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 the diversity both in terms of the situations that must be faced and of the data used to extract conclusions. This complex and changing context imposes diverse necessities and restrictions that must be fulfilled by the management tools in order to result useful. 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 and forecasting for network dimensioning.**

I. Introduction

Network Management tasks are currently characterized by the diversity both in terms of the situations that must be faced and of the data used to extract conclusions. Furthermore, the huge amount of data generated in modern computer networks that is processed and persisted during Network Management activities must be optimized to improve the scalability of the 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 can result insufficient if their hypothesis are not satisfied or if the ideal deployment scenarios are not in accordance with those under analysis. For example, the encryption of the transmitted information and other legal and privacy aspects concerning this data limit some state-of-the-art solutions – *e.g.* intrusion detection systems that rely on Deep Packet Inspection (DPI) techniques.

To face theses challenges, we propose the application of Functional Data Analysis (FDA) techniques. FDA considers random variables which are functions, thus immersing them in spaces of infinite dimension. The good results of the application of FDA to other problems that deal with data such as weather forecasting and some economical studies, motivate the exploration of the applicability of FDA in the area of Network Management.

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

II. Related work

In this section, we present related works that cover different tasks of capital importance in Network Management. Specifically, we describe solutions, tools and methods used to face different processes that conform some of the activities of the different functional areas of Network Management. For all of them, we comment their limits and highlight different aspects that potentially point to FDA as a "toolbox" that could improve current solutions.

Regarding to network performance measurement, in [1] authors propose a new metric with 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, so it can be widespread to almost every scenario.

Following with network modeling, 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 distribution of throughput is used to detect certain anomalous behaviors. 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] authors present statistical network models using Gaussian processes. Particularly, 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 an hypothesis that cannot be directly assumed in general [5], [2]. FDA techniques do not suffer from this problem, as they no require assumptions relative to the marginal distribution of the considered parameters.

Authors in [6], [7] propose the application of some filtering and preprocessing techniques to network data. Those solutions share 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 it is quite restrictive – *e.g.* Wavelets. Nevertheless, the central ideas of those works pinpoint to the gaining 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 base of several machine learning and data mining techniques that take advantage of some properties of sets when their dimension increases. For instance, Support Vector Machines (SVM) are a well-known example that has been successfully applied to diverse problems related to Network Management activities. In [8], authors explore the results that SVMs provide in anomaly detection, management of Quality of Experiences (QoE) and QoE prediction.

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 to the analysis and a base to use all the advanced features that FDA encompass.

III. FDA and Network Management

FDA is a set of techniques that embed statistical methods in functional spaces. In this section, we formally define several elements and applications with an extensive description of use cases that highligh 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, so the first step when using functional approaches is to interpolate and – if necessary – smooth this data *with any global approximation technique*. In the literature, B-splines are a common election due to their properties [11], although other representations are totally admissible if the structure of data is well preserved. In general, we will represent the set of functions that conform the selected basis as $\{B_k(t)\}_{t\in\mathbb{R}_+,k\in\mathbb{Z}}$ and the coefficients giving the projections of the observations with respect the basis as $\{\beta_k\}_{k\in\mathbb{Z}}$. Thus, given an observation $\{X_t\}_{t\in\mathbb{R}_+}$, we can represent that observation as $\{X_t\} = \sum_{j\in\mathbb{Z}} \beta_j B_j(t),\ t \in \mathbb{R}_+$. In practice, the representation is truncated so the representation is given by the expression: $\{X_t\} = [\sum_{j\in\mathbb{J}} \beta_j B_j(t)] + \epsilon(\mathbb{J}, \{B_j\}),\ t \in \mathbb{R}_+$ with $\mathbb{J}$ a finite set of indexes and ϵ a term of error dependent of both the set of indexes and the selected basis. This representation presents several advantages. First, 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. Second, it provides a robust estimation of the derivatives of the model. Finally, it provides a mean to select the components that contain the more representative information of the model as we show below.

FDA allows the development of compact expressions of network parameters (packets, flows, bytes, active IP directions,...) represented as a function of other parameter or 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 shows the result of the interpolation of throughput data using third grade B-splines. To obtain this representation, we use a set of sampled points and later, we evaluate the resulting curve for each time point.

