Functional Data Analysis: A step forward in Network Management

David Muelas*, Jorge E. López de Vergara*, José R. Berrendero†
*Departamento de Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior
†Departamento de Matemáticas, Facultad de Ciencias
Universidad Autónoma de Madrid
Email: {dav.muelas, jorge.lopez_vergara, joser.berrendero}@uam.es

***Abstract*—Network Management tasks are currently characterized by their diversity both in terms of the situations that must be faced and the data used to reach conclusions. This complex and changing context imposes diverse needs and restrictions that must be covered by management tools in order for them to be useful. In order to face current challenges, we propose the application of Functional Data Analysis (FDA) techniques in the different functional areas of Network Management. FDA can be applied to network data compression, definition of baselines, anomaly detection, or traffic classification as well as forecasting for network dimensioning.**

I. Introduction

Network Management tasks are currently characterized by their diversity both in terms of the situations that must be faced and the data used to reach conclusions. Furthermore, the huge amount of data generated by modern computer networks that is processed and persisted during Network Management activities must be optimized in order to improve the scalability of solutions. These processes can enrich the conclusions of several analytical tasks (*e.g.* anomaly detection) in the era of Big Data if we apply suitable analysis processes. Classical methods may be insufficient if their hypotheses are not satisfied or if the ideal deployment scenarios are not in accordance with those being analyzed. For example, the encryption of transmitted information and other legal and privacy-related aspects concerning this data limit some state-of-the-art solutions – *e.g.* intrusion detection systems that rely on Deep Packet Inspection (DPI) techniques.

In order to face these challenges, we propose the application of Functional Data Analysis (FDA) techniques. FDA considers random variables which are functions, thus embedding them in infinite dimensional spaces. The positive results obtained through the application of FDA to problems that share some characteristics with those of Network Management (*e.g.* weather forecasting and certain economic studies) would justify further exploration into the applicability of FDA in this area.

The rest of this paper is structured as follows. Firstly, we review existing solutions, which present some limitations due to the evolution of the structure and dynamics of computer networks. Secondly, we provide a brief formal discussion of some elements of FDA and several particular applications in the field of network data analysis that have overcome some of the limitations of classical solutions. Finally, we offer some conclusions about this exploratory study and comment on the scope for future work in this field.

II. Related work

In this section, we present related work that covers different Network Management tasks. Specifically, we describe solutions, tools and methods used to face different processes that make up some of the activities of the different functional areas of Network Management. In all cases, we comment on their limits and highlight different aspects that potentially point to FDA as a "toolbox" that could improve current solutions.

Regarding network performance measurement, in [1] the authors propose a new metric with a reduced computational cost that condenses significant information when applied to Data Center monitoring. This approach highlights some of the principles included in our solution, but is restricted to a particular context. One of the advantages of FDA is that few *a priori* assumptions are made, and thus it can be extended to almost every scenario.

Continuing with network modeling, the authors in [2] propose the use of α-stable distributions to model traffic in low-aggregation points. The deviation of some of the parameters of the throughput distribution is used to detect certain anomalous behavior. The main problem with this proposal is the high computational requirements for the estimation of the parameters that define a particular element of this family of distributions. Thus, the deployment of this approach may be unfeasible in many contexts. In [3], [4] the authors present statistical network models using Gaussian processes. In particular, the solution proposed in [3] is oriented to link capacity planning inside a network by inference on the busy hour. The methodology described in [4] is oriented to the detection of sustained changes in load utilization. The Gaussianity of traffic load is the base of these and other models, but it is a hypothesis that cannot be directly assumed in general [5], [2]. FDA techniques do not suffer from this problem, as they do not require assumptions related to the marginal distribution of the parameters considered.

The authors in [6], [7] propose the application of some filtering and preprocessing techniques to network data. Those solutions have some similarities with functional representation and functional PCA, which will be introduced in the following sections. A key difference between these approaches and FDA techniques is that the former do not make use of a primary common representation of flows in terms of any type of basis or this is quite restrictive – *e.g.* Wavelets. Nevertheless, the central ideas of these studies point to the gains that a functional treatment of network parameters entails.

Although the idea of using functional random variables that are defined in infinite dimensional spaces seems to be

self-defeating, it is the basis of several machine learning and data mining techniques that take advantage of some properties of sets when their dimension increases. For instance, the Support Vector Machine (SVM) is a well-known example that has been successfully applied to several problems related to Network Management activities. For instance, in [8], the authors explored the results that SVMs provide in anomaly detection, management of Quality of Experience (QoE) and QoE prediction.

Some other methods directly oriented to the discrimination of anomalous behavior are described in [9]. These approaches use Network Behavior Analysis (NBA) techniques to detect and classify patterns that might indicate the presence of any type of anomaly. NBA can be seen from the point of view of FDA as a set of functions that describe the network state, providing formal soundness for the analysis and a basis for the use of all the advanced features that FDA includes.

III. FDA and Network Management

In this section, we formally define several elements and applications through an extensive description of use cases that highlight the main advantages of the functional approach in certain Network Management activities. For the sake of brevity and clarity, we will omit some of the formal aspects in the following discussion. Further information about FDA may be found in [10], which is a general study of current techniques and possibilities in this field.

A. Data representation

In empirical experiments, it is not possible to obtain measurements in a continuous manner, and therefore the first step when using functional approaches is to interpolate and – if necessary – smooth this data. In the literature, B-splines are a common choice due to their properties [11], although other representations are totally admissible if the structure of the data is well preserved. In general, we will represent the set of functions that make up the selected basis as $\{B_k(t)\}_{t\in\mathbb{I},k\in\mathbb{Z}}$ with $\mathbb{I}$ a real interval, and the coefficients that give the projections of the observations to respect the basis as $\{\beta_k\}_{k\in\mathbb{Z}}$. In practice, the representation is truncated so the representation is given by the expression $\{X_t\} = [\sum_{j\in\mathbb{J}} \beta_j B_j(t)] + \epsilon(\mathbb{J}, \{B_j\}),\ t \in \mathbb{I}$ with $\mathbb{J}$ a finite set of indexes and ϵ a term of error dependent on both the set of indexes and the selected basis. This representation has several advantages. Firstly, the amount of data required to describe the evolution of the process is drastically reduced as the number of temporal points is much bigger than the number of components selected. Secondly, it involves a robust estimation of the derivatives of the model. Finally, it provides means to select the components that contain the most representative information of the model as we show below.

FDA allows for the development of compact expressions of network parameters (packets, flows, bytes, active IP addresses,...) represented as a function of a set of parameters – *e.g.* time series, if they are represented as functions of time. This is of particular interest in the case of defining baselines [12], as it provides a continuous time approximation. Additionally, we can use surfaces or curves describing the joint behavior of an arbitrary number of parameters. This feature is essential for network managers, as it is necessary to detect some types of anomalies (*e.g.* some DDoS attacks [13]).

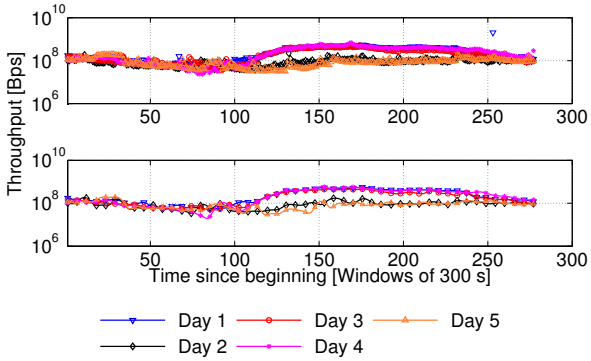

Figure 1. Third grade B-splines representation for 5 days of throughput registers.

Figure 1 shows the result of the interpolation of throughput data from an educational network using third grade B-splines. This interpolation reduced the amount of data required to store the network state by a factor of three. It is worth noting that this factor can be increased if a greater error term is admissible. In order to obtain this representation, we used a set of sampled points and then we evaluated the resulting curve for each point.

Given the characteristics of functional representation, the mapping of network parameters to functional elements entails a first-level compression. This aspect of the application of FDA to Network Management tasks is of interest when we consider scalability issues, as well as providing a first step to the application of other FDA techniques.

B. Functional PCA

Functional PCA allows for the selection of the projection directions that maximize variance. As functional PCA uses a previous representation in terms of a certain functional basis, it does not lead to a semantic obfuscation which is one of the main problems when applying PCA.

