Cyber Risk Analysis, with a Focus on Extremes

Marie Kratz

ESSEC Business School

CREAR ESSEC Research Center on Risk

http://crear.essec.edu

kratz@essec.edu

Workshop on Insurance and Financial Mathematics:
Cyber Risk and Insurance

Hannover, June 08, 2023

Agenda

1 Cyber Risk Landscape

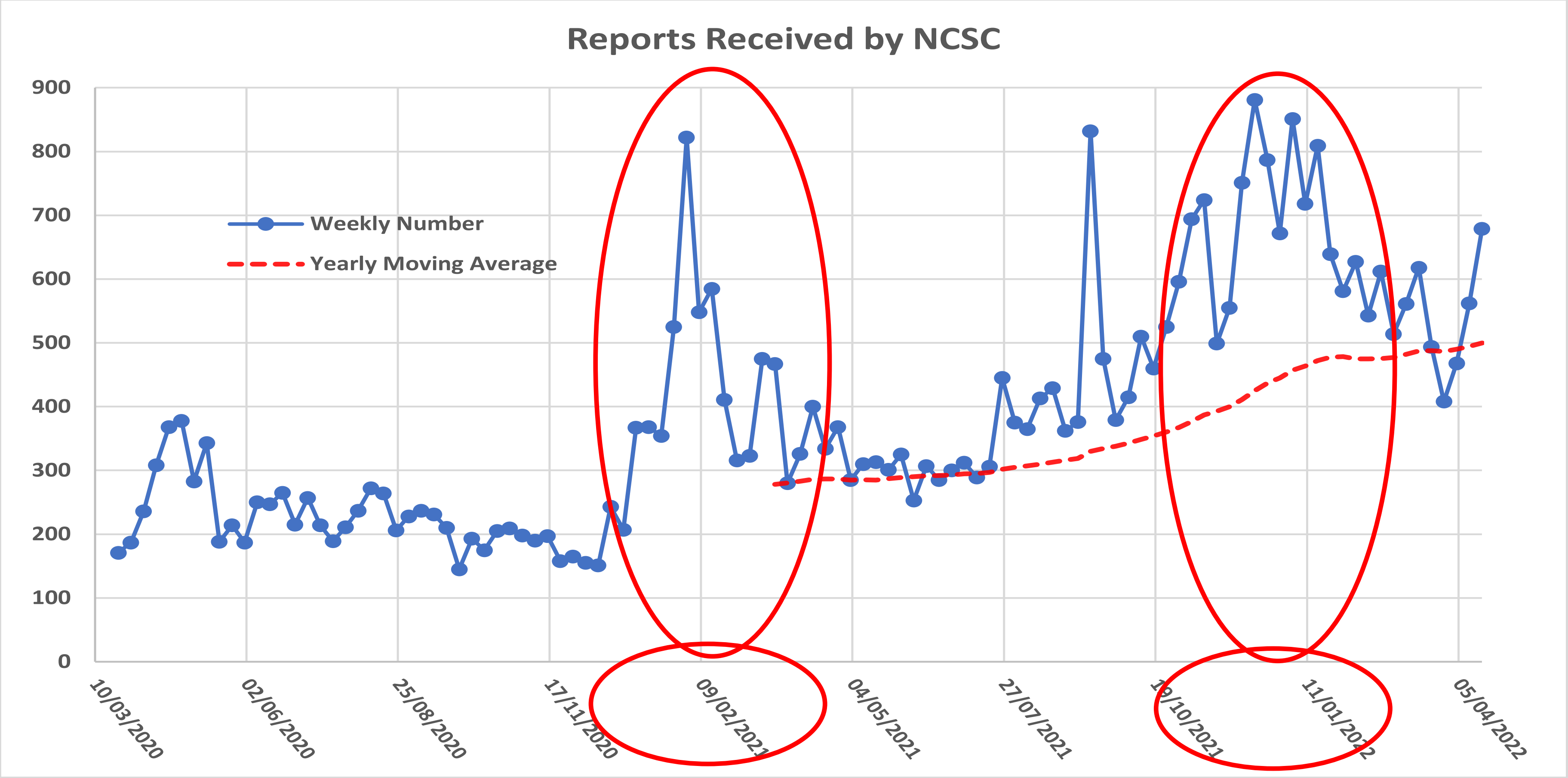
2 Statistical Analysis: The GN Cyber Complaints Database

