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

Jordan Frecon[1], Romain Fontugne[2,3], Gustavo Didier[4],
Nelly Pustelnik[1], Kensuke Fukuda[2] and Patrice Abry[1,2,3]

[1] CNRS, Physics Department, ENS Lyon, France, firstname.lastname@ens-lyon.fr
[2] The National Institute of Informatics, Tokyo, Japan, firstname@nii.ac.jp
[3] Japanese French Laboratory for Informatics, Tokyo, Japan,

[4] Mathematics Department, Tulane University, New Orleans, USA, gdidier@tulane.edu

ICASSP, Shanghai, 25th March 2016

Anomaly detection in Internet traffic

- **IP packet:**

HEADER (20 bytes)	**PAYLOAD** (variable size)
– source address – destination address ...	– data

- **Network monitoring** $\leftarrow$ statistical modeling of Internet traffic

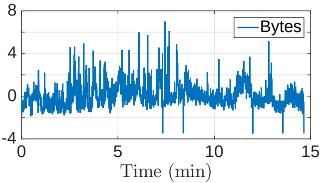
Anomaly detection in Internet traffic

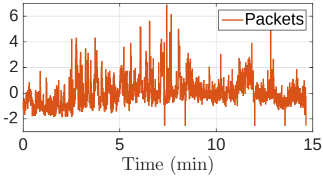
- **IP packet:**

HEADER (20 bytes)	**PAYLOAD** (variable size)
– source address – destination address ...	– data

- **Network monitoring** $\leftarrow$ statistical modeling of **Internet traffic**

1. **Aggregated time serie:** $\begin{cases} \#\text{Packets received / time bins ?} \\ \#\text{Bytes received / time bins ?} \end{cases}$





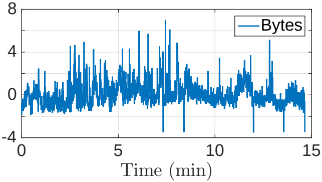
Anomaly detection in Internet traffic

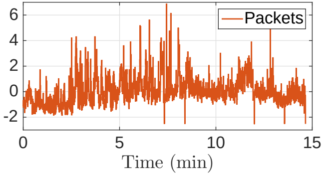
HEADER (20 bytes)	**PAYLOAD** (variable size)
– source address – destination address ...	– data

- **IP packet:** header and payload as shown above.

- **Network monitoring** $\leftarrow$ **statistical modeling** of Internet traffic

  1. **Aggregated time serie:** $\begin{cases} \#\text{Packets received / time bins ?} \\ \#\text{Bytes received / time bins ?} \end{cases}$
  2. **Statistical modeling:** self-similar process





Statistical modeling of Internet traffic

- **Self-similarity:** time scales from *ms* (ethernet dynamic) to hours (human dynamic) [Abry et al, 2002]

- **Univ. fractional Gaussian noise** $\rightarrow \underbrace{\text{Hurst parameter } H \in [0,1]}_{\text{quantify relation accross scales}}$

Past work [Borgnat et al, 2009]

2 univariate analyses

$$\underbrace{\begin{pmatrix} Y^{\text{Byt}} \\ Y^{\text{Pkt}} \end{pmatrix} \begin{matrix} \leftarrow H^{\text{Byt}} \\ \leftarrow H^{\text{Pkt}} \end{matrix}}_{\text{2 independent fGn}}$$

Anomaly if $|H^{\text{Byt}} - H^{\text{Pkt}}| \gg 0$

Contribution

1 bivariate analysis

$$\underbrace{\begin{pmatrix} Y^{\text{Byt}} \\ Y^{\text{Pkt}} \end{pmatrix} = W \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \begin{matrix} \leftarrow H_1 \\ \leftarrow H_2 \end{matrix}}_{\text{mixture of 2 correlated fGn}}$$

Anomaly if $|H_1 - H_2| \gg 0$

Operator Fractional Gaussian Noise

- **2 correlated fGn** $(X_1, X_2)^\top$**:**

$$\Sigma_{X_p,X_{p'}}(s) = \frac{\sigma_{x_p}\sigma_{x_{p'}}\rho_{x_p,x_{p'}}}{2}\left(|s-1|^{H_p+H_{p'}} - 2|s|^{H_p+H_{p'}} + |s+1|^{H_p+H_{p'}}\right)$$

$$\Sigma_X(0) = \begin{pmatrix} \sigma_{x_1}^2 & \sigma_{x_1}\sigma_{x_2}\rho_x \\ \sigma_{x_1}\sigma_{x_2}\rho_x & \sigma_{x_2}^2 \end{pmatrix}$$

