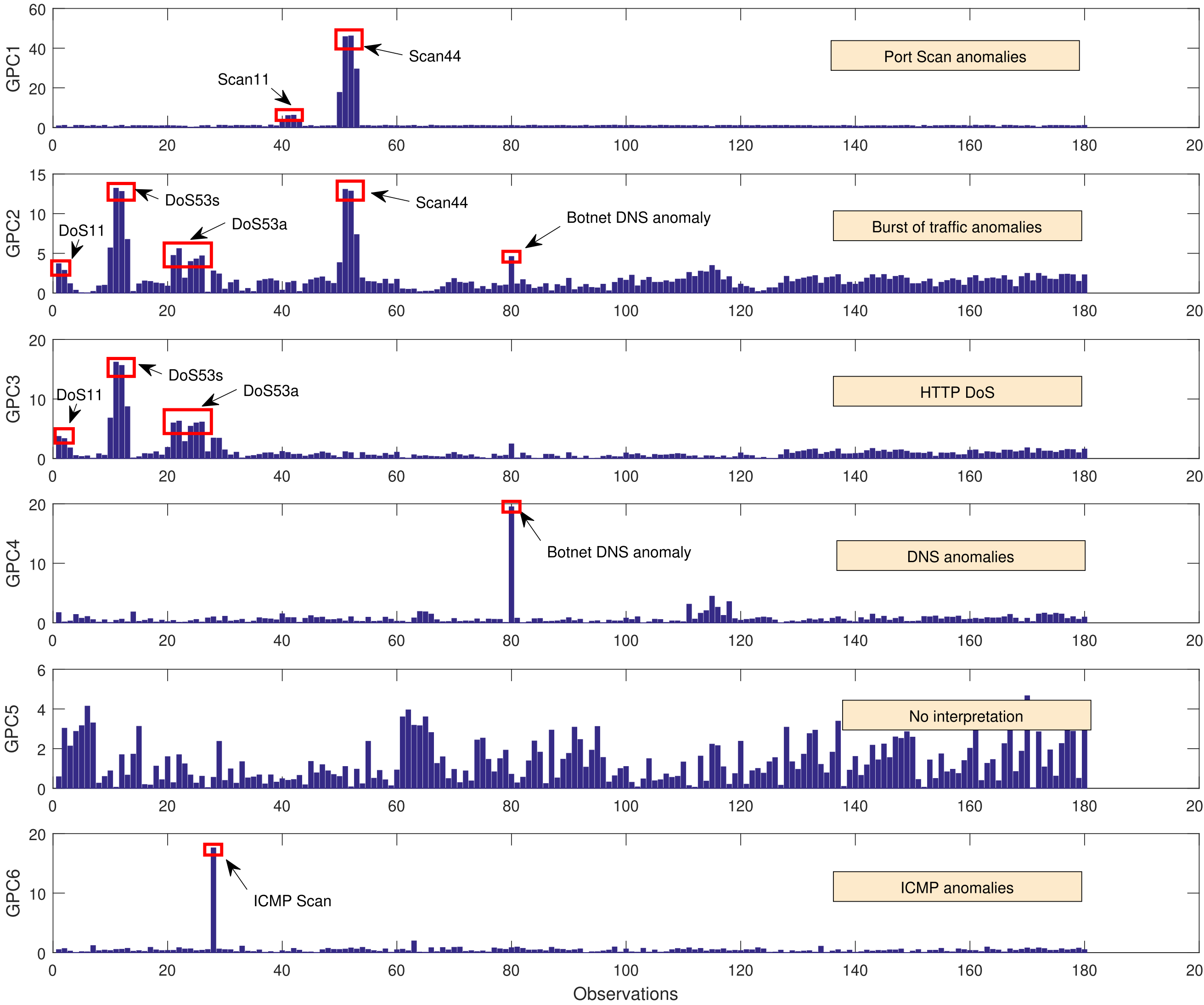


FIGURE 12. Evolution of scores in the different GPCA Groups for the UGR'16 trace.

able to see that DoS53s is detected with a higher level of scores, as the traffic volume generated in this attack is higher than in DoS11 and DoS53a.

- DNS anomaly generated by the Neris Botnet. This anomaly is directly pointed out at observation #80 by GPC3 (DNS anomalies), but it is also detected by GPC2 (bursts of traffic anomalies). Yet, the detection level in the case of GPC2 is more reduced than in GPC3, mainly because GPC2 only considers TCP traffic.
- ICMP anomaly. This anomaly is detected by GPC6 at observation #28.

