Cyber-insurance: an actuarial approach

Olivier Lopez, Sébastien Farkas (with Maud Thomas)
Chairwoman : Caroline Hillairet

Sorbonne Université, ISUP,
UMR CNRS 8001,
Laboratoire de Probabilités, Statistique et Modélisation,
with the support of Fondation du Risque and AXA Research Fund

18 mars 2019

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Cyber-risk and insurance

- CESIN (annual barometer for cyber-security):
  - 80% of organisations were targeted by at least one cyber-attack in 2018,
  - with damages (like business interruption) more significant than in 2017.

- Cyber-insurance (Opinion Way for CESIN in 2017):
  - 40% of organisations had subscribed (25% in 2016),
  - 37% were in the process of subscribing or willing to (17% in 2016).

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

How to evaluate risks for cyber-insurance ?

- Basics of insurance pricing: if we want to equalize (in expectation) the engagements of the policyholder (premium π) and of the insurer,

$$\pi = E[N]E[X],$$

  where N is the number of claims ($E[N]$ ="frequency") and X is the cost of a claim ("severity"), assuming that X and N are independent.
- If we want to perform reserving, we must understand the whole distribution of N and X.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Adverse selection, moral hazard

- Asymmetry of information: (potential) policyholders have better knowledge of their risk than the insurer.
- More precisely :
  - People who already experienced the risk are more likely to buy insurance.
  - Existence of potential changes of behaviors after subscription (the policyholder replaces a "physical" protection by an insurance protection).
- Also absence of reporting of some claims.

Specificities of cyber-risk

- With cyber-risk, these phenomena are aggravated and are more complex than for traditional risks.
- Examples:
  - absence of reporting of small claims (because they are undetected or with small impact)...
  - ... but also potential absence of reporting of large claims (because of reputation issues).

Specificities of cyber-risk

- With cyber-risk, these phenomena are aggravated and are more complex than for traditional risks.
- Examples:
  - absence of reporting of small claims (because they are undetected or with small impact)...
  - ... but also potential absence of reporting of large claims (because of reputation issues).
- Changes through time in the reporting behavior, due to regulation and evolutions in the perception of the risk.

Specificities of cyber-risk

- With cyber-risk, these phenomena are aggravated and are more complex than for traditional risks.
- Examples:
  - absence of reporting of small claims (because they are undetected or with small impact)...
  - ... but also potential absence of reporting of large claims (because of reputation issues).
- Changes through time in the reporting behavior, due to regulation and evolutions in the perception of the risk.
- Missing information is particularly important because of the relative lack of data on a new risk.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Data

- Internal data must be completed by external data because of the novelty of the risk.
- Problem: the number of registered events in the database depends on the exposure.
- In other words, increase of the number of cyber events may be caused by the increase of entities reporting, but not by a true evolution of the risk.

Data

- Internal data must be completed by external data because of the novelty of the risk.
- Problem: the number of registered events in the database depends on the exposure.
- In other words, increase of the number of cyber events may be caused by the increase of entities reporting, but not by a true evolution of the risk.
- Exposure is hard to track for public databases.

Outline

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Outline

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Presentation of the PRC database

- Aim to raise awareness about privacy issues
- Chronology of data breaches maintained from 2005
- Focus on the number of records affected: can be known or unknown
- Gathering events information from multiple sources:

Source	Share in PRC	Share in PRC known
Nonprofit organization	39%	37%
US GA - HIPAA	27%	36%
US GA - State	22%	17%
Media	12%	10%
Total	8860	6641

- 25% of data breaches have an unknown number of records

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Data contained in the PRC database

Exposition data	Organization's name
	Organization's type
	Organization's geographic position
Event data	Release's source
	Release's date
	Breach's type
	Number of records affected
	Description of the event

Data contained in the PRC database

Type of organization
Banking and Insurance
Retail (including online)
Education
Government
Healthcare
Business (Other)
Nonprofit organization

Type of breach
Hacking or Malware
Physical Loss
Portable Device
Stationary Device
Unintended Disclosure
Payment Card Fraud
Insider

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Cost of data breaches

Main costs:

- Incident investigation cost
- Customer notification and regulatory sanctions cost
- Post data breach response
- Loss business

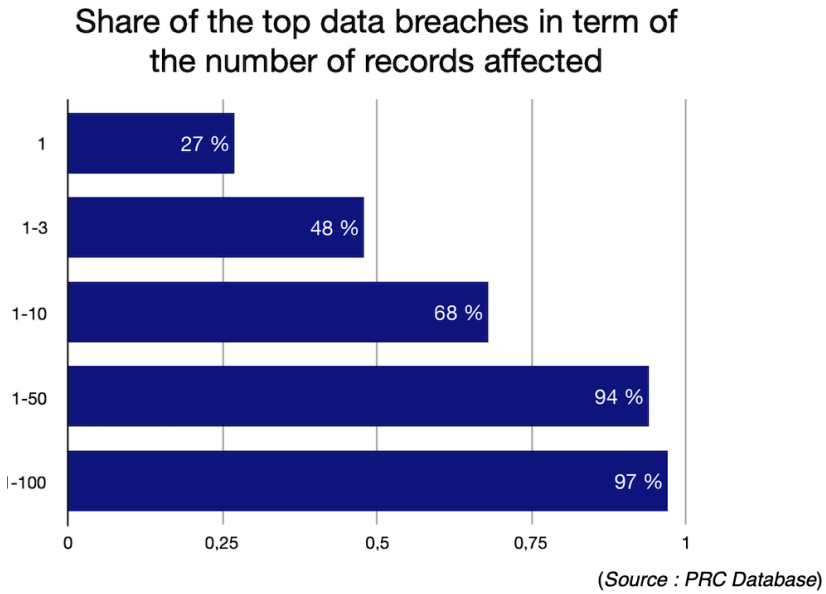
Assessment of costs:

- Ponemon Institute LLC
  - Data collection
  - Yearly report: *"Cost of a data breach study"* (CODB)
- Jacobs' formula
  - Based on 2013 and 2014 CODB reports
  - $log(Costs) = 7.68 + 0.76 \times log(Records)$

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Volatility of cyber-events

- An important disparity exists between cyber-events.
- Illustration on the PRC database:



(*Source : PRC Database*)

- Also necessity to plan for events even worse than what was already observed.