- **Condition of existence:**

$$g(H_1, H_2, \rho_x) \equiv \Gamma(2H_1+1)\Gamma(2H_2+1)\sin(\pi H_1)\sin(\pi H_2) - \rho_x^2\Gamma(H_1+H_2+1)^2\sin^2(\pi(H_1+H_2)/2) > 0.$$

- **Mixing:** $\begin{pmatrix} Y^{\mathrm{Byt}} \\ Y^{\mathrm{Pkt}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1+\gamma^2}} & \frac{\beta}{\sqrt{1+\beta^2}} \\ \frac{-\gamma}{\sqrt{1+\gamma^2}} & \frac{1}{\sqrt{1+\beta^2}} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$

Objective: estimate $\Theta = (H_1, H_2, \rho_x, \sigma_{x_1}, \sigma_{x_2}, \beta, \gamma)$

Wavelet spectrum

- **Wavelet coefficients:** $D_{\mathrm{y_p}}(j,k) = \int_{\mathbb{R}} \psi_{j,k}(t) Y_p(t) \mathrm{d}t$

$$\text{where} \quad \psi_{j,k}(t) = 2^{-j(1/2-\mu)} \underbrace{\psi_0(2^{-j/2}t - k)}_{\text{dilation and translation of } \psi_0}$$

and Fractional integration parameter μ (default: $\mu = 0$)

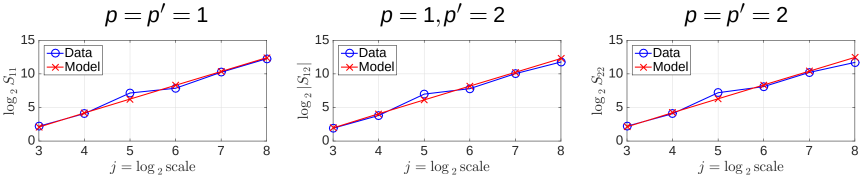
- **Wavelet spectrum:** $(E_{p,p'}(\Theta))_j = \mathbb{E} D_{y_p}(j,k) D_{y_{p'}}(j,k)^*$

$$= + \alpha_{p,p'}^{(1,1)}(\beta,\gamma)\, \sigma_{\mathrm{x_1}}^2 \eta_{j,H_1} 2^{j(2H_1+1+2\mu)}$$
$$+ \alpha_{p,p'}^{(1,2)}(\beta,\gamma)\, \rho_{\mathrm{x}} \sigma_{\mathrm{x_1}} \sigma_{\mathrm{x_2}} \eta_{j,\frac{H_1+H_2}{2}} 2^{j(H_1+H_2+1+2\mu)}$$
$$+ \alpha_{p,p'}^{(2,2)}(\beta,\gamma)\, \sigma_{\mathrm{x_2}}^2 \eta_{j,H_2} 2^{j(2H_2+1+2\mu)}$$

$\rightarrow$ Empirical estimate: $(S_{p,p'})_j = \frac{2^j}{N} \sum_{k=1}^{N/2^j} D_{y_p}(j,k) D_{y_{p'}}(j,k)^*$

Estimate Θ such as $(E_{p,p'}(\Theta))_j$ fits $(S_{p,p'})_j$ jointly for all j

Non-linear Wavelet Regression Problem



$$\widehat{\Theta} = \arg\min_{\Theta \in \mathcal{Q}_0} \underbrace{\sum_{p,p'=1}^{2} \sum_{j=j_1}^{j_2} \Big(\log_2 \big| \underbrace{(S_{p,p'})_j}_{\text{Data}} \big| - \log_2 \big| \underbrace{(E_{p,p'}(\Theta))_j}_{\text{Model}} \big| \Big)^2}_{C(\theta)}$$

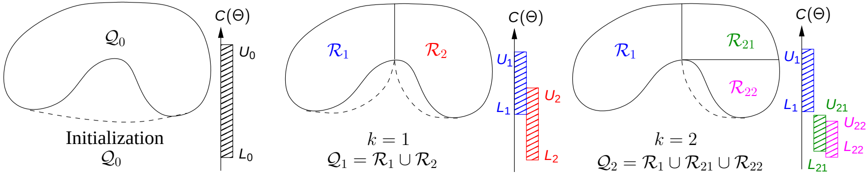
Non-convex optimization problem $\longrightarrow$ Branch & Bound algorithm

[Hansen, 1980]

Branch & Bound Algorithm

$$\widehat{\Theta} = \arg\min_{\Theta \in \mathcal{Q}_0} C(\Theta) \quad , \text{ non-convex criterion C and search space } \mathcal{Q}_0$$

Step 0   Step 1   Step 2 ...