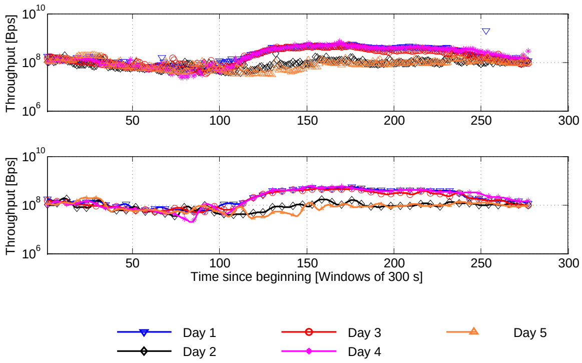

Figure 1. Third grade B-splines representation for 5 days of throughput registers, Educational network

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 results interesting when we consider scalability issues, at the same time that provides a first step to apply other FDA techniques.

B. Functional PCA

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

To test these ideas in the field of Network Management, we apply Functional PCA to the registers that were considered above, extended with further observations. We use 20 daily observations interpolated as previously described. The computation of Functional PCA on this set of observations has been obtained using the package `fda` of `R`. Figure 2 represents the first 6 harmonics – that is, the number of components needed to cover the 95% of the original variance. With respect to the compression, the result allows the reduction in a factor of 10 of the number of components needed to represent the data. Notice that the first principal component, highlighted in the figure, represents a scaled *approximation* of the dynamic of the observations – omitted for the shake of clarity, given that they are similar to those represented in Figure 1. With the consideration of additional principal components, we enrich the representation with *details* that cover a higher proportion of the observed variability.

As a result of these properties, this method allows the reduction of the volume of the data that must be persisted with a criterion based on the variability structure. Compared to other alternatives as those commented in Section II, PCA harmonics represent a meaningful decomposition of the observations.

Functional PCA provides several advantages in problems derived of certain Network Management activities. First of all, it provides a second-level compression of data, as we can select a subset of the principal components controlling variability information losses. Furthermore, changes in a certain harmonic indicate different types of changes in the measured parameter depending on the variability covered by such harmonic. This fact motivates a novel vision of anomalies and other variations of network dynamics, linking them to the behavior of the func-

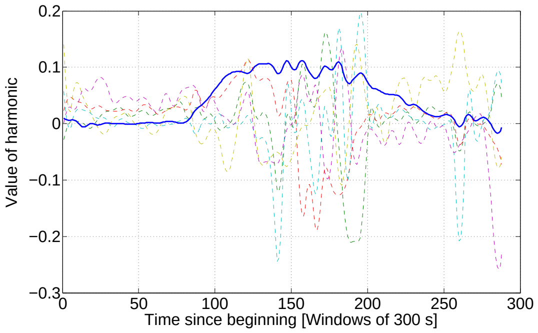

Figure 2. Harmonics covering 95% of the original variance – 6 components – after Functional PCA on throughput registers, Educational network. The first component, covering 79.27% of the original variance is highlighted

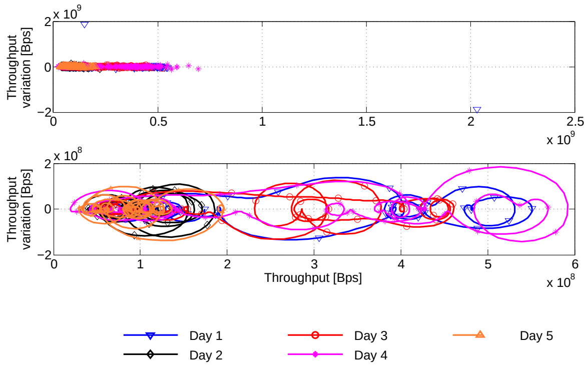

Figure 3. Comparative view of phase-plane plot obtained with point estimation (finite difference method), and analytical derivation based on B-spline representation

tional principal components. Finally, the semantic information of this decomposition allows to define sensitive baselines in terms of the evolution of the harmonics.

