Unsupervised Anomaly Detection in Internet Traffic

Philippe Owezarski
LAAS-CNRS, Toulouse, France
RC/SARA

Shortcomings in nowadays network security

Network security is based on some PREVIOUS KNOWLEDGE:

- Signature-based: detect the attacks THAT WE KNOW
- Anomaly detection: detect DIFFERENCES from WHAT WE KNOW

HOW STABLE-in-time is this PREVIOUS KNOWLEDGE?

- Network attacks are a moving target: new attacks are constantly emerging, and the birth-rate is increasing
- New services and applications modify normal-operation profiles

We depend TOO-MUCH on the PREVIOUS KNOWLEDGE:

- This knowledge is difficult and expensive to obtain
- Long periods of VULNERABILITY (e.g. weeks) between a new attack and the construction of a new signature
- Current network security is REACTIVE, and as such WE ARE ALWAYS ONE STEP BEHIND THE ATTACKERS!!!

Research direction for security

- Unsupervised clustering for detecting and characterizing classes of anomalies without relying on previous knowledge, signatures, statistical training or labeled traffic
- Automatic production of filtering rules (→ firewalls, filtering equipments, …)
- Discrimination between legitimate vs. Illegitimate anomalies
  - Root cause analysis
- Automatic mitigation of attacks vs. Reporting to network/security administrator

Unsupervised Detection of Network Attacks

A COMPLEMENTARY approach: Clustering for Detection of Attacks

"ATTACKS are STATISTICALLY DIFFERENT from normal traffic"

Benefits of clustering-based detection

(+) **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-defense**

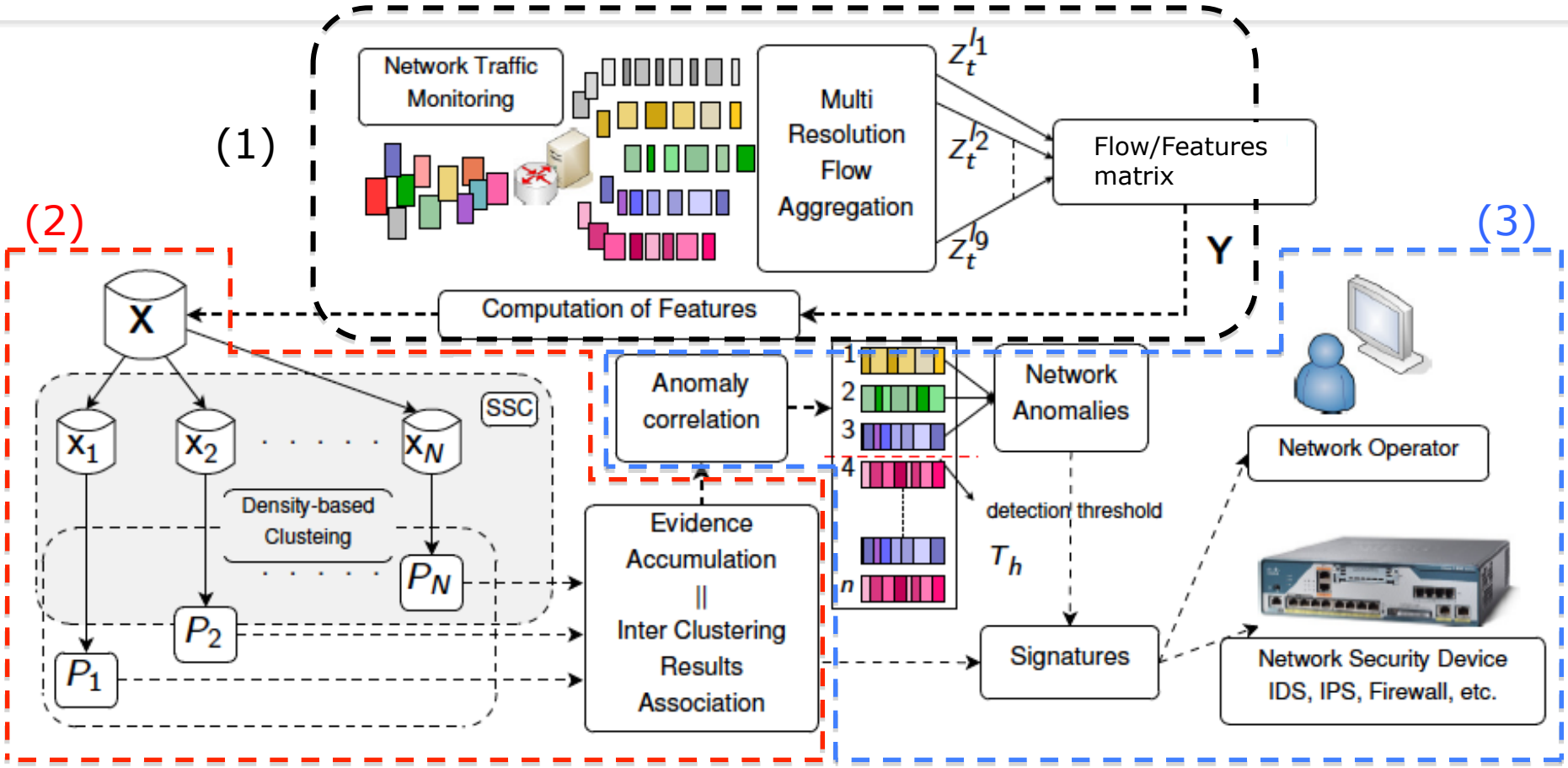
…but clustering for network security is CHALLENGING

