Detecting and Coloring Anomalies in real cellular network using Principle Component Analysis

Y. Segal, D. Vilenchik, and O. Hadar

Communication Systems Engineering Department,
Ben Gurion University of the Negev (BGU),
Beer-Sheva, 84105,
Israel[1]

***Abstract*— Anomaly detection in a communication network is a powerful tool for predicting faults, detecting network sabotage attempts and learning user profiles, for marketing purposes and quality of services improvements. In this article, we convert the unsupervised data mining learning problem into a supervised classification problem. We will propose three methods for creating an associative anomaly within a given commercial traffic data database and demonstrate how, using the Principle Component Analysis (PCA) algorithm, we can detect the network anomaly behavior and classify between a regular data stream and a data stream that deviates from a routine, at the IP network layer level. Although the PCA method was used in the past for the task of anomaly detection, there are very few examples where such tasks were performed on real traffic data that was collected and shared by a commercial company.**

The article presents three interesting innovations: The first one is the use of an up-to-date database produced by the users of an international communications company. The dataset for the data mining algorithm retrieved from a data center which monitors and collects low-level network transportation log streams from all over the world. The second innovation is the ability to enable the labeling of several types of anomalies, from untagged datasets, by organizing and prearranging the database. The third innovation is the abilities, not only to detect the anomaly but also, to coloring the anomaly type. I.e., identification, classification and labeling some forms of the abnormality.

Keywords— Anomaly detection; PCA; Data Mining; Machine learning;

I. INTRODUCTION

Anomaly detection which is based on Network traffic analysis tools are the foundation stones for network upgrades, protecting against cyber-attacks, and are a marketing tool for analyzing user profiles. Many heuristics can serve as starting points for filtering out data that flows at extremely high speeds. Analysis of network traffic is the most effective means of reducing search within the amount of information required for further analysis. Business companies use network traffic testing tools as the primary means of their solution architecture for intelligence and law enforcement bodies that monitor national internet services providers (ISP). It is also a significant focus on the solution concept of companies that offer optimization and advertising solutions based on network transportation.

Traffic anomaly detection has received a great deal of attention in the research literature. While there has been some work that leverages data structures to find heavy-hitters [[1],[2], most papers have utilized statistical-analysis techniques to detect outliers in traffic time series. Numerous methods have been evaluated, including wavelets [3], moving average variants, Fourier transforms [4],[5], Kalman filters [6], and PCA [7]. Early work in this area often analyzed data from a single link [3], whereas more recent papers have shown promising results by examining network-wide measurements [8]. With such a large body of work, it becomes increasingly important to be able to compare presented approaches. While there have been a few papers that analyzed a subset of the statistical-analysis techniques [4],[5], researchers have only very recently begun investigating how data-reduction technologies impact the ability to detect traffic anomalies [9]. Much in the same way that early papers on traffic anomaly detectors had a limited scope, this new line of work has analyzed the impact of only one form of data-reduction [10], on only one type of traffic anomaly [11], or analyzed data from a small number of links [12].

We are focusing on unsupervised techniques for big cellular data set. Our observation vectors have 97 different parameters. In the literature, various strategies proposed for dimensionality reduction [13]. The actual dimensionality reduction methods can classify into two classes: Feature extraction and Feature selection. Feature selection aims to seek optimally or suboptimal subsets of the original features [14], by preserving the main information carried by the collected complete data, to facilitate future analysis for high-dimensional problems. Another approach is the opposite approach, instead of reducing the dimensionality, Breiman [15] suggested to increase the dimensionality by adding many functions of the predictor variables. Two outstanding examples of work in this direction are the AmitGeman method [16] and support vector machines [17]. In feature extraction model [18], the original features in the measurement space initially transformed into a new dimension-reduced space via some specified transformation. Significant characteristics determined in the new axis.

Viswanath et al. [19] used PCA to classify Facebook users as either “normal” or “anomalous” (user considered anomalous

[1] This work was supported by the Israel Innovation Authority (Formerly the Office of the Chief Scientist and MATIMOP).

if its behavior was tagged as such by Facebook). Other papers that applied PCA successfully for anomaly detection include [20],[21],[22],[23],[24].

The ability to enable the labeling of several types of anomalies, from untagged datasets presented in some other works such as[25]. In [25] the validation data is split into two sets, one set that represents nominal data, and the other that represents potentially anomalous data. In some instances, benign anomalies may appear in the validation of nominally categorized data where there was no prior suspicion of them. In our case, we are adding external knowledge such as geographical location or period which allows us to classify the data without mixing between anomalies and regular sets.

Our study in this article identifies and evaluates three main challenges: (1) Identifying anomalies from logs of real network traffic. (2) Development of new statistical algorithms to identify anomalies that are adapted to the unique problem. (3) Verification of the quality of results by breaking the data into normal and the rest according to some parameter: cell congestion, time rather than statistical methods only.

II. ANOMALY DETECTION TECHNICS

The article deals with two main challenges: The first one is that there is no definition of what an anomaly is, no training sets for anomalies. In practice the data is unlabeled. The second challenge is handling big-data stream, off-line and certainly in an online situation is a complicated technological challenge. The techniques for identifying anomalies can be divided into two types: Techniques, which are unsupervised and assume that most of the database observations represent normal or normal cases. For example, cluster analysis techniques can be used to characterize typical representation. A representation that does not belong to any cluster defined as an anomaly. Supervised techniques in which database observations were pre-categorized for "normal" or "abnormal" observations. In this case, computational learning methods can use for categorized training, which enables the classification of new observation that we have not encountered in the learning process.