3 Probabilistic Modelling

4 Conclusion

The Lockdowns during the COVID Pandemics Have Exacerbated Attacks

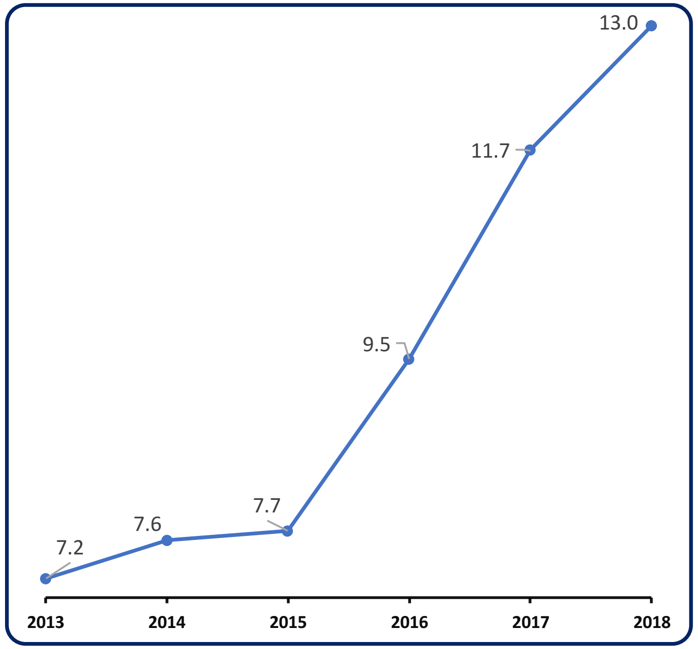
Swiss National Cyber Security Centre NCSC - Announcements per week



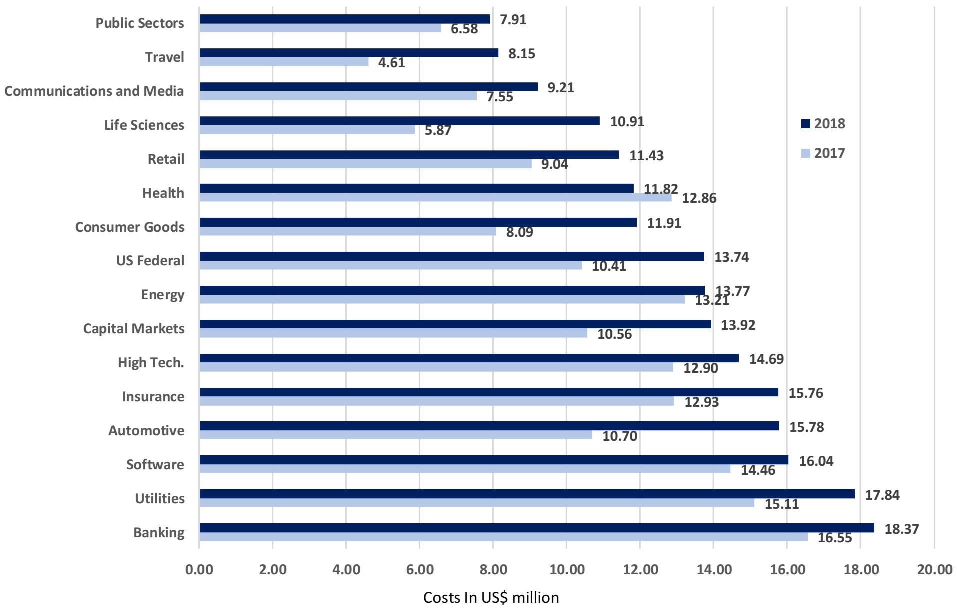
Source: https://www.ncsc.admin.ch/ncsc/en/home/aktuell/aktuelle-zahlen.html

Cyber Crime Costs are on Exponential Growth with Worsening Financial Consequences

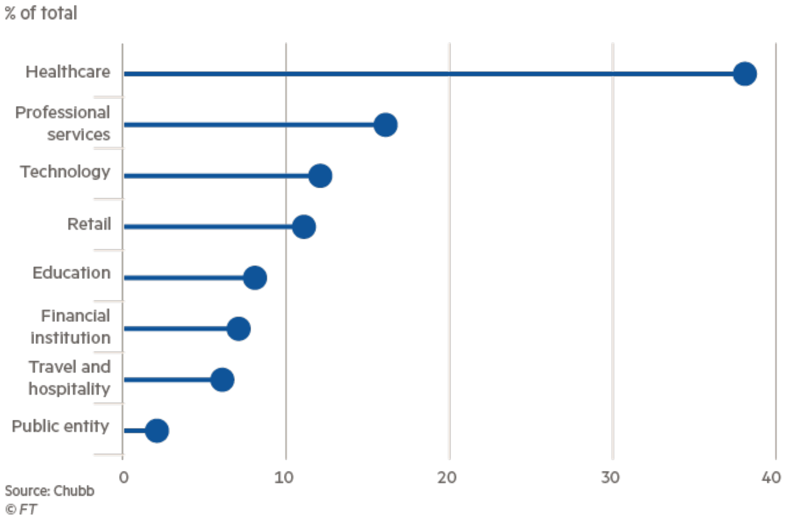
Average cost of cyber crime over 6 years ($M)