In order to test these ideas in the field of Network Management, we applied Functional PCA to the registers that were considered above, extended with further observations. We used 20 daily observations interpolated as previously described. Figure 2 represents the first 6 harmonics – *i.e.* the number of components needed to cover 95% of the original variance. Regarding compression, the result allows the number of needed to represent the data to be reduced by a factor of ten. It should be noted that the first principal component, highlighted in the figure, represents a scaled *approximation* of the dynamic of the observations – omitted for the sake of brevity, given that they are similar to those represented in Figure 1. The consideration of additional principal components enriches the representation with *details* that cover a higher proportion of the observed variability.

Functional PCA has several advantages for problems derived from certain Network Management activities. First of all, it entails a second-level compression of data, as we can select a subset of the principal components controlling variability information losses. Furthermore, compared to other alternatives such as those commented in Section II, PCA harmonics represent a meaningful decomposition of the observations: changes in a certain harmonic indicate different types of changes in the measured parameter, depending on

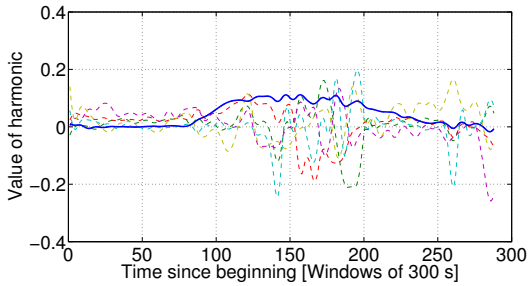

Figure 2. Harmonics covering 95% of the original variance – 6 components – after Functional PCA on throughput registers. The first component, covering 79.27% of the original variance is highlighted

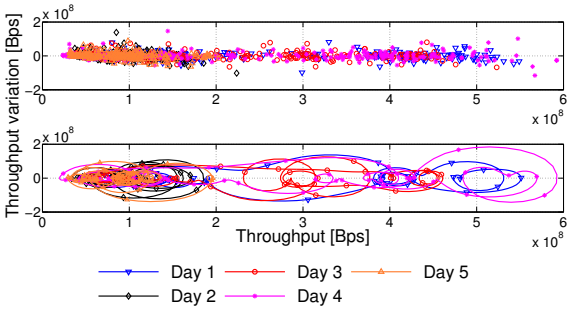

Figure 3. Comparative view of phase-plane plot obtained with numerical and analytical derivation based on B-spline representation.

the variability covered by such a harmonic. This fact means a new vision of anomalies and other variations of network dynamics, linking them to the behavior of the functional principal components. Finally, the semantic information of this decomposition permits the definition of baselines in terms of the evolution of the harmonics.

C. Phase-plane analysis

Phase-plane analysis describes the temporal evolution of a system making use of the relation between the value of a function and the associated value of its derivative. This representation of a system allows for several analytical processes, such as the study of the stability of a certain system. This joint evolution can provide enhanced results when considering dynamical aspects of functions in processes oriented to homogeneity studies or clustering of curves.

If we consider a certain functional observation X_t, we can obtain a phase-plane plot using *(a)* numerical estimations (*i.e.* with finite difference methods) or (b) analytical derivation using the functional representation previously described. Figure 3 shows two phase-plane plot of the previously mentioned set of throughput registers. The first one is numerically obtained, using a first-order finite difference method; while the second is derived from an analytical differentiation applied to the functional representation previously obtained.

Since not only the value of parameters, but also speed changes are important in several Network Management tasks, phase-plane analysis is an approach that can be useful in various decision making processes. This representation is useful for the visual detection of abnormal events and provides extended information about the evolution of the network state. Moreover, the inclusion of several parameters in this analysis using multivariate functions permits the study of joint relations between different magnitudes and their derivatives and it can

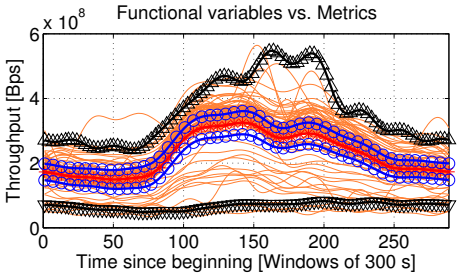

Figure 4. Example of depth region using an extended set of throughput registers.Mean curve: red asterisks. Mean confidence interval: blue circles. Depth region: black triangles.

be used to characterize different events in network dynamics by means of clustering and classification methods for curves and surfaces.