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

(-) **difficult to perform feature selection** for clustering

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

Unsupervised network anomaly detection and characterization



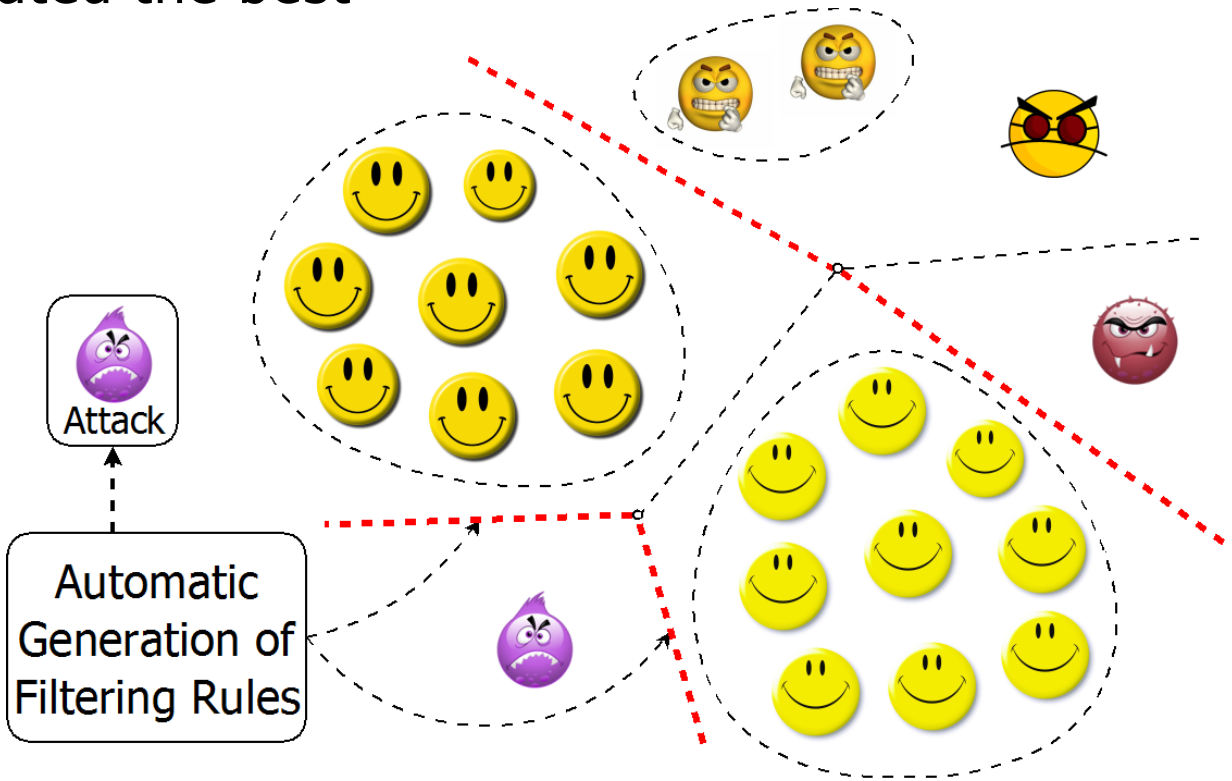
(1) Multi-reso., flow aggreg., Change-detection & Attribute building

