# Eigen-Value based Multivariate Selfsimilarity Analysis in Internet Traffic: A case study

Romain Fontugne
*Internet Intiative Japan (IIJ)*
*Tokyo, (JP)*

Patrice Abry
*CNRS, ENS de Lyon, LPENSL,*
*UMR5672, 69342, Lyon cedex 07, France*

Kensuke Fukuda
*National Institute of Informatics (NII)*
*Tokyo, (JP)*

Kenjiro Cho
*Internet Intiative Japan (IIJ)*
*Tokyo, (JP)*

Gustavo Didier
*Math Department,*
*Tulane University, New Orleans (USA)*

Herwig Wendt
*Université de Toulouse*
*CNRS, IRIT, Toulouse (FR)*

*Abstract*—**Internet traffic modeling and analysis is critical for network design and for cybersecurity. Internet time series are well characterized by scalefree temporal dynamics. However, scalefree analysis remained so far univariate, applied independently to directional counts of either bytes or packets while challenges in cybersecurity naturally call for multivariate analysis. Elaborating on recent theoretical developments on eigenvalue-based multivariate selfsimilarity analysis, this work provides evidence, for the first time, of multivariate selfsimilarity in 17 years of Internet traffic data from the MAWI repository. It discusses the potential use of multivariate selfsimilarity for regular background traffic characterization and anomaly detection.**

*Index Terms*—**component, formatting, style, styling, insert.**

## I. Introduction

**Context.** Ever since the seminal works [1]–[3], it has been abundantly documented that Internet traffic time series are well characterized by scalefree temporal dynamics, and can be accurately modeled by selfsimilarity and long-memory [4]–[6]. However, scalefree analysis of Internet traffic has so far remained univariate (one time series at a time), whereas network monitoring and cybersecurity challenges call for multivariate analysis. This work thus puts forth the first multivariate selfsimilarity analysis of Internet traffic time series.

**Related work.** Internet traffic is classically analyzed via the study of time series consisting of aggregated counts of either IP (Internet Protocol) packets or bytes. After the fundamental works [1]–[3], it has been shown that their temporal dynamics display robust scalefree dynamics over two broad ranges of fine and coarse scales, well separated by the *Round-Trip-Time* [6], [7]. Wavelet (or multiscale) analysis were shown to permit the robust estimation of the scaling exponents and, thus, the characterization of Internet traffic [5], [8].

A classical issue involved in analyzing Internet temporal dynamics is that most traces, if not all, consist of mixtures of background (or normal) traffic possibly mixed up with massive anomalous traffic, be it malicious or simply associated with specific random events. To disentangle the statistics of the background traffic from those of anomalous traffic, original strategies were devised, following approaches from video flow analysis [9], [10]. Based on random projections, an original trace is *hashed* into several surrogates traces, and further used to compute robust statistics [7], [11]. Combined with random projections, wavelet tools have permitted the robust assessment of scalefree dynamics that could hence be correctly associated with either queuing mechanisms at coarse scales [3], [6] or with technological protocols at fine scales [6]. These combined tools also provided anomalous traffic detection based on deviations from background traffic scalefree properties [7], [11], [12]. This further led to the investigation of the existence of multifractal properties at fine scales [4], [6], [13], [14]. However, due to lack of existing theoretical and practical tools, selfsimilarity analysis in Internet traffic has so far been applied independently either to packet or byte count time series, to incoming or outgoing traffic. This led to the question of whether scalefree dynamics were to be associated with packet or byte counts, or were identical for incoming or outgoing traffics [7], [11]. These issues triggered ongoing debates not well framed in univariate analysis. Recently, multivariate selfsimilarity was defined and theoretically studied in [15]–[17]. Furthermore, wavelet-based tools based on the eigenvalues of the (wavelet) spectrum were devised to permit the theoretically robust and practically efficient assessment of multivariate selfsimilarity, and the estimation of the corresponding scaling exponents [18], [19]. This work constitutes the first attempt at applying these tools to multivariate Internet traffic.

