Mobile Networks and Applications manuscript No.
(will be inserted by the editor)

Facing Network Management Challenges with Functional Data Analysis: Techniques & Opportunities

David Muelas^{1,*} · Jorge E. López de Vergara^{1} · José R. Berrendero^{2} · Javier Ramos^{1} · Javier Aracil^{1}

Received: 15th November 2015. Revised: 2nd April 2016. Revised: 19th April 2016. Accepted: 20th April 2016.
The final publication is available at Springer via `http://dx.doi.org/10.1007/s11036-016-0733-5`.

Abstract Current fixed and mobile networks' behavior is rapidly changing, which calls for flexible monitoring approaches to avoid loosing track with such a fast evolutionary pace. Due to the many challenges that this scenario is posing to network managers, we propose the exploration of Functional Data Analysis (FDA) techniques as a mean to easily deal with network management and analysis issues. Specifically, we describe and evaluate several FDA methods with applications to network measurement preprocessing and clustering, bandwidth allocation, and anomaly and outlier detection. Our work focuses on how these FDA-based tools serve to improve the outcomes of traffic data mining and analysis, providing easy-to-understand and comprehensive outputs for network managers. We present the results that we have obtained from real case studies in the Spanish Academic network using throughput time series, comparing them with other alternatives of the state of the art. With this comparative, we have qualitatively and quantitatively evaluated the advantages of FDA-methods in the networking area.

Keywords Network Management · Functional Data Analysis · Traffic Modeling · Baselines · Capacity Planning · Anomaly Detection

CR Subject Classification C.2.3 Network Operations · C.4 Performance of Systems · G.3 Probability and Statistics

1 Introduction

Nowadays, network management is suffering an important transformation as a result of the evolution of both the users' requirements and the deployed technologies. The use of new communication services and infrastructures is changing the approaches that Internet Service Providers (ISPs) follow to maintain and monitor their networks. This fact, which is inherent to rapidly changing network dynamics, entails that traditional measurement and analysis methods may easily become not flexible and adaptable enough. Thus, approaches based on particular statistical assumptions, such as concrete marginal distributions or stationary processes, are useless in deployment scenarios where measurements present a different behavior —*e.g.*, data Gaussianity is the base of many anomaly detection systems and capacity and bandwidth allocation methods, but we note that this is not the case in many scenarios as reported in [30,38].

Furthermore, the design of fixed and mobile network solutions that reduce both the CAPEX and OPEX and better suit the clients' requirements —*e.g.*, such as Self-Organizing Networks (SONs) [7], Software-Defined Networks (SDN), or future cellular networks [2,31]— can suffer from the application of management approaches that do not exploit their capabilities. For such architectures, the resources (*e.g.*, bandwidth) can be allocated in a very flexible manner and the consumers' habits change rapidly. Hence, the usage of fine-grained baselines can improve current network management solu-

^{1}HPCN Research Group, Departamento de Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior · ^{2}Departamento de Matemáticas, Facultad de Ciencias.
Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, 28049 Madrid (Spain).
E-mail: {dav.muelas, jorge.lopez_vergara, joser.berrendero, javier.ramos, javier.aracil}@uam.es.
*Corresponding author.

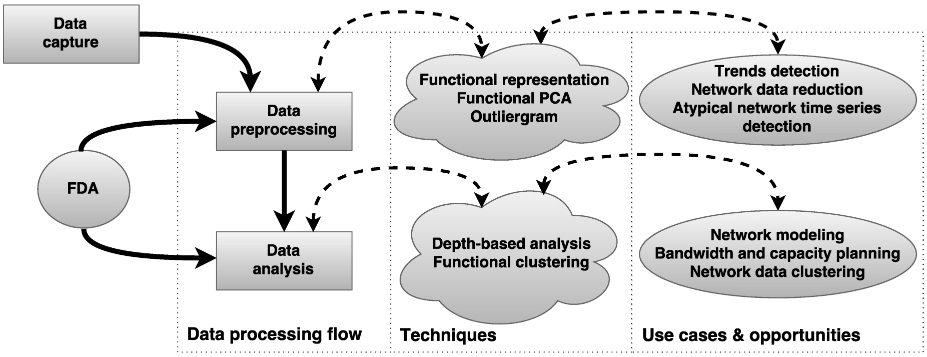

Fig. 1 Conceptual diagram of our proposal.

tions which are mainly based on static and coarsely windowed thresholds [15].

Other aspects, such as network data anonymity and the proliferation of encrypted protocols, limit current network management techniques. For example, monitoring systems that rely on Deep Packet Inspection (DPI) [3] are becoming totally useless as encrypted traffic nowadays represents more than 70% of the total Internet traffic[1]. Moreover, when network data privacy is mandatory, such techniques are not an option.

Based on the previous statements, we focus on computational methods that *(i)* do not relay on statistical assumptions to ensure adaptability to heterogeneous and evolutionary contexts, in particular those related to Gaussianity; *(ii)* help to fine tune management policies to the evolution of networks with time, even in presence of non-stationarity; and *(iii)* enhance the analysis of aggregated measurements that do not require to deal with sensitive data, by improving the detection of patterns in time series. Our final objective is to provide network managers with solutions that alleviate the manual inspection of data and provide visual results, which are easier to interpret.

To this end, we contribute with the application of Functional Data Analysis (FDA) [24,35] to different traditional management tasks. FDA considers random variables which are functions, hence studying the trajectories of stochastic processes as realizations of such random variables. As a consequence, FDA extends classic statistical tools to infinite dimensional spaces. In the network management research field, there is a huge variety of operational and performance measurements that can be considered as functional data [9] as they can be (at least theoretically) taken in a continuous manner —*e.g.* time series [28] or density functions [27].

The strength of such methods are evaluated by considering several use cases that represent current network management challenges. To better assess such use cases, we have used real throughput time series obtained from the Spanish Academic network and the available implementations of FDA methods. Hence, we illustrate their applicability to network data analysis following an out-of-the-box approach —that is, without any kind of tuning. Additionally, the employed dataset and the developed code is available under request, for the sake of reproducibility of our results and also for illustrative purposes.

Figure 1 summarizes the conceptual structure of our work: we link typical network management tasks to FDA methods that fulfill the previously mentioned conditions. In this manner, we show how to cope with network data preprocessing and analysis in the functional scope and highlight the main advantages of this approach. To do so, the rest of this paper is organized as follows. In Section 2 we describe several FDA techniques, and we frame them throughout all the network analysis stages —we describe some formal aspects and point to network management applications that can benefit from them. Next, Section 3 compiles several real case studies that reveal the improvements of the application of functional techniques in network analysis. After presenting the case studies, in Section 4 we discuss the key findings and their applicability to existing network management developments. Finally, Section 5 presents the conclusions and other research lines that can be addressed in the future.