C. Phase-plane analysis

Phase-plane analysis describe 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 several analytical processes, such as the study of the stability of a certain system. Furthermore, this joint evolution can provide enhanced results when considering dynamical aspects of functions in analysis processes related to homogeneity studies or clustering of curves.

If we consider a certain functional observation $f(t)$, 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.

Since not only the value of parameters, but also speed changes are important in several tasks of Network Management, phase-plane analysis is an approach that can be useful in various processes of decision making.

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 on the functional representation previously obtained. Notice that some points are not detectable in the latter the abnormal points observed in the former are supressed in the interpolation step.

This representation is useful for visual detection of abnormal events and provides extended information about the evolution of the network state. Additionally, the inclusion of several parameters in this analysis using multivariate functions allows the study of joint relations between different magnitudes and their derivatives and it can be used to characterize different events in network dynamics by means of clustering and classification methods for curves and surfaces.

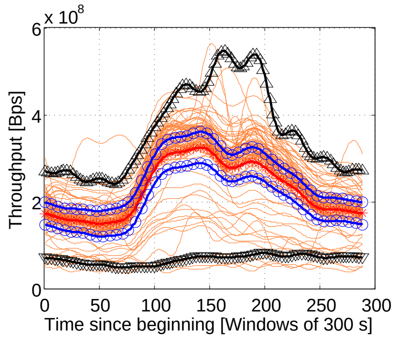

Figure 4. Example of depth region using an extended set of throughput registers, Educational network. Mean curve: red asterisks. Mean confidence interval: blue circles. Depth region: black triangles.

D. Functional depth

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

Although many depth notions are described in the literature of FDA, in the following we will always consider the definition included in [14], given by the expression $MS_{n,H}(x) = \min\{SL_n(x), IL_n(x)\}$ where $SL_n(x) = \frac{1}{n\lambda(I)}\sum_{t=1}^{n}\lambda\{t \in I : x(t) \leq x_i(t)\}$ and $IL_n(x) = \frac{1}{n\lambda(I)}\sum_{t=1}^{n}\lambda\{t \in I : x(t) \geq x_i(t)\}$ with λ the Lebesgue measure.

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 observation n is dominated and dominating other elements of the set of observations. Additionally, there are some proposals of multivariate functional depth which allow to apply depth-based concepts not only for curves, but also for multivariate functions and surfaces.

Order statistics give robustness when analyzing the typical behavior of a network as a result of the isolate character of outliers and abnormal values. In Figure 4, we show a depth region based on the definition given above. This region covers the 80% of observations of the set of observations that we have considered along this work. The lines marked with the black triangles correspond to the curves that delimit that region. The mean curve is represented in red with asterisks, while the blue lines wit circles indicate the confidence interval of the mean at each point with $\alpha = 5\%$. The depth region leave inside this values, providing frontiers derived from the data structure.

The appearance of network infrastructures that allow both dynamic configuration of rules and resource deployment (*e.g.* Software Defined Networking or the Application-based Network Operations (ABNO) architecture [15]) points to the establishment of baselines that take into account network behavior at each time frame. Depth-based metrics are good candidates to 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 to be detected [13].

E. Functional homogeneity

Given two samples, a natural question is 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, providing some promising results. Different homogeneity measurements have been considered in the FDA literature. Those are based in computational approaches to obtain estimations of the distribution of the statistics that are used to test if 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.* application that generates that traffic) must be studied, in order to overcome limitations derived from data encryption. Forecasting based on functional regression is, additionally, 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 an evaluation of the gaining that FDA entails in several areas of Network Management activities. Our study has provided an initial exploration of this branch of Statistics, and the description of several FDA techniques and their application to Network Management activities.

Firstly, we have considered data representation and functional PCA, which provide some means of data preprocessing. Among other advantages, they provide a certain level of compression, as a result of the reduction of the components that are needed to represent data in terms of a suitable functional basis; and a semantic decomposition of parameters that enriches network dynamic interpretation. We have considered other analytical techniques, namely phase-plane analysis, and 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 indicates the direction of an interesting step forward for network researchers and practitioners. This first contact of FDA with Network Management tasks 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 Ministry of Economy and Competitiveness under the project PackTrack (TEC2012-33754).

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] 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.