**Goals, contributions and outline.** The goal of this work is to report the first 4-variate analysis of selfsimilarity in Internet traces. The contributions are i) to study the extent to which robust random projection-based scalefree analysis extends to the proposed eigen-wavelet multivariate selfsimilarity analysis; and ii) to investigate the potential benefits of the proposed eigen-wavelet multivariate selfsimilarity analysis, as compared to classical bivariate analysis, for Internet traffic modeling and anomalous traffic detection. To that end, the eigen-wavelet-based multivariate selfsimilarity analysis is compared to the classical bivariate analysis (Section II). These tools are then applied to Internet traffic traces obtained from the MAWI repository spanning the years 2007-2023, described in Sec-

PA gratefully acknowledges the National Institute of Informatics for a yearly visiting researcher position, that significantly supported this work.

tion III. The relevance of random projections to perform a robust-to-anomalies analysis of multivariate selfsimilarity in Internet traffic is detailed in Section IV. The robust assessment of multivariate selfsimilarity in 4-variate Internet traces is reported in Section V, where its potential for Internet time series modeling and anomaly detection is discussed.

## II. Multivariate Selfsimilarity Analysis

**Multivariate Wavelet Spectrum.** Let $\psi$ denote the so-called mother wavelet, characterized by its number of vanishing moments $N_\psi$, itself a positive integer such that $\forall n = 0, \ldots, N_\psi - 1$, $\int_{\mathbb{R}} t^k \psi(t) dt \equiv 0$ and $\int_{\mathbb{R}} t^{N_\psi} \psi(t) dt \neq 0$ [20].

For a $M$-variate signal $\underline{X}(t) = (X_1(t), \ldots, X_M(t))$, the discrete wavelet transform vector coefficients are defined as $D_{\underline{X}}(j,k) = d_{X_1}(j,k), \ldots, d_{X_M}(j,k))$, where the univariate wavelet coefficients are computed independently for each component as $d_{X_m}(j,k) = 2^{-j/2} \langle \psi_{j,k} | X_m \rangle$, with $\{\psi_{j,k}(t) = 2^{-j/2} \psi(2^{-j}t - k)\}_{(j,k) \in \mathbb{Z}^2}$ the collection of dilated and translated templates of the mother wavelet.

The multivariate wavelet spectrum is defined as the set of covariance matrices $S(2^j)$ of $D_{\underline{X}}(2^j, k)$, computed at each scale $2^j$, as $(2^j) \triangleq \frac{1}{n_j} \sum_{k=1}^{n_j} D_{\underline{X}}(2^j,k) D_{\underline{X}}(2^j,k)^*$, where $^*$ denotes matrix transposition, and $n_j$ the number of wavelet coefficients available at scale $2^j$.

**Univariate and Bivariate Selfsimilarity Analysis.** Univariate selfsimilarity analysis amounts to assuming that the diagonal entries of $S$ display power-law behavior across scales $2^j$, each controlled by a scaling exponent associated with univariate selfsimilarity parameters. This leads to linear relations in terms of log-log representations:

$$\log_2 S_{m,m}(2^j) \simeq \log_2 \sigma_m^2 + j(2H_m^U + 1). \tag{1}$$

The off-diagonal entries $S_{m,m'}(2^j)$ $(m' \neq m)$ of the wavelet spectrum account for the pairwise cross-temporal dynamical dependencies amongst components, and thus provide classical bivariate analysis. Bivariate selfsimilarity thus translates into off-diagonal $S_{m,m'}(2^j)$ displaying power-law behavior across scales $2^j$, controlled by scaling exponents $H_{m,m'}$:

$$\log_2 |S_{m,m'}(2^j)| \simeq \log_2 |\sigma_{m,m'}| + j(2H_{m,m'} + 1). \tag{2}$$

These $H_{m,m'}$ bring new insights on cross-temporal dynamics when $H_{m,m'}$ departs from $(H_m^U + H_{m'}^U)/2$ [21].

This classical multivariate wavelet analysis leads to the definition of the wavelet coherence function, which can be read as a pairwise, scale-dependent, correlation coefficient [21]:

$$C_{m,m'}(2^j) = \frac{S_{m,m'}(2^j)}{\sqrt{S_{m,m}(2^j) S_{m',m'}(2^j)}}. \tag{3}$$

When $H_{m,m'} = (H_m^U + H_{m'}^U)/2$, $C_{m,m'}(2^j)$ does not depend on scale and simplifies into the classical overall correlation coefficient $C_{m,m'}(2^j) = \sigma_{m,m'} / \sqrt{\sigma_m^2 \sigma_{m'}^2}$.

