From supervised to unsupervised anomaly detection in Internet Traffic

Philippe Owezarski
LAAS-CNRS, Toulouse, France
owe@laas.fr

Outline

- Anomalies: definition and problematics
- Traffic characteristics - Anomaly detection issues
- Supervised anomaly detection
  - A detailed example
- Unsupervised anomaly detection
  - A detailed example
- Conclusion

Motivation

- Traffic anomalies (on a link)
  - One or several occurrences that change the way traffic is flowing in the network

- Consequences
  - Performance decrease
  - QoS degradation

Existing work

- Several projects on traffic anomalies detection arised in the past
  - They rely in general on simple statistics on traffic characteristics
    - But they lack by a bad knowledge on traffic characteristics

    → Limited efficiency

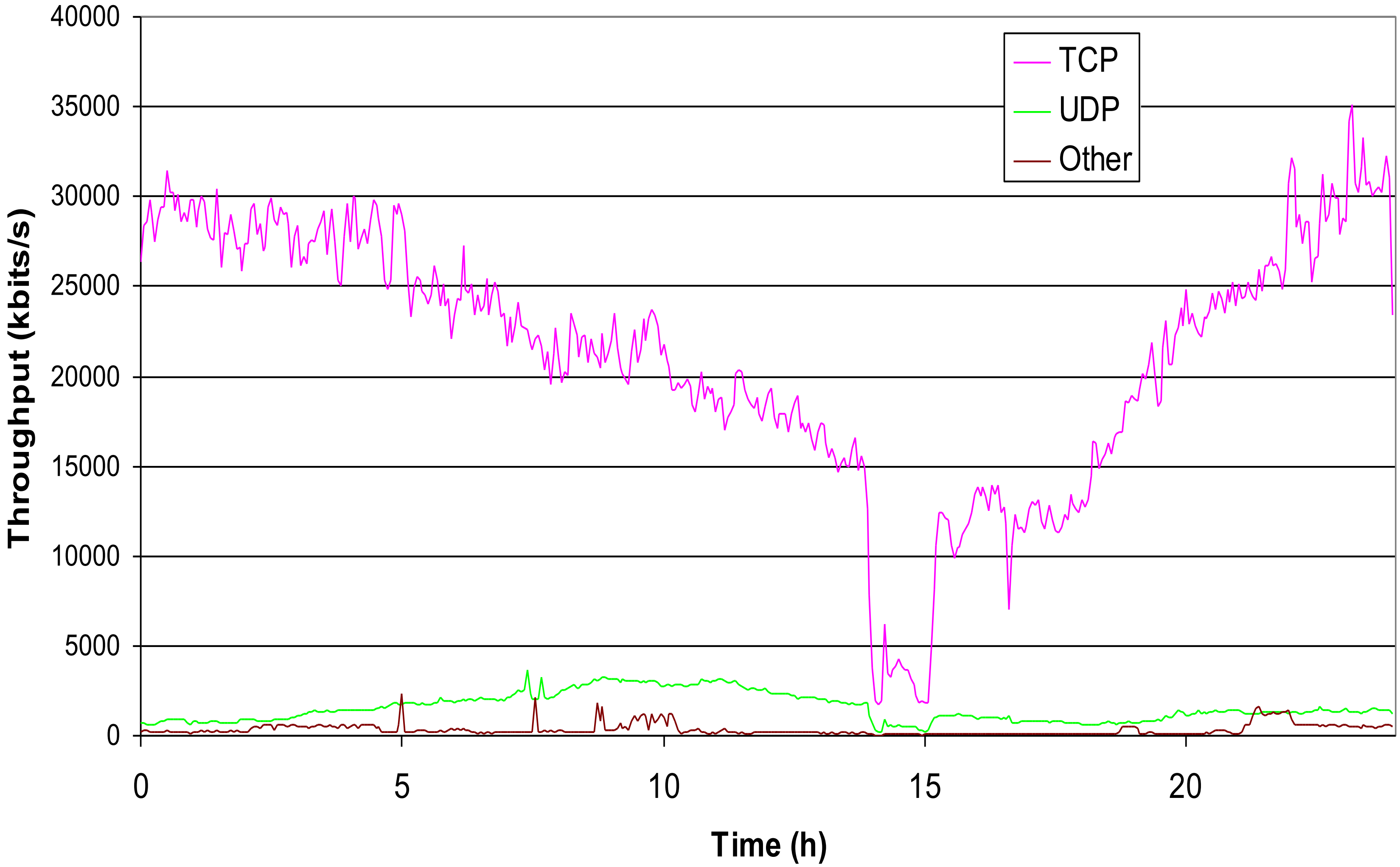
Known traffic characteristics

- Non Gaussian, non Poisson statistics
- Long Range Dependence (LRD), Strong correlations
- Traffic can look different according to the granularity of observation