VII. DISCUSSION

Through the previous two examples, we observe how both workflows (PCA and GPCA) can detect anomalies. We claim that the GPCA workflow is a good candidate to complement PCA, mainly because it is more natural for security analysts for the following reasons:

a) It allows incorporating expert knowledge in the model. We have done it with a weight associated with every feature, but alternative models can be used.

b) It allows prioritising the analysis by selecting specific groups based on the amount of variance of the components and/or expert knowledge. In the case of PCA, note that the analyst can only prioritise the anomalies to be studied according to the Tscore level (*e.g.*, Fig. 9), which could not be the best indicator of the real relevance of incidents.

c) Once the interpretation of groups is done in GPCA, the analysis of additional anomalies is straightforward, while in PCA a new (diagnosis) analysis is needed per any new anomaly that appears.

d) In some cases, GPCA will be able to detect new anomalies that remain hidden to PCA because they are small and/or sustained. If these anomalies fall in the groups prioritised in GPCA, they will be signalled with a higher probability. This is the case of IRC traffic in case study I and Scan11 in case study II, which remained hidden in PCA, while it can be identified in GPCA.

VIII. CONCLUSION

In this paper, we propose an exploratory anomaly detection methodology based on the Group-wise Principal Component Analysis (GPCA) method. This methodology has shown to be powerful and easy to understand by security practitioners without strong knowledge on multivariate analysis. It can also be combined with expert knowledge, allowing the analyst to tune the system according to her experience. The application of the approach is illustrated with two case studies. We believe this method is a useful addition to the security analyst toolbox.

REFERENCES

[1] VERIZONE. (2018). *Data Breach Investigations Report*. [Online]. Available: https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf

[2] D. Kish, L. Pingree, J. Heng, D. Gardner, A. Litan, P. Carpenter, E. Ahlm, S. Deshpande, R. Contu, and E. Kim "Market insight: Security market transformation disrupted by the emergence of smart, pervasive and efficient security," *Gartner*, Feb. 2017.

[3] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Comput. Secur.*, vol. 28, nos. 1–2, pp. 18–28, Feb./Mar. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404808000692

[4] A. Kanaoka and E. Okamoto, "Multivariate statistical analysis of network traffic for intrusion detection," in *Proc. 14th. Int. Workshop Database Expert Syst. Appl. (DEXA)*, Sep. 2003, pp. 1–5.

[5] M. L. Shyu, S. C. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," in *Proc. IEEE Found. New Directions Data Mining Workshop, Conjunct. 3rd IEEE Int. Conf. Data Mining (ICDM)*, 2003, pp. 171–179.

[6] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 2002.

[7] J. E. Jackson, *A User's Guide to Principal Components*. Hoboken, NJ, USA: Wiley, 2003.

[8] J. Camacho, G. Maciá-Fernández, J. Díaz-Verdejo, and P. García-Teodoro, "Tackling the big data 4 vs for anomaly detection," in *Proc. IEEE INFOCOM*, no. 1, Apr./May 2014, pp. 500–505.

[9] T. Kourti and J. F. MacGregor, "Multivariate SPC methods for process and product monitoring," *J. Qual. Technol.*, vol. 28, no. 4, pp. 409–428, 1996.

[10] C. F. Alcala and S. J. Qin, "Analysis and generalization of fault diagnosis methods for process monitoring," *J. Process Control*, vol. 21, no. 3, pp. 322–330, 2011.

[11] J. Camacho, P. García-Teodoro, and G. Maciá-Fernández, "Traffic monitoring and diagnosis with multivariate statistical network monitoring: A case study," in *Proc. IEEE Secur. Privacy Workshop (SPW)*, May 2017, pp. 241–246.