Extreme value theory

Generalized Pareto Distribution

A random variable X with Generalized Pareto Distribution of parameters (γ, σ) is characterized by its survival function

$$S_{\gamma,\sigma}(x) = \mathbb{P}(X \geq x) = \begin{cases} \frac{1}{\left(1+\frac{x\gamma}{\sigma}\right)^{1/\gamma}} & , \quad \gamma \neq 0 \\ \exp\left(-\frac{x}{\sigma}\right) & , \quad \gamma = 0 \end{cases} .$$

Extreme value theory

Generalized Pareto Distribution

A random variable X with Generalized Pareto Distribution of parameters (γ, σ) is characterized by its survival function

$$S_{\gamma,\sigma}(x) = \mathbb{P}(X \geq x) = \begin{cases} \frac{1}{\left(1+\frac{x\gamma}{\sigma}\right)^{1/\gamma}} & , \quad \gamma \neq 0 \\ \exp\left(-\frac{x}{\sigma}\right) & , \quad \gamma = 0 \end{cases}.$$

Approximation beyond a threshold:
Pickands (1975): for a random variable Y, let $S_{u_n}(y) = \mathbb{P}(Y \geq y | Y \geq u_n)$, there exists (γ, σ_n) such that

$$\lim_n |S_{u_n}(y) - S_{\gamma,\sigma_n}(y)| = 0,$$

where u_n tends towards $\tau_S = \sup\{x : S(x) > 0\}$.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Analysis on extreme cyber events


Edwards B., Hofmeyr S. and Forrest S. (2016), "Hype and heavy tails: A closer look at data breaches." *Journal of Cybersecurity*, vol. **2** (2057-2085), pp. 3-14.


Eling M. and Loperfido N. (2017), "Data breaches: Goodness of fit, pricing, and risk measurement." *Insurance: Mathematics and Economics*, vol. **75** (0167-6687), pp. 126-136.

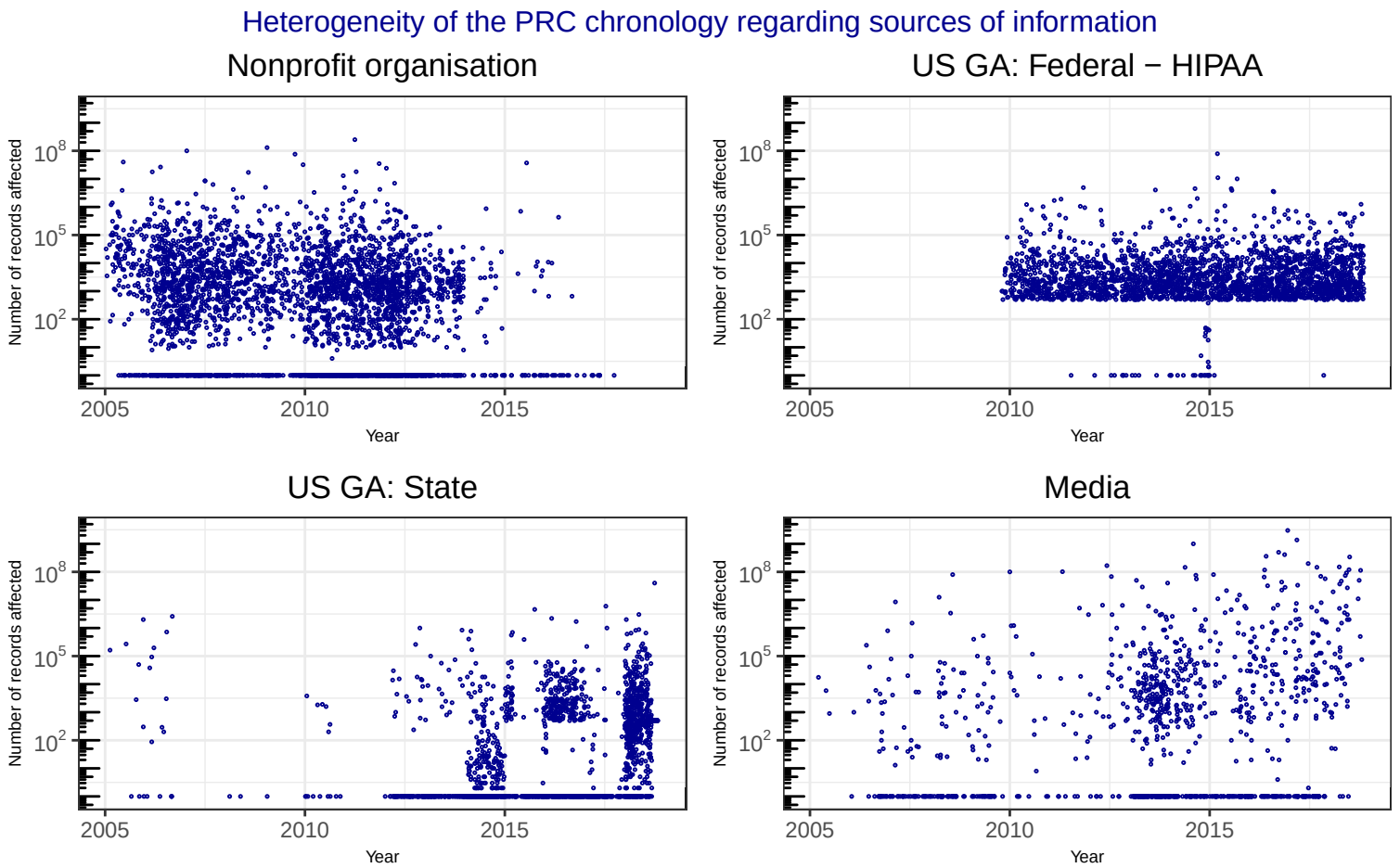

Wheatley S., Maillart T. and Sornette D. (2016), "The extreme risk of personal data breaches and the erosion of privacy." *European Physical Journal B*, vol. **89** (1434-6036), pp. 7.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Outline