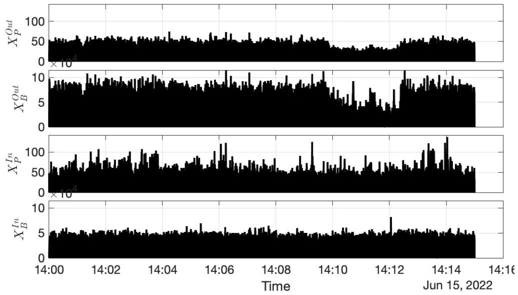

Figure 1. **Internet Traffic.** 4-Variate time series for an arbitrary day.

**Eigen-Wavelet-based Multivariate Selfsimilarity Analysis.** The classical analysis described above studies the behavior of each of the entries of $S$ as a functions of the scales $2^j$, one after the other. Therefore, it remains in essence pairwise, and hence, restrictive in the construction of bivariate selfsimilarity analysis. To better account jointly for multivariate cross-dependencies and rich cross-temporal dynamics, an alternative *eigen-wavelet*-based multivariate selfsimilarity analysis was recently proposed [18], [19]. It reverses the paradigm of classical analysis: First, it considers the $M$ components together by computing the eigenvalues $(\lambda_1(2^j), \ldots, \lambda_M(2^j))$, of $S(2^j)$ independently at each scale ; Second, it assumes that the eigenvalues $\lambda_m(2^j)$ of the wavelet spectrum $S(2^j)$ asymptotically behave as power laws with respect to the scales $2^j$, with scaling exponent $2H_m^M - 1$. This leads to linear relations in log-log representations [18], [19], [22]:

$$\log_2 |\lambda_m(2^j)| \simeq \log_2 \lambda_m^0 + j(2H_m^M + 1). \tag{4}$$

The set of eigen-functions $\{\log_2 |\lambda_m(2^j)|\}$, $m = 1, \ldots, M$, thus constitutes a new and original tool to investigate jointly, or in a truly multivariate way, the multiscale statistics of a multivariate signal $\underline{X}$. It complements multiscale univariate analysis via the functions $\{\log_2 S_{m,m}(2^j)\}$ and bivariate analysis via the functions $\{\log_2 |S_{m,m'}(2^j)|\}$ and $C_{m,m'}(2^j)$.

## III. Mawi Dataset

MAWI archive, http://mawi.wide.ad.jp, consists of a unique Internet backbone traffic repository, collecting traces every day, from 14:00 to 14:15 (Japanese Standard Time), from 2001 till today. Packet header traces are anonymized and traces made publicly available [23]. Each recording comprises several hundreds of millions of IP packets, implying an inter-arrival time of the order of microseconds.

Data used here were collected at the *samplepoint-F* transit link of the MAWI network, connecting several Japanese research institutes and universities to the Internet. For the present case study, traces recorded the 15th (arbitrary choice) of each month, from 2007 to 2023 were analyzed, consisting of 17×12 count times series. Each recording is split into four time series, consisting of Pkt or Byte counts, either entering (in) or leaving (out) the MAWI network. Each trace is thus 4-variate: $\underline{X} = (X_1, X_2, X_3, X_4) = (X_{Pkt}^{Out}, X_{Byte}^{Out}, X_{Pkt}^{In}, X_{Byte}^{In})$. Counts are aggregated on a fine temporal grid: $T_s = 0.125$ milliseconds, resulting in sample size $N = 7200000$ for each of the time series. Fig. 1 illustrates these four time series for an arbitrary day.

## IV. Random projections

**Random projection-based robust analysis.** A key feature of Internet traces is that they consist of marked point processes,

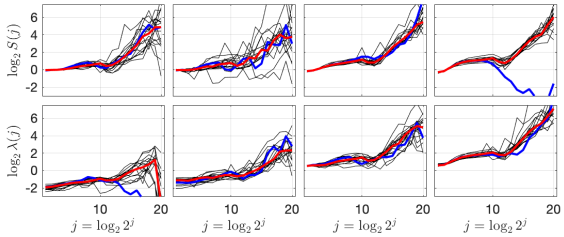

Figure 2. **Univariate/Multivariate selfsimilarity Analysis.** Functions $\{\log_2 S_{m,m}(2^j)\}_{m=1,\ldots,4}$ (top plots), and eigen-functions $\{\log_2 |\lambda_m(2^j)|\}_{m=1,\ldots,4}$ (bottom plots). Blue, black and red lines correspond to estimates obtained respectively, from the full trace $\underline{X}$, from each of the $2^L = 16$ sub-traces $\underline{X}_l$, and as the median across sub-traces estimates. Date: 2003/06/03.