2 A review of some FDA techniques

In this section, we introduce how a functional approach can be used for the analysis of network measurements. To do so, we describe several techniques that will be empirically evaluated later in Section 3. We follow a usual data-flow, considering data preprocessing techniques in the functional environment first, and then, some methods that can help to better understand network dynamics.

[1] `https://www.sandvine.com/trends/encryption.html`

Our review of FDA focuses on techniques that accomplish the objectives highlighted in Section 1. Hence, it is not intended to extensively cover all the current results in the FDA field but to synthesize a set of methods that are later evaluated in the network management scope. For the sake of brevity, our description omits some formal aspects of those methods. For further information about formal aspects beyond the scope of our present work, we refer to [9,24], which are two recent FDA surveys with a broad scope, including theoretical and applied results, and to [34,35], which include further mathematical aspects of FDA and information about implementations in R and MatLab.

2.1 Functional representation

Functional data present high-dimension, since they are related to the trajectories of continuous-time stochastic processes. To cope with such data, two main approaches have been used in the FDA literature. Some works and techniques consider functional sampled data that can be directly obtained from measurements, whereas some others require functional representations using expansions with respect to a functional basis. We note that following the latter approach entails a first data preprocessing step, which will be described here adapted to the particular case of network measurements.

During network monitoring, measurements are obtained as a discrete set of values with a certain granularity. Consequently, the first step is to interpolate observations with a technique that globally minimizes a suitable error function, in terms of projections onto a certain functional basis —which can be either inferred from the observations or fixed to be any well-known family, such as B-Splines or Fourier basis. In general, we represent the family of functions in the selected functional basis as $\{B_k(t)\}_{t\in\mathbb{T},k\in\mathbb{Z}}$, with $\mathbb{T}$ an interval in $\mathbb{R}$. The projections obtained from functional observations with respect to the selected functional basis are denoted as $\{\beta_k\}_{k\in\mathbb{Z}}$. Then, if we consider a certain observation $\{X(t)\}_{t\in\mathbb{T}}$, its functional representation in terms of the selected functional basis is given by the expression in Eq. 1:

$$\{X(t)\} = \sum_{j\in\mathbb{Z}} \beta_j B_j(t),\ t\in\mathbb{T} \tag{1}$$

Nonetheless, it is not possible to computationally consider all the elements in this expression, so it is necessary to truncate the series. A certain error term corresponds to this truncation so that the final functional representation of the observation is given by Eq. 2:

$$\{X(t)\} = [\sum_{j\in\mathbb{J}} \beta_j B_j(t)] + \epsilon(\mathbb{J},\{B_j\}),\ t\in\mathbb{T} \tag{2}$$

where $\mathbb{J}$ is the finite index set and ϵ is the error term, which is dependent on both the selected index set and the specific functional basis.

This representation presents several advantages. On the one hand, it is possible to drastically reduce the needed data to represent a certain process. By adequately adjusting the cardinal of $\mathbb{J}$, we can compress data with some losses related to the term $\epsilon(\mathbb{J},\{B_j\})$. On the other hand, this representation makes it possible to robustly obtain the derivatives of the process trajectories. As observations are represented via a linear combination of functions, we can explicitly obtain their derivatives as shown in Eq. 3:

$$\frac{d}{dt}\{X(t)\} = \sum_{j\in\mathbb{Z}} \beta_j \frac{d}{dt} B_j(t),\ t\in\mathbb{T} \tag{3}$$

This process is of particular interest in certain analysis (*e.g.*, network anomaly detection or clustering, as shown in Section 3.3) that considers not only the magnitude value but also its variation rate. Additionally, the joint analysis of a function and its derivatives is related to the study of the stability of dynamical systems, which is of evident applicability in network modeling and characterization.

Furthermore, this representation allows us to evaluate and select linear combinations of the functional components that provide the most representative model information. Using such an approach, we can further reduce the data volume necessary to persist the observations by keeping a reduced functional basis that optimally represents them in terms of the explained variance. This functional consideration of measurements reduces the necessary volume of data to persist the network behavior as it will be shown in Section 3.2. Functional representation can be used to define highly detailed baselines [15], as we can obtain with it continuous-time robust estimations of the network typical behavior. Additionally, FDA can also be applied to handle other types of data (*e.g.*, Empirical Cumulative Distribution Functions (ECDFs) of network flow characteristics [27]) and not only time series.

FDA techniques are also valuable for the study of multivariate functions —that is, functions taking values in $\mathbb{R}^m$. Interestingly, that means that we can represent the network state by using $f:\mathbb{R}\rightarrow\mathbb{R}^m$, which links sets of variables in the form of multivariate curves. Such multivariate analysis can ease the detection of certain events that require the consideration of several network performance parameters —*e.g.*, Denial of Service attacks as presented in [26].

2.2 Functional Principal Component Analysis

Functional Principal Component Analysis (FPCA) [35] is a transformation of the functional basis that is used to represent the observations. FPCA selects combinations of the original functional basis with variance-based criteria, which allows for an optimal representation of data. It is performed by projecting the original basis on a different space to maximize the explained variance while minimizing the correlation between the components. This provides good visualization despite of the information losses derived from the selection of only a subset of the components.

FPCA is conceptually equivalent to Principal Component Analysis (PCA) in finite dimension spaces. Nonetheless, as we are using a previous representation in terms of a certain functional basis, there is not any semantic obfuscation of the resulting components; on the contrary, an optimal basis to represent the observations is obtained. We recall that in the FDA context, instead of multivariate variable values we have function values $X_i(t)$. That is, the discrete index of each dimension of the multivariate variable is changed by a "continuous index" t. Additionally, the inner products that appear in the PCA definition for finite dimension vectors must be replaced by L^2 inner products, so if we denote the FPCA weights with ξ we get:

$$\int \xi x = \int \xi(t) X(t) dt$$

Hence, the weights ξ are now functions with values $\xi_j(t)$. The scores corresponding to each principal component are given by Eq. 4:

$$f_i = \int \xi x_i = \int \xi(s) x_i(t) dt \tag{4}$$

In the first FPCA step, the weight function $\xi_1(s)$ is chosen to maximize the quantity in Eq. 5:

$$\frac{\sum_i f_{i_1}^2}{N} = \frac{\sum_i \int (\xi_1 x_i)^2}{N}, \tag{5}$$

where N is the sample size and we are assuming data $x_1, \ldots, x_n$ are centered. Additionally, all the weight functions are orthonormal, that is, they must satisfy the restrictions in Eq. 6:

$$\begin{cases} \int \xi_j(t)^2 dt = 1, \forall\ j \\ \int \xi_k \xi_m = 0,\ \forall\ k < m \end{cases} \tag{6}$$

In this manner, each function ξ_j define the most important mode of variation. Note that the weight functions are defined only up to sign change.