[12] J. Camacho, J. M. García-Giménez, N. M. Fuentes-García, and G. Maciá-Fernández, "Multivariate big data analysis for intrusion detection: 5 steps from the haystack to the needle," 2019, *arXiv:1906.11976*. [Online]. Available: https://arxiv.org/abs/1906.11976

[13] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 4, pp. 219–230, Oct. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1030194.1015492

[14] J. Camacho, A. Pérez-Villegas, P. García-Teodoro, and G. Maciá-Fernández, "PCA-based multivariate statistical network monitoring for anomaly detection," *Comput. Secur.*, vol. 59, pp. 118–137, Jun. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404816300116

[15] R. Theron, R. Magán-Carrión, J. Camacho, and G. Maciá-Fernández, "Network-wide intrusion detection supported by multivariate analysis and interactive visualization," in *Proc. VizSec*, Phoenix, AZ, USA, Oct. 2017, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/8062198/?reload=true

[16] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the LASSO," *J. Comput. Graph. Statist.*, vol. 12, no. 3, pp. 531–547, 2003. [Online]. Available: http://oro.open.ac.uk/3949/

[17] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, Jun. 2006.

[18] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. London, U.K.: Chapman & Hall, 2015.

[19] J. Camacho, R. A. Rodríguez-Gómez, and E. Saccenti, "Group-wise principal component analysis for exploratory data analysis," *J. Comput. Graph. Statist.*, vol. 26, no. 3, pp. 501–512, 2017.

[20] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[21] R. Zuech, T. Khoshgoftaar, and R. Wald, "Intrusion detection and big heterogeneous data: A survey," *J. Big Data*, vol. 2, no. 3, pp. 1–41, Dec. 2015.

[22] L. Wang and R. Jones, "Big data analytics for network intrusion detection: A survey," *Int. J. Netw. Commun.*, vol. 7, no. 1, pp. 24–31, 2017.

[23] G. Chao, Y. Luo, and W. Ding, "Recent advances in supervised dimension reduction: A survey," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 1, pp. 341–358, 2019. [Online]. Available: https://www.mdpi.com/2504-4990/1/1/20

[24] M. H. Rehman, C. S. Liew, A. Abbas, P. P. Jayaraman, T. Y. Wah, and S. U. Khan, "Big data reduction methods: A survey," *Data Sci. Eng.*, vol. 1, no. 4, pp. 265–284, 2016.

[25] W. Meng, W. Li, C. Su, J. Zhou, and R. Lu, "Enhancing trust management for wireless intrusion detection via traffic sampling in the era of big data," *IEEE Access*, vol. 6, pp. 7234–7243, 2018.

[26] P. Dahiya and D. K. Srivastava, "Network intrusion detection in big dataset using spark," *Procedia Comput. Sci.*, vol. 132, pp. 253–262, 2018.

[27] K. K. Vasan and B. Surendiran, "Dimensionality reduction using principal component analysis for network intrusion detection," *Perspect. Sci.*, vol. 8, pp. 510–512, Sep. 2016.

[28] T. Zhang and B. Yang, "Dimension reduction for big data," *Statist. Interface*, vol. 11, no. 2, pp. 295–306, 2018.

[29] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for traffic anomaly detection," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 35, no. 1, pp. 109–120, Jun. 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1269899.1254895

[30] A. Lakhina, M. Crovella, and C. Diot, "Characterization of network-wide anomalies in traffic flows," in *Proc. 4th ACM SIGCOMM Conf. Internet Meas. (IMC)*, vol. 6, 2004, pp. 201–206. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1028788.1028813

[31] P. Nomikos and J. F. Macgregor, "Monitoring batch processes using multiway principal component analysis," *AIChE J.*, vol. 40, no. 8, pp. 1361–1375, 1994.

[32] N. Tracy, J. Young, and R. Mason, "Multivariate control charts for individual observations," *J. Quality Technol.*, vol. 24, no. 2, pp. 88–95, 1992.

[33] M. Aiello, M. Mongelli, E. Cambiaso, and G. Papaleo, "Profiling DNS tunneling attacks with PCA and mutual information," *Logic J. IGPL*, vol. 24, no. 6, pp. 957–970, 2016.

[34] G. Fernandes, Jr., L. F. Carvalho, J. J. P. C. Rodrigues, and M. L. Proença, "Network anomaly detection using IP flows with principal component analysis and ant colony optimization," *J. Netw. Comput. Appl.*, vol. 64, pp. 1–11, Apr. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804516000618

[35] D. Jiang, C. Yao, Z. Xu, and W. Qin, "Multi-scale anomaly detection for high-speed network traffic," *Trans. Emerg. Telecommun. Technol.*, vol. 26, no. 3, pp. 308–317, 2015.

[36] Z. Chen, C. K. Yeo, B. S. Lee, and C. T. Lau, "Detection of network anomalies using improved-MSPCA with sketches," *Comput. Secur.*, vol. 65, pp. 314–328, Mar. 2017.

[37] K. Peng, V. C. M. Leung, and Q. Huang, "Clustering approach based on mini batch kmeans for intrusion detection system over big data," *IEEE Access*, vol. 6, pp. 11897–11906, 2018.

[38] F. Salo, A. B. Nassif, and A. Essex, "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection," *Comput. Netw.*, vol. 148, pp. 164–175, Jan. 2019.

[39] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, pp. 331–349, Aug. 1979.

[40] M. Fuentes-García, G. Maciá-Fernández, and J. Camacho, "Evaluation of diagnosis methods in PCA-based multivariate statistical process control," *Chemometrics Intell. Lab. Syst.*, vol. 172, pp. 194–210, Jan. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743917302046?via%3Dihub

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

[42] J. Camacho, "Missing-data theory in the context of exploratory data analysis," *Chemometrics Intell. Lab. Syst.*, vol. 103, pp. 8–18, Aug. 2010.

[43] *VAST Challenge 2012*. [Online]. Available: http://www.vacommunity.org/VAST+Challenge+2012

[44] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, and R. Therón, "UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs," *Comput. Secur.*, vol. 73, pp. 411–424, Mar. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404817302353

[45] *Malware Capture Project*. Accessed: Jul. 1, 2019. [Online]. Available: https://mcfp.weebly.com/ctu-malware-capture-botnet-42.html

JOSÉ CAMACHO received the degree in computer science from the University of Granada, Spain, in 2003, and the Ph.D. degree from the Technical University of Valencia, in 2007. He is currently an Associate Professor with the Department of Signal Theory, Telematics and Communication and a Researcher with the Information and Communication Technologies Research Centre, University of Granada. His research interests include exploratory data analysis, anomaly detection and optimization with multivariate techniques applied to data of very different nature, including manufacturing processes, chemometrics, and communication networks. He is especially interested in the use of exploratory data analysis to big data. His Ph.D. was awarded with the second Rosina Ribalta Prize to the best Ph.D. projects in the field of information and communication technologies (ICT) from the EPSON Foundation, and with the D. L. Massart Award in Chemometrics from the Belgian Chemometrics Society.

ROBERTO THERÓN received the Diploma degree in computer science from the University of Salamanca, the B.S. degree from the University of A Coruña, the B.S. degree in communication studies, the B.A. degree in humanities from the University of Salamanca, and the Ph.D. degree from the Research Group Robotics, University of Salamanca. His Ph.D. thesis was on parallel calculation of the configuration space for redundant robots. He is currently the Manager of the VisUSAL Group (within the Recognized Research Group GRIAL), University of Salamanca, which focuses on the combination of approaches from computer science, statistics, graphic design, and information visualization to obtaining an adequate understanding of complex data sets. He has authored over 100 articles in international journals and conferences. In recent years, he has been involved in developing advanced visualization tools for multidimensional data, such as genetics or paleo-climate data. In the field of visual analytics, he develops productive collaborations with groups and institutions internationally recognized as the Laboratory of Climate Sciences and the Environment, France, or the Austrian Academy of Sciences, Austria. He received the Extraordinary Doctoral Award for his Ph.D. thesis.

JOSÉ M. GARCÍA-GIMÉNEZ received the B.Sc. and M.Sc. degrees in telecommunication engineering from the University of Granada, Spain, in 2016 and 2018, respectively, where he is currently with the Department of Signal Theory, Telematics and Communications. His research interests include anomaly detection and network security, specifically on the use of data analysis techniques for network security.

GABRIEL MACIÁ-FERNÁNDEZ is currently an Associate Professor with the Department of Signal Theory, Telematics and Communications, University of Granada, Spain, where he belongs with the Network Engineering and Security (NESG) Research Group. His research interests include systems and network security, with special emphasis on intrusion detection, reliable protocol design, penetration testing techniques, network information leakage, and denial of service.

PEDRO GARCÍA-TEODORO is currently a Full Professor with the Department of Signal Theory, Telematics and Communications, University of Granada, Spain, where he is also the Head of the Research Group Network Security and Engineering Group (NESG). His current research interests include computer and network security, especially focused on anomaly-based intrusion detection and denial of service attacks.