with the 5-tuple mark consisting of the Internet Protocol, the Source and destination IP addresses and ports. A random projection (or sketch) of the trace consists in applying a $k$-universal hash function $h$ [24], taking values in an alphabet of size $2^L$, to any of the four last attributes, referred to as $A$. A sketch procedure thus splits an original IP trace $\underline{X}$ into $2^L$ surrogate-traces, $\underline{X}_l$, each consisting of all packets with identical sketch output $h(A)$, thus preserving flow structure (packets belonging to a same flow are assigned to the same sub-trace). A key point is that the $M = 4$ components are all jointly subjected to the same random projection so that the multivariate statistical structure of the original trace is preserved in each of the sub-traces.

The intuition is that when there is no anomaly, all surrogate traces are statistically equivalent, up to a multiplicative constant. Conversely, whenever present, an anomaly is projected on a single sub-trace, thus exhibiting outlier statistics. Robust estimation stems from applying a median across surrogate traces, hence providing a reference point for normal traffic with little sensitivity to anomalies.

Selecting different hash keys $A$ leads to different projections. The present work used both Source IP and Destination IP addresses as obvious choices. It is observed that equivalent robust characterizations stem from these two different choices, or from the use of different hash functions (not shown here).

**Robust multiscale characterizations.** Figs. 2 and 3 illustrate, for different days, the robust multiscale analysis achieved by combining random projections ($L = 4$) with multivariate wavelet analysis.

For the sake of pedagogy, let us focus first on Fig. 2. The blue lines correspond to the functions $\log_2 S_{m,m}(2^j)$ (top) and $\log_2 |\lambda_m(2^j)|$ (bottom) estimated from the full trace $\underline{X}$ before random projection. They can vary significantly from one component to the other or across different days. These estimates are, indeed, significantly affected by the anomalous part of traffic and do not represent a robust characterization of the temporal dynamics of Internet traffic. The $2^L$ black lines correspond to these same functions estimated from each single sub-trace $\underline{X}_l$. For one same day and one same component, some sub-trace behaviors may significantly differ from the majority or of typical behaviors. These sub-traces with different behaviors likely correspond to anomalous traffic, whereas those collapsing on a same typical behavior can be associated with background normal traffic. The red lines result from computing a median, hence a robust statistic, independently at each scale, across these $2^L$ (black) sub-traces. It can be seen that the blue and the red lines differ, often significantly. These median (red) functions thus provide a robust characterization of the evolution across scales of the statistics of the full trace $\underline{X}$, as opposed to the blue ones, which are significantly biased by anomalous traffic. The same observations and conclusions can be drawn from the analysis of Fig. 3, which reports either the functions $\{\log_2 |S_{m,m'}(2^j)|\}$ (lower triangle) or $C_{m,m'}(2^j)$ (upper triangle).

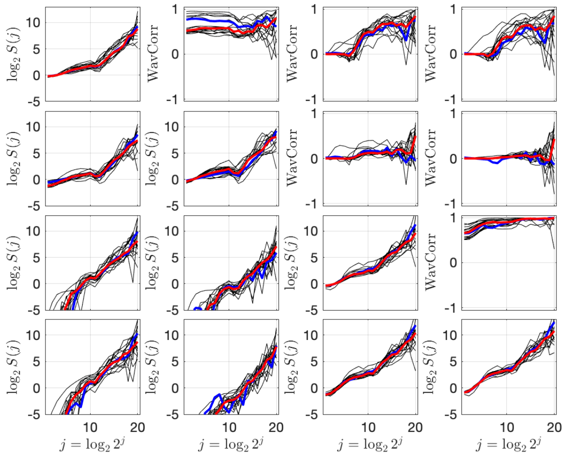

Figure 3. **Bivariate selfsimilarity analysis and wavelet coherence function.** Functions $\{\log_2 S_{m,m}(2^j)\}_{m=1,\ldots,4}$ (diagonal plots), $\{\log_2 |S_{m,m'}(2^j)|\}_{m=1,\ldots,4,m'<m}$ (lower triangle) and wavelet coherence function $\{C_{m,m'}(2^j)\}_{m=1,\ldots,4,m'>m}$ (upper triangle). Blue, black and red lines correspond to estimates obtained respectively, from the full trace $\underline{X}$, from each of the $2^L = 16$ sub-traces $\underline{X}_l$, as the median across sub-traces estimates. Date: 2007/02/15.