We will use the PCA method which trained on normal behavior and identifies deviations from this behavior. We are showing characteristics that best explain the normal behavior. PCA will do this by projecting on a base with a smaller or the same dimension on which we will perform statistical analyzes.

Now we are going to explain the PCA model. The first principal component (PC) is defined to be the direction (unit vector) $\boldsymbol{V_1} \in \mathbb{R}^{\boldsymbol{p}}$ in which the variance of x is maximal. The variance of x in direction v is given by the expression $\boldsymbol{v^T \Sigma v}$. Therefore $\boldsymbol{V_1} = \boldsymbol{argmax_{V}}_{\in \mathbb{R}^p} \in \boldsymbol{v^T \Sigma v}$. The latter is the Rayleigh Quotient definition of the largest eigenvalue of a matrix, therefore $\boldsymbol{V_1}$ is the leading eigenvector of $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda_1 = v_1^T \Sigma v_1}$ is the variance explained by $\boldsymbol{V_1}$. The remaining PCs are defined in a similar way and together they form an orthonormal basis of $\mathbb{R}^{\boldsymbol{p}}$. The sample PCs $\widehat{\boldsymbol{v}}_{\boldsymbol{1}},\ldots,\widehat{\boldsymbol{v}}_{\boldsymbol{p}}$ are the eigenvectors of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$. Under various reasonable assumptions it was proven that the principal components $\boldsymbol{V_1},\ldots,\boldsymbol{V_p}$ converge to the sample ones $\widehat{\boldsymbol{v}}_{\boldsymbol{1}},\ldots,\widehat{\boldsymbol{v}}_{\boldsymbol{p}}$[26], [27]. We assume that this is true in our case, and we justify it by the fact that we are in the "fixed p large n" regime, where the ratio p/n tends to 0.

III. CREATING AN ANOMALY DATABASE

This research deals with the study of traffic of a cellular communication network to discover anomaly based on traffic data. A cellular network contains many access points to the Internet. Designated routers serve as a bridge between the Internet and the cellular data flow. These routers regularly monitored so that the traffic information through them is centralized into an information center, allowing a holistic, international view of the behavior of network traffic.

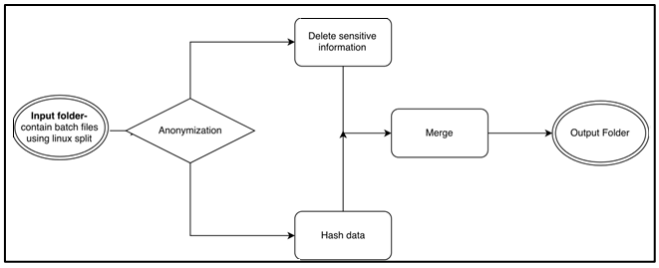


Figure 1:Block diagram to preserve data confidentiality

Naturally, this information center (which based on the Log Center) generates significant data at the rate of tens of gigabytes per second. It should emphasize that the stream of information and information content is not constant and changes according to use. Therefore, we averaged each measured parameter, separately, in time units. Such as averaging over an hour of HTTP request size. Another problem we had to deal with was maintaining anonymity and confidentiality. The cellular networks traffic logs contain private user information. There is a need for log anonymization platform scalable for big-data. As a result, we defined a batch based anonymization tool.

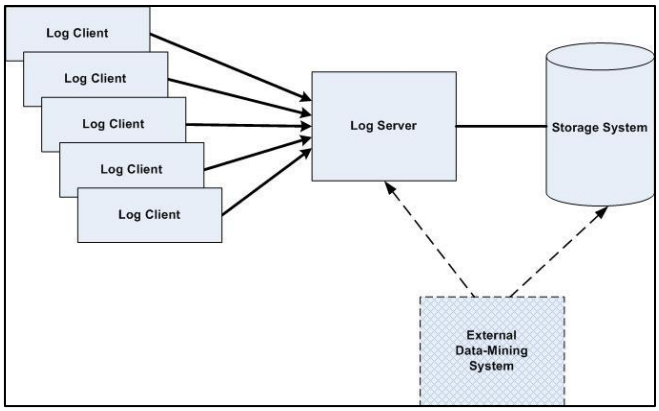


Figure 2: Data center architecture

The database fields divided into three types: Anonymous fields- Those fields used as is; Fields that reveal user information- Those fields have been deleted; Fields that can be used but still have indirect information about the user and therefore have a low risk of user exposure. For those fields, we used at the beginning the well-known PBKDF2 anonymization algorithm. PBKDF2 is very secure and used for protecting password on almost every server. The drawback is that the algorithm is slow. It makes the anonymization process to be

prolonged. Therefore, currently, we are working with SHA-256 due to resource constraints. It takes one minute to anonymize a log of 1 Gb while almost a day with PBKDF2.

Our database includes fields of the complete set of transaction log records and their formats, but the transaction log fields in any specific geographical location depend on its local configuration. HTTPS records contain data per connection, as transactions not identified. HTTPS and HTTP Tunneled transactions records include only fields that captured during their limited processing (e.g., timing, data amounts IP addresses, etc.)