D. Functional depth

Depth measures in FDA are useful as they provide a notion of the *relative position* of elements in the set of observations. As several attempts to define and adapt L-statistics to functional data have appeared, depth measures have become a key element in constructing some statistics that require a certain order of the sample space.

Although many depth notions are described in the literature of FDA, we will now consider the definition included in [14]. This definition is widely used, due to the low computational cost of this expression and its intuitive meaning. Roughly speaking, this functional depth is based on the fraction of 'time' that the a certain observation is dominated by and dominating other elements of the set of observations. Additionally, there are some proposals of multivariate functional depth [15], which allow us to apply depth-based concepts not only for curves, but also for multivariate functions and surfaces.

In Figure 4, we show a depth region based on the definition given above. This region covers 80% of observations of the set that we have considered throughout this study. The lines marked with black triangles correspond to the curves that delimit that region. The mean curve is represented in red with asterisks, while the blue lines with circles indicate the confidence interval of the mean at each point with $\alpha = 5\%$. It should be noted that the depth region includes this interval, providing borders derived from the data structure. Thus, these order statistics provide robust approaches for the analysis of the typical behavior of a network, as a result of the isolation of outliers and abnormal values.

The appearance of network infrastructures that permit both dynamic configuration of rules and resource deployment (*e.g.* Software Defined Networking or the Application-based Network Operations (ABNO) architecture [16]) points to the establishment of baselines that take into account network behavior in each time frame. Depth-based metrics are good candidates for the definition of such baselines, and the multivariate definitions support the joint consideration of a set of parameters, which is interesting as some events required the monitoring of several characteristics in order to be detected [13].

E. Functional homogeneity

Given two samples, it is natural to ask when the observations from the two samples are realizations of the same stochastic process. In classical statistics, there are many methods that can be used to test the homogeneity of two samples (*e.g.* χ^2 test). In the field of FDA, there are some recent proposals to face this problem which have provided some promising results [17]. Those are based on computational approaches to obtain estimations of the distribution of the statistics that are used to test whether or not the functions come from the same model.

The use of functional homogeneity statistics in the study of Network Management data is a natural approach to test the representativeness of a certain set of parameters. Taking into account these measurements, it is possible to define algorithms to detect parameters that characterize the typical state of a network considering the whole evolution of the parameters and not only certain statistical summaries – *e.g.* means or medians. Additionally, functional homogeneity tests can be applied in order to detect changes with different aims. They can help to detect anomalous events and sustained trend changes using the divergence between the evolution of the observed values.

F. Other FDA-based techniques

FDA includes other techniques, such as functional clustering, classification and forecasting (*e.g.* regression models with functional or scalar response, among other prediction tools). The application of curve clustering and classification to the identification of certain characteristics of network traffic (*e.g.* an application that generates that traffic) must be studied, in order to overcome the limitations derived from data encryption. Furthermore, forecasting based on functional regression is of particular interest, as it could impact on dynamical planning in SDNs and other flexible network systems such as Cloud infrastructures.

IV. Conclusions

We have presented a preliminary evaluation of the gains that FDA entails in several areas of Network Management activities. Our study has provided an initial exploration of this branch of Statistics, and a description of several FDA techniques and their application to Network Management activities.

Firstly, we have considered data representation and functional PCA. Among other advantages, these data preprocessing steps provide a certain level of compression, as a result of the reduction of the components that are needed to represent data; and a semantic decomposition of parameters that enriches network dynamic interpretation. We have considered other analytical techniques, namely phase-plane analysis, functional depth and homogeneity. These instruments provide a starting point for the characterization and detection abnormal behaviors; and robust approaches that consider curves or surfaces as a whole taking into account the joint behavior of an arbitrary number of parameters.

This initial exploration of FDA applied to Network Management tasks shows the way towards an interesting step forward for network researchers and practitioners. This first contact of FDA with Network Management must be continued with further evaluation and with the development of advanced methods that take advantage of the strengths of FDA.

Acknowledgment

This work has been partially supported by the Spanish Ministries of Economy and Competitiveness (PackTrack, TEC2012-33754), and of Science and Innovation (MTM2013-44045-P).