This is the adaptation of the usual derivation of PCA to the functional context. Nevertheless, in the functional environment we can see the principal components as the basis functions that approximate the curve as closely as possible.

Some additional restrictions must be imposed when solving the optimization problem. Otherwise, results could be degenerated, as the maximization of the explained variance could not perform well with noisy data. To prevent this situation, FPCA usually *(i)* includes some penalties in the optimization problem, or *(ii)* considers smoothed versions of data.

The principal components can be interpreted as details of the original observations linked to certain variance levels. As a result, they represent different modes of variation of the sample, which is a richer decomposition when compared to other data reduction methods that provide only filtered or reduced outputs. Furthermore, as we will illustrate in Section 3, the study of the observations' coefficients can help to detect clusters in the sample, which proves the advantages of this decomposition.

To complete the FPCA description, we further pinpoint the opportunities that it offers for network analysis. The relation between principal components and certain variance levels is also useful to detect anomalous events and anomalous observations —as usually they are characterized by abrupt changes in certain statistical parameters, such as departures from mean. FPCA paves the way for a novel categorization of anomalies that takes into account the behavior of several principal components. Additionally, the reduction of variance improves capacity planning solutions in scenarios where dynamic resource allocation procedures appear —we will take advantage of this fact in sections 3.4 and 3.5. With this technique, it is possible to control the proportion of the variance that is taken into account, providing a continuous-time methodology to define resource consumption baselines.

2.3 Functional depth and depth-based analysis

Functional depth measures provide ways to determine the relative position of observations into the sample, from the center outwards. They are useful to extend concepts such as centrality measures and order statistics to functional data. Recently, the FDA community has proposed a huge variety of functional depth definitions, each of them taking into account different observations' centrality aspects [22,42]. Additionally, some depth measures have been proposed to cope with multivariate functional data [8,10], which opens the gate to multi-factorial centrality considerations of network measurements —*e.g.* multiple network flow characteristics.

A complete review of the different functional depth alternatives is beyond the scope of this work. Therefore, for the sake of brevity and with illustrative purposes, we consider one of the half-region depth measures in [23], defined with the expression in Eq. 7:

$$MS_{n,H}(x) = \min\{SL_n(x), IL_n(x)\} \tag{7}$$

where

$$\begin{aligned} SL_n(x) &= \tfrac{1}{n\lambda(\mathbb{T})} \sum_{i=1}^{n} \lambda\{t \in \mathbb{T} : x(t) \leq x_i(t)\} \\ IL_n(x) &= \tfrac{1}{n\lambda(\mathbb{T})} \sum_{i=1}^{n} \lambda\{t \in \mathbb{T} : x(t) \geq x_i(t)\} \end{aligned} \tag{8}$$

and λ is the Lebesgue measure on $\mathbb{R}$. This definition is quite popular, as it has a low computational cost and an intuitive interpretation. It makes the observations to be ordered using the minimum of the proportion of time that they are in the hypograph ($SL_n(x)$) or epigraph ($IL_n(x)$) of other observations, which ranks their centrality.

Depth-based analysis is a robust alternative for network data analysis. As it will be shown in Section 3, the isolation of anomalous observations constitutes a suitable methodology for improving results when outliers or high variance are present in the data under analysis. Regarding network measurement time series, current directions in network dynamic resources allocation (*e.g.*, bandwidth) and the flexibility of novel network infrastructures (*e.g.* Software-Defined Networking (SDN), Application-Based Network Operations (ABNO) [1] or 5G cellular networks [2]) can be optimized if we consider a finer grain or even continuous time baselines. Depth measures can help to robustly define such baselines as they define regions that cover a certain proportion of the observations. Furthermore, this approach characterizes the network behavior during a whole period (*e.g.*, a day) instead of using statistical summaries or windowed analysis —as it does not require to test the stationarity of stochastic processes.

Other functions, such as Cumulative Distribution Functions (CDFs) can be robustly estimated and analyzed by using a depth-based methodology [27]. Moreover, the definition of bands based on the extension of the concepts of centiles to the functional environment can enrich certain analysis, as we exemplify in sections 3.4 and 3.5. On the other hand, multivariate depth measures can evaluate centrality of observations in terms of several dimensions (*e.g.* bandwidth and flow concurrence), which is absolutely necessary to detect some events such as SYN flooding attacks [26].

2.4 Shape outlier detection

Outlier detection is a key activity during data mining processes, as inference results can suffer from important deviations if anomalous observations are considered during those processes. In the functional environment, different attributes can lead to mark certain observation as atypical —*e.g.*, amplitude, variance or frequency. As in the case of functional depth, outlier detection has recently attracted much attention in the FDA community, but there is not a well-established methodology to cope with this matter yet. For example, some recent works regarding this field make use of different functional depth notions to sort out observations which differ from the usual pattern of the sample. This is the case of [11], where authors evaluate several functional depths and define an algorithm to exclude atypical observations. Additionally, such methods have also been extended to cope with multivariate functions [16]. While these alternatives seem to be promising for network analysis tasks, in what follows we focus on shape outliers. Such outliers are particularly interesting to detect and extract anomalous network events from measurements which are commonly difficult to detect otherwise —*e.g.*, detection of daily observations with atypical throughput patterns that do not change the maximum nor minimum values.

In [4], authors present the outliergram, a method to detect shape outliers in terms of two centrality measures —that is, indicators of the position of a particular observation in the sample. They consider the modified band depth (MBD_n) [22] and the modified epigraph index —which we have denoted as SL_n in Eq. 8. They prove that there exists a relation between the values of SL_n and MBD_n given by a quadratic equation which can be explicitly calculated. This relation allows projecting the observations in a two dimensional space using the value provided by each centrality measure —that is, each observation is represented by the point defined by (MDB_n, SL_n) in $\mathbb{R}^2$. As a second stage to detect the shape outliers, the algorithm uses the distribution of the distance between (MDB_n, SL_n) and the exact parabola defined by the quadratic relation of both measures. Hence, observations with a typical shape have projections which lay in the proximity of the parabola, while the corresponding to shape outliers are relatively far from it —which allows defining a confidence interval to discriminate the atypical observations.

3 Use cases: functional analysis of network time series

After reviewing FDA concepts, in this section we present different uses cases that show the applicability of FDA techniques on real data obtained from the Spanish Academic network. These use cases are representative in the typical agenda of a network manager. Namely, we consider the reduction and clustering of measurements, the characterization of the usual network behavior, bandwidth and capacity planning in non-stationary scenarios and the detection of atypical days. Throughout this section, we compare the results of some well-known management methods with the corresponding ones obtained by applying a functional approach, showing the advantages of the use of FDA.