IV. EXPERIMENTS AND RESULTS

This section describes our PCA model, methodology and software for detecting and coloring the traffic anomaly by manipulate the same database in three major ways.

A. Time-period traffic analysis

The first method for discovering anomaly based on different time-period traffic analysis. The information divided into three-periods categories: Night\Early Morning, Morning and Evening from all geographical locations. The motivation was to examine whether traffic congestion can discover based on the assumption that each time profile has a unique pattern. Based on the observed time profile, we injected vector information belonging to other time profiles, and tried to discover them as an anomaly.

Initially, the time profiles tested naively, and elementary statistical parameters such as mean and standard deviation were measured to characterize each period by mean and standard deviation of its bytes stream volume. Sample results presented in Table 1.

Table 1: Elementary parameters from some data sets examples

Log File	Records	Date	Time	Hours Recorded	Mean	Std. dev.
#1 (1GB)	890,650	Sunday, 08.05.16	Night\Early Morning 00:00-07:59	6	35,105.823	700,800
#2 (20GB)	20,131,028	Tuesday, 06.12.16	Morning 08:00-11:59	1	36,531.608	639,683.435
#3 (27GB)	28,676,280	Monday, 05.12.16	Evening, 21:00-23:59	3	40,428.878	820,315.637

Table 1 demonstrates the fact that an attempt to classify periods via first and second order statistical characteristics does not allow proper classification. The average plus the variance of each period creates an overlap that does not allow sufficient separation.

Since the naive method of detecting the anomaly of different time periods is not relevant, we used the PCA method to identify an anomaly in datasets gathered in one period and reached the system at a different time.

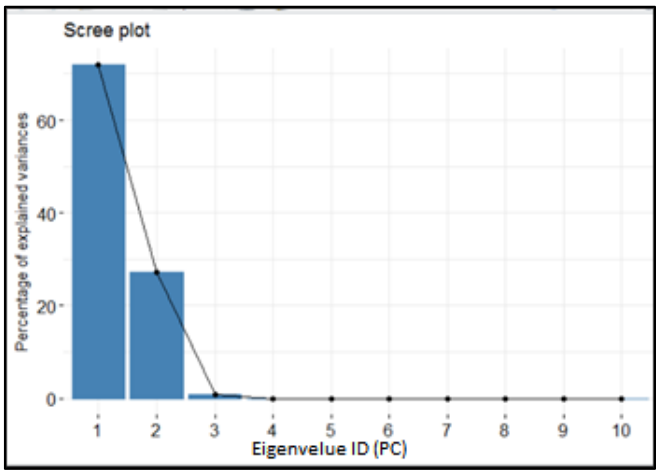


Figure 3: Time-periods major PCs

As a conclusion from Figure 3, the PC effect is negligibly starting from the fourth eigenvalue. Table 2 presents some of the eigenvectors components for each relevant eigenvalue and associates them to the original dataset components.

Table 2: Eigenvectors and time-periods components association table

PC3	PC2	PC1	Original Vector Components
2.60E-06	-7.69E-07	7.99E-08	RESPONSE_CODE
-0.50005	0.006683	-0.00453	RESPONSE_ORIG_SIZE
-0.49974	0.006682	-0.00453	DOWNLOAD_DATA_SIZE
-8.35E-10	-2.65E-09	7.30E-10	COMPRESSION_LVL
-1.95E-07	4.41E-08	-8.84E-09	CONTENT_TYPE
-9.77E-05	-1.32E-05	8.88E-07	UPSTREAM_SIZE
-0.49975	0.006689	-0.00453	DOWNSTREAM_SIZE

To colored anomaly, compare to the Weekday Night Hours (WNH) dataset, we transformed all time-periods datasets into a new PCA space. We used the Normal State Transformation Matrix (NSTM), calculated by performing principal components analysis on the WNH dataset. We extracted the two independent eigenvectors and performed the projection of the datasets of all time-periods on a single shared two-dimensional graph (Figure 4). We received a reduction of the dimension of information from a space of 97 dimensions to a 2-dimensional space that allows us to present a point of view that represents the distribution in a state without anomaly (Night hours). Similarly, we carried out the information with Weekend hours, and Evening hours and these points were marked in Yellow and Red respectively.

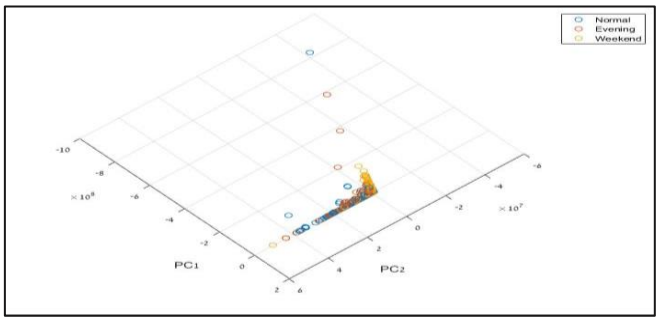


Figure 4: Colored time-periods classes

After transforming each time-period dataset, separately, by the NSTM, we figured, per time-period dataset, the average absolute value of projections, in the direction of each eigenvector, individually. The following graph (Figure 5) shows the average absolute values of the projection on each PC.