References

[1] K. Xu, F. Wang, and H. Wang, "Lightweight and Informative Traffic Metrics for Data Center Monitoring," *Journal of Network and Systems Management*, vol. 20, no. 2, pp. 226–243, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10922-011-9200-6

[2] F. Simmross-Wattenberg, J. Asensio-Pérez, P. Casaseca-de-la Higuera, M. Martín-Fernández, I. Dimitriadis, and C. Alberola-López, "Anomaly detection in network traffic based on statistical inference and alpha-stable modeling," *Dependable and Secure Computing, IEEE Transactions on*, vol. 8, no. 4, pp. 494–509, July 2011.

[3] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, and S. López-Buedo", "Characterization of the busy-hour traffic of IP networks based on their intrinsic features," *Computer Networks*, vol. 55, no. 9, pp. 2111 – 2125, 2011.

[4] F. Mata, J. L. García-Dorado, and J. Aracil, "Detection of traffic changes in large-scale backbone networks: The case of the Spanish academic network," *Computer Networks*, vol. 56, no. 2, pp. 686 – 702, 2012.

[5] R. De O Schmidt, R. Sadre, N. Melnikov, J. Schönwälder, and A. Pras, "Linking network usage patterns to traffic gaussianity fit," in *Networking Conference, 2014 IFIP*, June 2014, pp. 1–9.

[6] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," *SIGMETRICS Perform. Eval. Rev.*, vol. 32, no. 1, pp. 61–72, Jun. 2004. [Online]. Available: http://doi.acm.org/10.1145/1012888.1005697

[7] J. L. García-Dorado, J. Aracil, J. A. Hernández, and J. E. López de Vergara, "A queueing equivalent thresholding method for thinning traffic captures," in *Network Operations and Management Symposium, 2008. NOMS 2008. IEEE*, April 2008, pp. 176–183.

[8] M. Nassar, O. Dabbebi, R. Badonnel, and O. Festor, "Risk management in VoIP infrastructures using support vector machines," in *Network and Service Management (CNSM), 2010 International Conference on*, Oct 2010, pp. 48–55.

[9] S. Saad, I. Traore, A. Ghorbani, B. Sayed, D. Zhao, W. Lu, J. Felix, and P. Hakimian, "Detecting P2P botnets through network behavior analysis and machine learning," in *Privacy, Security and Trust (PST), 2011 Ninth Annual International Conference on*, July 2011, pp. 174–180.

[10] A. Cuevas, "A partial overview of the theory of statistics with functional data," *Journal of Statistical Planning and Inference*, vol. 147, no. 0, pp. 1 – 23, 2014.

[11] P. H. Eilers and B. D. Marx, "Flexible smoothing with B-splines and penalties," *Statistical science*, pp. 89–102, 1996.

[12] L. H. Gibeli, G. D. Breda, R. S. Miani, B. B. Zarpelão, and L. de Souza Mendes, "Construction of baselines for VoIP traffic management on open MANs," *International Journal of Network Management*, vol. 23, no. 2, pp. 137–153, 2013. [Online]. Available: http://dx.doi.org/10.1002/nem.1820

[13] V. Moreno, P. M. Santiago del Río, J. Ramos, D. Muelas, J. L. García-Dorado, F. J. Gómez-Arribas, and J. Aracil, "Multi-granular, multi-purpose and multi-Gb/s monitoring on off-the-shelf systems," *International Journal of Network Management*, vol. 24, no. 4, pp. 221–234, 2014. [Online]. Available: http://dx.doi.org/10.1002/nem.1861

[14] S. López-Pintado and J. Romo, "A half-region depth for functional data," *Comput. Stat. Data Anal.*, vol. 55, no. 4, pp. 1679–1695, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.csda.2010.10.024

[15] G. Claeskens, M. Hubert, L. Slaets, and K. Vakili, "Multivariate functional halfspace depth," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 411–423, 2014.

[16] A. Aguado, V. López, J. Marhuenda, J.-P. Fernández-Palacios *et al.*, "ABNO: a feasible SDN approach for multi-vendor IP and optical networks," in *Optical Fiber Communication Conference*. Optical Society of America, 2014, pp. Th3I–5.

[17] R. J. F. Díaz, R. E. Lillo, and J. Romo, "Homogeneity test for functional data based on depth measures," Tech. Rep., 2014.