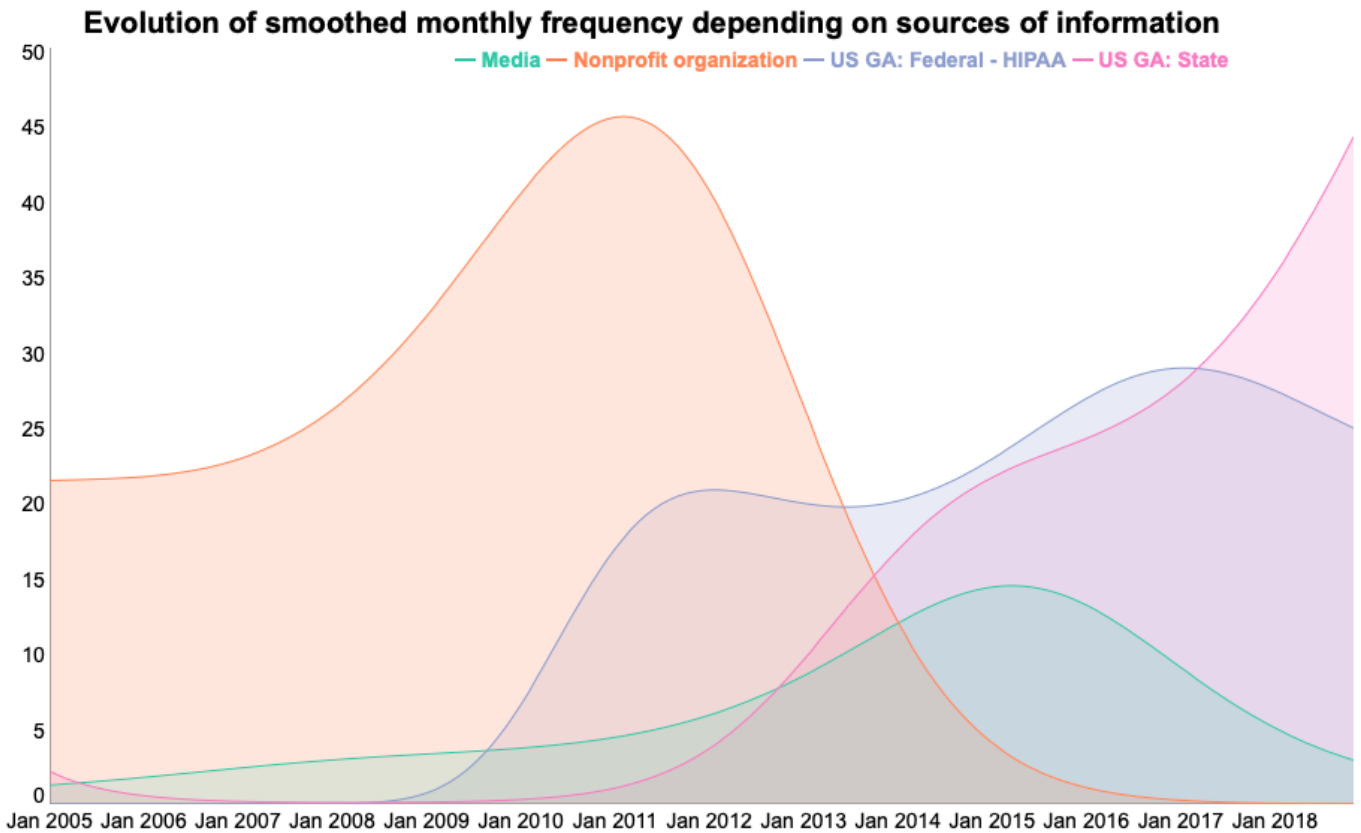
SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Heterogeneity of the PRC chronology



Evolution of the frequency

- The heterogeneity caused by the evolution of data collection makes difficult temporal evolution.



Frequency analysis

- How many times a company have been affected by a data breach ?
- Work with count data
- The 8860 events have affected 7737 different companies

Total	1	2	3	4	5	6	7	8	9	10	11	12
7737	6929	634	103	41	12	6	7	3	1	0	0	1

Issues:

- 0-truncated data because of the PRC chronology can be viewed as a claim database
- 1-inflated data: a lot of companies have been only affected one time

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

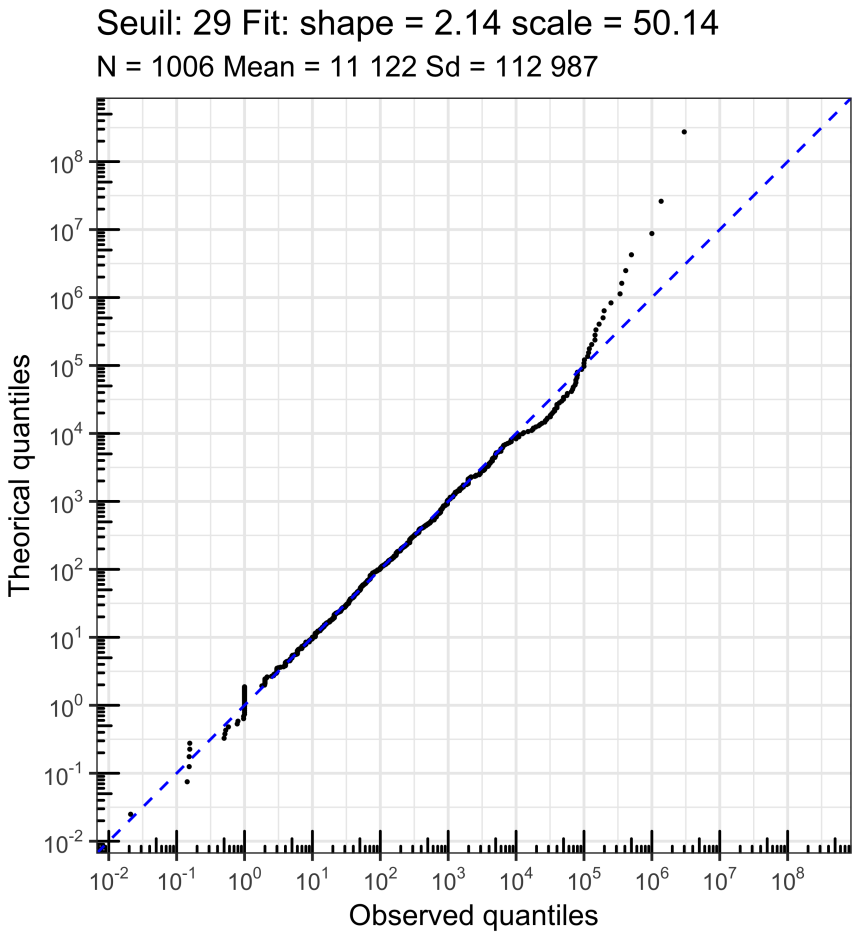
Hypotheses and results

- Restriction with count data above 2
- 8 years window of data because of the instability of sources
- Risk factor: the type of organization
- We deduce yearly rate

Type of organization	λ
Unknown	0.27
Education	0.23
Banking and Insurance	0.15
Business (Other)	0.13
Retail (including online)	0.13
Health	0.10
Government	0.07
Nonprofit organization	0.01

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

QQ-plots


Seuil: 29 Fit: shape = 2.14 scale = 50.14
N = 1006 Mean = 11 122 Sd = 112 987
Theorical quantiles
Observed quantiles
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5} 10^{6} 10^{7} 10^{8}

Clustering And Regression Trees (CART)