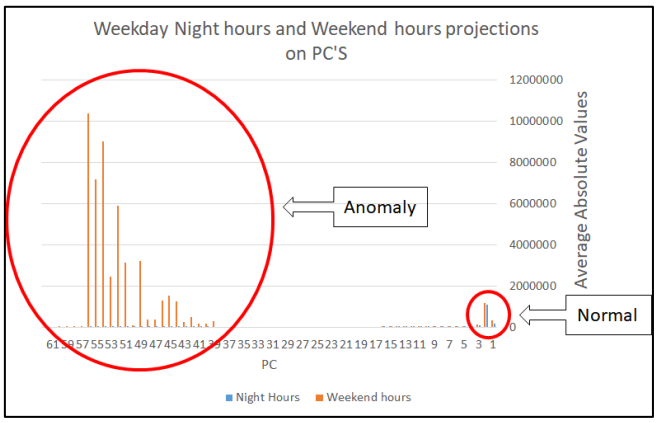


Figure 5: Comparing average absolute values of time-periods projections on all PCs

Significant PCs for Weekday Night hours (Blue color) in importance order: 1. PC1, 2. PC2, 3. PC3

Significant PCs for Weekend hours (Orange color) in importance order: 1. PC56, 2. PC54, 3. PC55

With PCA model and NSTM, we rotate the original dataset axis system so that the eigenvectors become the basis of the new axial system. The PC column in Table 3 indicates the sequence number of the most relevant eigenvectors which the dataset vectors projected on it (most relevant PC's are the PCs with the highest average absolute values after projections into the PC direction). Our dataset eigenvectors are in dimension 97, meaning that each eigenvector has 97 components that can be interpret as eigenvector weights. In Table 3, the weight column presents the X highest eigenvector weights per PC. A projection toward a PC is a linear combination (inner product) between original dataset vectors and eigenvector weights. Therefore, the eigenvector weights can interpret as the importance of the dataset vectors components (before projection). The last column in Table 3 connects between the PCA space and the real dataset log components. It allows us to interpret more efficient our log data and to characterize the most relevant features that have the highest influence on the data transportation during different time-periods.

When examining anomaly at different times of the week, it is easy to see that the distribution of the evening and morning hours is almost identical. But when compared to the weekend we got an extreme deviation, when in fact all significant PCs that belong to the "normal traffic" dataset are not substantial in the weekend traffic. The significant PC's for the weekend hours focused mainly on watching the video, and moreover, it was noticeable that most of the video views had been interrupted (indicating a traffic load).

Table 3: Eigenvectors components weights interpretation

PC	Weight	Original dataset columns Description
1	0.998	The payload size only (from cache)
	0.049	The estimated connection bandwidth at session beginning
	0.004	The size of response from WEB without the headers
	0.004	Size of original response on WEB containing headers
	0.004	Size of the response data on the RAN side, including headers.
2	0.998	The estimated connection bandwidth at session beginning
	0.049	The payload size only (from cache)
3	0.500	The size of response from WEB without the headers
	0.500	Size of original response on WEB containing headers
	0.499	Size of the response data on the RAN side, including headers.
54	0.514	When a Media file or Software download file is requested in range requests, this field holds the full resource length
	0.437	The number of times the video stopped playing
	0.410	The time it took the video to start playing in milliseconds
	0.261	This field indicates the method by which the file was processed for Multi-Level Transcoding and Dynamic Rate Adaptation
55	0.659	The stalls average time in milliseconds
	0.409	When a Media file or Software download file is requested in range requests, this field holds the full resource length
	0.398	This field indicates the method by which the file was processed for Multi-Level Transcoding and Dynamic Rate Adaptation
56	0.700	The stalls average time in milliseconds
	0.591	When a Media file or Software download file is requested in range requests, this field holds the full resource length
	0.287	The time it took the video to start playing in milliseconds

B. Congestion traffic analysis

Second data structure: classification by congestion fields. The database contains some columns describing the level of transportation load. That refers to three levels of the number of Bytes per second passing through the examined network junction. 0 - low load level, 1 - medium load level and 2 - high load level.

After computing the PCA model on the low-level congestion dataset (level 0), sorting them and selecting the 3 with the highest eigenvalues, we extracted the three independent eigenvectors (the ones that belong to the three highest eigenvalues) and performed the projection of the datasets of all levels on a single shared three-dimensional graph. If to be more precise, our 97x3 matrix operator transformed each vector that belongs to level 0 into a three-dimensional vector and colored them as a blue dot in the graph. We received a reduction of the dimension of information from a space of 97 dimensions to a 3-dimensional space that allows us to present a point of view that represents the distribution in a state without anomaly. Similarly, we carried out the information with a congestion level 1, and a congestion level 2 and these points were marked in purple and red respectively.

Since the base is three dimensions we could display the results in a 3D graph, and we obtained the following results:

To colored anomaly, compare to the conjunction 0 dataset principle components (PCs), we transformed all conjunctions levels (0,1,2) into a new PCA space. We used the Normal State Transformation Matrix (NSTM), obtained by performing principal components analysis on the level 0 conjunction dataset. After transforming each conjunction level dataset, separately, by the NSTM, we calculated, per conjunction level dataset, the average absolute value of projections, in the direction of each eigenvector, separately. The following graph shows the average conjunction of the projection on each PC.