These investigations lead to the first major contribution of this work: Extending results already obtained for univariate selfsimilarity analysis, the combination of random projections with multivariate wavelet analysis permits us to obtain a characterization of multivariate selfsimilarity of the background normal traffic, that is robust to the anomalous traffic. This holds both for the bivariate selfsimilarity analysis (functions $\{\log_2 |S_{m,m'}(2^j)|\}$ and $C_{m,m'}(2^j)$) and for the truly multivariate analysis selfsimilarity analysis (functions $\log_2 |\lambda_m(2^j)|$). This could not be taken for granted as computing eigenvalues consists of a highly nonlinear operation.

## V. 4-variate SelfSimilarity in Internet traffic

The present section focuses on the analysis, interpretation and use of these robust characterizations. All results reported below and discussed are illustrated by means of graphics associated with different days, chosen for illustration and pedagogical purposes. Plots for each of the days analyzed in the present case study are available upon request. Their analysis yields essentially identical conclusions.

**Univariate selfsimilarity analysis.** Fig. 2 (top plots) and Fig. 3 (diagonal plots) report, for two different days, univariate

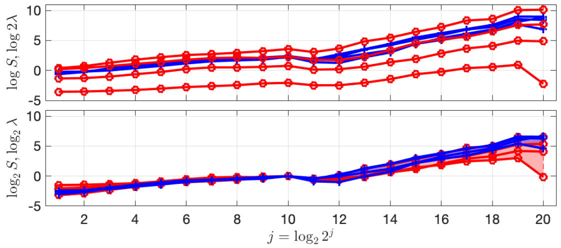

Figure 4. **Multivariate vs. selfsimilarity analysis.** Day: 2007/06/15. Top: $\log_2 S_{m,m}(2^j)_{m=1,\dots,4}$ (blue) and $\{\log_2 |\lambda_m(2^j)|\}_{m=1,\dots,4}$ (red). Bottom: $\log_2 S_{m,m}(2^j)_{m=1,\dots,4}$ and $\{\log_2 |\lambda_m(2^j)|\}_{m=1,\dots,4}$ with forced value 0 at octave $j = 10$ to ease comparisons of scaling behavior.

functions $\log_2 S_{m,m}(2^j)$ and confirm earlier findings [6], [7], [13]: Internet time series present remarkable scalefree dynamics in two distincts ranges of scales, separated by a characteristic time-scale of roughly $T_s * 2^{12} \simeq 0.5$s, highly reminiscent of the typical Round-Trip-Time (RTT) in Internet Protocol (IP) [6], [13]. The coarse scale (CS) range spans 7 octaves (2 decades), from $2^{13}T_s \simeq 1\text{s} \leq 2^j T_s \leq 2^{19} T_s \simeq 2\text{min}$ and can be associated with selfsimilarity related to queuing mechanisms for Internet content files, with heavy tail file sizes [1], [6]. The fine scale (FS) range also spans 7 octaves, from $2^3 T_s \simeq 1\text{ms} \leq 2^j T_s \leq 2^{10} T_s \simeq 0.25$s and can be related IP technological mechanisms [6], [13].

The next paragraphs investigate whether scalefree dynamics are also present in the multivariate and cross-temporal dynamics of Internet time series.