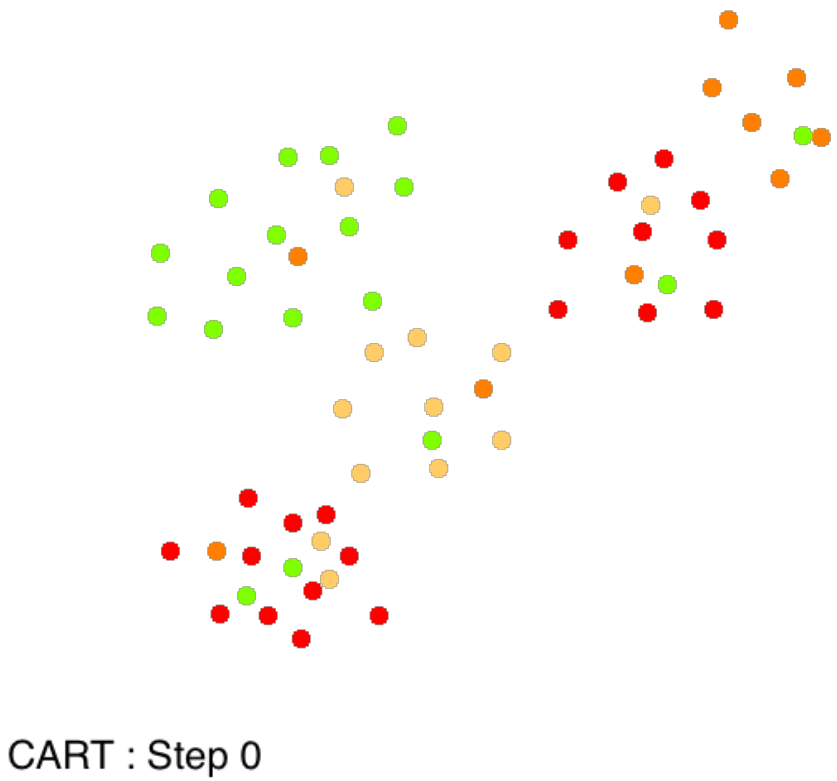
- Introduced by Breiman (1984). Many extensions: Su (2004), Hothorn (2006), Loh (2014),...
- Consider a random variable Y (here the "size" of the breach) and X some covariates.
- Regression trees:
  - combining clustering with regression (that is evaluation of the impact of covariates on a variable).
  - regression trees aim to estimate a function $m(x)$ (characterizing the distribution of Y when $X = x$, for example $m(x) = E[Y|X = x]$) by

$$\hat{m}(x) = \sum_{j=1}^{K} m_j R_j(x),$$

where R_j are called a "rule," that is $R_j(x) = 0$ or 1, and, for all x, only one $R_j(x)$ is nonzero.

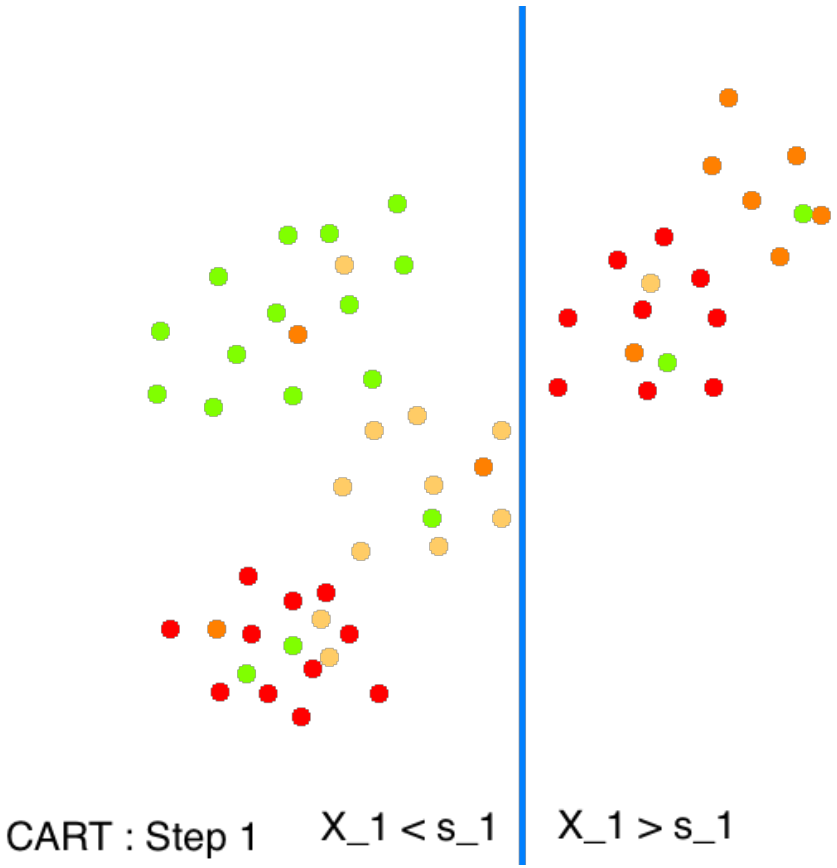
SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