To evaluate the latter, we have used a set of network throughput measurements corresponding to 546 consecutive days in a node of the Academic Spanish network. Each day comprises 288 equally spaced observations —that is, one sample every 300 s. To obtain our results, we have used the R implementations included in packages fda [36] and fda.usc [12]. We have used those implementations, as our evaluation is not focused on computational performance nor resource consumption, but on usefulness and validity of a functional network data analysis.

3.1 Network data processing

Once we have obtained network measurements from a certain point of presence, the first data preprocessing step in the functional environment is to obtain a representation in terms of a certain basis. In our case, the selected representation features a number of terms equal to the number of observations of each element (that is, 288 samples corresponding to the 5-minutes intervals in a day) of second grade B-Splines without penalization nor data (pre)smoothing—this corresponds to the tested setup with the best behavior in our data using the the fda package for R. Furthermore, when using this functional representation we have also explicitly obtained the first order derivatives by applying the expression in Eq. 3, to explore the information that can be retrieved from them during throughput time series mining.

Next, we have applied FPCA (both to the original data functional representation and its derivatives) to obtain an optimal representation of observations with a reduced basis. Note that in the previous step, we have considered a huge amount of terms to evaluate the error term that FPCA generates. Nonetheless, the compression factor of the first functional representation may be increased in case a higher error term is acceptable. The explained variance analysis leads to a representation with 30 principal components —as it explains more than a 99% of variance.

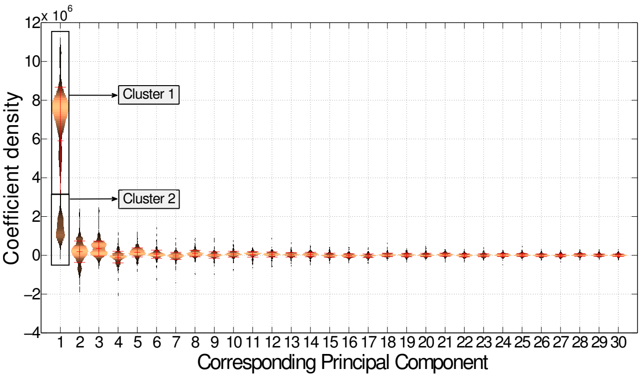

Fig. 2 Coefficient density for each Principal Component

After selecting the basis with the first 30 principal components, we have obtained the coefficients for each observation. The behavior of such coefficients is shown in Figure 2, where we distinguish the estimated coefficient density for each principal component. Interestingly, if we consider the density associated with the first principal component, we can discriminate two well-differentiated clusters (labeled in the figure), which correspond to working and non-working days, respectively.

This method reduces the available information and introduces some error in the punctual values of the reconstructed time series. To assess the FPCA performance, we have analyzed the residuals (that is, the differences between observations and estimations) and obtained the punctual relative error values. Figure 3 presents the survival functions of such a metric for each observed point along a day, which illustrates the statistical behavior of the punctual error for all the daily observations. In this figure, we highlight the median survival function, and the ones covering the 5% and 95% of observations. We note that this functional evaluation of the relative error provides a complete characterization of the FPCA residuals.

We now focus on the characterization of central and extreme observations in terms of depth-based rankings. In what follows, we consider a functional representation with only 15 functional principal components. This restriction introduces a stronger data regularization, and hence minimizes random and atypical perturbations which are not desirable when characterizing centrality in network throughput measurements. Figure 4 summarizes the main results of our depth-based analysis, and highlights several noticeable curves with different depth values. We note that the two previously detected clusters may compromise the half-region depth

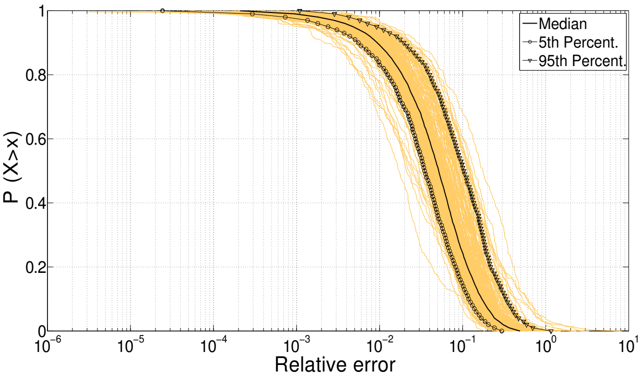

Fig. 3 Survival functions of relative error between observations and recovered curves after applying FPCA, 30 components.

behavior —as it is an overall depth measure. However, the obtained results suit the case studies presented below, so for the sake of simplicity we omit finer processing —*e.g.*, alternative depth measures or factorial analysis.

To visually compare the behaviors of such noticeable curves and of the sample set, we have included the entire original observations in the figure in light orange without markers. To compare depth-based results with other centrality measures, we have also included the sample mean function —in black without markers. Outliers and the previously identified clusters cause a bad representation of the network typical behavior —as we have considered the estimation using all the observations, and the mean is not a robust centrality measure. We have also included the deepest observation of our sample as an alternative centrality measure —it is equivalent to the sample median. To compare the basis restriction effect (which improves the representation of the centrality measures) we show both the original observation and the estimation —red with diamonds, and blue with squares, respectively. Both of them represent the network usual behavior better than the mean function as they suffer from lower distortions by non-usual patterns.

Moreover, we have considered the depth-based ranking of observations to define thresholds for extreme values. We have included in Figure 4 the behavior of the time series with the minimum depth value both in the epigraph (green with asterisks) and in the hypograph (green with crosses) of the deepest function. Additionally, we have constructed curves that punctually minimize the depth value. Specifically, in Figure 4 we represent curves that leave out the 5% of the most extreme values of the observations.

3.2 Network data reduction

There are some previous works that have addressed the reduction of data requirements in the scope of network monitoring. For example, some data preprocessing techniques that can be understood as FDA precursors are those included in [13, 18]. Authors in both works use multi-resolution analysis based on wavelets to compress network measurement. They provide a statistical evaluation of the properties of such compression method, obtaining interesting results. Formally, multiresolution analysis provides a functional representation of data, making use of a specific functional basis. As we explained in Section 2, this is usually the first step when using FDA techniques. As a consequence, we are proposing a general setup that includes the results in those works. In [19] authors apply Principal Component Analysis (PCA) on throughput records to obtain *eigenflows* that represent different variance levels of the observations. The idea is similar to that of FPCA we introduced in Section 2, but it makes no use of a previous data representation in terms of a functional basis. This aspect makes it difficult to interpret the meaning of each *eigenflow*, as this method does not provide a semantic intuition of the information structure which is being used. Remarkably, that proposal points towards the advantages of the consideration of some network measurements as functional data.

Our results prove that FPCA is feasible as a data reduction technique during network measurements time series analysis. By selecting only the first 30 functional principal components, the number of data elements required to reconstruct the original observations is less than a 16% of the original data. This data reduction provides good global estimations of data (the median and 95$^{\text{th}}$ percentile of the mean absolute percentage error (MAPE) is less than 7.5% and 15%, respectively) and punctual error is below 10% in most cases —this is the median of the 95$^{\text{th}}$ percentile punctual relative error, as shown in Figure 3.

When compared to the previously mentioned methods (*i.e.*, PCA and wavelets), these error values are very promising. In the same experimental setup, FPCA outperforms PCA for extreme values (that is, it keeps the 95$^{\text{th}}$ percentile of MAPE lower than PCA) and provides estimations with similar errors in the rest of the cases. Furthermore, it obtains better results than the other methods when the data volume is drastically reduced to 1% of the original data (which is in the order of the recommendation in RFC 1857 [20] for data lasting more than a year) reducing the MAPE values in a range from 7 to 54%.

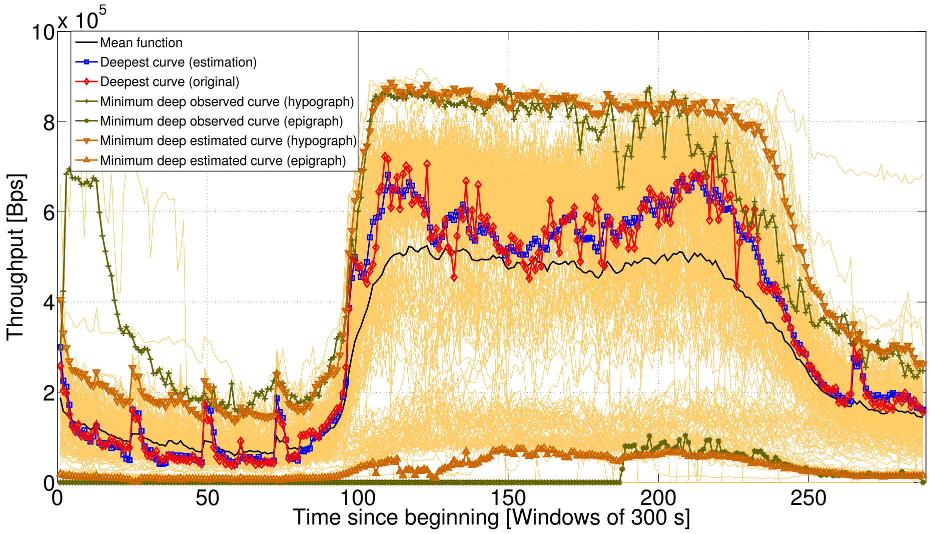

Fig. 4 Summary of our depth-based analysis results.

3.3 Network data clustering

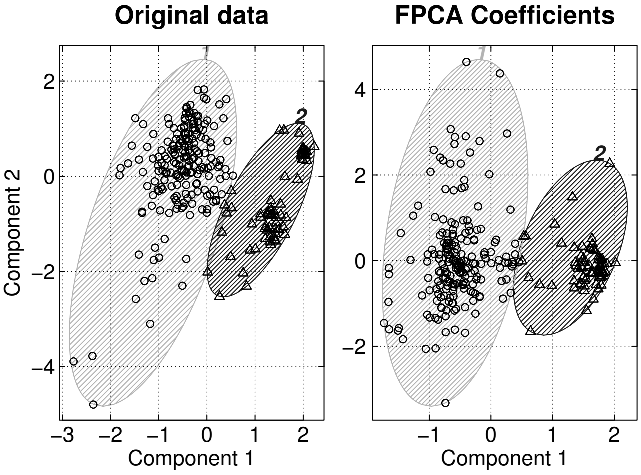

Fig. 5 Comparison of observation clustering using original data, and the first functional principal components of both the throughput time series functional representation and their derivatives. The representation is obtained using the CLUSPLOT tool and includes the clusters' spanning ellipses.

Following with the FPCA representation, we have studied the two clusters that we detected when using the coefficient with respect to the first functional principal component. The analysis of such problem indicates that the difference in the behavior of each cluster makes the problem easily separable, and that the average value of each curve is determinant when assigning it to one of the clusters. Remarkably, using only that single projection we have been able to obtain the same assignment that the one provided by K-means algorithm when fed with all the values of the daily throughput curves —which illustrates the potential of this functional approach in feature selection. For the sake of brevity, we omit further performance comparisons between other clustering algorithms in this work: for those interested in this matter, we point to [17], where authors have surveyed several functional clustering algorithms on well-known problems.

We have also included the information we have retrieved from the curves' derivatives. To do so, we have also considered their coefficient with respect to their first functional principal component. While the addition of this information does not change the assignment of each curve to a cluster, it improves the inter-group separation: Figure 5 includes the representation provided by CLUSPLOT [33] for the clusters defined from the original data and from the coefficients with respect to the first functional principal components of both the throughput time series functional representation and their derivatives. This representation shows the better differentiation of classes when using a suitable FPCA-based reduced set of features from the observations and their derivatives.

These results provide a new approach for Network Behavior Analysis (NBA). For example, the proposals in [37,40], can be considered from the point of view of FDA as the analysis of a set of functions that describes the network state. Those proposals are based on pattern detection to discriminate anomalous behaviors that could indicate intrusions or other malicious actions. Hence, the application of functional feature selection and clustering can improve, as shown in our example, the discrimination among different behavioral groups —therefore, providing a more complete and formally consistent framework to face this type of studies.

3.4 Network modeling and characterization

So far, depth-based analysis provides a set of central and extreme curves that are suitable to characterize the network behavior. Such curves provide a high-dimensional definition of the usual network patterns, beside of the consideration of marginal traffic distributions —hence cutting out the hypothesis about such distributions required in other state-of-the-art approaches. Furthermore, most of the existent methods also assume that the underlying stochastic processes are stationary during certain periods of observation (*e.g.*, during 15 minutes [25]), while the results derived from functional methods allow to study measurements during more complex and meaningful periods —*e.g.*, a whole day as in our case. In what follows, we qualitative compare the characteristics of the results in some previous works devoted to univariate or multivariate network modeling and characterization, with those obtained with a depth-based functional approach.

In [38], α-stable distributions are proposed to study network throughput in low aggregation points. Additionally, authors study the perturbations in the distribution parameters to link them to certain anomalous events. On the other hand, other previous works such as [14, 25] consider Gaussian processes to model network behavior. Specifically, [14] is devoted to capacity planning based on the characterization of the busy hour, and in [25], authors describe a methodology to detect sustained changes in network load. Both works require a Gaussian fit of traffic load, which is a hypothesis that sometimes is not met —*e.g.*, [30, 38] include some situations where Gaussian models do not fit in the observations.