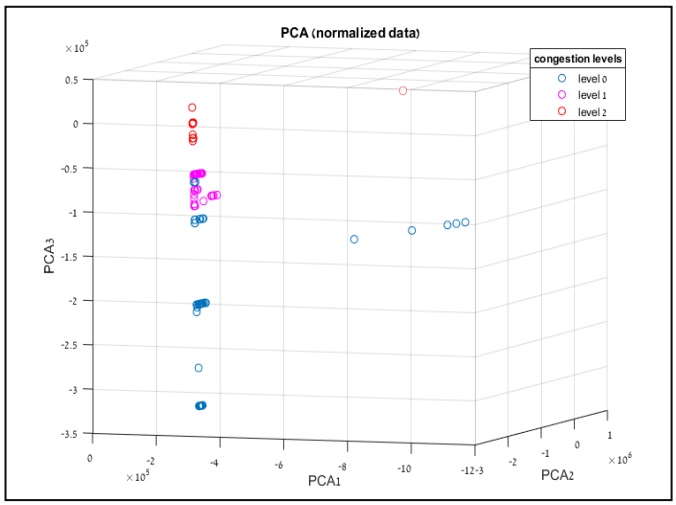


Figure 6:Congestion Levels via PCA

PC-s significant for conj1 (Level 0 - Blue color) in importance order: 1. PC17, 2. PC41, 3. PC38. PC-s are significant for conj2 (Level 1 - Orange color) in importance order: 1. PC41, 2. PC38, 3. PC32

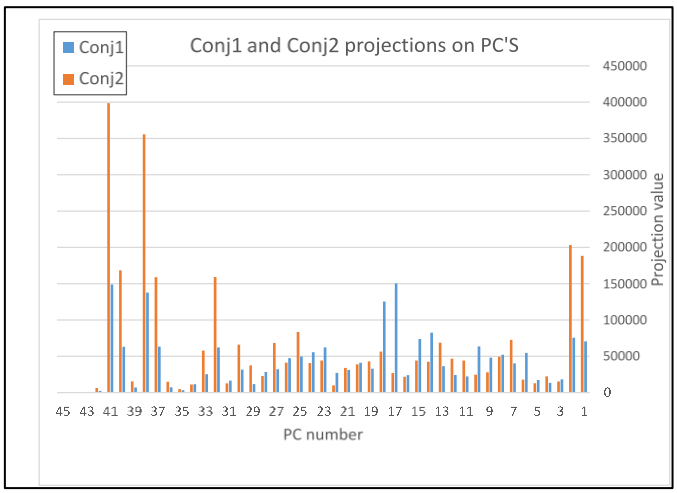


Figure 7: Comparing average absolute values of congestion projections on all PCs.

The PCA method allows us to distinguish between different conjunctions levels by performing a linear transformation of a new incoming measurement vector to the trained PC's space. If the new vector components (after transformation) will present in its component 41, 28 and 32 values which are significantly higher compare to its other components, then we know that there is an abnormal state and the reason for the anomaly is that we have moved from level 0 to congestion level 1.

It is important to emphasize that the level of congestion does not represent a single parameter whose value has exceeded a specific threshold value, that can interpret as a sole conjunction criterion. The conjunction criteria is a linear combination of 97 different measurements (components), each of which can be at its normal values range. Only the linear combination indicates an increase in the level of congestion. Therefore, a naïve and manual attempt to detected and recognize an anomaly in a vector of 97 dimensions is in the range of difficult to the point of impossible. The PCA method allows us to lower the vector dimension and also introduces interpretation that can detect and recognize congestion anomalies in low-level network transportation.

C. Geographical traffic analysis

The third data structure deals with geographical location. Routers that spread all over the world collected data stream flow from anonymous internet domains (150 different domains - one column per domain, each line is one hour aggregated bytes flow). Those datasets contain three months transportation log data. It divided into three continents groups: Africa, North America, and South America. The aim was to reveal information coming from one mainland within another mainland (for example, learning about the African continent, injecting vectors from the North American continent, and coloring such vectors as anomalies).

Remark: The original database was 97 dimensions and at a size that required analysis with big-data tools such as SPARC and HADOOP. One of the ways to reduce big-data is the use of preliminary network traffic expert knowledge. Therefore, based on the expert's guidance, which explained that hour resolution and domains transportation load is enough to detect a geographic anomaly, we performed preliminary processing on the original 97-dimensional database. Instead of doing machine learning with heavy-duty distributed cloud processing power, we conducted pre-processing utility that reduced the data into several tens of gigabytes without compromising the quality and ability to detect and classify anomalies. We extracted only columns with domain loads (The domain names converted to symbols for to preserve user confidentiality). Additionally, we reduced our dataset from 97 to 10 dimensions by selecting the top ten domains (classified by traffic average) on each of the three continents.

Figure 8 expose the common variance between African samples (X-axis) and North American samples (Y-axis) as obtained by the Canonical Component Analysis (CCA) operation. (The correlation coefficient is 0.82). It can understand that there is a great deal of commonality between the two sources of information and therefore we cannot expect to identify anomaly naively (as it presented with time-periods or conjunctions level).