A few words about the algorithm

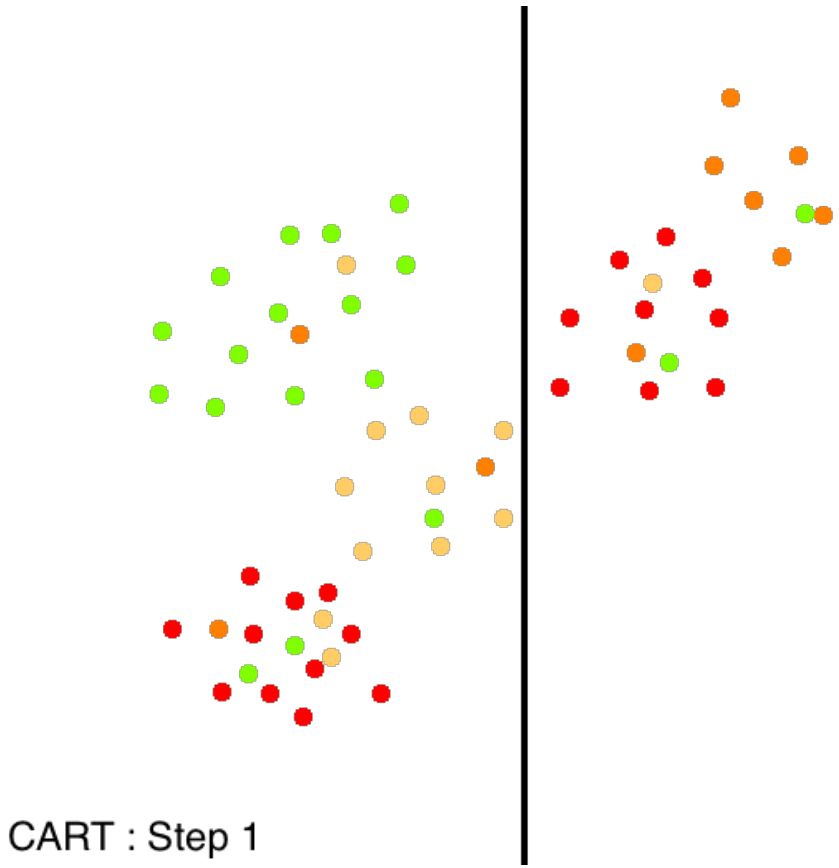


CART : Step 0

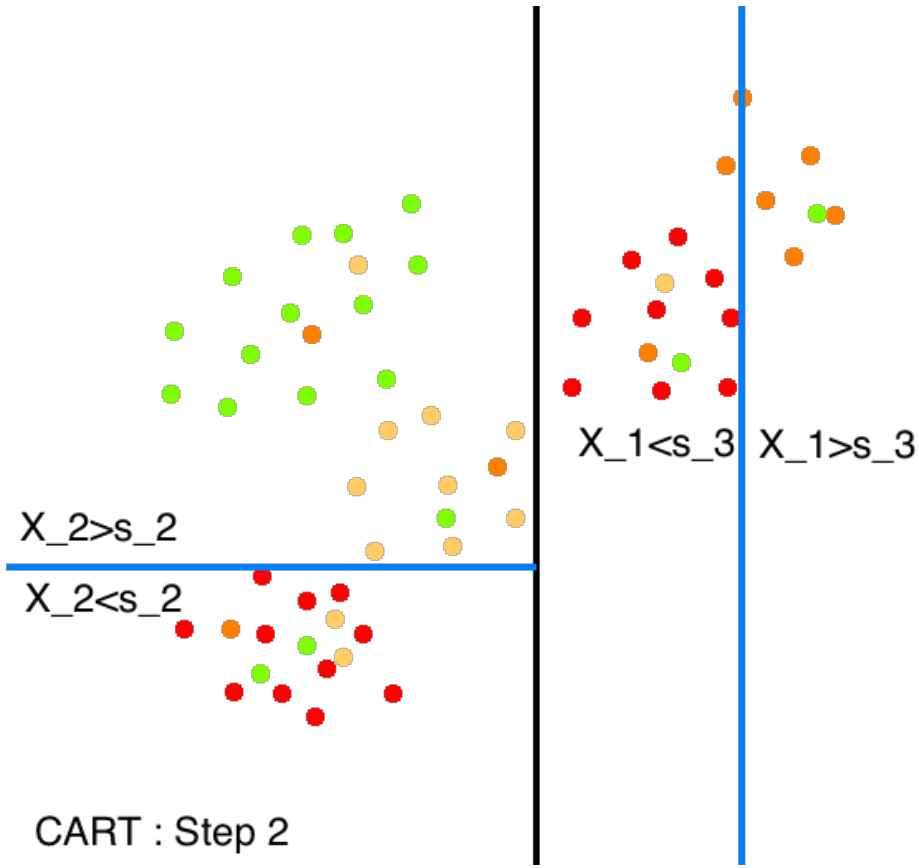
A few words about the algorithm



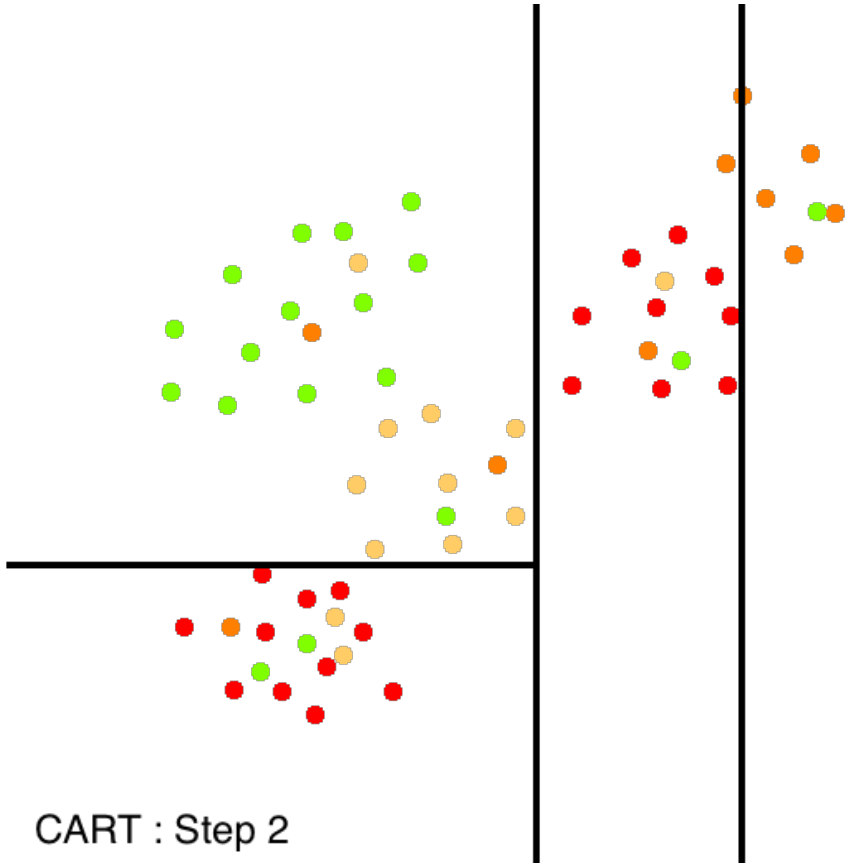
A few words about the algorithm



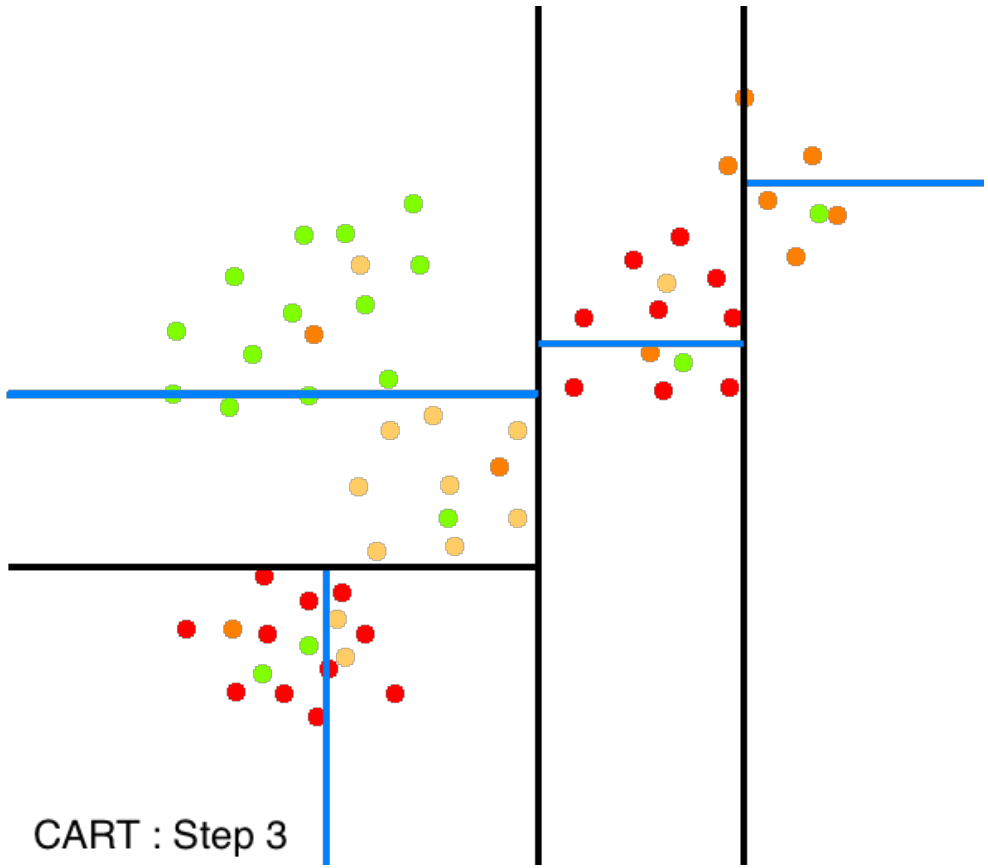
A few words about the algorithm



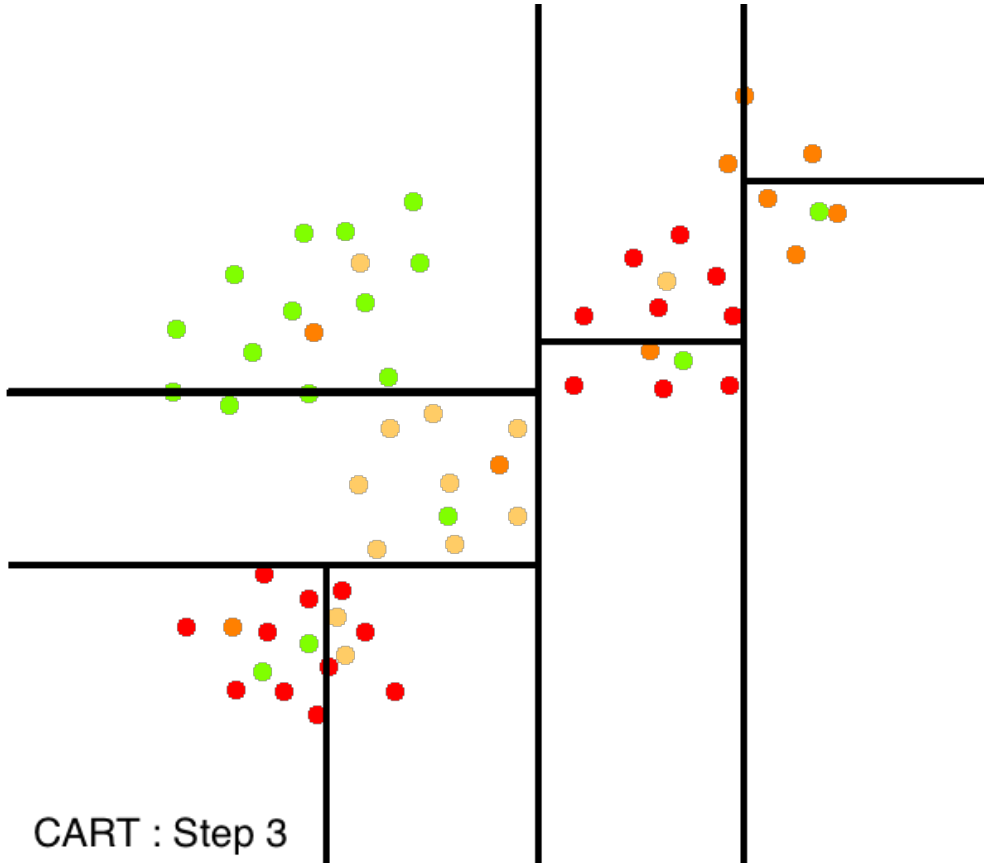
A few words about the algorithm



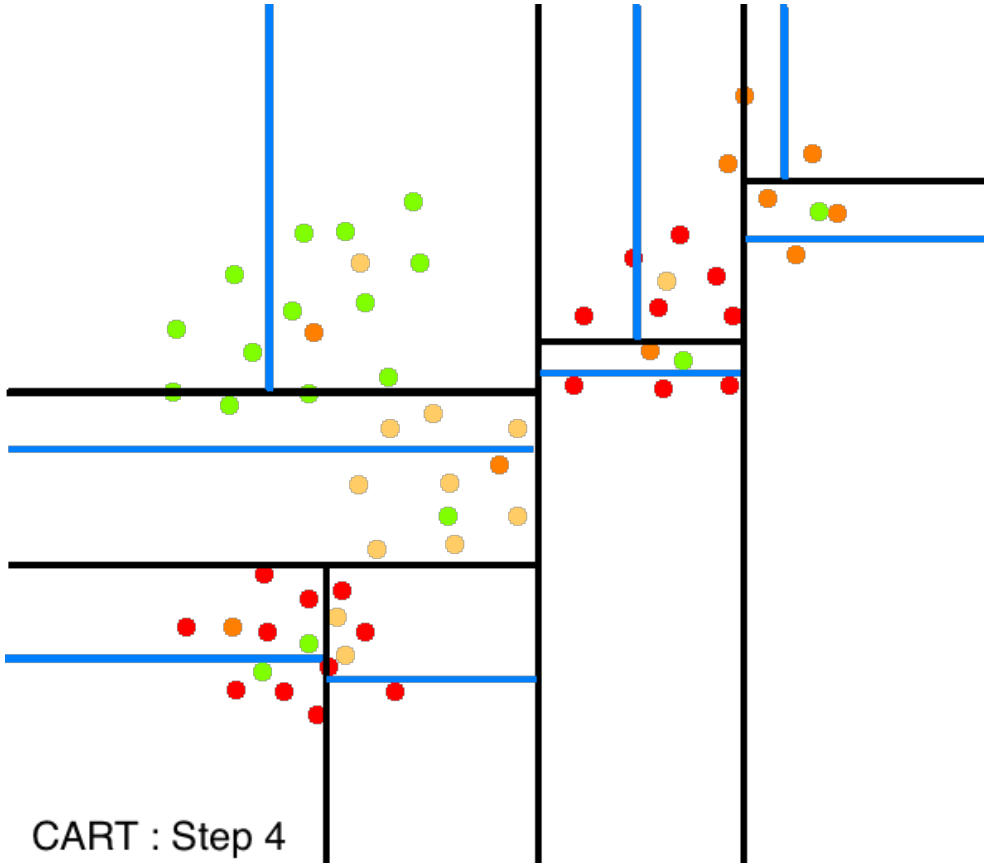
A few words about the algorithm



A few words about the algorithm



A few words about the algorithm



The splitting rule

- Most classical version of regression trees use a "quadratic" splitting rule, that is, we look for a split such that, the following quantity is minimal:

$$\sum_{j=1}^{K} \sum_{i:R_j(X_i)=1} (Y_i - \bar{Y}_j)^2,$$

  where $\bar{Y}_j$ is the mean-value of members of group R_j.
- Such a procedure produces an estimation of the target $m(x) = E[Y|X = x]$.