- Bounding over the convex relaxation of $\mathcal{Q}_0$

- Split $\mathcal{Q}_0$ into $\mathcal{R}_1$ & $\mathcal{R}_2$
- Bounding over the convex relaxation of $\mathcal{R}_1$ & $\mathcal{R}_2$

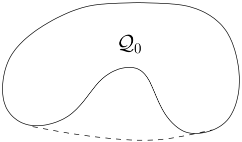
- Split $\mathcal{R}_2$ into $\mathcal{R}_{21}$ & $\mathcal{R}_{22}$
- Bounding over the convex relaxation of $\mathcal{R}_{21}$ & $\mathcal{R}_{22}$
- Discard $\mathcal{R}_1$

Branch & Bound Algorithm - Initalization

$$\mathcal{Q}_0 = \Big\{ \Theta = (H_1, H_2, \rho_x, \sigma_{x_1}, \sigma_{x_2}, \beta, \gamma) \in \mathbb{R}^7 \,|\, \Theta \in [0,1]^3 \times [0, \sigma_{\max}]^2 \times [-1,1]^2, \underbrace{g(H_1, H_2, \rho_x) > 0}_{\text{non convex}}, H_1 \leq H_2 \Big\}.$$

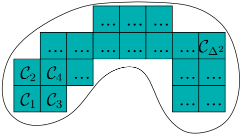
$g(h_1, h_2, \rho_x) > 0 \iff (X_1, X_2)$ correctly defined $g(h_1, h_2, \rho_x) > 0$ must not be relaxed !

Proposed solution:



Initialization
$\mathcal{Q}_0$

Inner convex relaxation $\Longrightarrow$

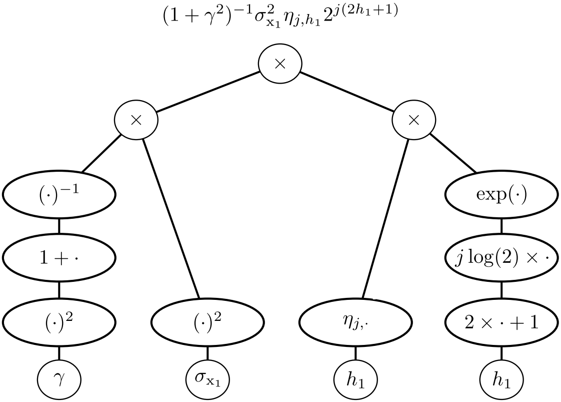


Initialization
$\mathcal{S}_0 = \cup_{i=1}^{\Delta^2} \mathcal{C}_i$

Branch & Bound Algorithm - Bounding

- **Bounding:** Interval arithmetic [Moore, 1966]

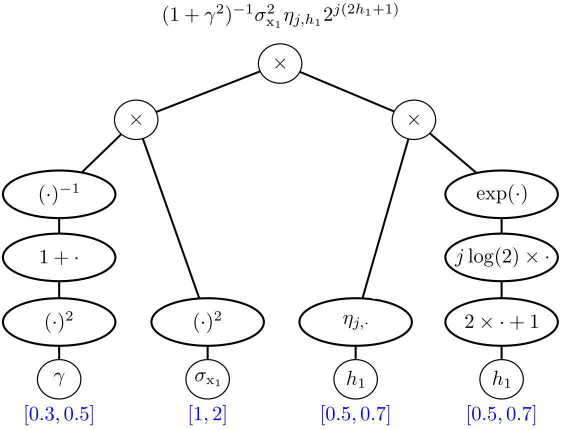
For example, for $j = 2$:



Branch & Bound Algorithm - Bounding

- **Bounding:** Interval arithmetic [Moore, 1966]