Nonetheless, the previous approaches do not match the three key points that we have depicted for network monitoring and analysis methods. First, they require the marginal traffic distributions to follow some specific distributions (namely, α-stable and Gaussian), which is a strong hypothesis that prevents from extending this method to environments where this hypothesis is not met. Second, authors indicate that the computation of some of the parameters of such models is computationally expensive, which can limit the definition of flexible management policies —as the application of such methods to the study of time series requires considering stationary intervals, which can limit flexible deployments of such approaches if we take into account the claims in [41]. Finally, these methods provide either difficult to interpret or extremely simple outputs for network managers —as the interpretation of their results are related to statistical tests or to the meaning of non-intuitive statistical summaries. As shown, the results of depth-based analysis alleviate these flaws by fulfilling those three principles.

Table 1 Results of the bandwidth allocation experiments.

Training set (%)	**Underestimations (%)**
1	17.74 ± 0.85
5	4.08 ± 0.21
10	2.23 ± 0.11
15	1.87 ± 0.08
20	1.58 ± 0.06
25	1.57 ± 0.06
30	1.56 ± 0.05
35	1.46 ± 0.04
40	1.46 ± 0.04

3.5 Network bandwidth and capacity planning

Bandwidth and capacity planning is a capital matter in virtualized environments such as Virtual Networks and Virtual CPDs [6], and it is also considered as a distinguishing feature of the future 5G networks [2].

To evaluate the advantages of functional approaches during bandwidth and capacity planning, we follow a methodology similar to the one exposed in [29]. In that work, the authors discussed several methods to dynamically allocate bandwidth for tenants in a common physical network architecture. Some differences arise between that work and the analysis we have leaded: in our case, we have used time series of throughput with a 5-minute aggregation interval, whereas they used finer-grained measurements. Interestingly, they only considered traces lasting for 15 minutes, as their method required the throughput time series to be stationary. In our case, we have defined a bandwidth allocation limit based on the previously presented depth-bands for a period lasting a whole day.

To conduct our evaluation, we have split our measurements set in two groups —one of them to train the depth-based threshold and the other one to evaluate the bandwidth requirement prediction. We have accounted the number of points above the defined threshold, thus providing an estimation of the underestimations impact —in this case, we consider a depth band leaving outside the 2% of the most extreme observed values. Table 1 shows the mean results with a corresponding 95% confidence interval for 500 repetitions of such experiments considering different percentages of observations for the training phase.

Using our approach, the percentages of underestimations are comparable to those reported in [29]. We recall that the focus of that work is different to ours —they obtain bandwidth requirement estimations for short time intervals. Nonetheless, with our approach, we can decide tenants that can coexist in the same physical architecture in terms of their usual activity among a whole period. Additionally, we relax the hypothesis of the methods which are considered in [29], as we do not require the throughput values to be Gaussian nor stationary.

3.6 Outlier detection in network time series

Let us now show the results of outliergram tool application to our throughput observations. This tool produces representations like that in Figure 6, which illustrates the relation between the two depth measures that it considers for each observation. With such tool, we can easily detect shape outliers, as anomalous observations lay out the confidence interval inferred from the sample. The outliers are represented in Figure 7, and we can visually assess that they do present anomalous behaviors.

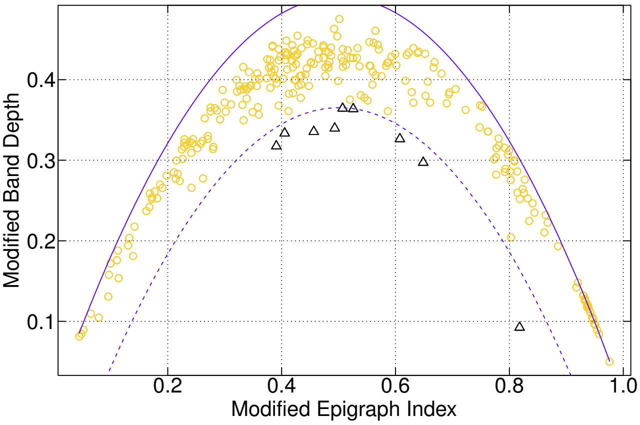

Fig. 6 Outliergram visualization of the projected observations. Black triangles correspond to shape outliers, while orange circles represent typical observations.

In this representation, we have highlighted throughput time series that are marked as shape outliers, but outliergram is also able to detect certain observations with atypical extreme values. There are several types of outliers that can be detected when using this approach:

- Observations which lay in the borders of the clusters we have previously detected, although they may not have extreme values in absolute terms.
- Observations which fluctuate from high values in some parts of the temporal domain to low values in other ones.
- Observations which abruptly fall during a certain period of time.

It is worth remarking that all of these types cause departures of centrality measures during inference processes if other techniques not as robust as those we have selected are applied. Hence, this FDA-based technique can improve results in later network data analysis; particularly with the two first types we have differentiated —given that to detect them it is necessary to consider the behavior of the whole observation and not only punctual values.

4 Discussion and application

According to the previous comparison of FDA and other well-known methods, the most remarkable findings and advantages follow:

- FDA techniques relax the hypothesis of network analysis state-of-the-art methods, thus providing more adaptable tools to cope with heterogeneous and changing environments.
- They allow considering network time series as a whole, which provides means to statistically study measurements taking into account their overall behavior.
- Additionally, they provide comprehensive and easy-to-understand data representations for network managers. That is, functional methods lead to straightforward visual outputs that highlight problems and trends without requiring further analysis.

Nonetheless, these advantages may be worthless if functional methods cannot be included in existent monitoring and management solutions. Fortunately, current tools follow some common design principles that simplify the introduction of these methods and provide several data sources that can be studied as functional data. In what follows, we briefly comment some recent approaches that highlight those principles —for further information and details about current trends, we refer to [5, 21].

Scap [32] is a stream-oriented system able to cope with high throughput rates. Taking into account their authors' claims, that system could be extended to use functional methods to improve its functionality and analytic capabilities. —*e.g.*, traffic capture online selection in terms of functional baselines. Scap is an example of the growing importance of aggregated data summaries (*e.g.*, values provided by SNMP, NetFlow records, etc.) to cope with the analysis of multi-Gb/s networks, as they reduce network analysis systems' computational demands.