The splitting rule

- Most classical version of regression trees use a "quadratic" splitting rule, that is, we look for a split such that, the following quantity is minimal:

$$\sum_{j=1}^{K} \sum_{i:R_j(X_i)=1} (Y_i - \bar{Y}_j)^2,$$

  where $\bar{Y}_j$ is the mean-value of members of group R_j.
- Such a procedure produces an estimation of the target $m(x) = E[Y|X = x]$.
- Potential problem: quadratic loss is very sensitive to high volatility.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

The splitting rule

- Most classical version of regression trees use a "quadratic" splitting rule, that is, we look for a split such that, the following quantity is minimal:

$$\sum_{j=1}^{K} \sum_{i:R_j(X_i)=1} (Y_i - \bar{Y}_j)^2,$$

  where $\bar{Y}_j$ is the mean-value of members of group R_j.
- Such a procedure produces an estimation of the target $m(x) = E[Y|X = x]$.
- Potential problem: quadratic loss is very sensitive to high volatility.
- Changing the splitting rule changes the target function and can also be used to improve robustness of the estimation.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Median trees and GPD Trees

- Replace the quadratic loss by the absolute loss leads to estimate the conditional median, that is the quantity $med(x)$ such that

$$\mathbb{P}(Y \leq med(x)|X = x) = \mathbb{P}(Y \geq med(x)|X = x) = 1/2.$$

- Median regression is less influenced by large observations.

Median trees and GPD Trees

- Replace the quadratic loss by the absolute loss leads to estimate the conditional median, that is the quantity $med(x)$ such that

$$\mathbb{P}(Y \leq med(x)|X = x) = \mathbb{P}(Y \geq med(x)|X = x) = 1/2.$$

- Median regression is less influenced by large observations.
- We used median regression trees to analyze the "center" of the distribution.

Median trees and GPD Trees

- Replace the quadratic loss by the absolute loss leads to estimate the conditional median, that is the quantity $med(x)$ such that

$$\mathbb{P}(Y \leq med(x)|X = x) = \mathbb{P}(Y \geq med(x)|X = x) = 1/2.$$

- Median regression is less influenced by large observations.
- We used median regression trees to analyze the "center" of the distribution.

- To look at the tail of the distribution, we use GPD-log likelihood as a splitting rule.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Median trees and GPD Trees

- Replace the quadratic loss by the absolute loss leads to estimate the conditional median, that is the quantity $med(x)$ such that

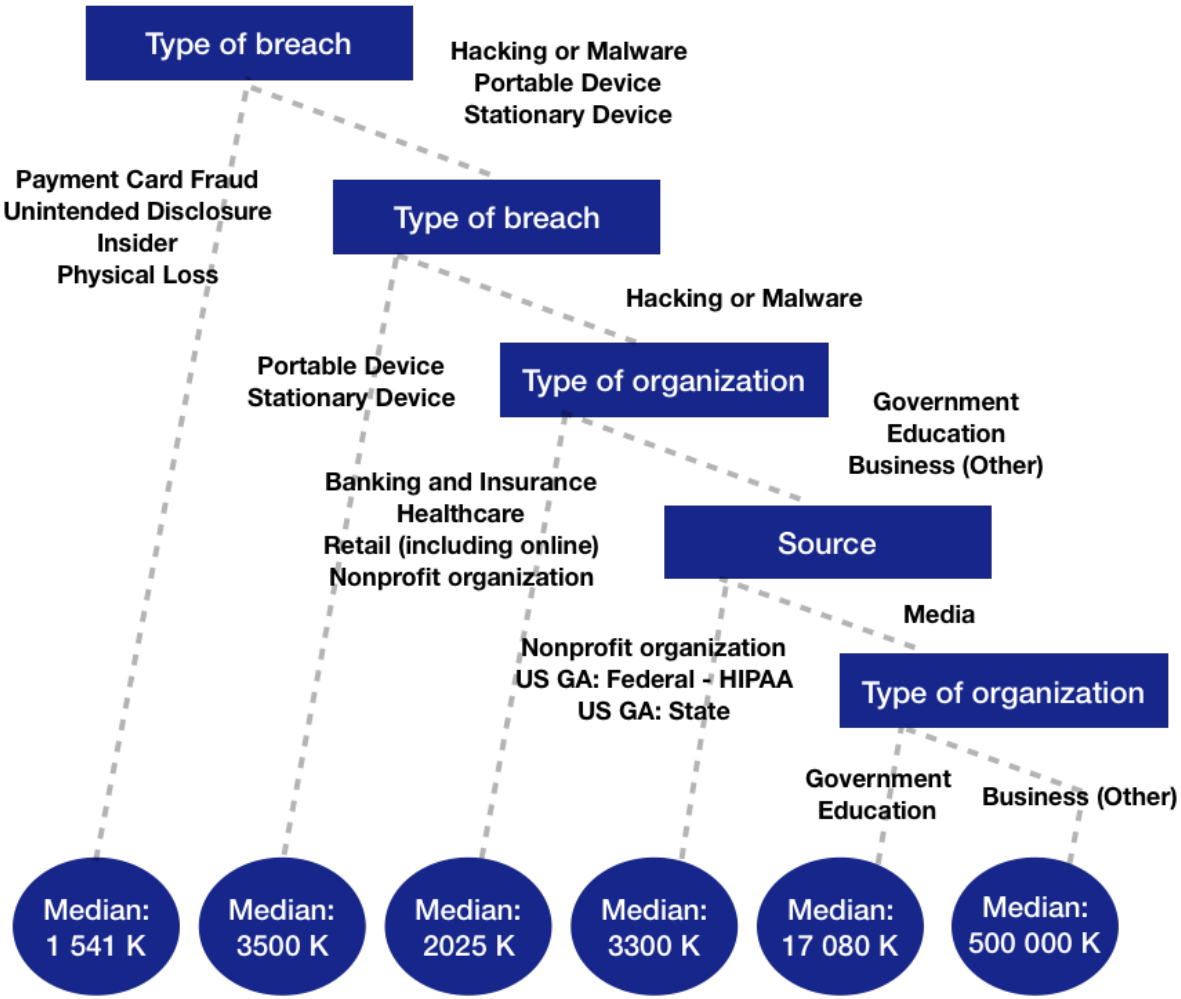
$$\mathbb{P}(Y \leq med(x)|X = x) = \mathbb{P}(Y \geq med(x)|X = x) = 1/2.$$

- Median regression is less influenced by large observations.
- We used median regression trees to analyze the "center" of the distribution.