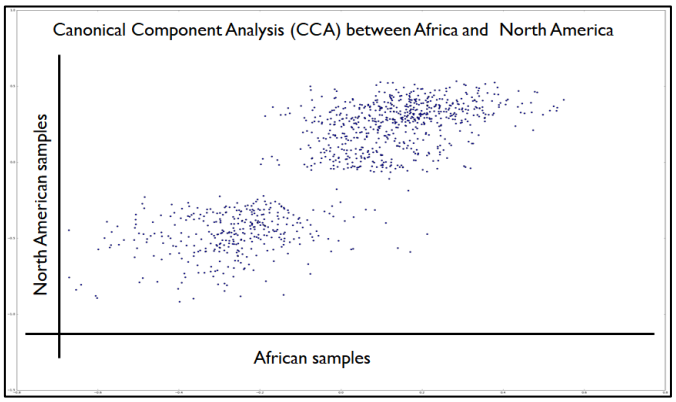


Figure 8: Canonical Component Analysis (CCA) between Africa and North America, correlation = 0.82

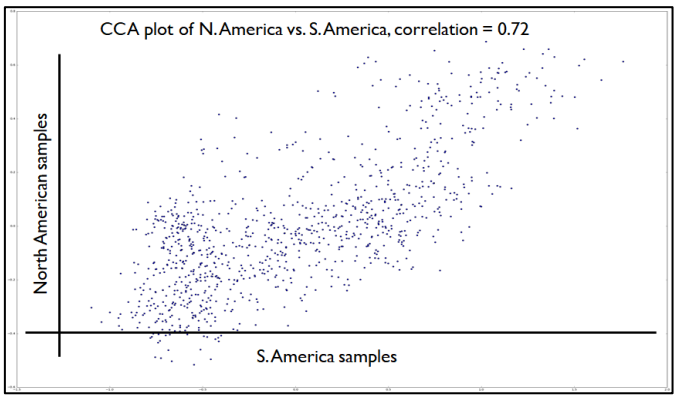


Figure 9: CCA of N. America vs. S. America, correlation = 0.72

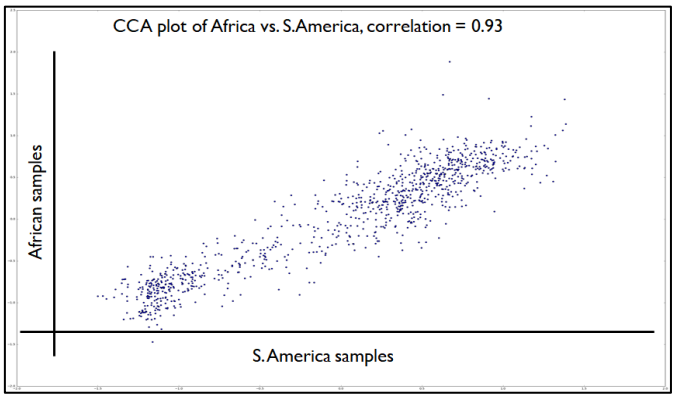


Figure 10: CCA plot of Africa vs. S.America, correlation = 0.93

The attempt to use the method used for time-period and congestion level, in a way that each class has other PCs that describe the specific type is inappropriate for the geographical case because here there is a strong correlation between the different PCs (see Figure 8, Figure 9, Figure 10). So, in the geographical situation, we look at the visual graph form, obtained after the projection. It can see that in the PC space each geographic region is placed elsewhere in the graph. And therefore, it is possible to perform separation using a linear regression line in the PC domain as a threshold between the different locations.

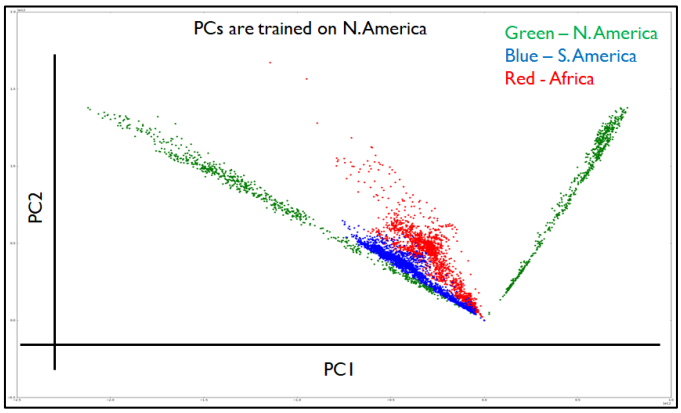


Figure 11: PCS that trained on N. America domains Explained variance: PC1=84%, PC2=13%

The geographical dataset is an example of analyzing different utilization mixtures with different locations and the ability to detect context (geographical) according to its pattern in the PCA space.

In Figure 11 PCs trained on N. America data as normal dataset. We found in the training set that two eigenvalues can explain most of the variance (Explained variance for PC1 is 84% and for PC2 is 13%). Of the two eigenvectors belonging to the most explanatory eigenvalues (above eigenvalues), we extracted the eigenvectors components with the highest weights. The domains that multiplied during the PCA transformation with those most upper weights are the most dominant domain in the North American continent - Domains marked as 0, 1 and 2 were dominant in the North American continent.