BlockMon [39] is another interesting example of novel monitoring tools. It is conceived as a modular and distributed system, providing users with a flexible and customizable framework to develop monitoring architectures that suit each particular scenario. Given its

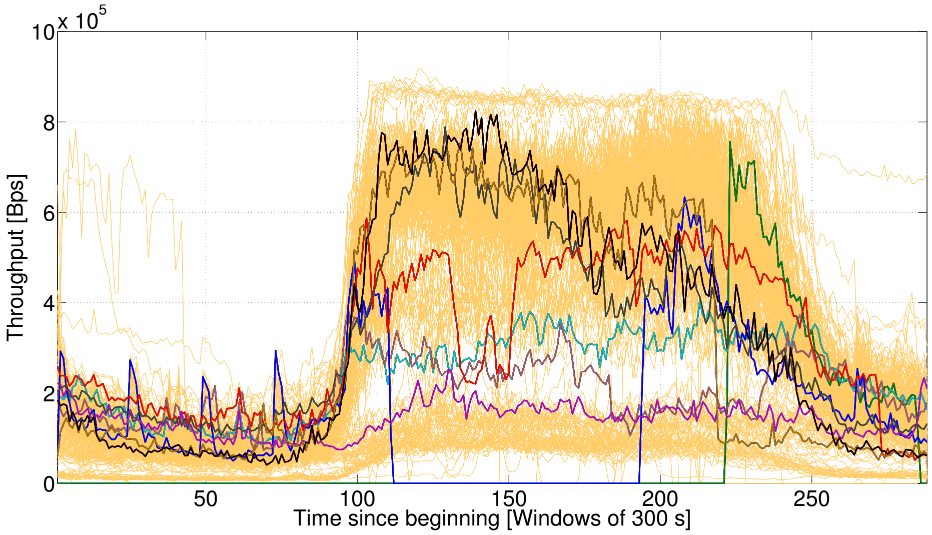

Fig. 7 Representation of the daily observations that outliergram identifies as shape outliers.

modular structure, BlockMon could be extended with FDA-based modules to provide advanced capabilities. For example, as we illustrate in Section 3, BlockMon could be complemented with functional data preprocessing techniques to produce enriched analysis and visualization outputs.

To end with, we mention M^{3}Omon, which is presented in [26]. M^{3}Omon is a monitoring framework that provides users with multi-granular data —specifically, aggregated time series, flow records, and raw network packets. Authors show the importance of simultaneous analysis of several data sources with different aggregation levels to effectively detect and completely understand network phenomena in high performance networks. With such data sources, we can make the best of functional-based methods to create a complete ecosystem of analytical applications. For instance, a capacity planning module can be easily implemented using the aggregated time series outputs. At the same time, FPCA-based data reduction can help to optimize storage requirements when using this framework.

5 Conclusions

This work constitutes a novel study of the FDA application in the network data analysis scope. Specifically, we have reviewed several FDA **techniques** that can be used to extract knowledge from network measurements. We have illustrated how FDA can be applied to different common network management tasks, comparing it with other state-of-the-art methods. In this light, we have considered several use cases with real network measurements (particularly, throughput times series), showing the **opportunities** that FDA-based techniques bring in network data analysis. The main advantages of FDA pave the way for the evolution of current techniques.

Regarding network data reduction, the functional representation and feature selection that we have applied provides good compression ratios with controlled information losses. Specifically, our evaluation has shown that FPCA estimations fairly represent the original observations using less than a 16% of the total amount of data. Using such a reduction, MAPE presented median and 95$^{\text{th}}$ percentile values below 7.5% and 16% respectively. Additionally, the median of the 95$^{\text{th}}$ percentile punctual relative error is below 10%. Concerning the clustering problem, we have compared the results of K-means algorithm with either the original observations or the FPCA projections of the data and its derivatives. The latter improves the group differentiation while reducing as well the input for the clustering method.

The evaluation of depth-based analysis has shown that it provides robust estimations of central and extreme network measurements behavior and it relaxes the hypothesis on marginal distributions of network time series. Furthermore, such estimations serve to define a continuous-time functional threshold for capacity planning. The obtained results are similar to those of other state-of-the-art methods, but without requiring the network time series to be stationary. Hence, depth-based analysis has proven useful for these tasks, especially when considering emerging network technologies that allow flexible resource allocations —such as SDNs, ABNO, SON, and 5G.

Finally, we have shown that some atypical time series might not present changes in their extreme values while still exhibit odd behavioral patterns. Therefore, shape outlier detection helps excluding such observations during inference in network analysis, which automates costly processes of data cleaning.

To sum up, FDA is a branch of statistics which can ease management tasks in emerging network infrastructures that are otherwise constrained by the application of classic statistics. Thus, we have presented to the Networking and Telematics community a methodology, assessing its usefulness and the opportunities it offers for network analysis. This work has focused on the foundations of the applicability of FDA to time series but it has not addressed other promising FDA techniques (*e.g.*, FDA-based forecasting and classification, functional homogeneity) that may also be applicable to a wide variety of network data and may unleash the true potential of FDA.

Acknowledgements This work has been partially supported by the Spanish Ministries of Economy and Competitiveness (PackTrack, TEC2012-33754; Tráfica, TEC2015-69417-C2-1-R), and of Science and Innovation (MTM2013-44045-P).