(2) Sub-Space Clustering and, evidence accumulation or Inter-Clustering Results Association

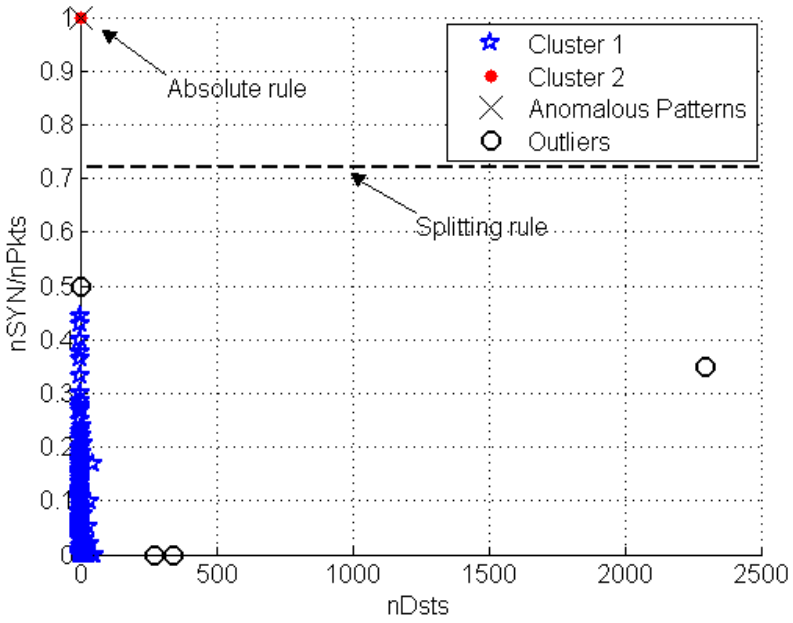
(3) Correlation & Characterization through filtering rules → signatures

Filtering rules for anomaly characterization

- Automatically produce a set of filtering rules f(Y) 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
- is isolated the best



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



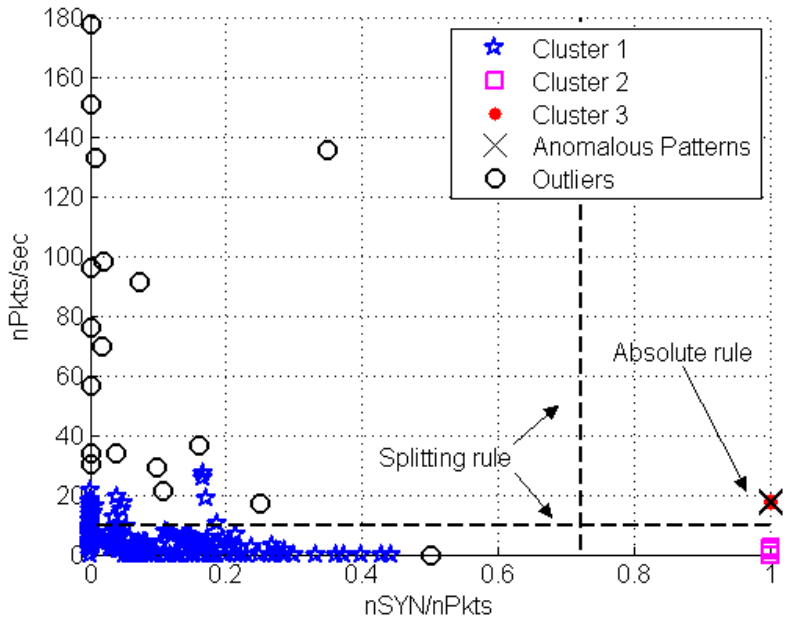


Illustration of clustering graphical results

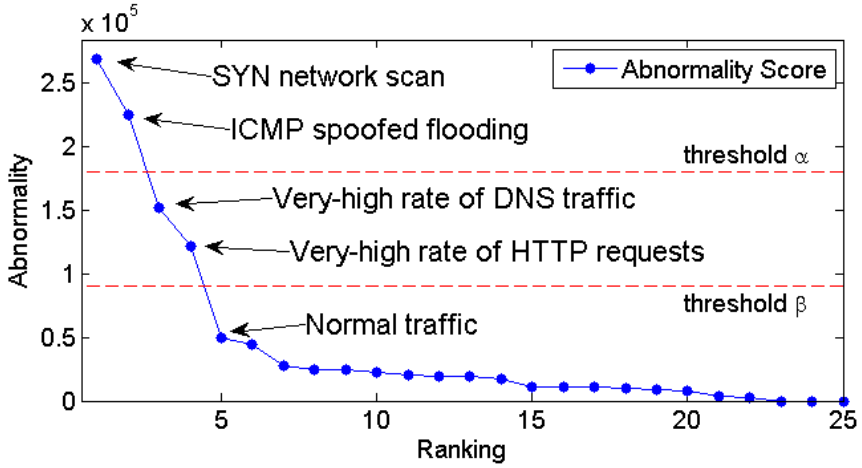
(a) SYN DDoS (1/2) (b) SYN DDoS (2/2)

Generated signature

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

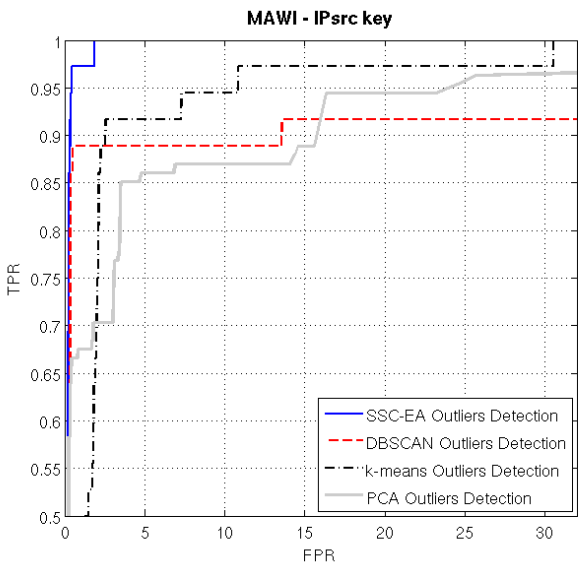
Anomaly classification in MAWI traffic

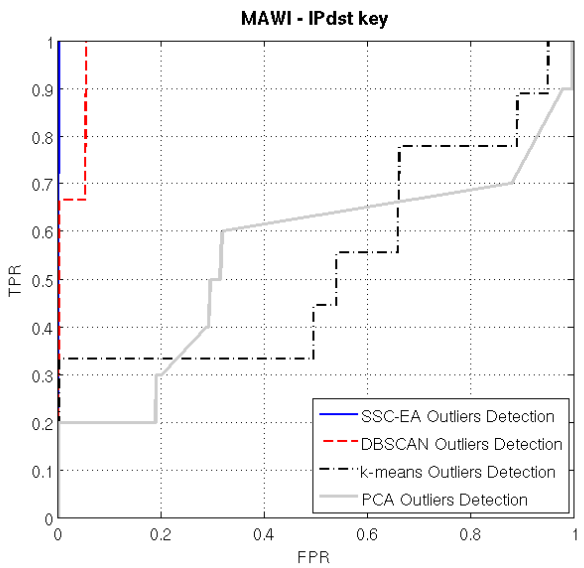
- Using an abnormality score (unupervised approach)

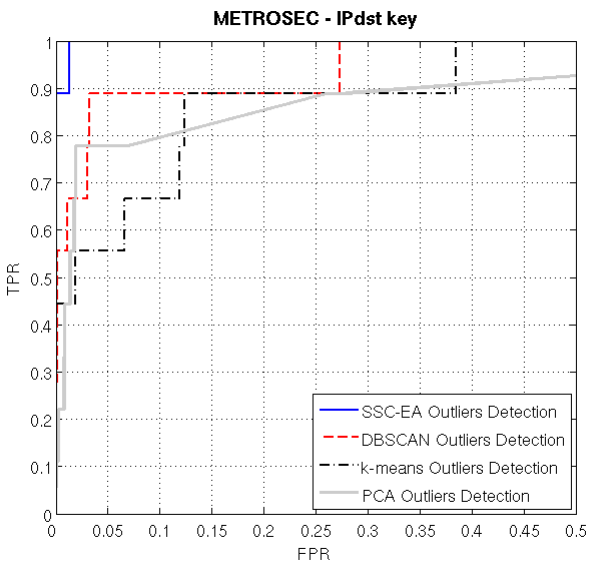


- Using signatures on flows in clusters (semi-supervised approach)

Comparison between ≠ unsupervised techniques





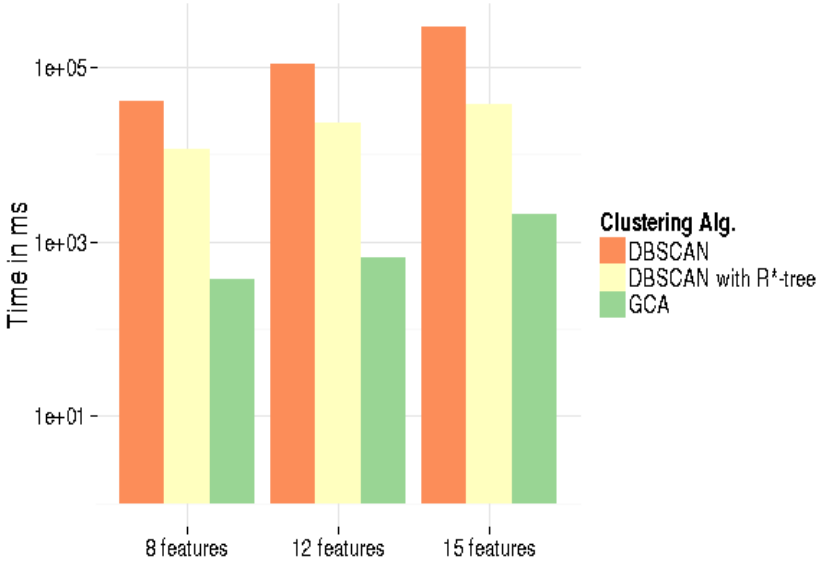


Comparison of detection performance of several detection algorithms

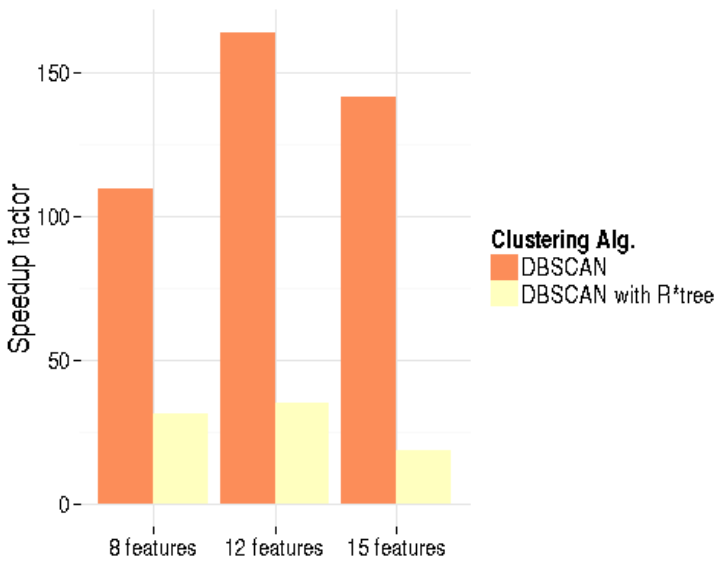
ROC (receiver Operating Characteristic) curves presenting True Positive Rate (TPR) vs. False positive rate (TPR)

Speed Performance

Evaluation performed with the ONTS dataset



Comparison of UNADA execution time



Gain in time using GCA

- Grid clustering can speed up the execution by a factor of 150
- Does it modify the detection performance of UNADA ?
- What is the minimum micro-slot that ORUNADA can deal with ?

Common ground truth for evaluation

- KDD'99 is still the only trustable ground truth for anomaly detector assessment
  - But it is getting old
- MAWILab relies on 4 detectors for labeling traffic
  - But some of the labels are controversial
- FP7 European ONTIC project targets to build a new ground truth
  - Based on the collected ONTS traffic dataset
  - Using MAWILab tool for providing a first set of labels
  - Organizing a Hackathon and opening it to the international community for correcting questionnable labels

Open issues

- Scalability (big data platforms, sampling, …)
- Main features selection
- Algorithm sensitivity
- Cluster classification
- Ground truth for evaluation

- Extend the ML approach in other security domains
  - Other algorithms

LAAS CNRS

Q & A