In addition to extracting the 150 dominant domains, the pre-processing utility allows us to produce another query on the geographical dataset. It enabled the extraction of traffic classification by 80 different types of communications protocols (HTTP, AAC, UDP, F4V, etc.). The protocols arranged in columns. Each table row is the amount of traffic per hour. (Each table is a different continent, each column in the table is a different protocol, each line is the amount of traffic at a given time.

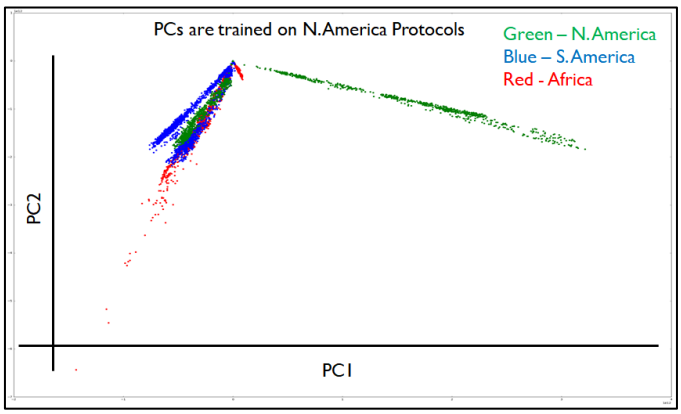
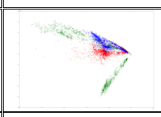
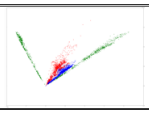
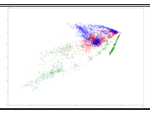
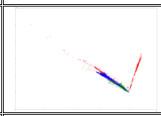
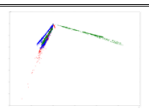
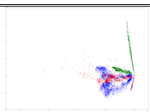


Figure 12: PCs are trained on N. America Protocols Explained variance: PC1=81%, PC2=13%

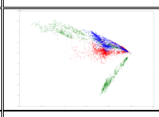
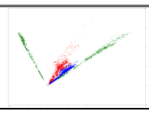
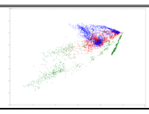
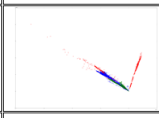
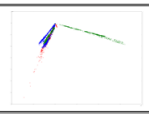
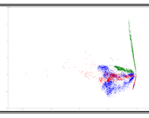
In Figure 12 PCs trained on N. America protocol as normal datasets (vector of 80 dimensions). After the PCA transformation, two eigenvalues in the training set can explain most of the variance (Explained variance: 84%,13%). From the two eigenvectors belonging to the most explanatory eigenvalues, we extracted, once again, the eigenvectors components with the highest weights. The protocols that multiplied during the PCA transformation with those most upper weights are the most dominant protocol in the North American continent - the top 5 features of PC1 are HTTP.Other, Image.WebP, HTTPS.Web Messaging, and Torrent.

Table 4: Summary of PCA geographical projections

	Trained on	Africa	N. America	S. America
Domains	N. America S. America Africa			
	Explained variance	85%, 7%	84%,13%	84%,12%
	Top Domains of PC1	0,1,2	0,1, 2	0,1
Protocols	N. America S. America Africa			
	Explained variance	76%,18%	81%,13%	76%,18%
	Top 5 features of PC1	*HTTPS.Web, Audio.AAC, Video.Facebook.CDN, Video.Google, HTTP.Other,*	*HTTP.Other, Image.WebP, HTTPS.Web, Messaging, Torrent*	*HTTPS.Web, Video.Facebook.CDN, Video.F4V, Facebook.CDN, UDP.Other*

The summary of the results of the PCA transformation of the cellular network transportation, in favor of the geographical investigation, by the cross-domain and by the cross-protocol queries, is summarized in Table 5.

Table 5: Summary of PCA geographical projections

	Trained on	Africa	N. America	S. America
Domains	N. America S. America Africa			
	Explained variance	85%, 7%	84%,13%	84%,12%
	Top Domains of PC1	0,1,2	0,1, 2	0,1
Protocols	N. America S. America Africa			
	Explained variance	76%,18%	81%,13%	76%,18%
	Top 5 features of PC1	*HTTPS.Web, Audio.AAC, Video.Facebook.CDN, Video.Google, HTTP.Other,*	*HTTP.Other, Image.WebP, HTTPS.Web, Messaging, Torrent*	*HTTPS.Web, Video.Facebook.CDN, Video.F4V, Facebook.CDN, UDP.Other*

V. RELIABILITY AND VALIDITY

The t-tests have used for verifying the accuracies. Statistical analyses are used to conclude if the accuracies taken with the proposed approach are significantly distinct from the others (whereas both the distribution of values were normal). The test for assessing whether the data come from normal distributions with unknown, but equal, variances is the Lilliefors test. Obtaining results by comparing the results produced by 100 trials (at each trial we used a different split of the data). Obtaining a test decision for the null hypothesis that the data comes from independent random samples from normal distributions with equal means and equal but unknown variances. Results show a statistical significant effect in performance (p-value < 0.05, Lilliefors test H=0).