References

1. Aguado, A., López, V., Marhuenda, J., Fernández-Palacios, J.P., et al.: ABNO: a feasible SDN approach for multi-vendor IP and optical networks. In: Optical Fiber Communication Conference, pp. Th3I–5. Optical Society of America (2014)
2. Andrews, J., Buzzi, S., Choi, W., Hanly, S., Lozano, A., Soong, A., Zhang, J.: What will 5G be? Selected Areas in Communications, IEEE Journal on **32**(6), 1065–1082 (2014)
3. Antonello, R., Fernandes, S., Kamienski, C., Sadok, D., Kelner, J., Gdor, I., Szab, G., Westholm, T.: Deep packet inspection tools and techniques in commodity platforms: Challenges and trends. Journal of Network and Computer Applications **35**(6), 1863 – 1878 (2012)
4. Arribas-Gil, A., Romo, J.: Shape outlier detection and visualization for functional data: the outliergram. Biostatistics **15**(4), 603–619 (2014)
5. Bajpai, V., Schönwälder, J.: A survey on internet performance measurement platforms and related standardization efforts. Communications Surveys & Tutorials, IEEE **17**(3), 1313–1341 (2015)
6. Bari, M.F., Boutaba, R., Esteves, R., Granville, L.Z., Podlesny, M., Rabbani, M.G., Zhang, Q., Zhani, M.F.: Data center network virtualization: A survey. IEEE Communications Surveys & Tutorials **15**(2), 909–928 (2013)
7. Chen, N., Rong, B., Mouaki, A., Li, W.: Self-organizing scheme based on NFV and SDN architecture for future heterogeneous networks. Mobile Networks and Applications **20**(4), 466–472 (2015)
8. Claeskens, G., Hubert, M., Slaets, L., Vakili, K.: Multivariate functional halfspace depth. Journal of the American Statistical Association **109**(505), 411–423 (2014)
9. Cuevas, A.: A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference **147**(0), 1 – 23 (2014)
10. Cuevas, A., Febrero, M., Fraiman, R.: Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics **22**(3), 481–496 (2007)
11. Febrero, M., Galeano, P., Gonzlez-Manteiga, W.: Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetrics **19**(4), 331–345 (2008)
12. Febrero-Bande, M., Oviedo de la Fuente, M.: Statistical computing in functional data analysis: the R package fda.usc. Journal of Statistical Software **51**(4), 1–28 (2012)
13. García-Dorado, J.L., Aracil, J., Hernández, J.A., López de Vergara, J.E.: A queueing equivalent thresholding method for thinning traffic captures. In: Network Operations and Management Symposium, 2008. NOMS 2008. IEEE, pp. 176–183 (2008)
14. García-Dorado, J.L., Hernández, J.A., Aracil, J., López de Vergara, J.E., López-Buedo, S.: Characterization of the busy-hour traffic of IP networks based on their intrinsic features. Computer Networks **55**(9), 2111 – 2125 (2011)
15. Gibeli, L.H., Breda, G.D., Miani, R.S., Zarpelão, B.B., de Souza Mendes, L.: Construction of baselines for VoIP traffic management on open MANs. International Journal of Network Management **23**(2), 137–153 (2013)
16. Hubert, M., Rousseeuw, P.J., Segaert, P.: Multivariate functional outlier detection. Statistical Methods & Applications **24**(2), 177–202 (2015)
17. Jacques, J., Preda, C.: Functional data clustering: a survey. Advances in Data Analysis and Classification **8**(3), 231–255 (2013)
18. Kyriakopoulos, K., Parish, D.: A live system for wavelet compression of high speed computer network measurements. In: S. Uhlig, K. Papagiannaki, O. Bonaventure (eds.) Passive and Active Network Measurement, *Lecture Notes in Computer Science*, vol. 4427, pp. 241–244. Springer Berlin Heidelberg (2007)
19. Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, E.D., Taft, N.: Structural analysis of network traffic flows. SIGMETRICS Perform. Eval. Rev. **32**(1), 61–72 (2004)
20. Lambert, M.: RFC 1857: A Model for Common Operational Statistics (1995)
21. Li, B., Springer, J., Bebis, G., Gunes, M.H.: A survey of network flow applications. Journal of Network and Computer Applications **36**(2), 567–581 (2013)
22. López-Pintado, S., Romo, J.: On the concept of depth for functional data. Journal of the American Statistical Association **104**(486), 718–734 (2009)
23. López-Pintado, S., Romo, J.: A half-region depth for functional data. Comput. Stat. Data Anal. **55**(4), 1679–1695 (2011)
24. Manteiga, W.G., Vieu, P.: Statistics for functional data. Computational Statistics & Data Analysis **51**(10), 4788 – 4792 (2007)
25. Mata, F., García-Dorado, J.L., Aracil, J.: Detection of traffic changes in large-scale backbone networks: The case of the Spanish academic network. Computer Networks **56**(2), 686 – 702 (2012)
26. Moreno, V., Santiago del Río, P.M., Ramos, J., Muelas, D., García-Dorado, J.L., Gómez-Arribas, F.J., Aracil, J.: Multi-granular, multi-purpose and multi-Gb/s monitoring on off-the-shelf systems. International Journal of Network Management **24**(4), 221–234 (2014)
27. Muelas, D., Gordo, M., García Dorado, J.L., López de Vergara, J.E.: Dictyogram: A statistical approach for the definition and visualization of network flow categories. In: 11th International Conference on Network and Service Management (CNSM 2015) (2015)

28. Muelas, D., López de Vergara, J.E., Berrendero, J.R.: Functional data analysis: A step forward in network management. In: Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on, pp. 882–885 (2015)
29. de O. Schmidt, R., van den Berg, H., Pras, A.: Measurement-based network link dimensioning. In: Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on, pp. 1071–1077 (2015)
30. de O. Schmidt, R., Sadre, R., Melnikov, N., Schönwälder, J., Pras, A.: Linking network usage patterns to traffic gaussianity fit. In: Networking Conference, 2014 IFIP, pp. 1–9 (2014)
31. Oh, E., Son, K., Krishnamachari, B.: Dynamic base station switching-on/off strategies for green cellular networks. Wireless Communications, IEEE Transactions on **12**(5), 2126–2136 (2013)
32. Papadogiannakis, A., Polychronakis, M., Markatos, E.P.: Scap: Stream-oriented network traffic capture and analysis for high-speed networks. In: Proceedings of the 2013 Conference on Internet Measurement Conference, IMC '13, pp. 441–454. ACM, New York, NY, USA (2013)
33. Pison, G., Struyf, A., Rousseeuw, P.J.: Displaying a clustering with CLUSPLOT. Computational Statistics & Data Analysis **30**(4), 381 – 392 (1999)
34. Ramsay, J., Hooker, G., Graves, S.: Functional Data Analysis with R and MATLAB. Springer New York (2009)
35. Ramsay, J., Silverman, B.: Functional Data Analysis. 1997. Springer, New York (1997)
36. Ramsay, J., Wickham, H., Graves, S., Hooker, G.: fda: Functional Data Analysis (2014). URL `http://CRAN.R-project.org/package=fda`. R package version 2.4.4
37. Saad, S., Traore, I., Ghorbani, A., Sayed, B., Zhao, D., Lu, W., Felix, J., Hakimian, P.: Detecting P2P botnets through network behavior analysis and machine learning. In: Privacy, Security and Trust (PST), 2011 Ninth Annual International Conference on, pp. 174–180 (2011)
38. Simmross-Wattenberg, F., Asensio-Pérez, J., Casaseca-de-la Higuera, P., Martín-Fernández, M., Dimitriadis, I., Alberola-López, C.: Anomaly detection in network traffic based on statistical inference and alpha-stable modeling. Dependable and Secure Computing, IEEE Transactions on **8**(4), 494–509 (2011)
39. Simoncelli, D., Dusi, M., Gringoli, F., Niccolini, S.: Stream-monitoring with BlockMon: convergence of network measurements and data analytics platforms. SIGCOMM Comput. Commun. Rev. **43**, 29–36 (2013)
40. Wei, T.E., Mao, C.H., Jeng, A., Lee, H.M., Wang, H.T., Wu, D.J.: Android malware detection via a latent network behavior analysis. In: Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on, pp. 1251–1258 (2012)
41. Xu, K., Wang, F., Wang, H.: Lightweight and Informative Traffic Metrics for Data Center Monitoring. Journal of Network and Systems Management **20**(2), 226–243 (2012)
42. Zuo, Y., Serfling, R.: General notions of statistical depth function. Annals of statistics **28**(2), 461–482 (2000)