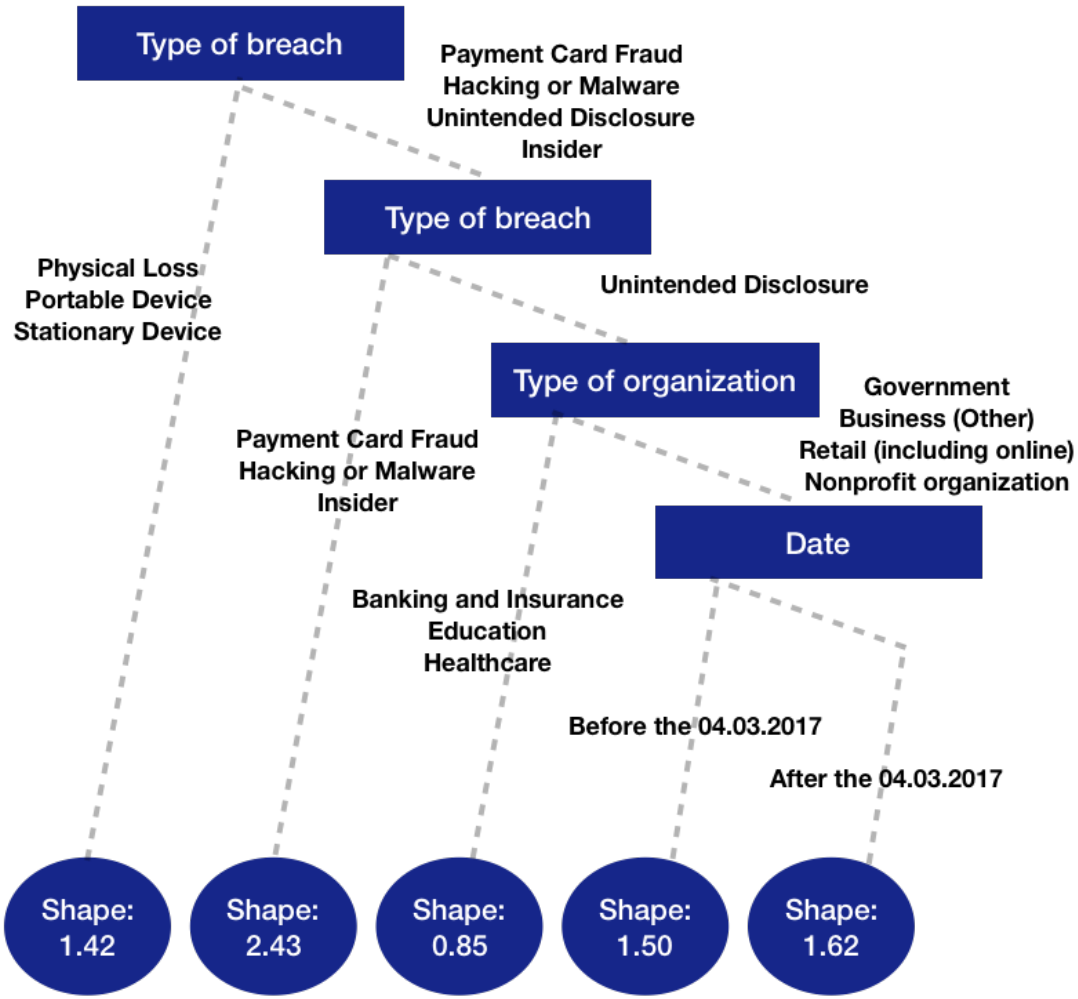
- To look at the tail of the distribution, we use GPD-log likelihood as a splitting rule.
- In this way, the tail of the distribution is approximated by a GPD, with parameters depending on x.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

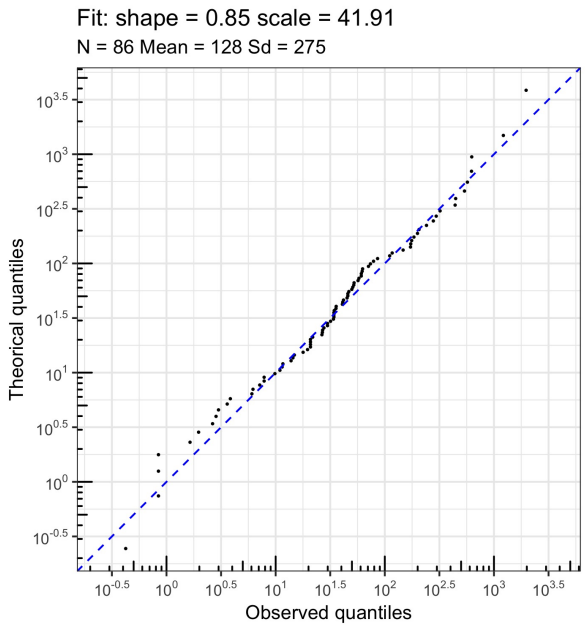
Median tree illustration

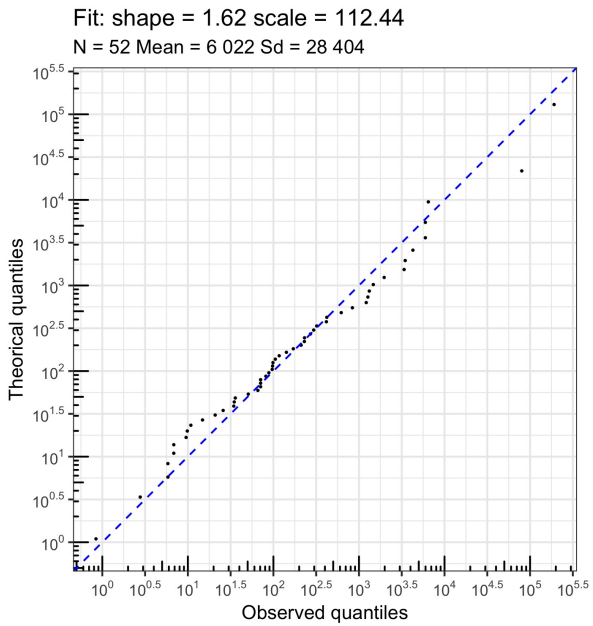


GPD tree illustration



QQ-plots on some GPD tree leafs





Conclusion

- The study we developed is illustrated on a real database, but the difficulties are common to cyber events dataset.
- Important difficulty with public databases: determining the exposure.

Conclusion

- The study we developed is illustrated on a real database, but the difficulties are common to cyber events dataset.
- Important difficulty with public databases: determining the exposure.
- We identified instability and heterogeneity in estimating the frequence due to a changing exposure.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Conclusion

- The study we developed is illustrated on a real database, but the difficulties are common to cyber events dataset.
- Important difficulty with public databases: determining the exposure.
- We identified instability and heterogeneity in estimating the frequence due to a changing exposure.
- A similar impact (less obvious) is identified on the severity...

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Conclusion

- The study we developed is illustrated on a real database, but the difficulties are common to cyber events dataset.
- Important difficulty with public databases: determining the exposure.
- We identified instability and heterogeneity in estimating the frequence due to a changing exposure.
- A similar impact (less obvious) is identified on the severity...
- ... although not so obvious for large events.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Thank you for your attention!

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257