VI. CONCLUSIONS AND FUTURE DIRECTIONS

In this article, we convert the unsupervised learning problem into a supervised classification problem. We proposed four methods for creating an associative anomaly within a given commercial traffic data database. We demonstrated how, using the PCA algorithm, we can detect the network anomaly behavior and classify between a regular data stream and a data stream that deviates from a routine, at the IP network layer level. The experiments we performed showed high and stable results, for example, it obtained that the detection and coloring of the time-period anomaly was PD = 90.2% and PF = 0.5%. and PD = 89.9% and PF = 1.5% for the detection of a geographical domains anomaly. Similar results obtained for the detection of anomalies in traffic congestions and for the geographical protocols anomalies.

The next direction that this study can take is the usage of advanced time series tools such as Facebook's Prophet tool. With time series tools, we expect to find trends and cycles in the dataset that will enable us to make an expectation forecast graph that any deviation from a predefined threshold around the forecasting graph will be defined as an anomaly.

VII. REFERENCES

[1] C. Estan, S. Savage, and G. Varghese, "Automatically inferring patterns of resource consumption in network traffic," in ACM SIGCOMM, (Karlsruhe, Germany), pp. 137–148, 2003.

[2] Y. Zhang, S. Singh, S. Sen, N. Duffield, and C. Lund, "Online identification of hierarchical heavy hitters: Algorithms, evaluation, and applications," in ACM Internet Measurement Conference, (Taormina, Sicily, Italy), pp. 101–114, 2004.

[3] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in ACM Internet Measurement Workshop, (Marseille, France), pp. 71–82, 2002.

[4] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, "Sketch-based change detection: Methods, evaluation, and applications," in ACM Internet Measurement Conference, (Miami Beach, FL, USA), pp. 234–247, 2003.

[5] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in ACM Internet Measurement Conference, (Berkeley, California, USA), October 2005.

[6] A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in ACM Internet Measurement Conference, (Berkeley, California, USA), October 2005.

[7] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in ACM SIGCOMM, (Philadelphia, Pennsylvania, USA), pp. 217–228, 2005.

[8] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in ACM SIGCOMM, (Portland, Oregon, USA), pp. 219–230, 2004.

[9] A. Soule, H. Ringberg, F. Silveira, J. Rexford, and C. Diot, "Detectability of traffic anomalies in two adjacent networks," Passive And Active Measurement Conference 2007.

[10] J. Mai, C.-N. Chuah, A. Sridharan, T. Ye, and H. Zang, "Is sampled data sufficient for anomaly detection?," in ACM Internet measurement Conference, (Rio de Janeriro, Brazil), pp. 165–176, 2006.

[11] J. Mai, A. Sridharan, C.-N. Chuah, H. Zang, and T. Ye, "Impact of packet sampling on portscan detection," IEEE Journal on Selected Areas in Communication, vol. 24, December 2006.

[12] D. Brauckhoff, B. Tellenbach, A. Wagner, M. May, and A. Lakhina, "Impact of packet sampling on anomaly detection metrics," in ACM Internet measurement Conference, (Rio de Janeriro, Brazil), pp. 159–164, 2006.

[13] I.K. Fodor, "A Survey of Dimension Reduction Techniques," Technical Report UCRL-ID-148494, Lawrence Livermore Nat'l Laboratory, Center for Applied Scientific Computing, June 2002.

[14] K.Z. Mao, "Identifying Critical Variables of Principal Components for Unsupervised Feature Selection," IEEE Trans. Systems, Man, and Cybernetics, Part B, vol. 35, pp. 339-344, 2005.

[15] L. Breiman, "Statistical Modeling: The Two Cultures," Statistical Science,vol. 16, no. 3, pp. 199-215, Aug. 2001.

[16] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," Neural Computation, vol. 9, no. 7, pp. 1545-1588, Oct. 1997.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," Machine Learning, vol. 46, pp. 389-422, 2002.

[18] A.R. Webb, Statistical Pattern Recognition, second ed. Wiley, 2002.

[19] B. Viswanath, M. Bashir, M. Crovella, S. Guha, K. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In 23rd USENIX Security Symposium (USENIX Security 14), pages 223–238, 2014.

[20] L.X. Bian, F. Crovella, M. Diot, C. Govindan, R. Iannaccone and A. Lakhina. Detection and Identification of Network Anomalies Using Sketch Subspaces. In Proc. of the 6th ACM SIGCOMM conference on Internet measurement, pages 147-152, 2006.

[21] A. Lakhina, M. Crovella, and C. Diot. Characterization of network-wide anomalies in traffic flows. In Proc. of the 4th ACM SIGCOMM Conference on Internet Measurement, pages 201–206, 2004.

[22] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. SIG-COMM Comput. Commun. Rev., 34(4):219–230, 2004.

[23] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. SIGCOMM Comput. Commun. Rev., 35(4):217–228, 2005.

[24] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. SIGMETRICS Perform. Eval. Rev., 32(1):61–72, 2004.

[25] R. A. Martin, M. Schwabacher, N. Oza, and A. Srivastava. Comparison Of Unsupervised Anomaly Detection Methods For Systems Health Management Using Space Shuttle Main Engine Data. Researchgate, 2007.

[26] T.W. Anderson. An introduction to multivariate statistical analysis. Wiley series in probability and mathematical statistics. Wiley, 2nd edition, 1984.

[27] R. J. Muirhead, Aspects of Multivariate Statistical Theory. Wiley, New York, 1982.