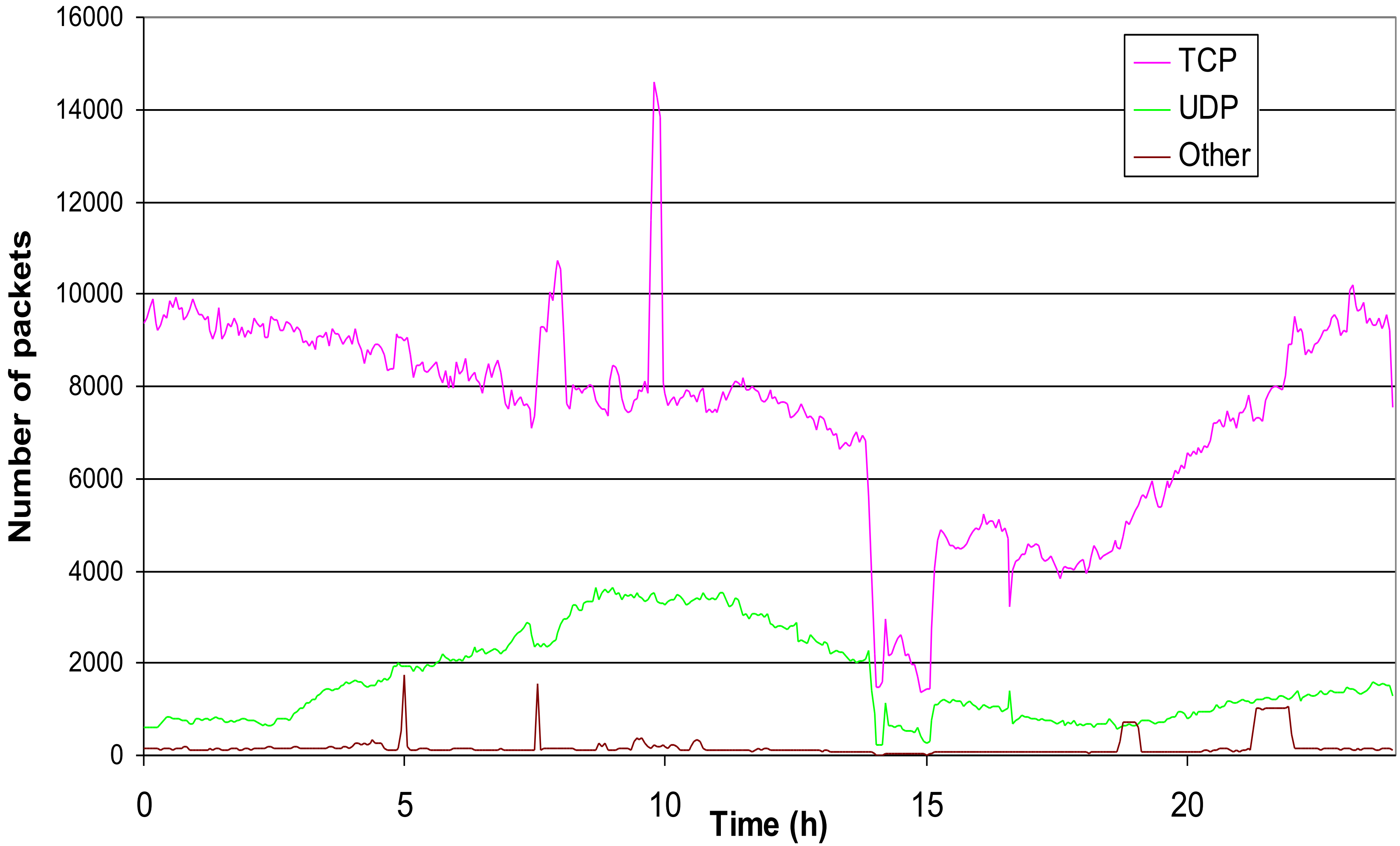
- And…

  …**Traffic is highly variable !**

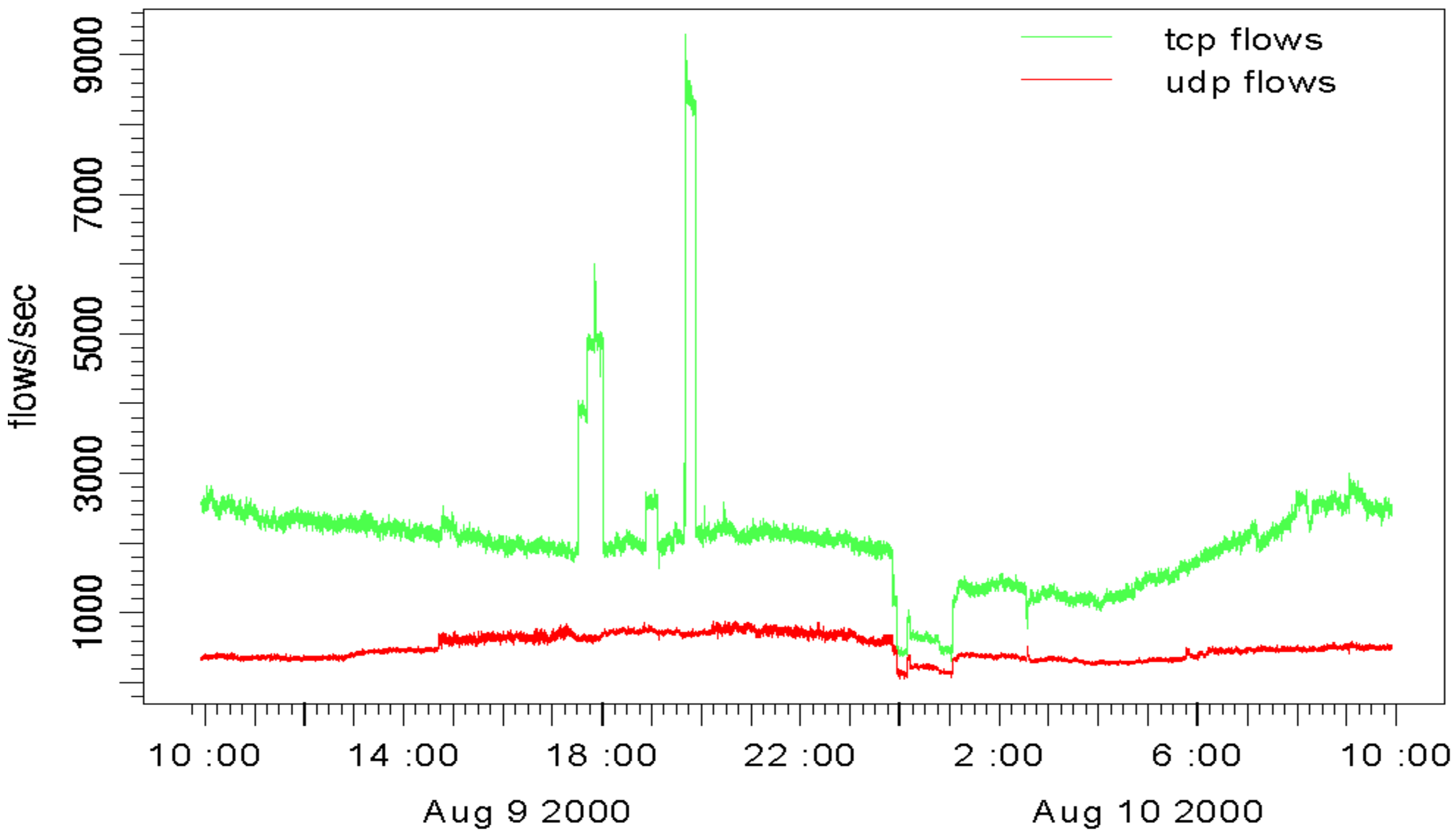
Link Utilization: bandwidth



Link utilization: packets



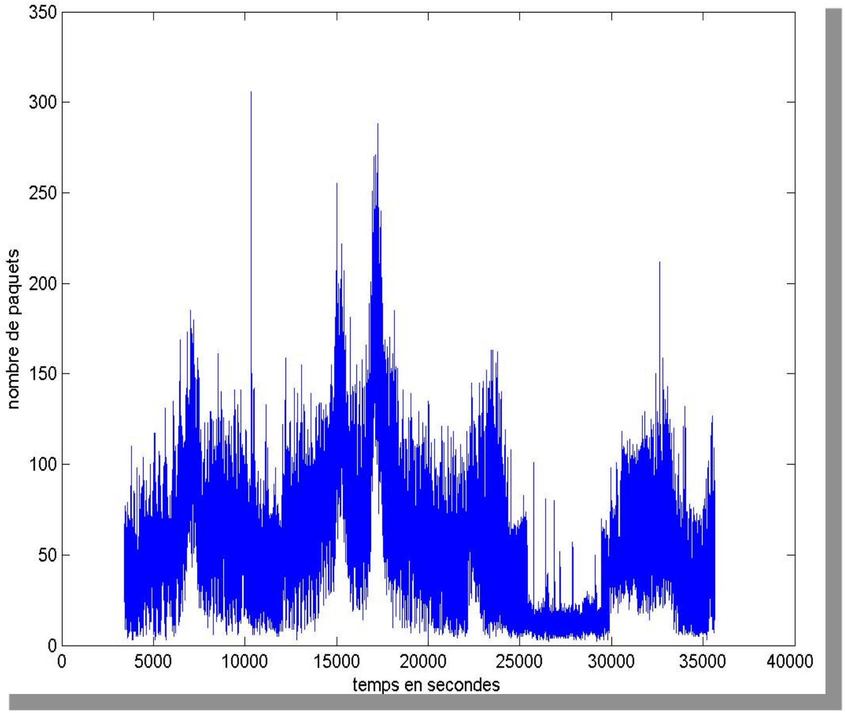
Link utilization: instantaneous flows



Profile based IDS issues

Traffic profiles in IDS do not consider such variability

False positive rate is high
→ Impossible to fix reliable thresholds



Temporal evolution of the number of TCP/SYN packets

A traffic profile cannot be based only on some averages (non Gaussian)
→ High level statistics are required

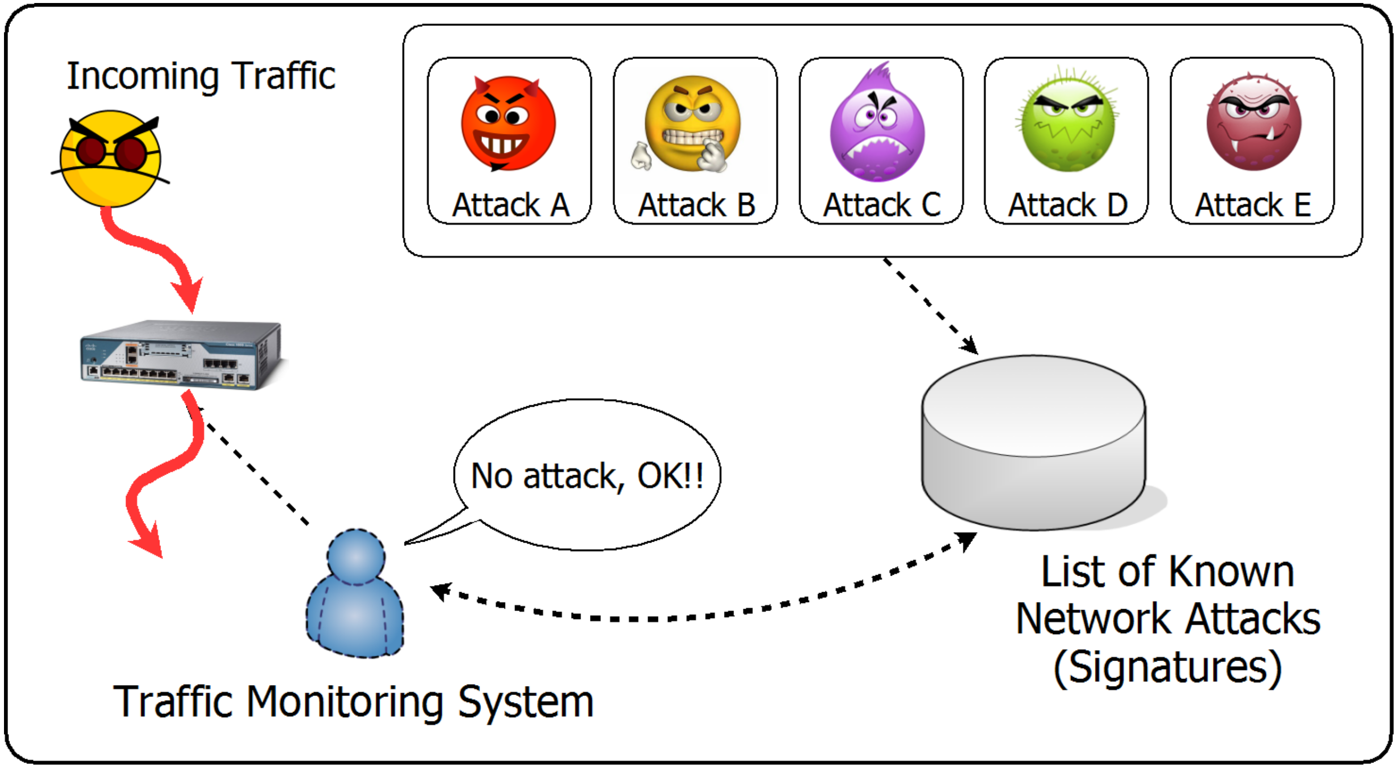
From supervised to unsupervised anomaly detection

Existing AD systems

- Current Anomaly Detection (AD) approaches are based on an "acquired knowledge" perspective
  - Signature based
  - Supervised approaches

Signature-based AD

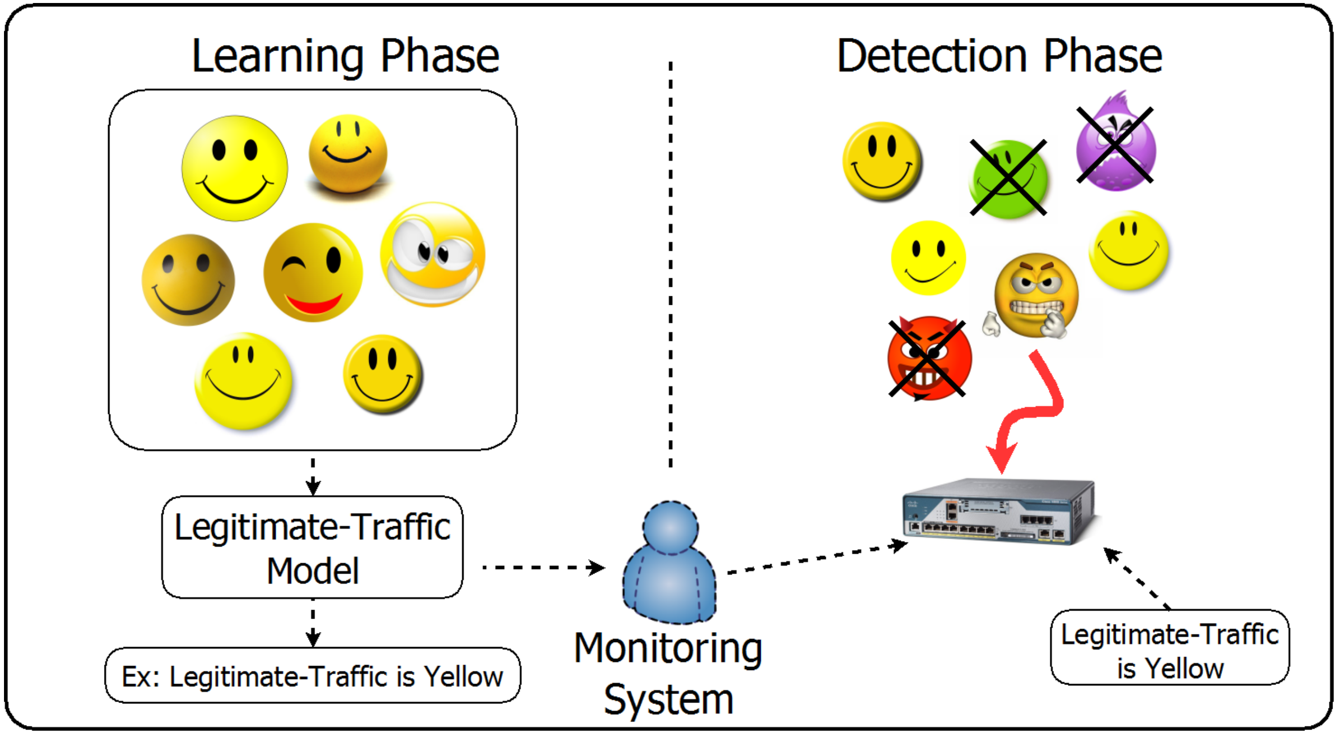
- Detect WHAT I ALREADY KNOW



(+) Highly effective to detect what it is programmed to alert on
(-) Cannot defend the network against unknown attacks
(-) Signatures are expensive to produce: human manual inspection

Supervised-learning-based AD

- Detect what is different from WHAT I KNOW



(+) It can detect new anomalies out-of the baseline
(-) Requires training on anomaly-free traffic
(-) Robust and adaptive models are difficult to conceive

Detailed example of supervidsed AD

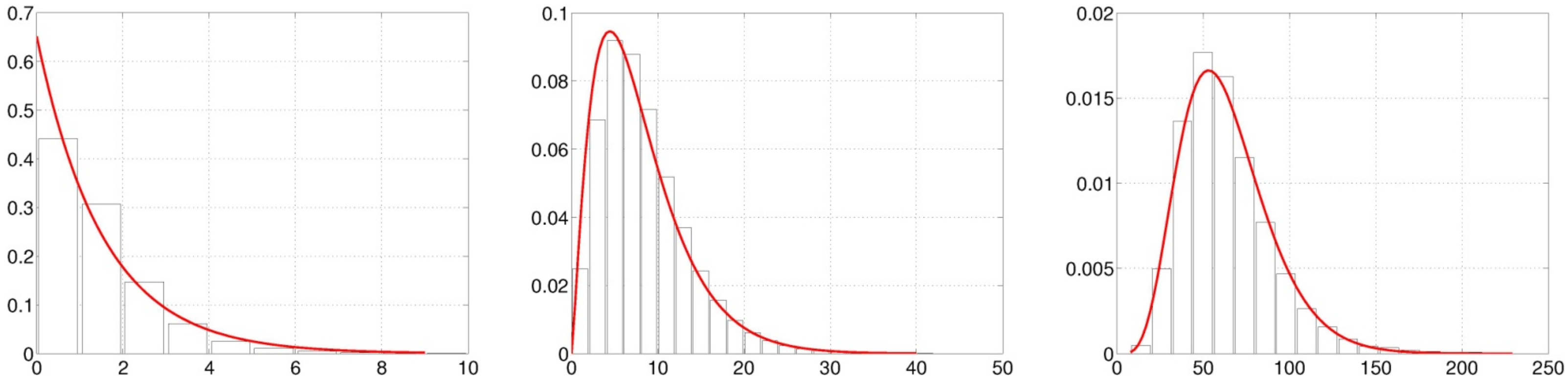
Internet Traffic

What model for a non Gaussian and long memory process ?

The Gamma-Farima model based AD approach

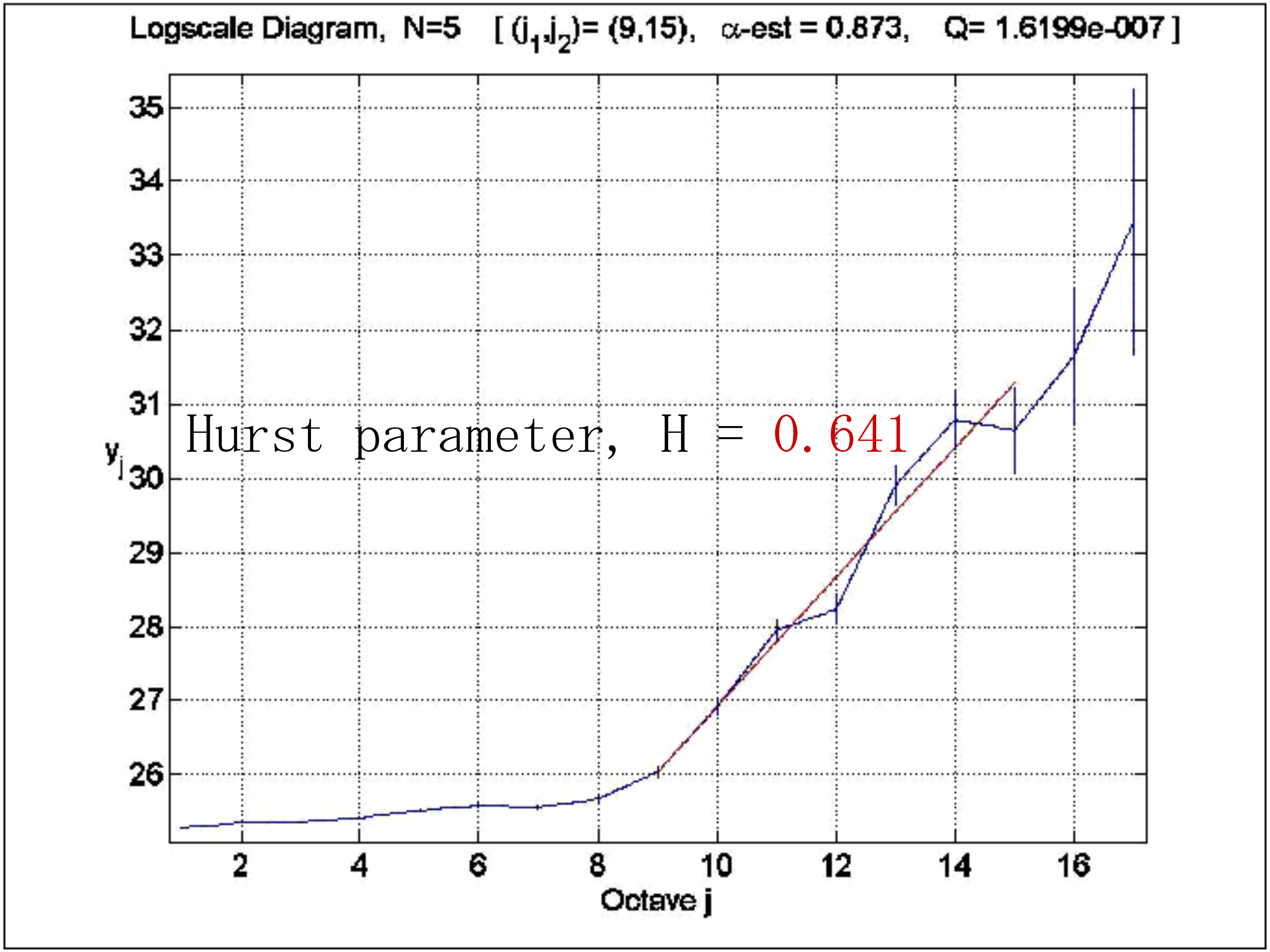
Marginal laws

- Distributions of empirical probabilities LBL-TCP-3

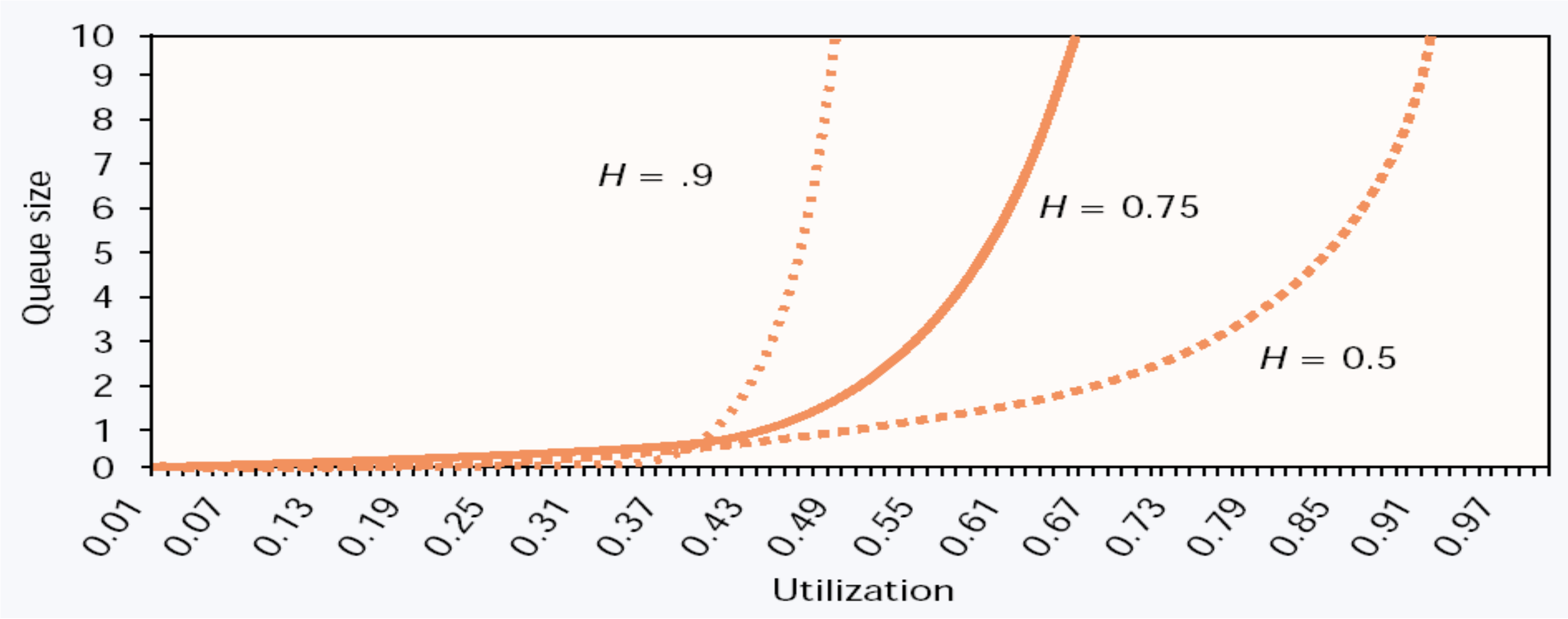


- Poisson model? Exponential law? Gaussian?
- What aggregation level to select?

Traffic Correlation (SRD and LRD)



Example : LRD and network performance



relation between LRD , network usage and queue sizes in routers

Non Gaussian with LRD model

Joint modelling of 1st and 2nd orders statistics

- Packet aggregated count process: $X_\Delta(k)$

$$X_\Delta(k) = \#\text{pkt during } [k\Delta, (k+1)\Delta]$$

or

- Bytes aggregated count process: $W_\Delta(k)$

$$W_\Delta(k) = \#\text{bytes during } [k\Delta, (k+1)\Delta]$$

- 1st. PDFs of marginals as gamma laws

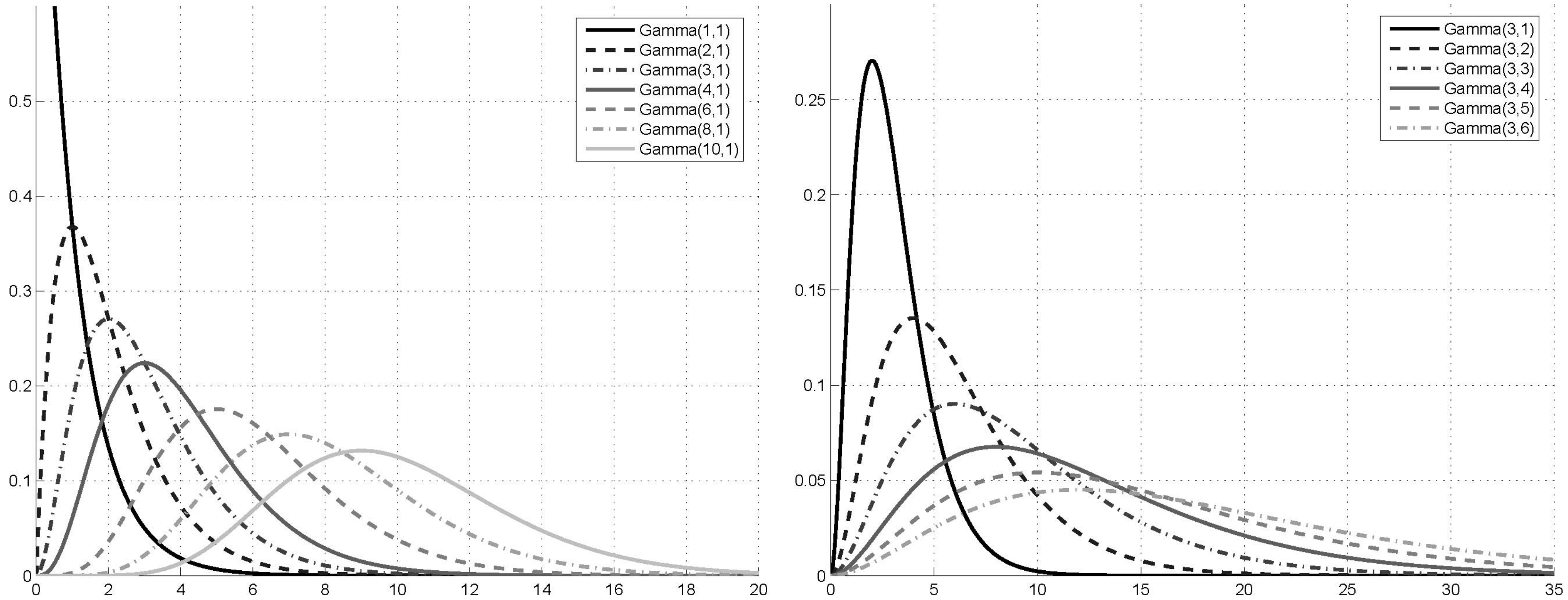
  Note: one fit for each Δ

- 2nd. Covariance (or spectrum) with LRD

  Covariance of a farima model

Gamma distributions

$$\Gamma_{\alpha,\beta}(x) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)$$



Shape parameter α : can model from Gaussian to exponential ;

$1/\alpha \approx$ distance to Gaussian

Scale parameter β : multiplicative factor

Long memory from a farima model

- Long range dependence

  covariance is a non-summable power-law → spectrum $f_{X_\Delta}(\nu)$:

  $f_{X_\Delta}(\nu) \sim C|\nu|^{-\gamma}$, $|\nu|\lozenge 0$, with $0<\gamma<1$

- Farima = fractionnaly integrated ARMA
  1. Fractional integration with parameter d→LRD with $\gamma=2d$
  2. Short range correlation of an ARMA(1, 1) →parameters θ, ϕ

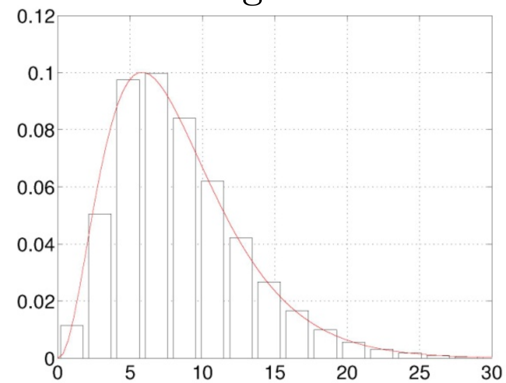
$$f_{X_\Delta}(\nu) = \sigma_\varepsilon^2 \left|1 - e^{-i2\pi\nu}\right|^{-2d} \frac{\left|1 - \theta e^{-i2\pi\nu}\right|^2}{\left|1 - \phi e^{-i2\pi\nu}\right|^2}$$

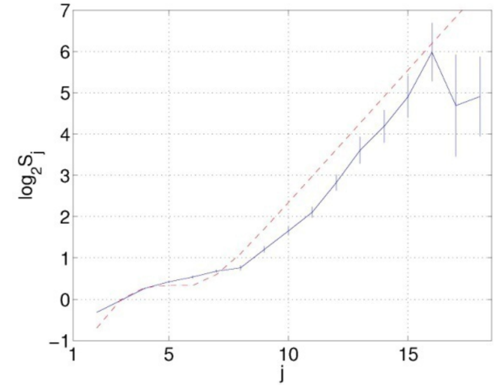
$\Gamma_{\alpha,\beta}$ – farima (ϕ, d, θ) fits

marginals

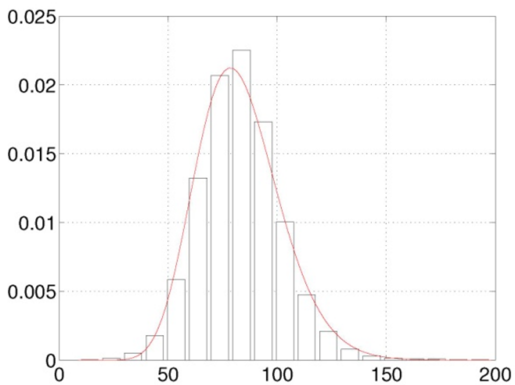
covariances

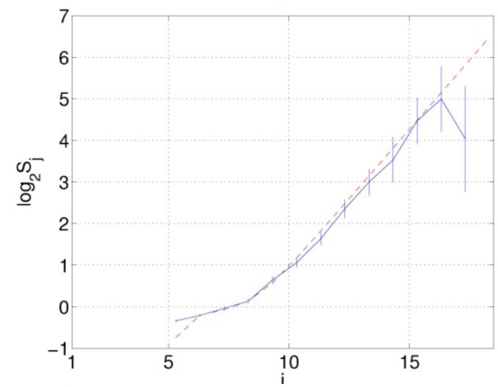
Δ=10ms



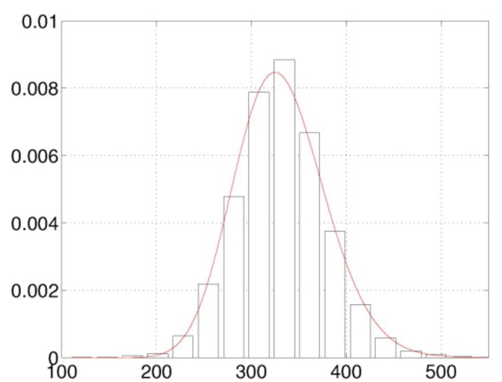


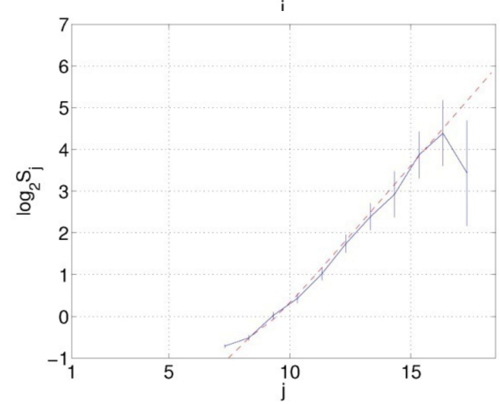
Δ=100ms





Δ=400ms



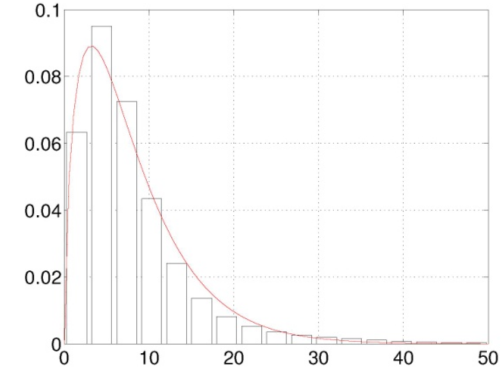


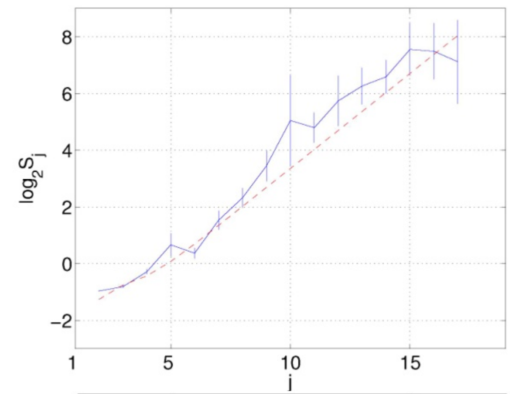
j=1 corresponds to 10 ms

$\Gamma_{\alpha,\beta}$ – farima (ϕ, d, θ) fits

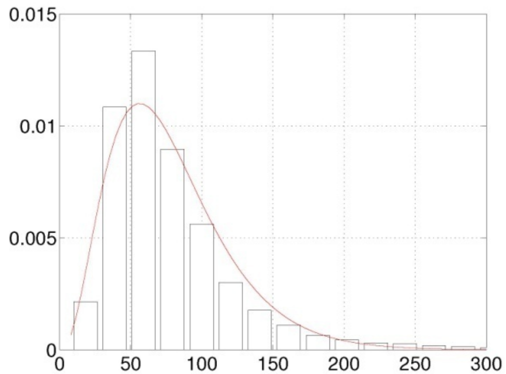
marginals | covariances

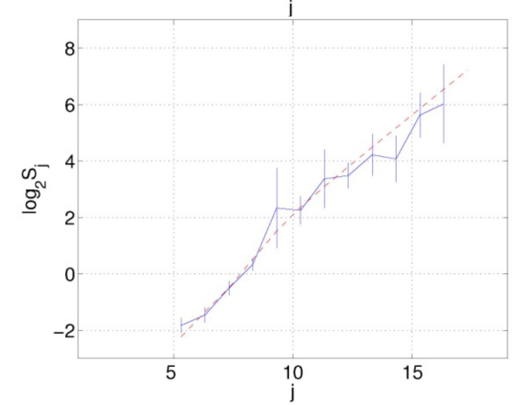
Δ=10ms



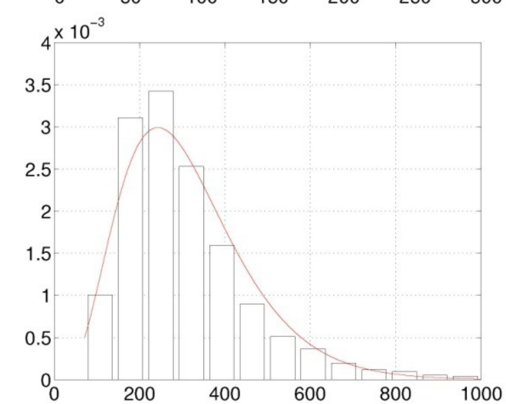


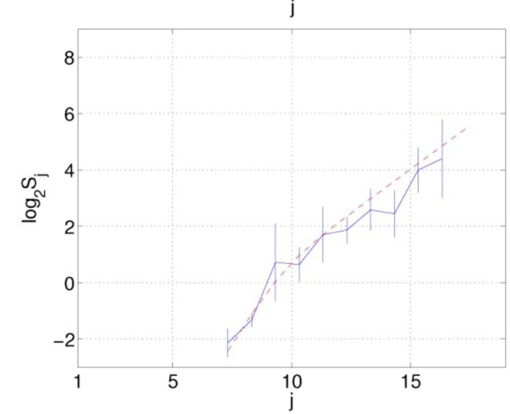
Δ=100ms





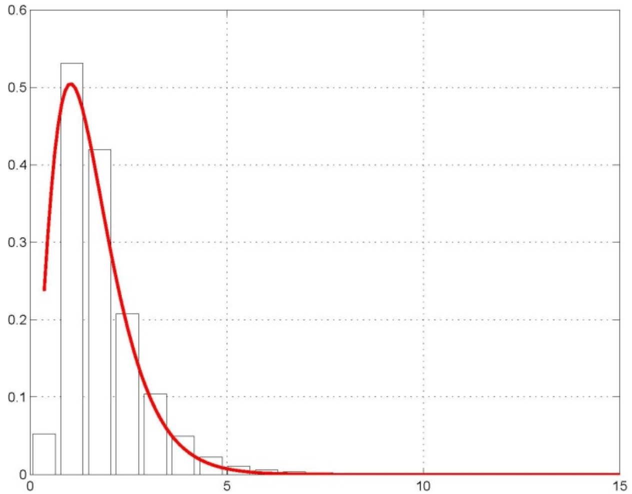
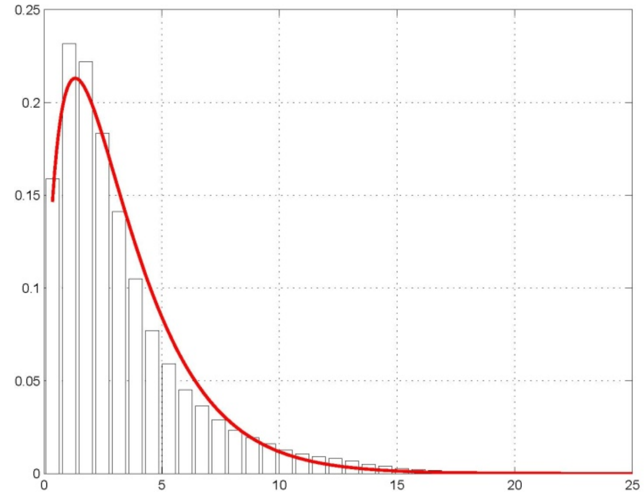
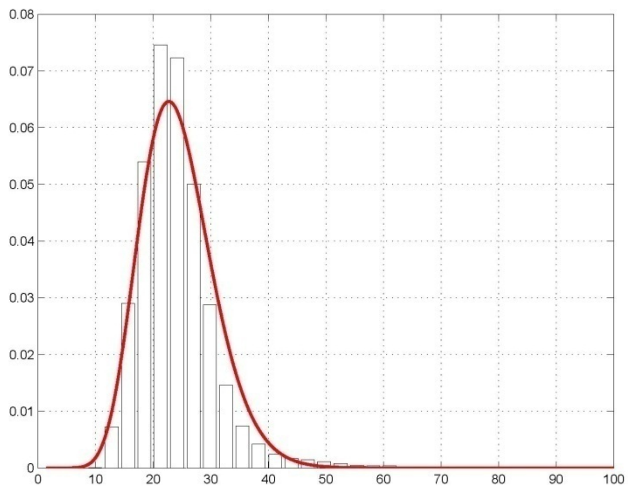
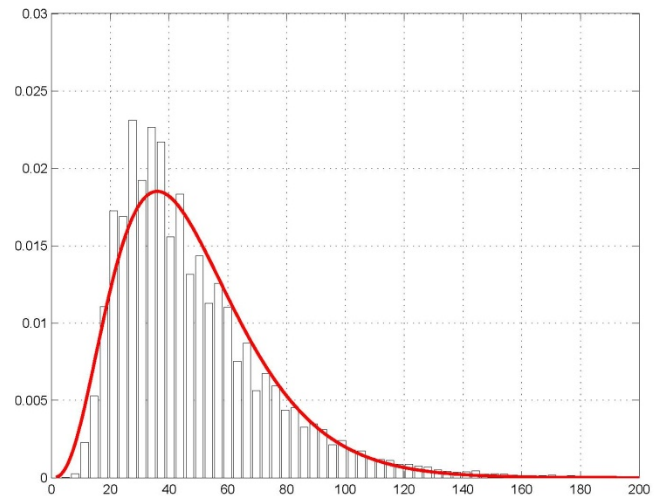
Δ=400ms



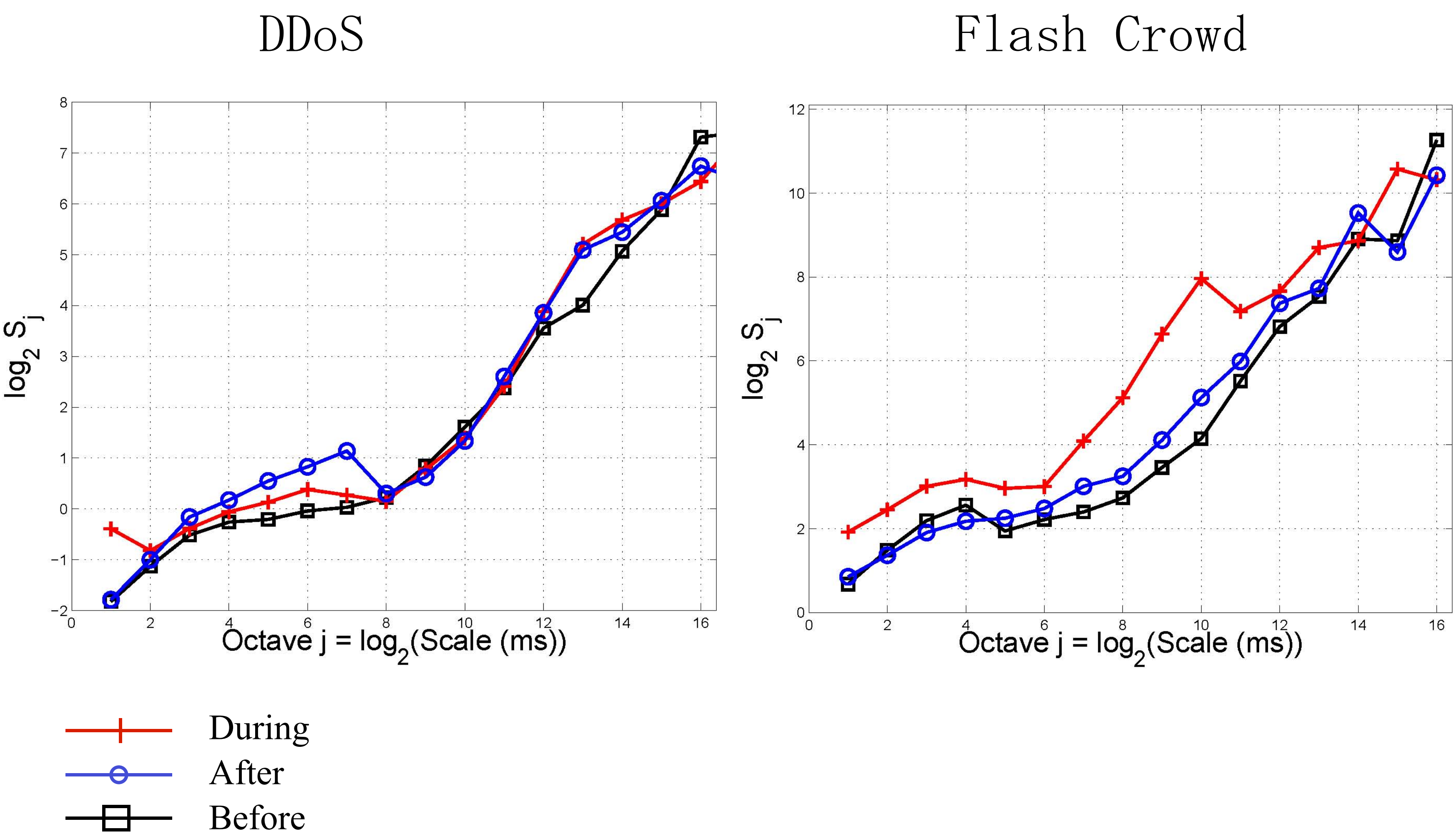


j=1 corresponds to 10 ms

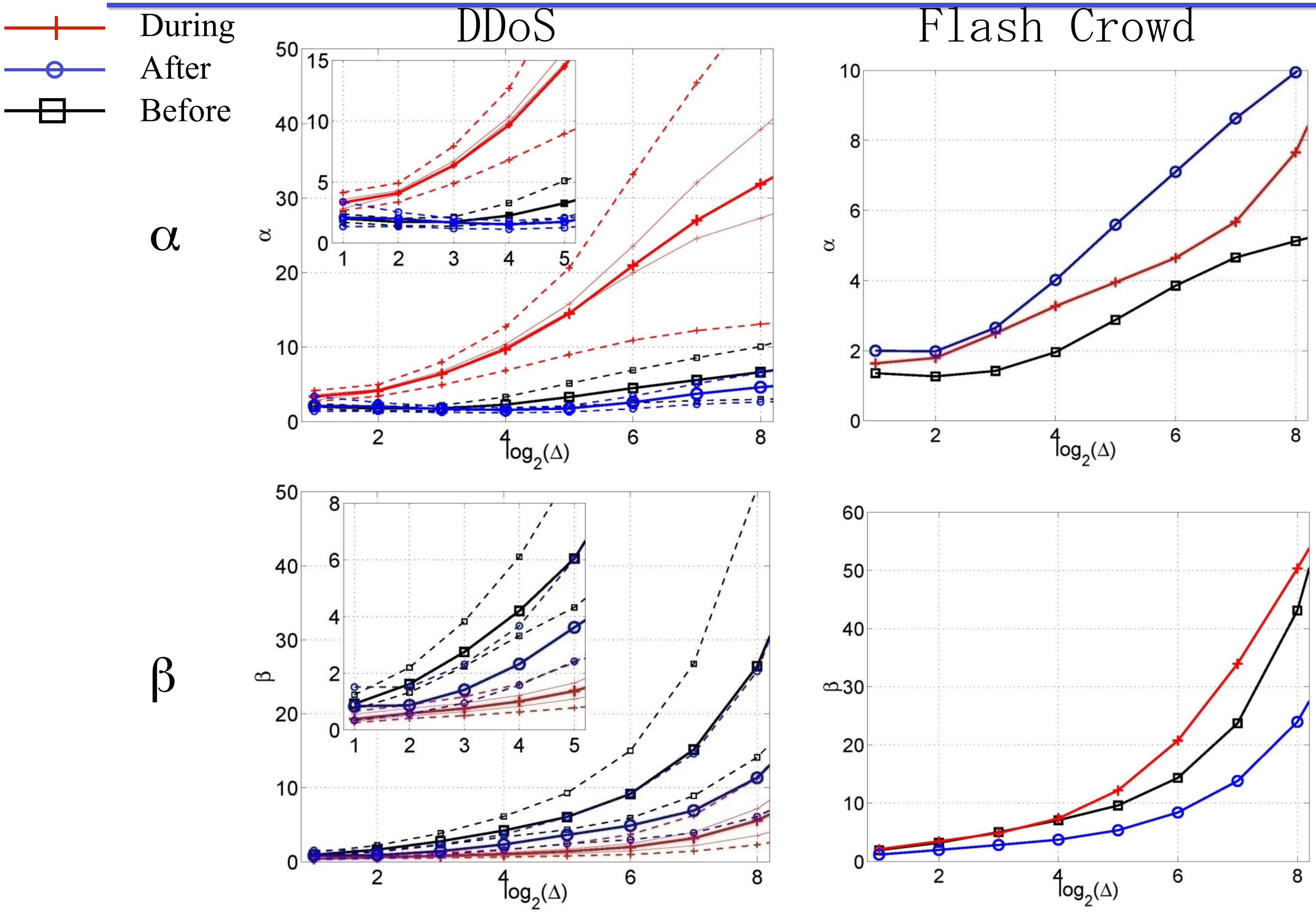
DDoS & FC: $\Gamma_{\alpha,\beta}$ marginal fits

	DDoS attack	Flash crowd
Δ=2ms		
Δ=32ms		

Logscale diagrams for DDoS & FC



Estimated α and β as a function of $\log_2\Delta$



DDoS impact on traffic

- α = shape parameter, $1/\alpha$ quantifies the gap with a Gaussian law
- β = scale parameter → decreases during DDoS attack

➔DDoS attack accelerates the convergence towards a Gaussian distribution of traces, and decreases the fluctuation scale around the average traffic

Partial conc. on Gamma-farima AD

- Model for characterizing Internet traffic which works with and without anomalies
- Some parameters change differently in the presence of a legitimate (flash crowd) or illegitimate (DDoS) anomaly

➔How to use such model for an efficient and robust profile based IDS?

Detection principles

- Select a reference window
- Segment the trace into sliding windows of duration T
- For a window at time I:
  - Aggregated trace at scales $\Delta=2j$, $j=1,\ldots,J$
  - Estimation of parameters : $\alpha_\Delta(I)$, $\beta_\Delta(I)$
  - Compute the distance to the reference, between I and R: D(I)
  - Selection of a threshold λ:
    - if $D(I) \geq \lambda$, $\Rightarrow$ anomaly

Selection of the best distance
(Basseville 89)

- Quadratic distance on parameters

$$D_\alpha(I) = \frac{1}{j}\sum_{j=1}^{J}(\alpha_{2j}(I) - \alpha_{2j}(R))^2$$

$$D_\beta(I) = \frac{1}{J}\sum_{j=1}^{J}(\beta_{2j}(I) - \beta_{2j}(R))^2$$

- Divergence of Kullback-Leibler; p1 and p2 are 2 p.d.f.

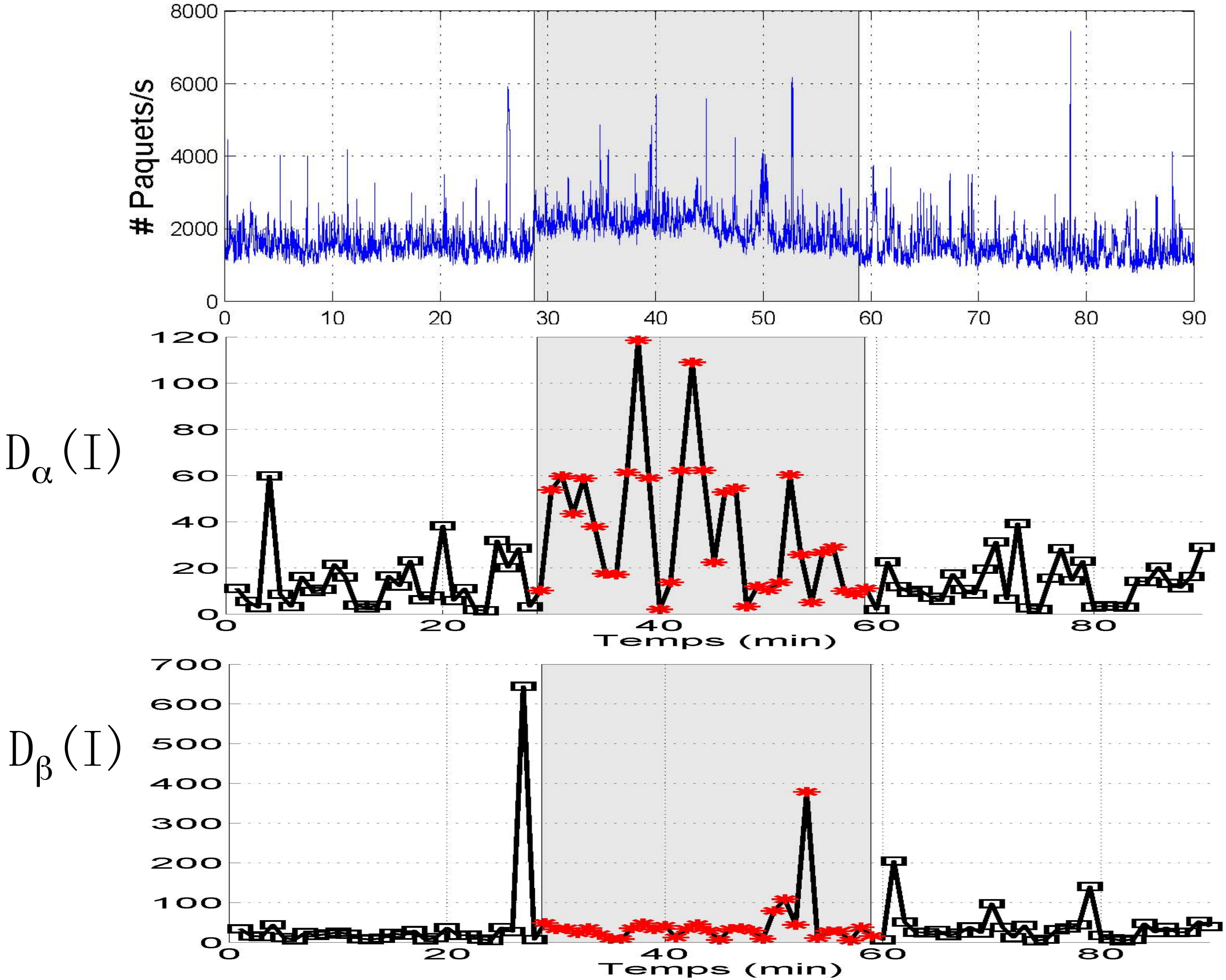
$$DK(p_1, p_2) = \int (p_1(x) - p_2(x)(\ln p_1(x) - \ln p_2(x))dx$$

giving a distance with one or two scales:

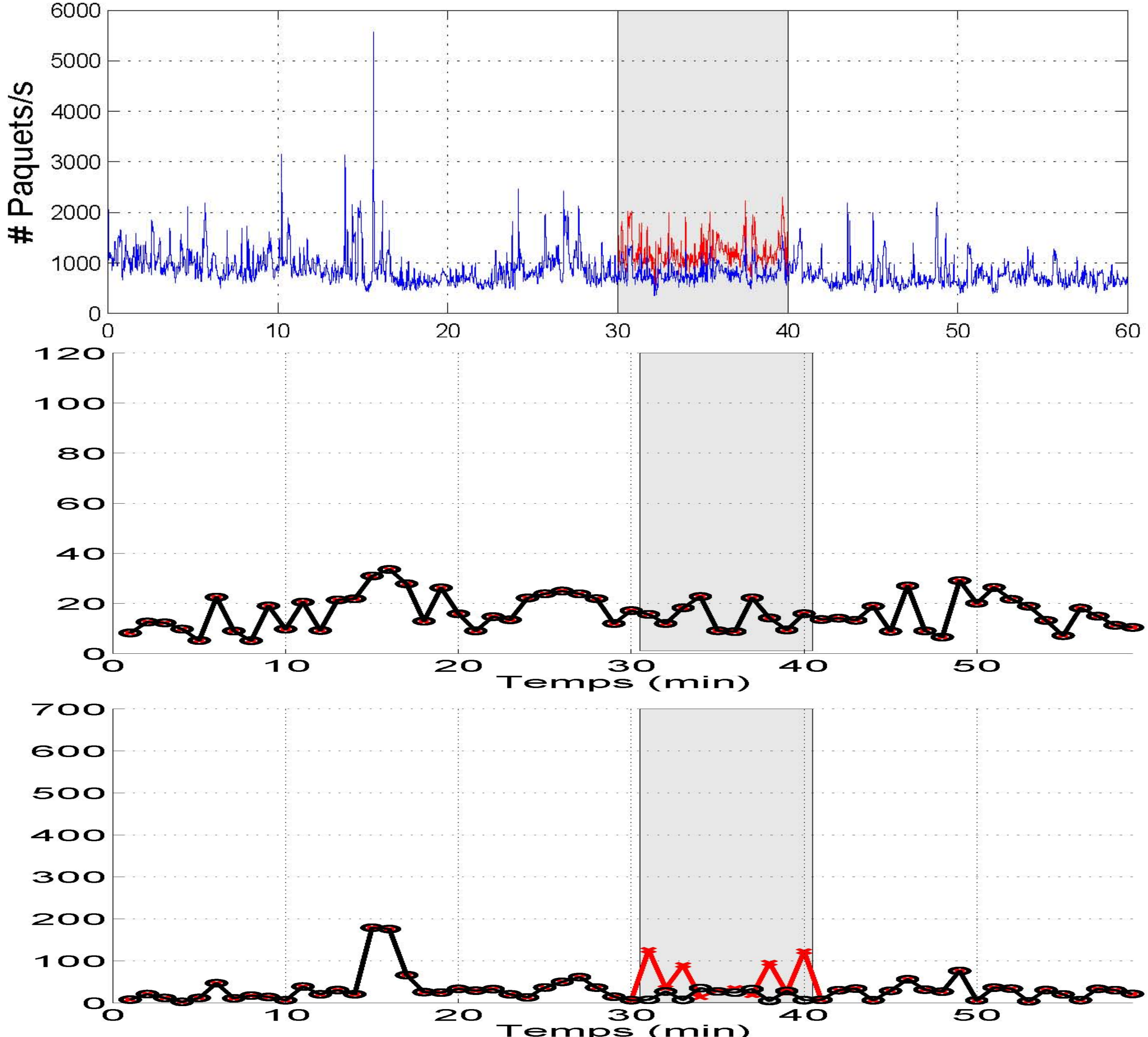
$$K_{\Delta}^{(1D)}(I) = DK(p_{\Delta,I}, p_{\Delta,R})$$

$$K_{\Delta,\Delta'}^{(2D)}(I) = DK(p_{\Delta,\Delta',I}, p_{\Delta,\Delta',R})$$

Ex. 1 : Denial of Service attack



Ex. 2: Multiplicative increase of traffic

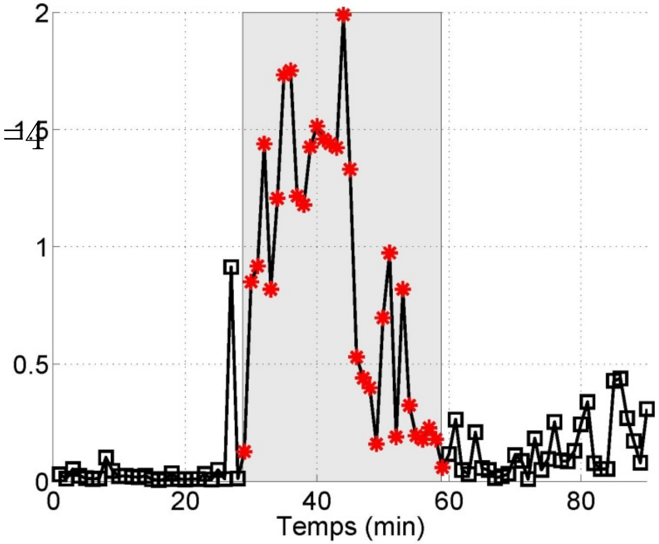


$D_\alpha(I)$

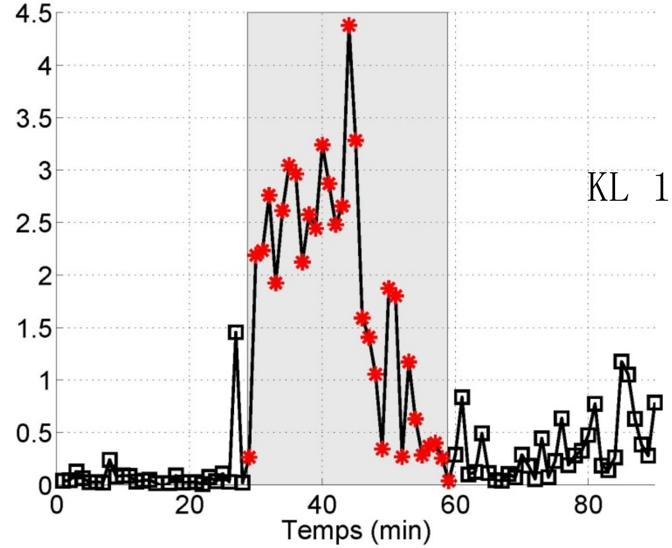
$D_\beta(I)$

Ex. 3: Comparison between distances

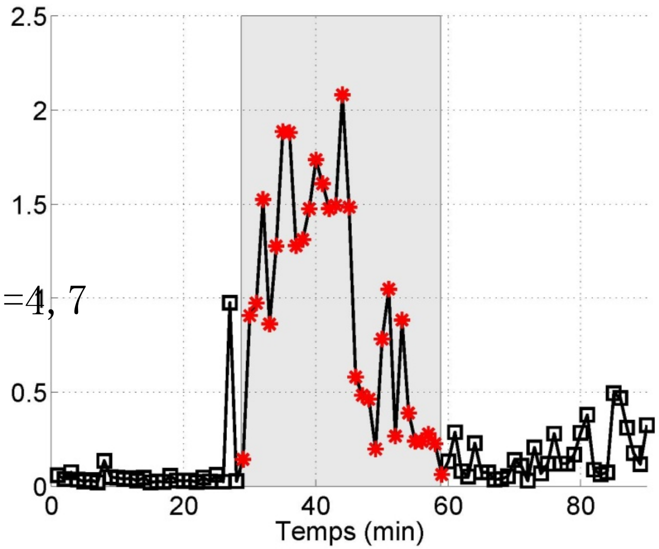
KL 1D, j=4



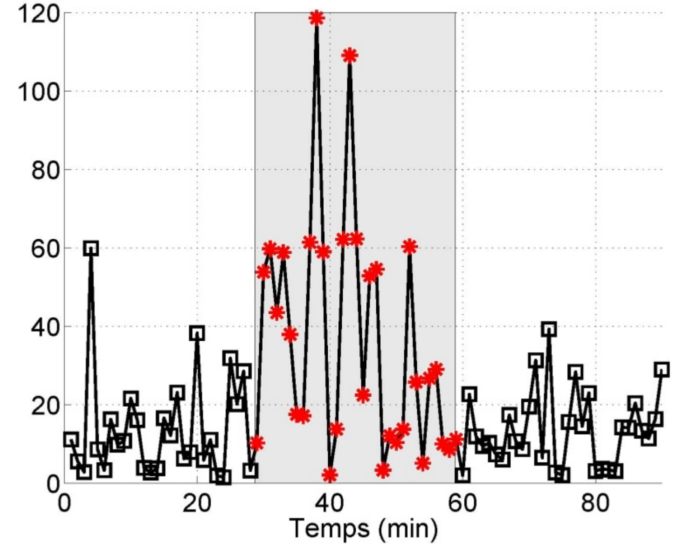
KL 1D, j=7



KL 2D, j=4, 7

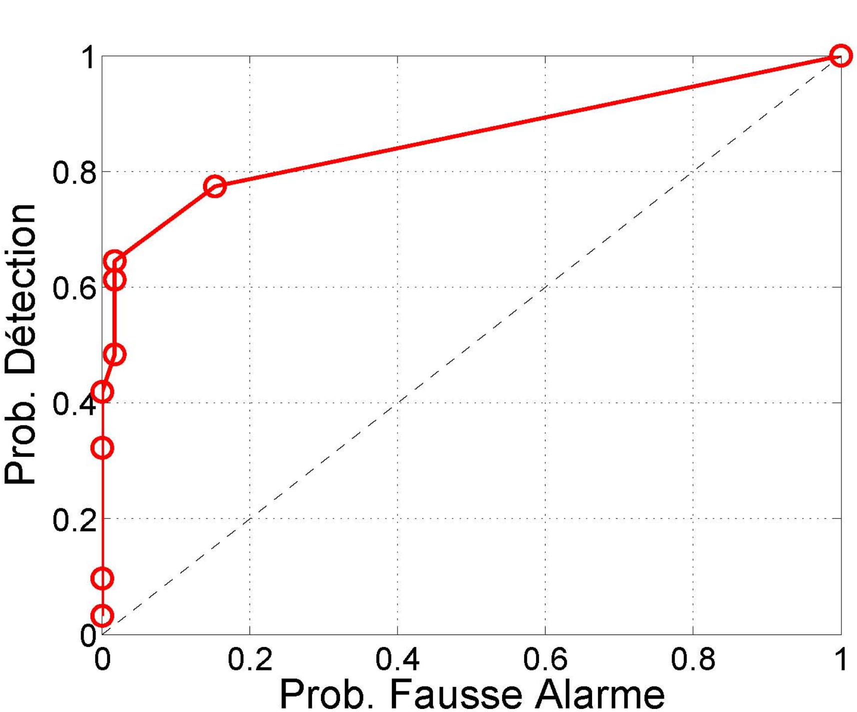


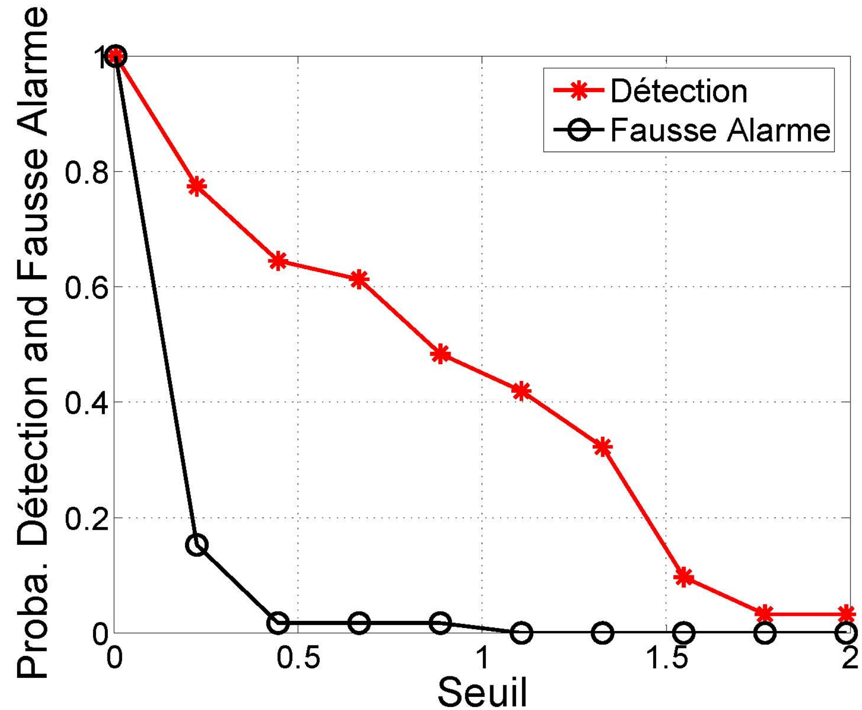
D_α



Statistical performance: ROC curves

- ROC curves: detection probability according to the fixed probability of false alarms
- $P_D = f(P_{FA})$ or $P_D = f(\lambda)$, $P_{FA} = f(\lambda)$





Conclusion on anomalies/attacks detection

- Parameters of the $\Gamma_{\alpha,\beta}$ – farima (ϕ, d, θ) model change differently depending on the type of anomaly
- Kullback- Leibler distance allows a robust detection of attacks, even when they represent less than 1% of the traffic (and is not sensitive to an artificial increase of the amount of traffic)

➔BUT: it is not possible to identify anomaly constituting packets / flows

➔Thresholds are difficult (impossible) to set

➔Classification of anomalies is limited

Unsupervised anomaly detection

From supervised to unsupervised AD

- Current Anomaly Detection (AD) approaches are based on an “acquired knowledge” perspective
  - signature based
  - Supervised approaches
- But
  - Network anomalies are a moving target
  - New attacks as well as new variants to already known attacks arise
  - New services and applications are constantly emerging
- And
  - Defense is reactive, often hand made, slow, costly
  - Network and system remain unprotected for long periods

From supervised to unsupervised AD

- Can we detect what we don't know in an evolving Internet ?
- Is current anomaly-detection perspective rich-enough to handle the problem ?
- Is it possible to manage the network security in a self aware basis to improve performance and reduce operation costs ?

➔unsupervised learning is the idea
  - For proactive security (e.g. 0d anomaly detection)
  - For autonomous defense system (cost reduction)

A detailed ex. of unsupervised AD

- Approach based on Clustering

- Benefits

(+) no previous knowledge: neither labeled data nor traffic signatures

(+) no need for traffic modeling or training (labeling traffic flows is difficult, time-consuming, and costly)

(+) can detect unknown traffic anomalies

(+) a major step towards self-aware monitoring

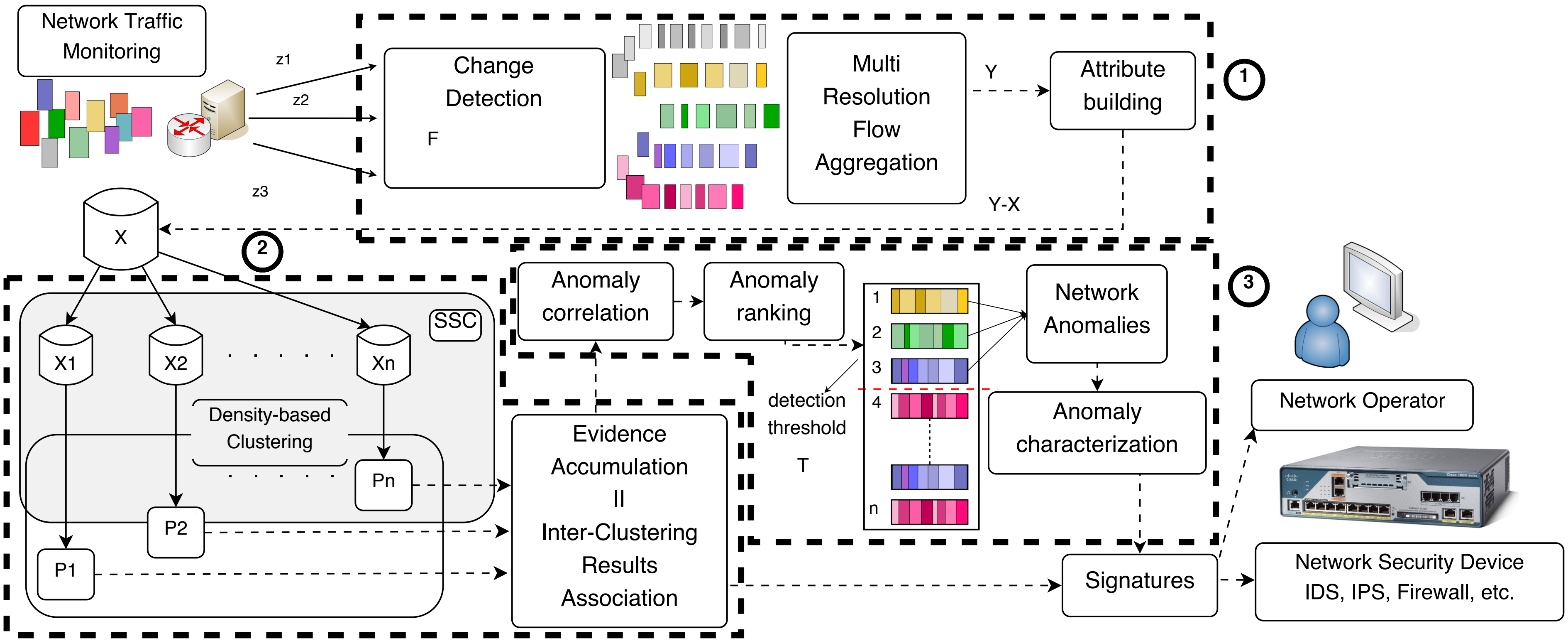
- Challenges with clustering

(-) lack of robustness: general clustering algorithms are sensitive to initialization, specification of number of clusters, etc.

(-) difficult to cluster high-dimensional data: structure-masking by irrelevant features, sparse spaces ("the curve of dimensionality")

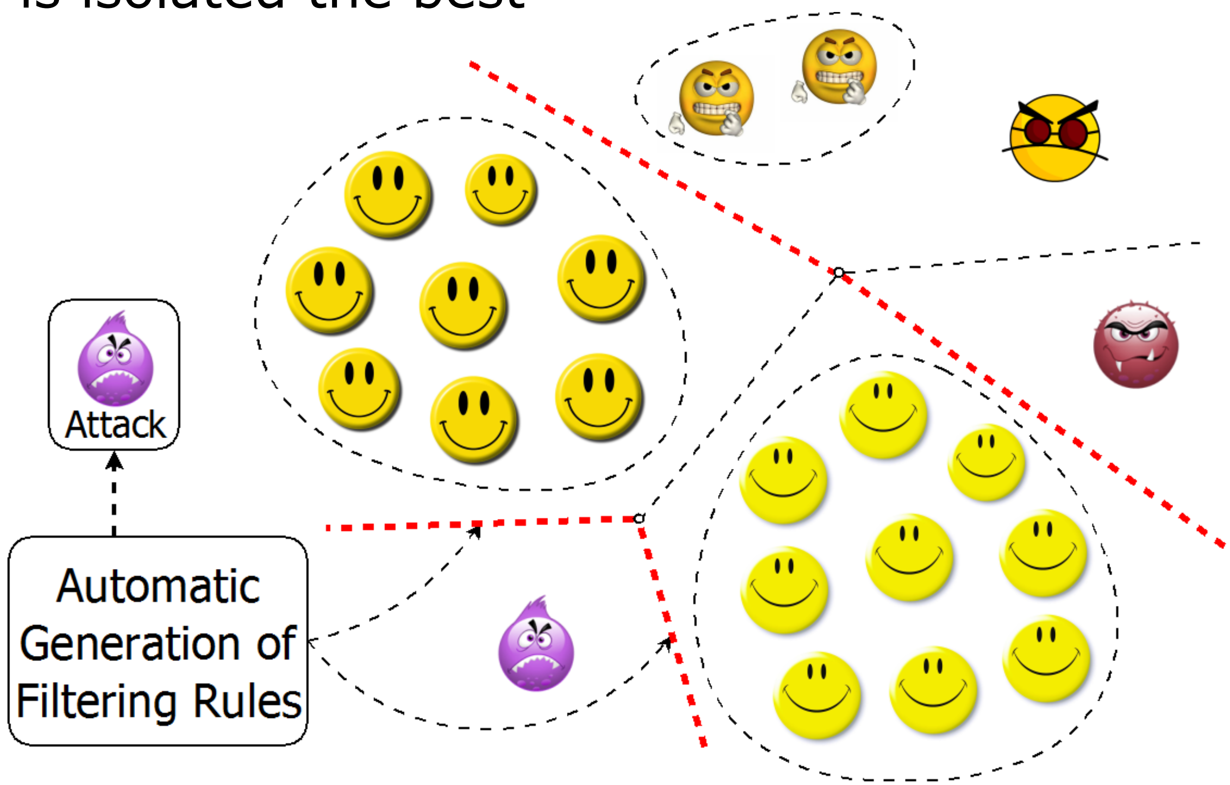
(-) clustering is used only for outliers detection

UNADA: Unsupervized Network AD & Characterization



Filtering rules for anomaly characterization

- Automatically produce a set of filtering rules to correctly isolate and characterize detected anomalous flows
- Select the “best” features to construct a signature of the anomaly, combining the top-K filtering rules
- In a nutshell, select those sub-spaces where anomalous traffic is isolated the best



Clustering for Traffic Analysis

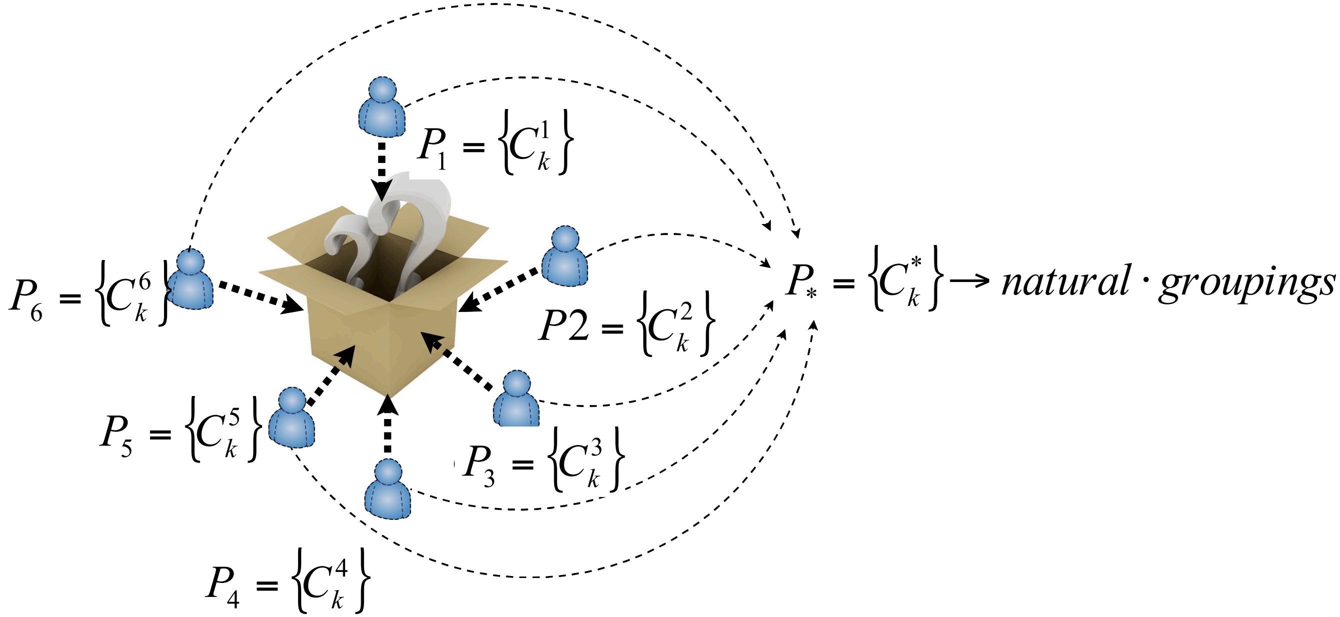
- Let $Y = \{y_1, . . . , y_n\}$ be a set of *n* flows captured at the network of analysis
- Each flow $y_i \in Y$ is described by a set of *m* traffic features: $x_i = (x_i(1), .., x_i(m)) \in \Re^m$
- $X = \{x_1, .., x_n\}$ is the complete matrix of features, referred to as the *feature space*

X is a black box

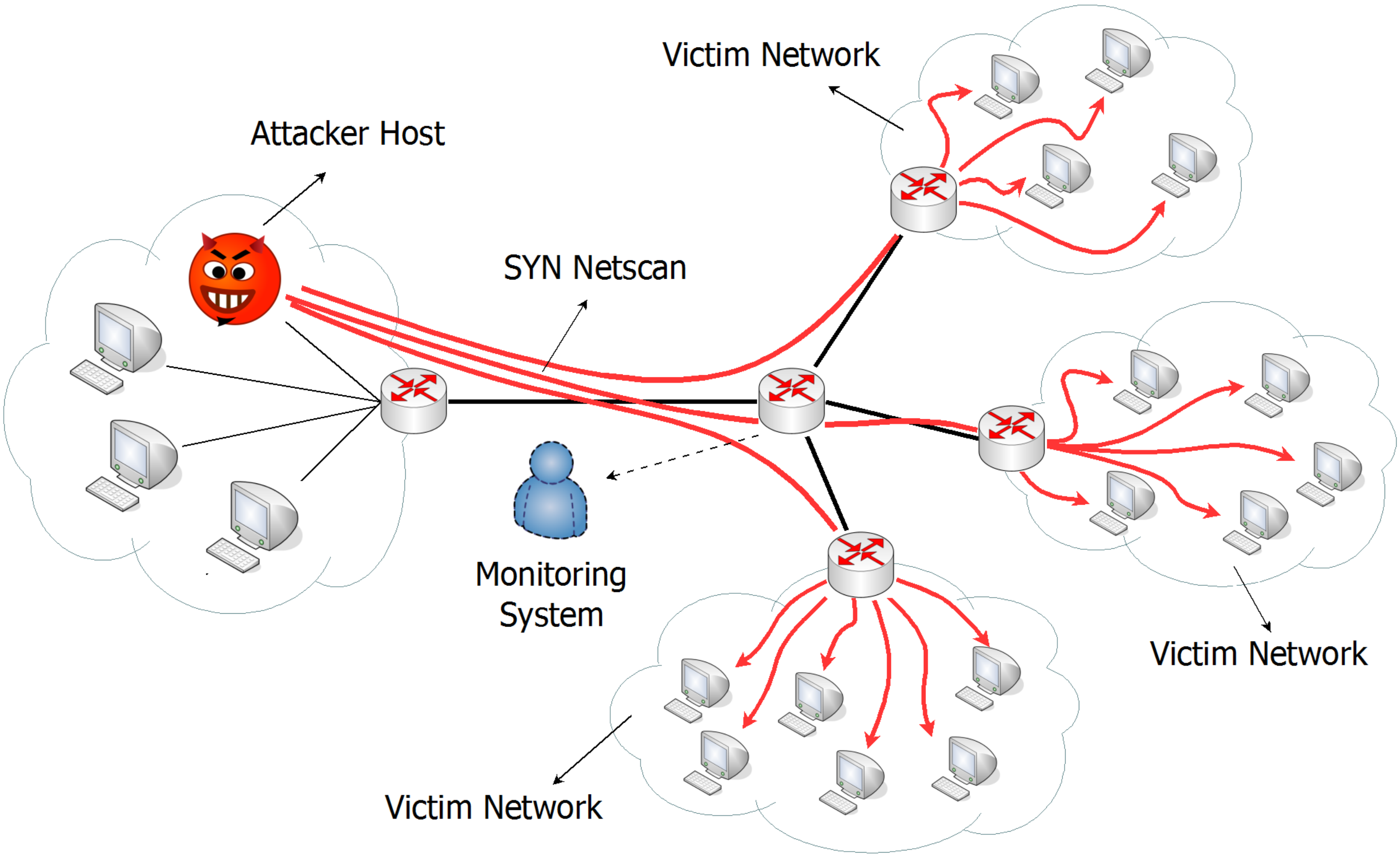
Retrieve natural groupings in X through clustering is challenging!!!

How to improve clustering robustness?

- Idea: combine the information provided by multiple partitions of X to "filter noise", easing the discovery of natural groupings
- How to produce multiple partitions? → Sub-Space Clustering
- Each sub-space $X_i \subset X$ is obtained by projecting X in *k* out of the *m* original dimensions. Density-based clustering (DBSCAN) at X_i



Example of evaluation scenario (emulated on LaasNetExp or ILAB.t)



detection of a SYN Distributed Denial of Service (DDoS) attack in MAWI traffic

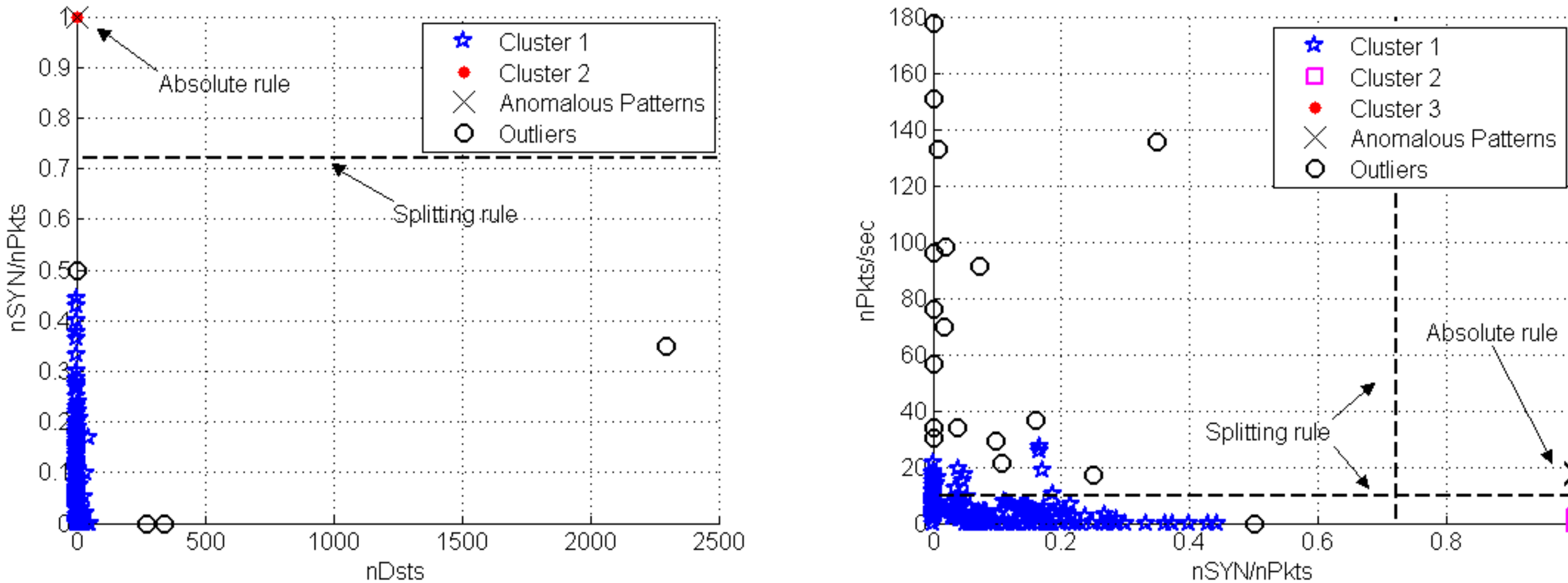


Illustration of clustering graphical results

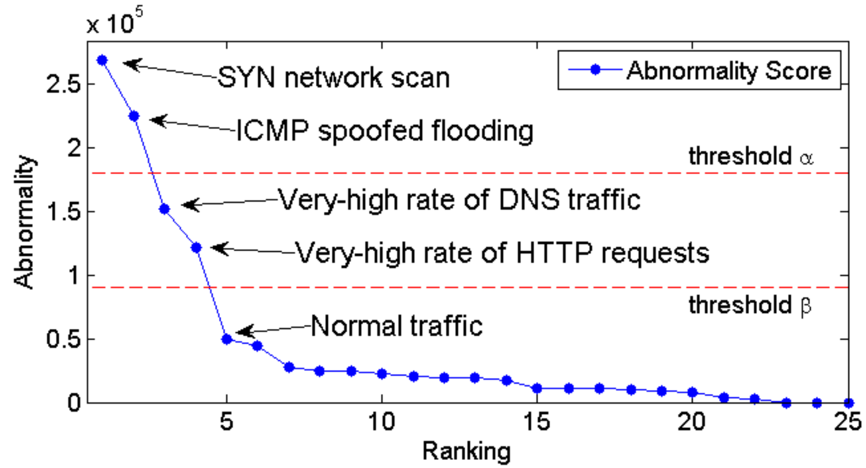
(a) SYN DDoS (1/2) (b) SYN DDoS (2/2)

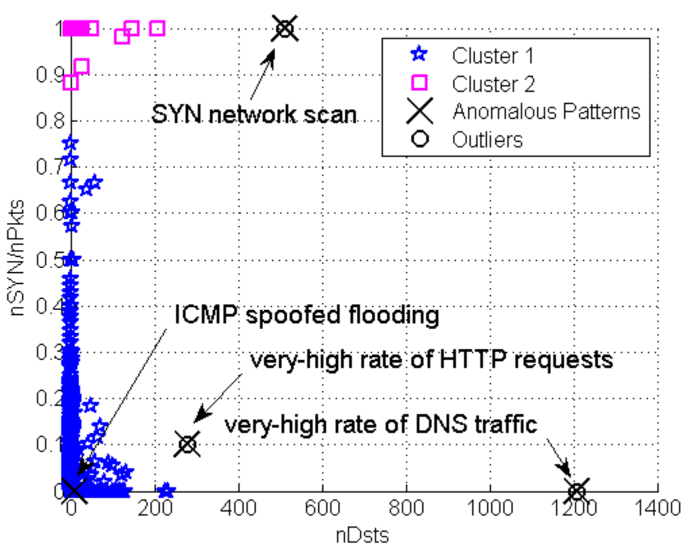
Generated signature

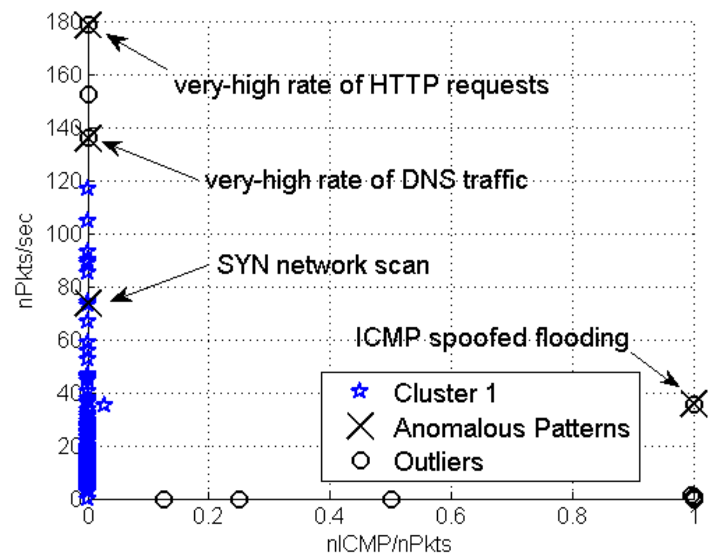
$(nDsts == 1) \wedge (nSYN/nPkts > \lambda_3) \wedge (nPkts/sec > \lambda_4) \wedge (nSrcs > \lambda_5)$

Attacks detection & characterization in MAWI traffic

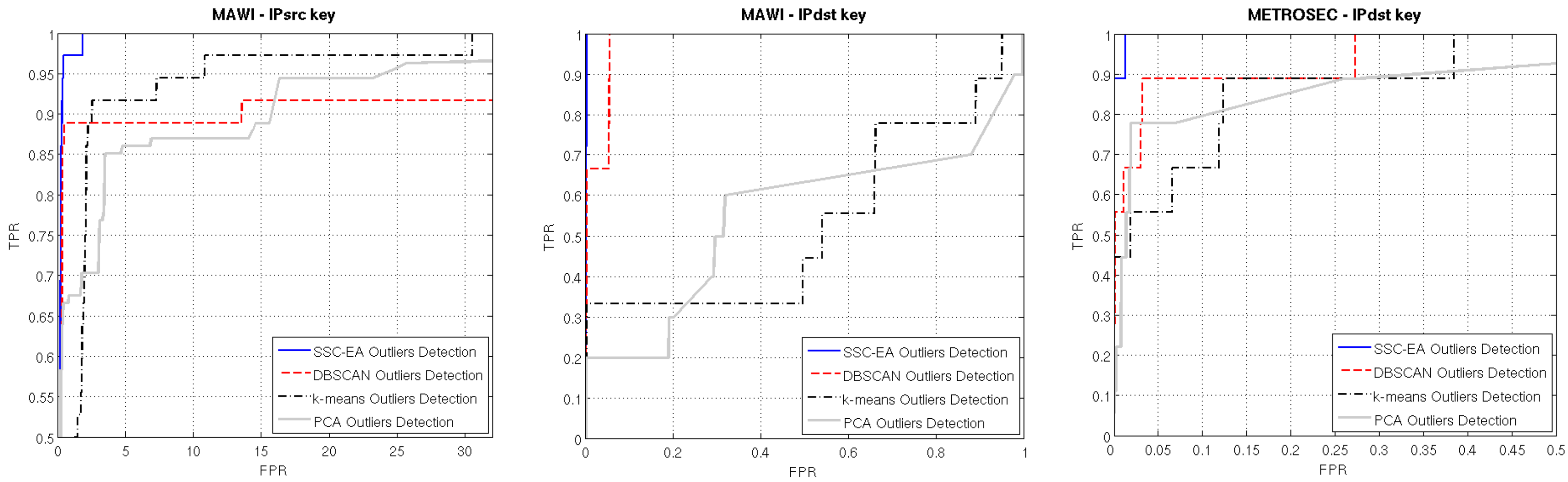
- Detect network attacks that are not the biggest elephant flows







Comparison between ≠ unsupervised techniques



Comparison of detection performance of several detection algorithms

ROC (receiver Operating Characteristic) curves presenting True Positive Rate (TPR) vs. False positive rate (TPR)

Conclusion on unsupervised AD

- Detection / classification reports of anomalies
- Reports are very complete in order to allow the automatic enforcement of countermeasures for the ML engine

(+) filtering rules ready to be exported towards security devices (e.g. Intrusion Detection Systems, Intrusion Protection Systems, Firewal, etc.)

Tutorial conclusion: keywords

- Botnets: main current threads on the Internet?
- Deep packet inspection / misuse detection
- Profile based detection
- Traffic characterization, analysis and modeling
- Supervised & unsupervised machine learning
- Distances
- Clustering

- +

Tutorial conclusion

- Supervised → unsupervised
  - Reducing the need of labeled traffic is paramount to achieve useful anomaly detectors
  - Gives methods for network Autonomy
  - Reduces management cost, and duration (limited hand made human interventions)
  - Allows 0day (unknown) anomaly detection
  - Network does not stay unprotected for a long period

→A way to adapt to botnet thread?

→A global trend in networks / networking

That’s all folks !