**Bivariate selfsimilarity analysis.** Prior to conducting any multivariate analysis, it is mandatory to asses the existence of cross-temporal dynamics in multivariate Internet time series. To that end, the wavelet coherence functions (Eq. 3) between pairs of components are reported in Fig. 3 (upper triangle plots). They show that, for each direction, byte and packet count time series, $(X_1, X_2) = (X^{Out}_{Pkt}, X^{Out}_{Byte})$ and $(X_3, X_4) = (X^{In}_{Pkt}, X^{In}_{Byte})$, display significant correlation levels, quasi-constant across the 20 available octaves. This shows that Packet and Byte time series share closely related temporal dynamics, across $\simeq 6$ decades of time scales (from milliseconds to several minutes). Interestingly, Packet time series $(X_1, X_3) = (X^{Out}_{Pkt}, X^{In}_{Pkt})$ from each direction also show significant correlation, but only across the coarse-scale range, i.e., for time scales above the RTT. This is likely due to the HTTP protocol that establishes bidirectional connections to regulate the flow of IP packets within connections depending on Acknowledgement of Receipt protocols. Packet time series are not correlated at fine scales, likely as packet injection protocols only depend on the technology at each end of the backbone and are thus not related. These significant correlations across scales between the components of Internet traces constitute the second contribution of this work, and motivate multivariate selfsimilarity analysis.

Fig. 3 (lower triangle plots) complements the analysis of cross-temporal dynamics. For the highly correlated pairs $(X^{Out}_{Pkt}, X^{Out}_{Byte})$ or $(X^{In}_{Pkt}, X^{In}_{Byte})$, the functions $\log_2 S_{m,m'}$ perfectly reproduce biscaling in the two ranges of coarse and fine scales. For $(X^{Out}_{Pkt}, X^{In}_{Pkt})$, cross-scalefree dynamics is observed at coarse scales only, in agreement with the fact that the pair is correlated at coarse scales only. This indicates a coupling of the In and Out directional time series via coarse-scale temporal dynamics of the packet count time series.

**Eigenvalue-based multivariate selfsimilarity analysis.** The eigen-wavelet-based multivariate analysis of the 4-variate Internet time series yields the function $\log_2 \Lambda_m(2^j)$ reported in Fig. 2 (bottom plots). Interestingly, these plots show that all four functions $\log_2 \Lambda_m(2^j)$, $m = 1, \dots M = 4$, display biscaling behavior, with scaling both at coarse scales and at fine scales, separated by the sole characteristic time scale of $\simeq 0.5$. It is important to underline that biscaling on each component, i.e., for each function $\log_2 S_{m,m}(2^j)$, does not at all automatically imply the same biscaling on all 4 functions eigen-function $\log_2 \Lambda_m(2^j)$. The fact that biscaling is observed on all 4 functions eigen-function $\log_2 \Lambda_m(2^j)$, even for $m = 1$ or $m = 2$, that is for the smallest eigenvalues, clearly indicates that biscaling is a truly multivariate statistical property of Internet time series, not only a univariate one. Fig. 4 superimposes the $\log_2 S_{m,m}$ (blue) and the $\{\log_2 |\lambda_m|\}$ (red). Bottom plots force all functions to take value 0 at octave $j = 10$. These plots show that, while scaling exponents (slopes of linear fits) estimated at fine scales from $\log_2 S_{m,m}$ (blue) would be mostly identical, they would differ when estimated from $\{\log_2 |\lambda_m|\}$ (red). For other days, the same observation can be reported for scaling at coarse scales.

The analysis clearly indicates that biscaling is deep-rooted in the joint or multivariate temporal dynamics of Internet traffic. It also shows that the eigen-wavelet-based multivariate selfsimilarity analysis extracts richer information that is embedded in their multivariate statistical structure, and thus better characterizes cross-temporal dynamics in Internet time series. This constitutes our third and most important contribution.

**Anomaly detection.** Repeating random projections, with different hash functions, can also be used for anomaly detection [7], [11]: For a chosen function, e.g., $\log_2 S_{1,1}(2^j)$, the 5 (out of $2^L = 16$) random projections $\log_2 S^l_{1,1}(2^j)$ that depart most (in $L^1$-norm) from the median are kept. This is repeated independently for 8 different hash-functions. Given that there exist at most $2^{32}$ IP addresses (in IPV4 protocol), the probability that a single address remains in the top-5 deviating projections after the use of 8 hash tables is extremely close to 0. Therefore, the fact that an IP address survives *by chance* in the intersection of the 8 random projections is a highly unlikely event, thus probably resulting from anomalous temporal dynamics in Internet traffic. This can be applied to any of the four either univariate $\log_2 S_{m,m}(2^j)$ or multivariate $\{\log_2 |\lambda_m(2^j)\}|$ functions.