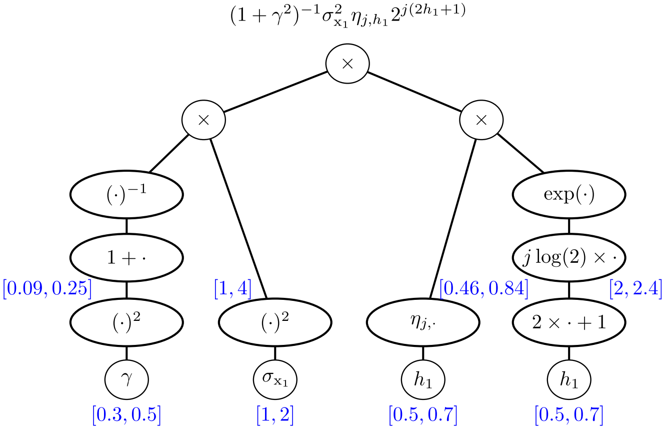
For example, for $j = 2$:



Branch & Bound Algorithm - Bounding

- **Bounding:** Interval arithmetic [Moore, 1966]

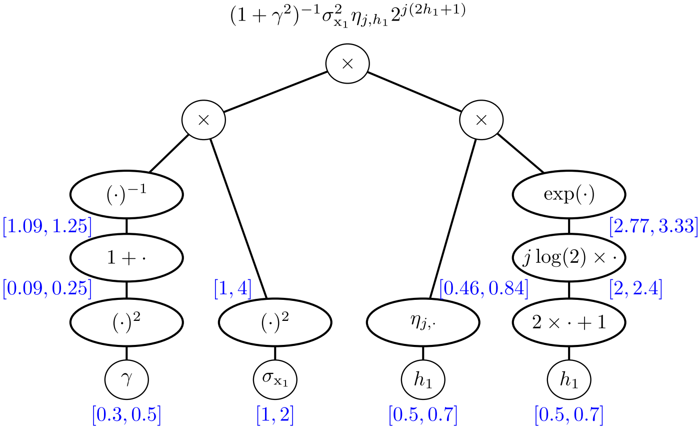
For example, for $j = 2$:



Branch & Bound Algorithm - Bounding

- **Bounding:** Interval arithmetic [Moore, 1966]

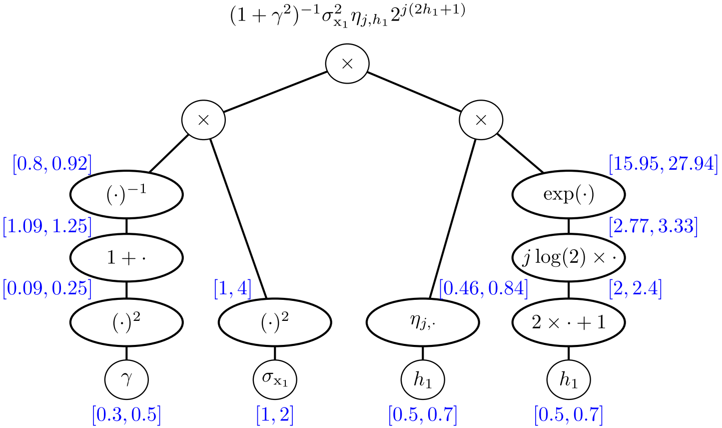
For example, for $j = 2$:



Branch & Bound Algorithm - Bounding

- **Bounding:** Interval arithmetic [Moore, 1966]

For example, for $j = 2$:

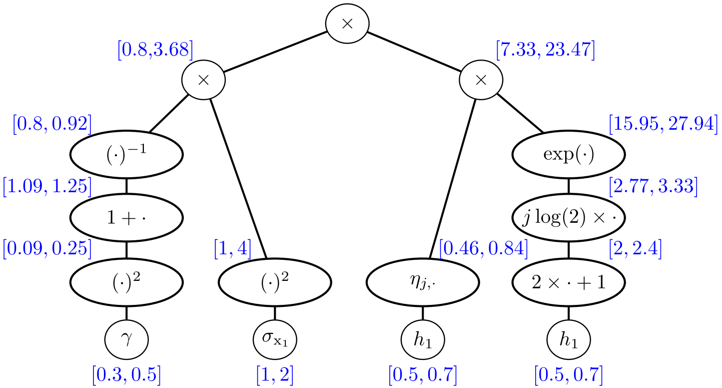


Branch & Bound Algorithm - Bounding

- **Bounding:** Interval arithmetic [Moore, 1966]

For example, for $j = 2$:

$(1+\gamma^2)^{-1}\sigma_{x_1}^2\eta_{j,h_1}2^{j(2h_1+1)}$

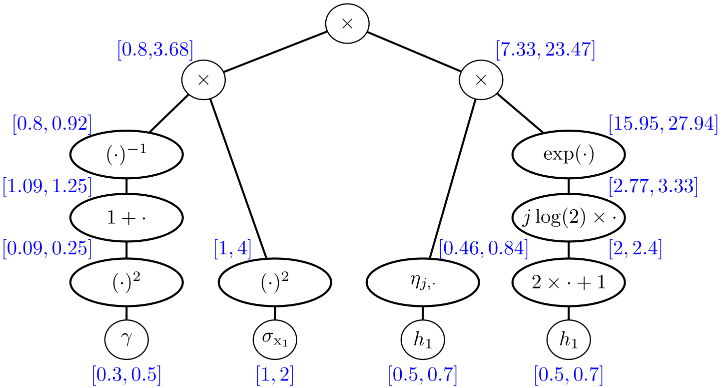


Branch & Bound Algorithm - Bounding

- **Bounding:** Interval arithmetic [Moore, 1966]

For example, for $j = 2$:

$$5.86 \leq (1+\gamma^2)^{-1}\sigma_{x_1}^2 \eta_{j,h_1} 2^{j(2h_1+1)} \leq 86.37$$

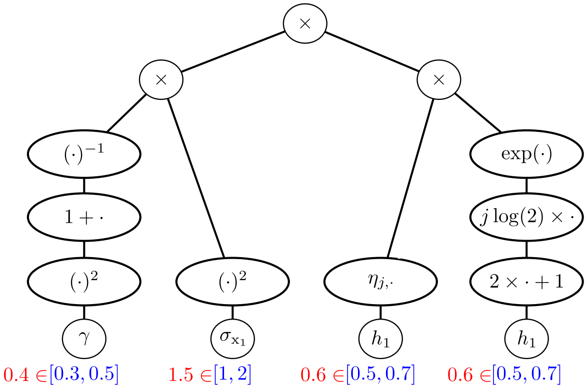


Branch & Bound Algorithm - Bounding

- **Bounding:** Interval arithmetic [Moore, 1966]

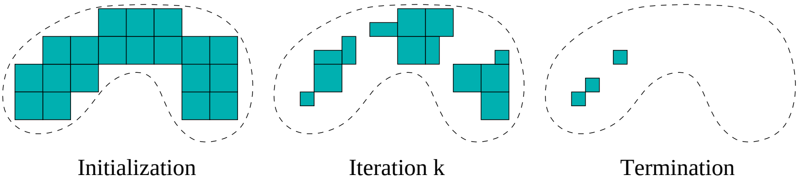
For example, for $j = 2$:

$$5.86 \leq (1+\gamma^2)^{-1}\sigma_{x_1}^2 \eta_{j,h_1} 2^{j(2h_1+1)} \cancel{\leq 86.37} \leq 26.50$$



Branch & Bound Algorithm - Termination

- **Evolution of the search space:**



- **Termination:** size of all regions $<$ size limit

- **Solution:** $\widehat{\Theta} = \mathrm{mid}$ (region with lowest upper bound)

Estimation Performance on Synthetic Data

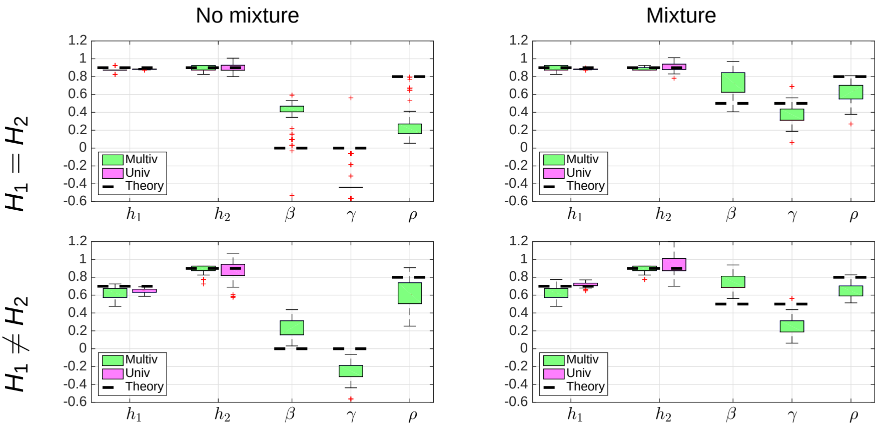


Figure: Four configurations potentially matching Internet Traffic data (only $N = 3600$ samples long).

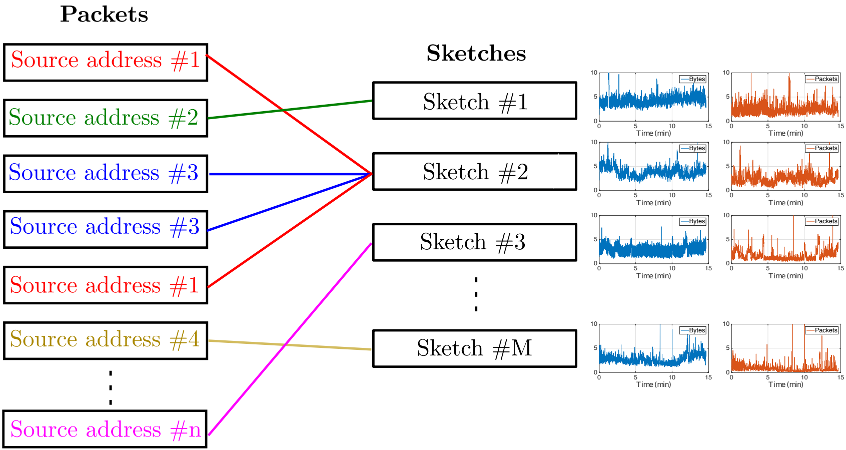
MAWI Database

- **MAWI repository:**
  - WIDE backbone network (Japan $\longleftrightarrow$ USA)
  - Daily collection of internet traces from 14:00 to 14:15 (JST)
  - Each trace $\sim$ 100 to 150 million IP packets
  - Packet 5-tuple and timestamps anonymized and publicly available

- **Anomaly detection:** aggregated Pkt and Byt counts
  1. How to construct a self-reference for normal traffic ?
  2. How to adjust to specificities of real-world data ?

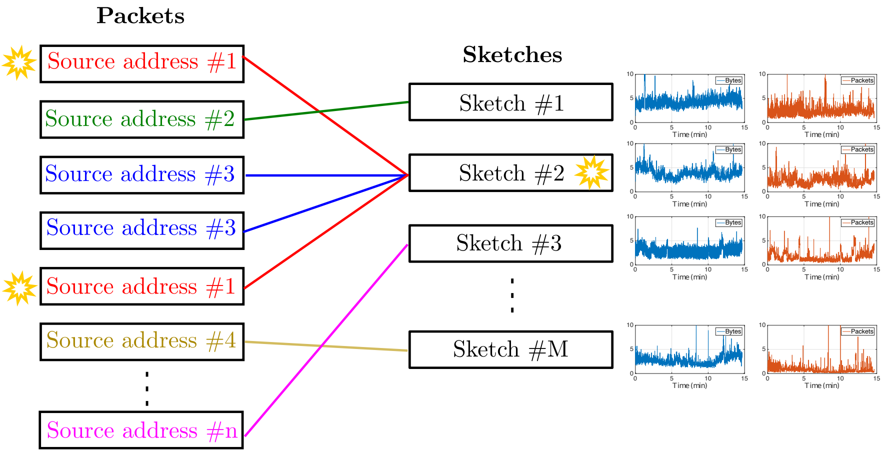
Reference for normal traffic

- **Random projections:** same source address → same sketch



Reference for normal traffic

- **Random projections:** same source address $\rightarrow$ same sketch



Role of fractional integration parameter μ

Biv.fGn requires $H_1, H_2 \in [0,1]$

- **Wavelet coefficients:** $D_{\mathrm{y_p}}(j,k) = \int_{\mathbb{R}} \psi_{j,k}(t) Y_p(t) \mathrm{d}t$

$$\text{where} \quad \psi_{j,k}(t) = 2^{-j(1/2-\mu_W)} \underbrace{\psi_0(2^{-j/2}t-k)}_{\text{dilation and translation of } \psi_0}$$

- **Wavelet spectrum:** $\mathbb{E} D_{y_p}(j,k) D_{y_{p'}}(j,k)^*$

$$\begin{aligned}(E_{p,p'}(\Theta))_j = &+ \alpha_{p,p'}^{(1,1)}(\beta,\gamma)\, \sigma_{\mathrm{x}_1}^2 \eta_{j,H_1} 2^{j(2H_1+1+2\mu_B)} \\ &+ \alpha_{p,p'}^{(1,2)}(\beta,\gamma)\, \rho_{\mathrm{x}} \sigma_{\mathrm{x}_1} \sigma_{\mathrm{x}_2} \eta_{j,\frac{H_1+H_2}{2}} 2^{j(H_1+H_2+1+2\mu_B)} \\ &+ \alpha_{p,p'}^{(2,2)}(\beta,\gamma)\, \sigma_{\mathrm{x}_2}^2 \eta_{j,H_2} 2^{j(2H_2+1+2\mu_B)}\end{aligned}$$

H_1 and H_2 shifted by $\mu_B - \mu_W$ into $[0,1]$

Application to MAWI database

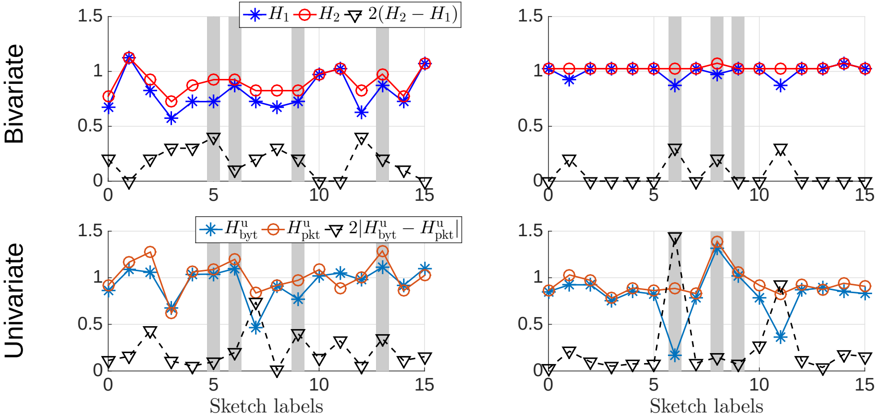
2008 data

(100% Trinocular anomalies detected)

(96% of Deny-of-Service attacks detected)

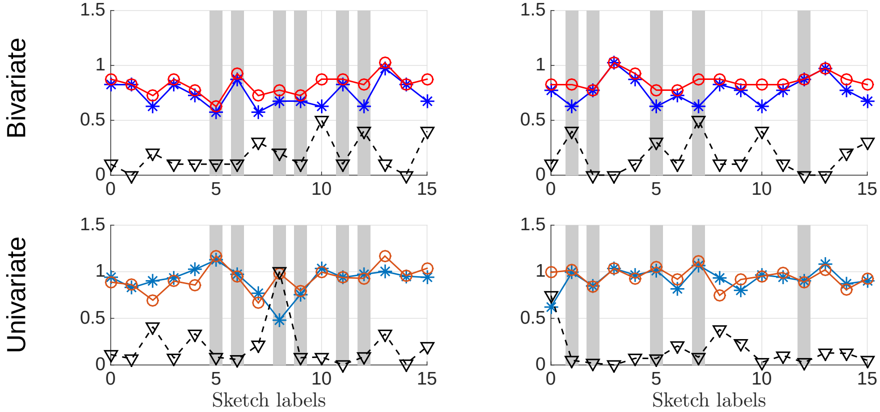
2009 data

(100% Trinocular anomalies detected)



Application to MAWI database

2014 data	2015 data
(100% Trinocular anomalies detected)	(100% Trinocular anomalies detected)
(100% of Heavy Hitter anomalies detected)	



Conclusion

1. **Joint scaling analysis of Packets and Bytes counts**

   $\longrightarrow$ Biv.OfGn

   $\longrightarrow$ Non-linear wavelet regression problem

2. **Branch & Bound algorithm**

   $\longrightarrow$ Toolbox available soon (http://perso.ens-lyon.fr/jordan.frecon/)

3. **Fractional integration parameter**

   $\longrightarrow$ permits to adjust to real world data

4. **Application to MAWI database**

   $\longrightarrow$ systematic detection of some anomalies