Average annualized cost by industry sector ($M)



Top Cyber Claims by Industry over the Past Decade



Source: Chubb
© FT

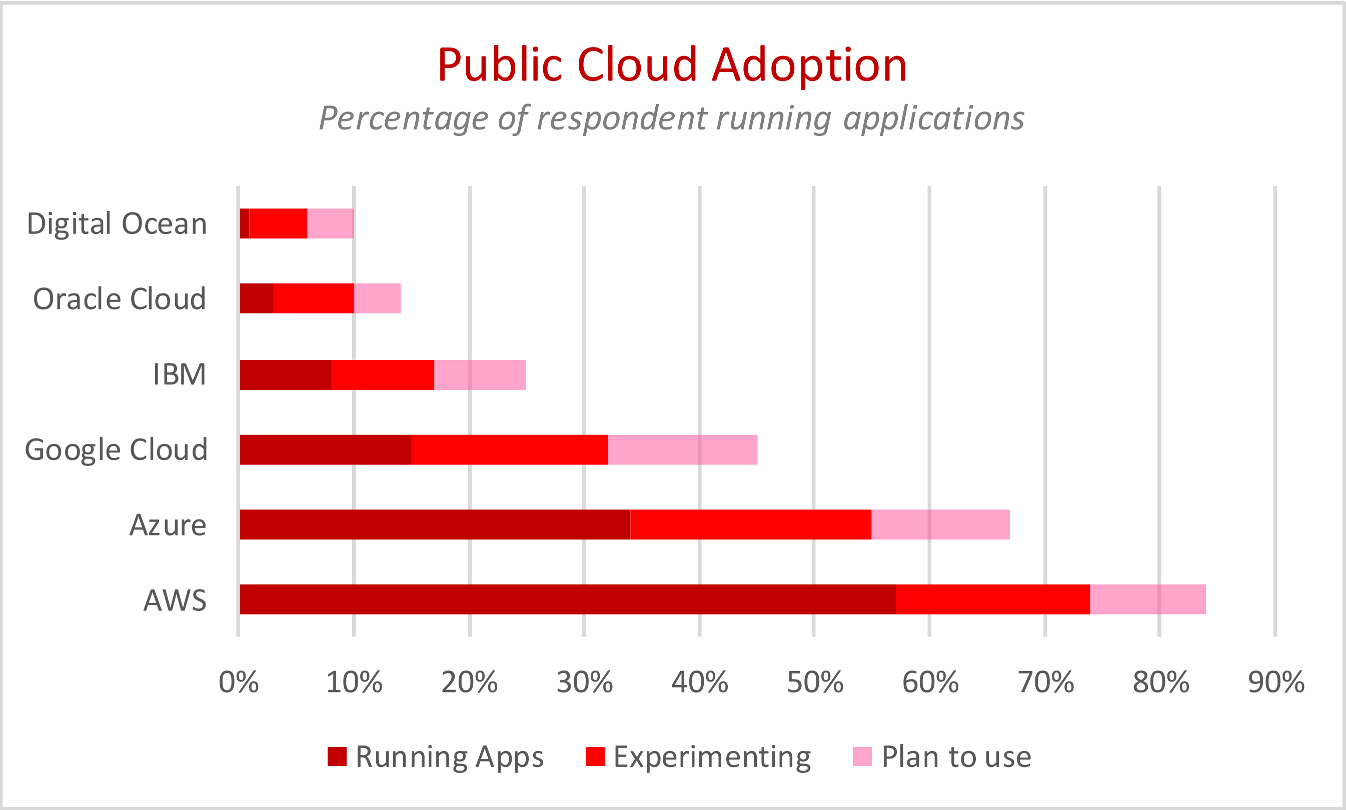
Source: Accenture, Survey from 355 Companies in 11 countries (Australia, Brazil, Canada, France, Germany, