Anomaly Detection with Extreme Value Theory

A. Siffer, P-A Fouque, A. Termier and C. Largouet

May 30, 2017

Contents

Context

General motivations

- Massive usage of the Internet

- Massive usage of the Internet
    - More and more vulnerabilities

WannaCry ransomware used in widespread attacks all over the world

- Massive usage of the Internet
  - More and more vulnerabilities
  - More and more threats

WannaCry ransomware used in widespread attacks all over the world

- Massive usage of the Internet
  - More and more vulnerabilities
  - More and more threats
- Awareness of the sensitive data and infrastructures

WannaCry ransomware used in widespread attacks all over the world

- Massive usage of the Internet
  - More and more vulnerabilities
  - More and more threats
- Awareness of the sensitive data and infrastructures

$\Rightarrow$ Network security : a major concern

A Solution

- IDS (Intrusion Detection System)
    - Monitor traffic
    - Detect attacks

A Solution

- IDS (Intrusion Detection System)
  - Monitor traffic
  - Detect attacks
- Current methods : rule-based
  - Work fine on common and well-known attacks
  - Cannot detect new attacks

A Solution

- IDS (Intrusion Detection System)
  - Monitor traffic
  - Detect attacks
- Current methods : rule-based
  - Work fine on common and well-known attacks
  - Cannot detect new attacks
- Emerging methods : anomaly-based
  - Use the network data to estimate a normal behavior
  - Apply algorithms to detect abnormal events (→ attacks)

- Basic scheme

data → ALGORITHM → alerts

Overview

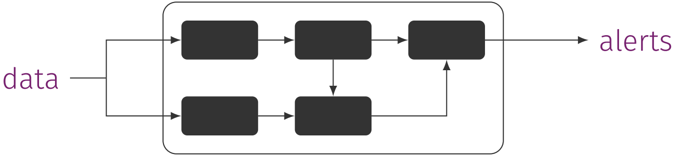
- Basic scheme

$$\text{data} \longrightarrow \boxed{\textsc{Algorithm}} \longrightarrow \text{alerts}$$

- Many ”standard” algorithms have been tested

- Basic scheme

- Many "standard" algorithms have been tested
- Complex pipelines are emerging (ensemble/hybrid techniques)



Inherent problem

- Algorithms are not magic
  - They give some information about data (scores)

Inherent problem

- Algorithms are not magic
  - They give some information about data (scores)
  - But the decision often rely on a human choice

```
if score>threshold then trigger alert
```

Inherent problem

- Algorithms are not magic
  - They give some information about data (scores)
  - But the decision often rely on a human choice

```
if score>threshold then trigger alert
```

- The thresholds are often hard-set
  - Expertise
  - Fine-tuning
  - Distribution assumption

Inherent problem

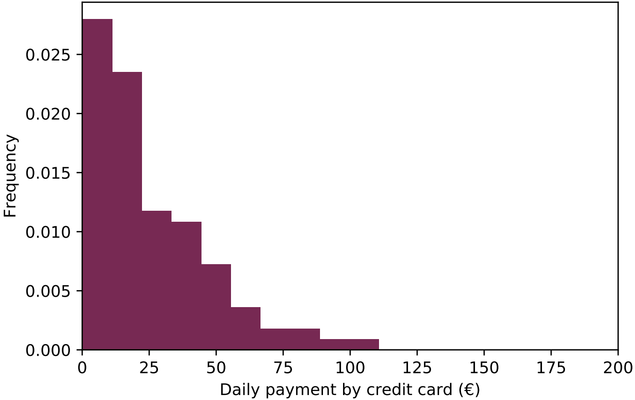
- Algorithms are not magic
  - They give some information about data (scores)
  - But the decision often rely on a human choice

```
if score>threshold then trigger alert
```

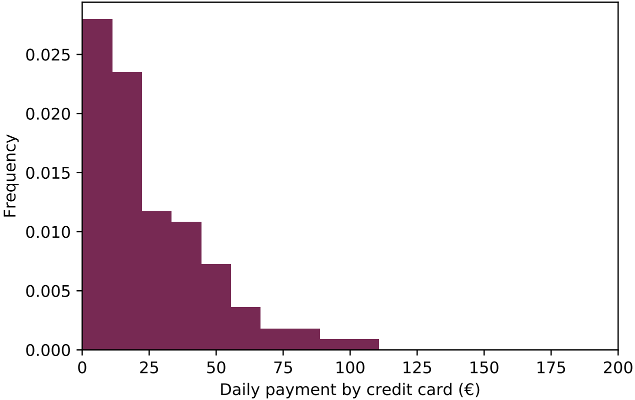
- The thresholds are often hard-set
  - Expertise
  - Fine-tuning
  - Distribution assumption
- **Our idea**: provide dynamic threshold with a probabilistic meaning

Providing better thresholds

My problem

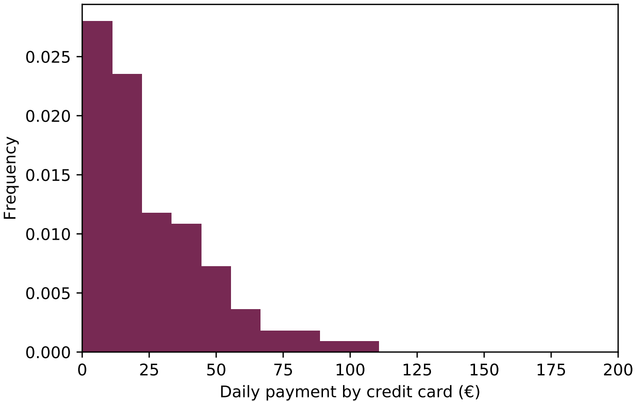


0.025
0.020
0.015
0.010
0.005
0.000
Frequency
0
25
50
75
100
125
150
175
200
Daily payment by credit card (€)

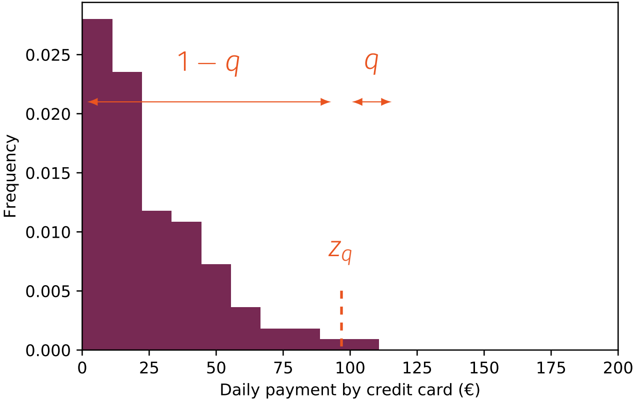


⊸ How to set z_q such that $\mathbb{P}(X€ > z_q) < q$?

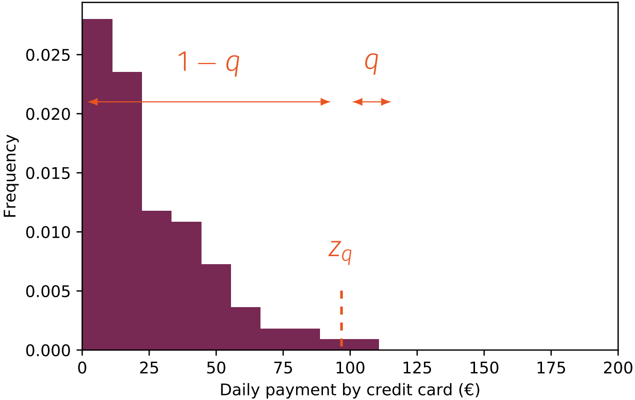
Solution 1: empirical approach



Solution 1: empirical approach

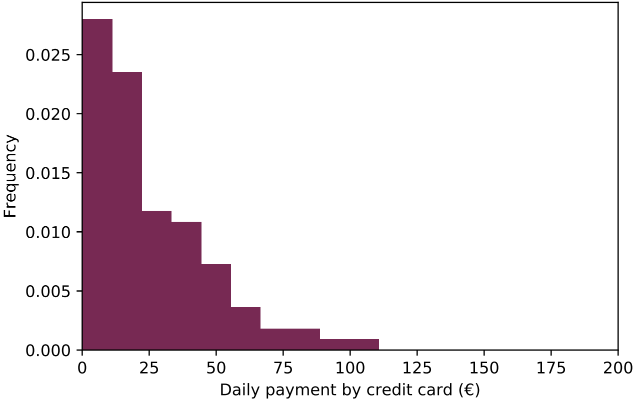


Solution 1: empirical approach

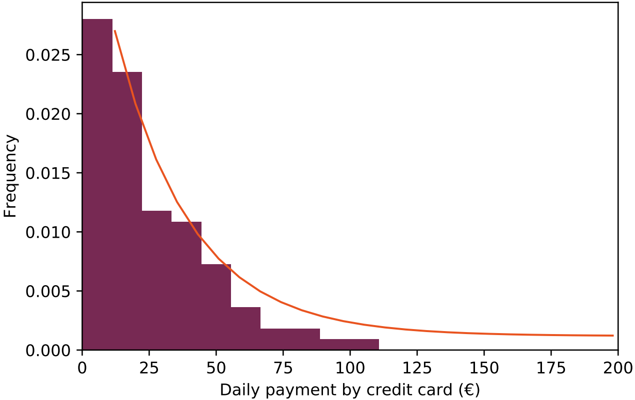


- Drawbacks: stuck in the interval, poor resolution

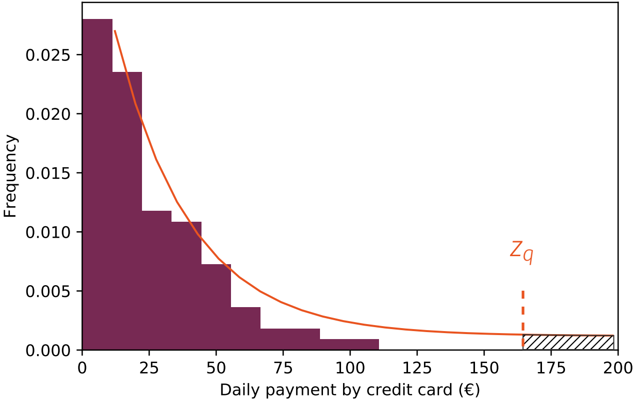
Solution 2: Standard Model



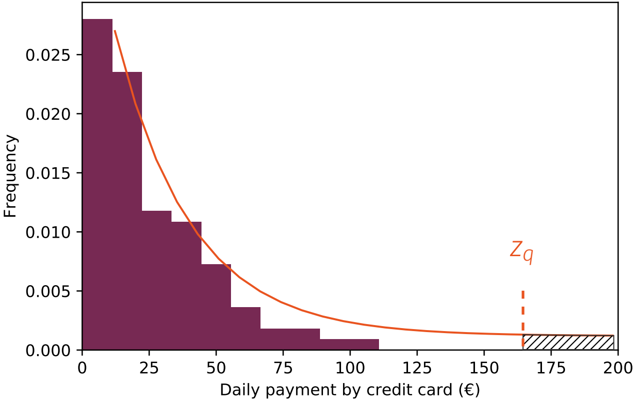
Solution 2: Standard Model



Solution 2: Standard Model

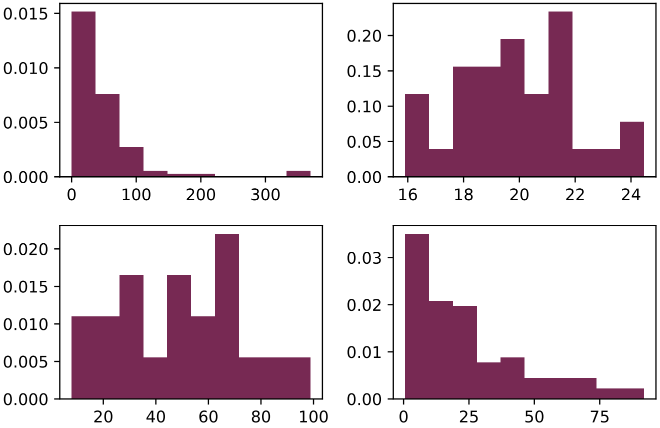


Solution 2: Standard Model



- Drawbacks: manual step, distribution assumption

Realities

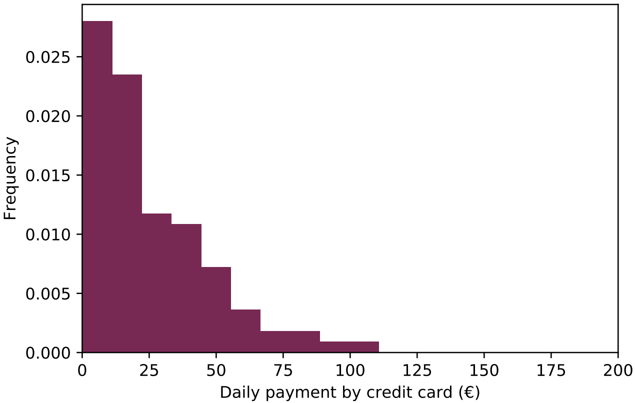


- Different clients and/or temporal drift

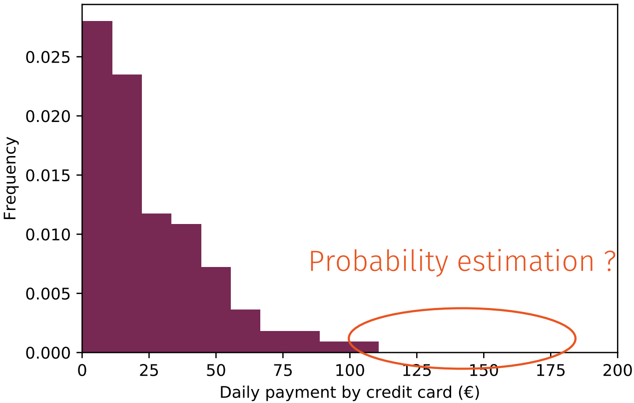
Results

Properties	**Empirical quantile**	**Standard model**
statistical guarantees	Yes	Yes
easy to adapt	Yes	No
high resolution	No	Yes

Inspection of extreme events



Inspection of extreme events



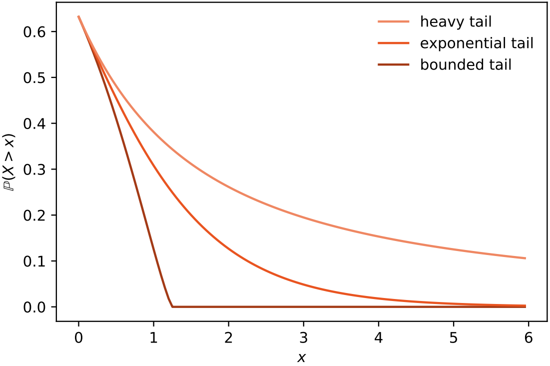
Extreme Value Theory

- Main result (Fisher-Tippett-Gnedenko, 1928)

  The extreme values of any distribution have nearly the same distribution (called Extreme Value Distribution)

- Main result (Fisher-Tippett-Gnedenko, 1928)

  The extreme values of any distribution have nearly the same distribution (called Extreme Value Distribution)



- Let $X_1, X_2, \ldots X_n$ a sequence of i.i.d. random variables with

$$S_n = \sum_{i=1}^{n} X_i \qquad M_n = \max_{1 \leq i \leq n} (X_i)$$

- Let $X_1, X_2, \ldots X_n$ a sequence of i.i.d. random variables with

$$S_n = \sum_{i=1}^{n} X_i \qquad M_n = \max_{1 \leq i \leq n} (X_i)$$

- Central Limit Theorem

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

An impressive analogy

- Let $X_1, X_2, \ldots X_n$ a sequence of i.i.d. random variables with

$$S_n = \sum_{i=1}^{n} X_i \qquad M_n = \max_{1 \leq i \leq n} (X_i)$$

- Central Limit Theorem

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

- FTG Theorem

$$\frac{M_n - a_n}{b_n} \xrightarrow{d} \mathrm{EVD}(\gamma)$$

- Second theorem of EVT (Pickands-Balkema-de Haan, 1974)

  The excesses over a high threshold follow a Generalized Pareto Distribution (with parameters γ, σ)

- Second theorem of EVT (Pickands-Balkema-de Haan, 1974)

  The excesses over a high threshold follow a Generalized Pareto Distribution (with parameters γ, σ)

- What does it imply ?
  - we have a model for extreme events
  - we can compute z_q for q as small as desired

How to use EVT

- Get some data $X_1, X_2 \ldots X_n$
- Set a high threshold t and retrieve the excesses $Y_j = X_{k_j} - t$ when $X_{k_j} > t$

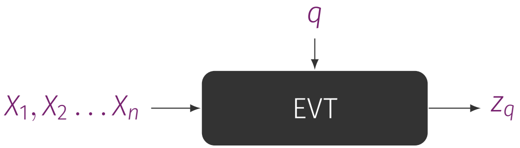
How to use EVT

- Get some data $X_1, X_2 \dots X_n$
- Set a high threshold t and retrieve the excesses $Y_j = X_{k_j} - t$ when $X_{k_j} > t$
- Fit a GPD to the Y_j ($\rightarrow$ find parameters γ, σ)

How to use EVT

- Get some data $X_1, X_2 \dots X_n$
- Set a high threshold t and retrieve the excesses $Y_j = X_{k_j} - t$ when $X_{k_j} > t$
- Fit a GPD to the Y_j ($\rightarrow$ find parameters γ, σ)
- Compute z_q such as $\mathbb{P}(X > z_q) < q$

- Get some data $X_1, X_2 \dots X_n$
- Set a high threshold t and retrieve the excesses $Y_j = X_{k_j} - t$ when $X_{k_j} > t$
- Fit a GPD to the Y_j ($\rightarrow$ find parameters γ, σ)
- Compute z_q such as $\mathbb{P}(X > z_q) < q$



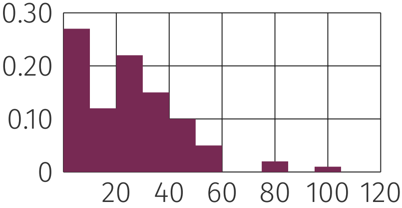
Finding anomalies in streams

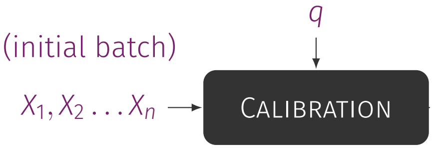
Streaming Peaks-Over-Threshold (SPOT) algorithm

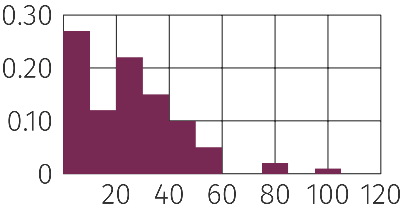
Streaming Peaks-Over-Threshold (SPOT) algorithm

(initial batch)

$X_1, X_2 \dots X_n$




q
(initial batch)
$X_1, X_2 \ldots X_n$
Calibration


0.30
0.20
0.10
0
20
40
60
80
100
120

Streaming Peaks-Over-Threshold (SPOT) algorithm

Streaming Peaks-Over-Threshold (SPOT) algorithm

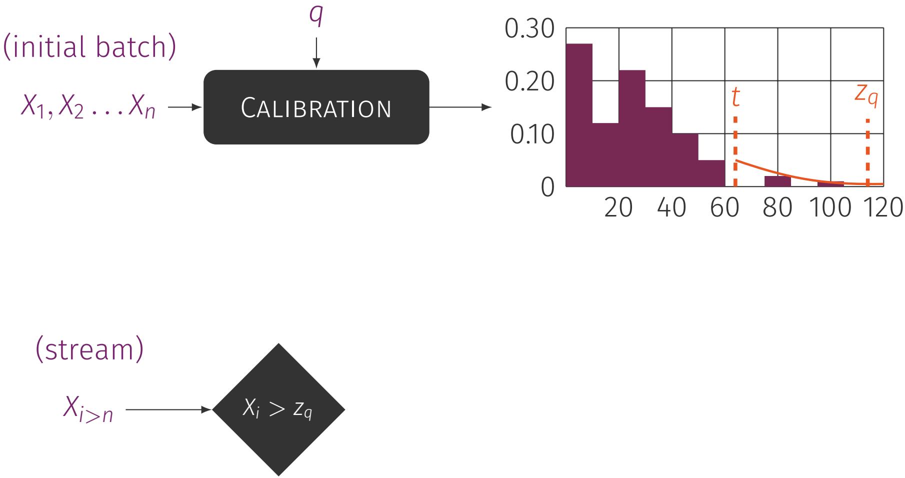
Streaming Peaks-Over-Threshold (SPOT) algorithm

Streaming Peaks-Over-Threshold (SPOT) algorithm

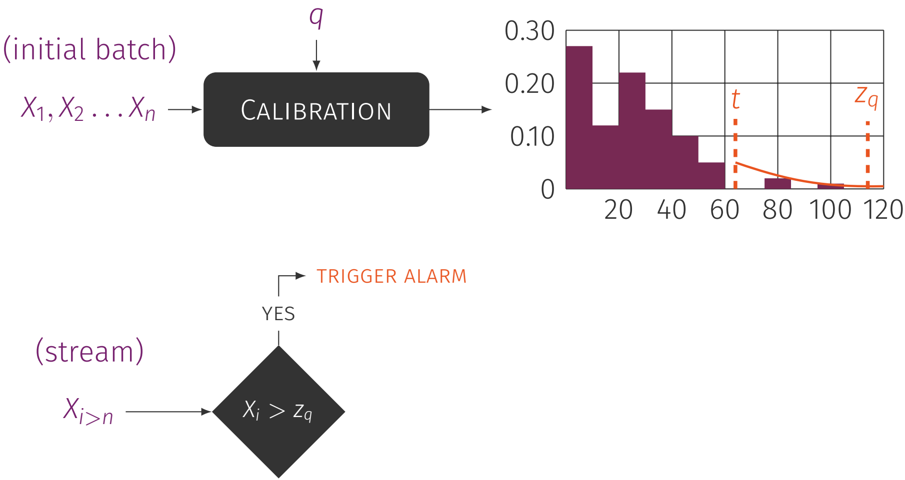
(stream)

$X_{i>n}$

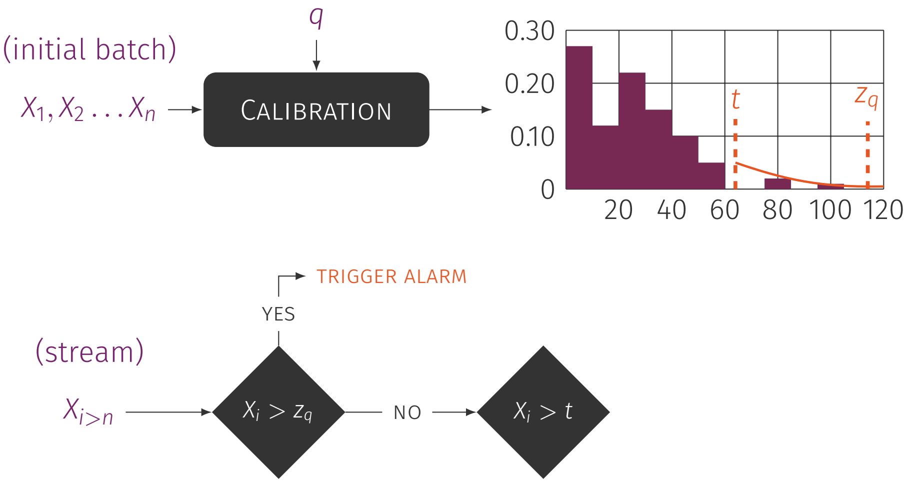
Streaming Peaks-Over-Threshold (SPOT) algorithm



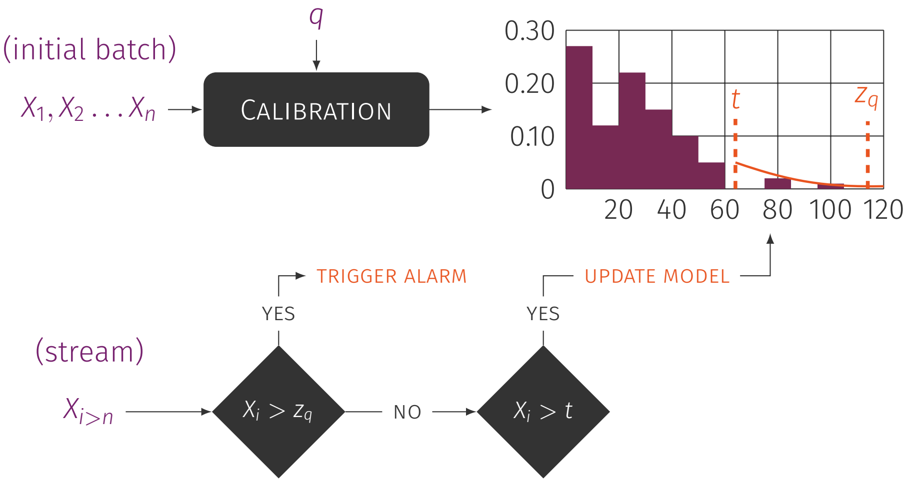
Streaming Peaks-Over-Threshold (SPOT) algorithm



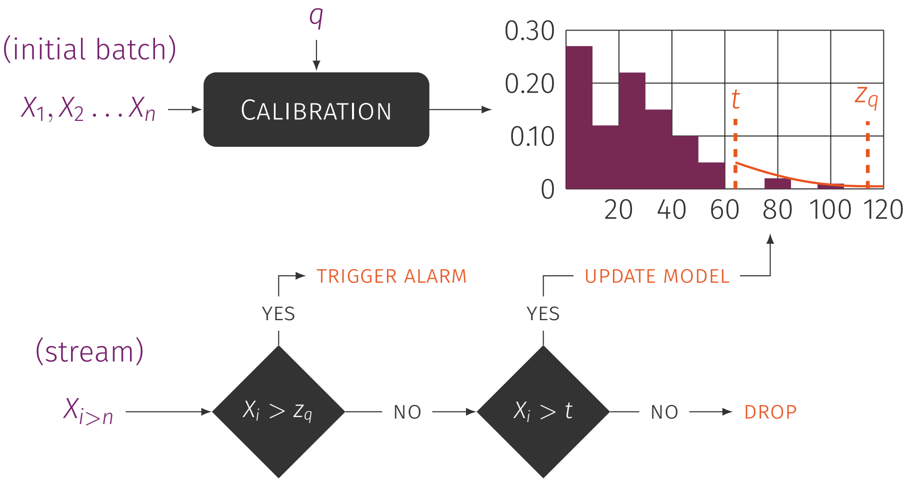
Streaming Peaks-Over-Threshold (SPOT) algorithm



Streaming Peaks-Over-Threshold (SPOT) algorithm



Streaming Peaks-Over-Threshold (SPOT) algorithm

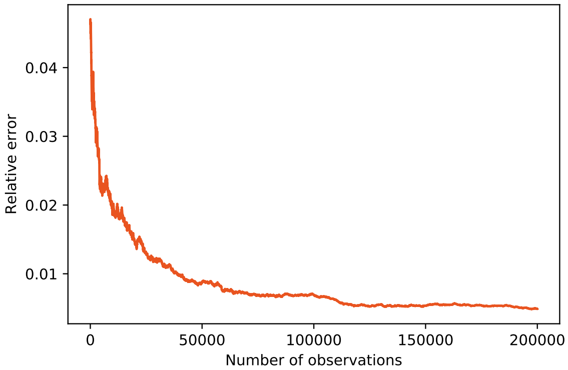


Can we trust that threshold z_q ?

- An example with ground truth : a Gaussian White Noise
  - 40 streams with 200 000 iid variables drawn from $\mathcal{N}(0, 1)$
  - $q = 10^{-3} \Rightarrow$ theoretical threshold $z_{th} \simeq 3.09$

Can we trust that threshold z_q ?

- An example with ground truth : a Gaussian White Noise
  - 40 streams with 200 000 iid variables drawn from $\mathcal{N}(0, 1)$
  - $q = 10^{-3} \Rightarrow$ theoretical threshold $z_{th} \simeq 3.09$
- Averaged relative error



Application to intrusion detection

- Lack of relevant public datasets to test the algorithms …

About the data

- Lack of relevant public datasets to test the algorithms …
- KDD99 ? See [McHugh 2000] and [Mahoney & Chan 2003]

About the data

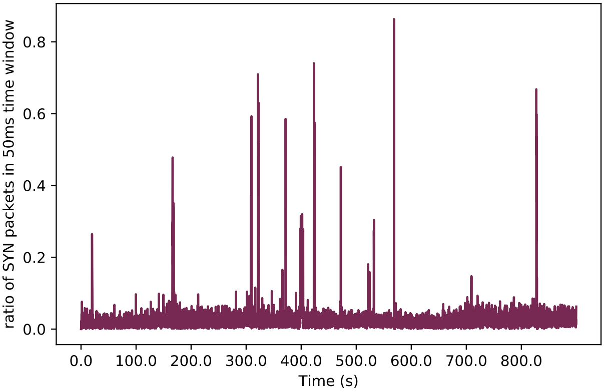
- Lack of relevant public datasets to test the algorithms ...
- KDD99 ? See [McHugh 2000] and [Mahoney & Chan 2003]
- We rather use MAWI
  - 15 min a day of real traffic (.pcap file)
  - Anomaly patterns given by the MAWILab [Fontugne *et al.* 2010] with taxonomy [Mazel et al. 2014]

About the data

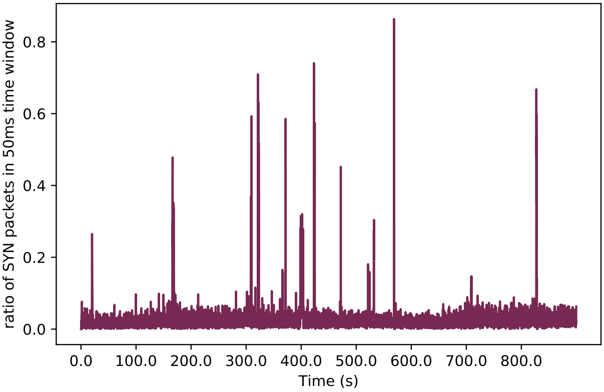
- Lack of relevant public datasets to test the algorithms …
- KDD99 ? See [McHugh 2000] and [Mahoney & Chan 2003]
- We rather use MAWI
  - 15 min a day of real traffic (.pcap file)
  - Anomaly patterns given by the MAWILab [Fontugne *et al.* 2010] with taxonomy [Mazel et al. 2014]
- Preprocessing step : raw .pcap $\rightarrow$ NetFlow format (only metadata)

- The ratio of SYN packets : relevant feature to detect network scan [Fernandes & Owezarski 2009]

- The ratio of SYN packets : relevant feature to detect network scan [Fernandes & Owezarski 2009]



- The ratio of SYN packets : relevant feature to detect network scan [Fernandes & Owezarski 2009]

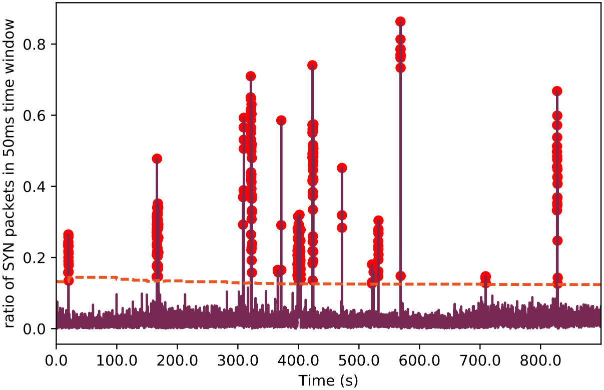


- Goal: find peaks

SPOT RESULTS

- Parameters : $q = 10^{-4}, n = 2000$ (from the previous day record)

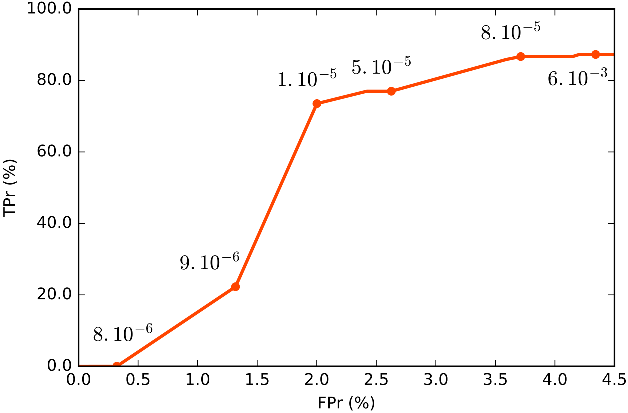
- Parameters : $q = 10^{-4}, n = 2000$ (from the previous day record)



Do we really flag scan attacks ?

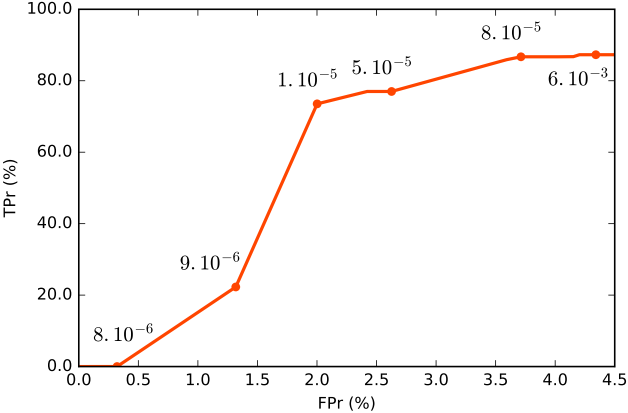
- The main parameter q: a False Positive regulator

- The main parameter q: a False Positive regulator



Do we really flag scan attacks ?

- The main parameter q: a False Positive regulator



- 86% of scan flows detected with less than 4% of FP

A more general framework

- A single main parameter q
  - With a probabilistic meaning $\rightarrow \mathbb{P}(X > z_q) < q$
  - False Positive regulator

SPOT Specifications

- A single main parameter q
  - With a probabilistic meaning $\rightarrow \mathbb{P}(X > z_q) < q$
  - False Positive regulator
- Stream capable
  - Incremental learning
  - Fast ($\sim$ 1000 values/s)
  - Low memory usage (only the excesses)

Other things ?

- SPOT
  - performs dynamic thresholding without distribution assumption
  - uses it to detect network anomalies

Other things ?

- SPOT
  - performs dynamic thresholding without distribution assumption
  - uses it to detect network anomalies
- But it could be adapted to

Other things ?

- SPOT
  - performs dynamic thresholding without distribution assumption
  - uses it to detect network anomalies
- But it could be adapted to
  - compute upper and lower thresholds

Other things ?

- SPOT
  - performs dynamic thresholding without distribution assumption
  - uses it to detect network anomalies
- But it could be adapted to
  - compute upper and lower thresholds
  - other fields

Other things ?

- SPOT
  - performs dynamic thresholding without distribution assumption
  - uses it to detect network anomalies
- But it could be adapted to
  - compute upper and lower thresholds
  - other fields
  - drifting contexts (with an additional parameter) $\rightarrow$ DSPOT

A recent example

A recent example

- Thursday the 9th of February 2017

A recent example

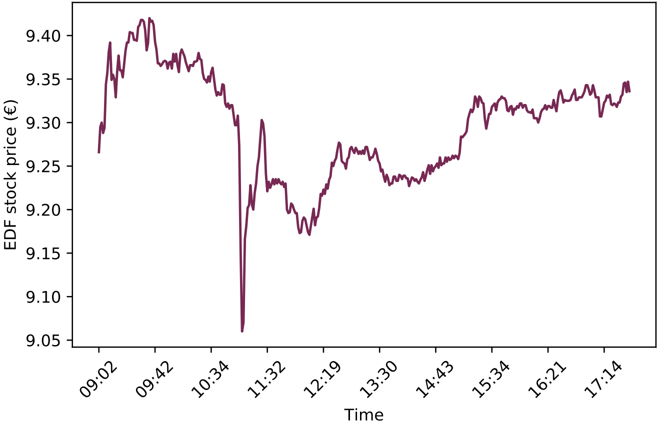
- Thursday the 9th of February 2017
  - **9h** : explosion at Flamanville nuclear plant

A recent example

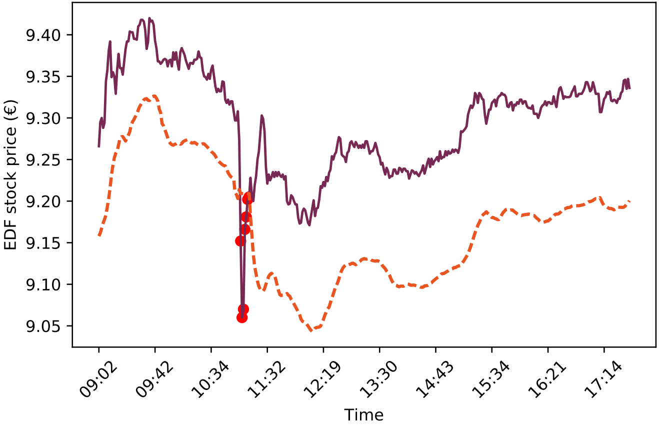
- Thursday the 9th of February 2017
    - **9h** : explosion at Flamanville nuclear plant
    - **11h** : official declaration of the incident by EDF

A recent example

- Thursday the 9th of February 2017
  - **9h** : explosion at Flamanville nuclear plant
  - **11h** : official declaration of the incident by EDF
- What about the EDF stock prices ?


9.40
9.35
9.30
9.25
9.20
9.15
9.10
9.05
EDF stock price (€)
09:02
09:42
10:34
11:32
12:19
13:30
14:43
15:34
16:21
17:14
Time

EDF stock prices



Conclusion

Conclusion

- Context: A great deal of work has been done to develop anomaly detection algorithms

Conclusion

- Context: A great deal of work has been done to develop anomaly detection algorithms
- Problem: Decision thresholds rely on either distribution assumption or expertise

Conclusion

- Context: A great deal of work has been done to develop anomaly detection algorithms
- Problem: Decision thresholds rely on either distribution assumption or expertise
- Our solution: Building dynamic threshold with a probabilistic meaning

Conclusion

- Context: A great deal of work has been done to develop anomaly detection algorithms
- Problem: Decision thresholds rely on either distribution assumption or expertise
- Our solution: Building dynamic threshold with a probabilistic meaning
  - Application to detect network anomalies

Conclusion

- Context: A great deal of work has been done to develop anomaly detection algorithms
- Problem: Decision thresholds rely on either distribution assumption or expertise
- Our solution: Building dynamic threshold with a probabilistic meaning
  - Application to detect network anomalies
  - But a general tool to monitor online time series in a blind way

Conclusion

- Context: A great deal of work has been done to develop anomaly detection algorithms
- Problem: Decision thresholds rely on either distribution assumption or expertise
- Our solution: Building dynamic threshold with a probabilistic meaning
  - Application to detect network anomalies
  - But a general tool to monitor online time series in a blind way
- Future: Adapt the method to higher dimensions