Most anomaly detection procedures involve an *energy* criterion: Anomalies must be *large* enough to induce a significant deviation in some statistics [25]. The univariate functions $\log_2 S_{m,m}$ fall in that category and can only detect large enough anomalies for either of the 4 times series $(X^{Out}_{Pkt}, X^{Out}_{Byte}, X^{In}_{Pkt}, X^{In}_{Byte})$. While the largest eigenvalue $\log_2 |\lambda_4|$ also belongs in that category, the three smaller

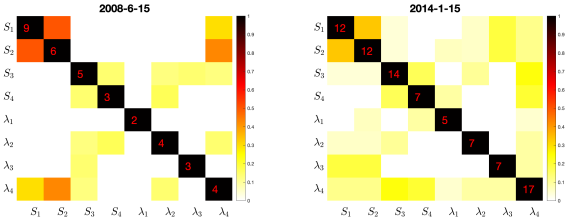

Figure 5. **Anomaly detection: Jaccard indices** of anomalies jointly detected by different functions ($L_m = \log_2 \lambda_m$, $S_m = \log_2 S_{m,m}$) for days with few (left, 2008/06/15) and many (right, 2014/01/15) anomalies.

eigenvalue functions are likely to detect anomalies with much smaller *energy*. As examples: on 2007/01/15, the smallest eigenvalue function $\log_2 |\lambda_1|$ permitted the detection of an anomaly that involved only 172 IP packets whereas most anomalies detected by the $\log_2 S_{m,m}$ involve several thousands of packets ; on 2007/02/15, $\log_2 |\lambda_2|$ permitted the detection of an anomaly that lasts only 149s whereas most anomalies detected by $\log_2 S_{m,m}(2^j)$ last the entire observation period (15min). Fig. 5 reports the Jaccard index (ratio of the sizes of intersection to union of two lists) for the joint detection of anomalies by two functions $\log_2 |\lambda_m|$ or $\log_2 S_{m,m}$. It shows that i) anomalies detected on packet times series are also seen on the byte time series in the same direction ; ii) anomalies detected by $\log_2 |\lambda_4|$ (largest eigenvalues) corresponds to the those seen on $\log_2 S_{m,m}$ ; conversely that iii) detections by $\log_2 |\lambda_1|, \log_2 |\lambda_2|$ (small eigenvalues) correspond to low volume anomalies, missed by univariate analysis $\log_2 S_{m,m}$ ; and, finally, that iv) the small eigenvalues $\log_2 |\lambda_1|$ and $\log_2 |\lambda_2|$ are sensitive to different types of low-volume anomalies.

## VI. Conclusions and perspectives

This work showed that random projections can be extended to (eigenvalue-based) multivariate selfsimilarity analysis to produce robust-to-anomaly statistical characterizations of multivariate temporal dynamics in Internet traffic. It further showed that scalefree dynamics in two independent ranges of fine and coarse scales, the biscaling regime, is deeply embedded in the multivariate statistics of Internet traffic, instead of being merely a univariate effect.

Because multivariate selfsimilarity analysis digs into the details of the joint multivariate structure of Internet time series, it allows for the detection of low-volume anomalies that are usually missed by univariate detection procedures – a key concern in cybersecurity. Multivariate selfsimilarity analysis tools will be made publicly available. The systematic longitudinal analysis across 17 years is under current investigation.

## References

[1] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic," *IEEE Trans. on Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[2] V. Paxson and S. Floyd, "Wide area traffic: The failure of poisson modeling," *IEEE Trans. on Networking*, vol. 4, no. 3, pp. 209–223, 1995.

[3] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of ethernet lan traffic at the source level," *IEEE Trans. on Networking*, vol. 5, no. 1, pp. 71–86, 1997.

[4] R.H. Riedi, M.S. Crouse V.J. Ribeiro, and R.G. Baraniuk, "A Multifractal Wavelet Model with Application to Network Traffic," *IEEE Trans. on Info. Theory, Special Issue "Multiscale Statistical Signal Analysis and its Applications"*, vol. 45, no. 3, pp. 992–1018, 1999.

[5] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "Multiscale network traffic analysis, modeling, and inference using wavelets, multifractals, and cascades," *IEEE Signal Process. Mag.*, vol. 3, no. 19, pp. 28–46, 2002.

[6] R. Fontugne, P. Abry, K. Fukuda, D. Veitch, K. Cho, P. Borgnat, and H. Wendt, "Scaling in internet traffic: a 14 year and 3 day longitudinal study, with multiscale analyses and random projections," *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 2152–2165, 2017.

[7] P. Borgnat, G. Dewaele, K. Fukuda, Abry P., and K. Cho, "Seven Years and One Day: Sketching the Evolution of Internet Traffic," *Proc. IEEE INFOCOM'09*, pp. 711–719, 2009.

[8] N. Hohn, D. Veitch, and P. Abry, "Cluster processes, a natural language for network traffic," *IEEE Trans. Signal Process., Special Issue "Signal Process. in Networking"*, vol. 8, no. 51, pp. 2229–2244, 2003.

[9] S. Muthukrishnan, "Data streams: Algorithms and applications," in *ACM SIAM Symp. Discrete Algorithms (SODA)*, 2003, p. 413.

[10] Balachander K., Subhabrata S., Yin Z., and Yan C., "Sketch-based change detection: Methods, evaluation, and applications," in *Proc. ACM SIGCOMM Conf. on Internet Measurement (IMC)*, 2003, pp. 234–247.

[11] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho, "Extracting hidden anomalies using sketch and non gaussian multiresolution statistical detection procedure," in *Proc. ACM SIGCOMM Workshop on Large Scale Attack Defense*, 2007, pp. 145–152.

[12] J. Frecon, R. Fontugne, G. Didier, N. Pustelnik, K. Fukuda, and P. Abry, "Non-linear regression for bivariate self-similarity identification: Application to anomaly detection in internet traffic based on a joint scaling analysis of packet and byte counts," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4184–4188.

[13] D. Veitch, N. Hohn, and P. Abry, "Multifractality in TCP/IP Traffic: the Case Against," *Computer Networks, Special Issue "Long-Range Dependent Traffic"*, vol. 48, no. 3, pp. 293–313, 2005.

[14] R. Fontugne, P. Abry, K. Fukuda, P. Borgnat, J. Mazel, H. Wendt, and D. Veitch, "Random projection and multiscale wavelet leader based anomaly detection and address identification in internet traffic," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2015.

[15] J. D. Mason and Y. Xiao, "Sample path properties of operator-self-similar Gaussian random fields," *Theory of Probability & Its Applications*, vol. 46, no. 1, pp. 58–78, 2002.

[16] G. Didier and V. Pipiras, "Integral representations and properties of operator fractional Brownian motions," *Bernoulli*, vol. 17, no. 1, pp. 1–33, 2011.

[17] P.-O. Amblard and J.-F. Coeurjolly, "Identification of the multivariate fractional Brownian motion," *IEEE Trans. Signal Proces.*, vol. 59, no. 11, pp. 5152–5168, 2011.

[18] P. Abry and G. Didier, "Wavelet eigenvalue regression for $n$-variate operator fractional Brownian motion," *J. Multivar. Anal.*, vol. 168, pp. 75–104, November 2018.

[19] Ch.-G. Lucas, G. Didier, H. Wendt, and P. Abry, "Multivariate selfsimilarity: Multiscale eigen-structures for selfsimilarity parameter estimation," *IEEE Trans. Signal Proc.*, 2024.

[20] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 1998.

[21] S. Combrexelle, H. Wendt, G. Didier, and P. Abry, "Multivariate scale-free dynamics: Testing fractal connectivity," in *IEEE Int. Conf. Acoust., Speech, and Signal Proces. (ICASSP)*, New Orleans, USA, March 2017.

[22] P. Abry and G. Didier, "Wavelet estimation for operator fractional Brownian motion," *Bernoulli*, vol. 24, no. 2, pp. 895–928, 2018.

[23] K. Cho, K. Mitsuya, and A. Kato, "Traffic data repository at the WIDE project," in *USENIX 2000 Annual Technical Conference: FREENIX Track*, June 2000, pp. 263–270.

[24] M. Thorup and Y. Zhang, "Tabulation Based 4-Universal Hashing with Applications to Second Moment Estimation," in *ACM SIAM Symp. Discrete Algorithms (SODA)*, 2004, pp. 615–624.

[25] R. Fontugne, P. Borgnat, P. Abry, and K Fukuda, "MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking," *Proc. ACM Co-NEXT*, 2010.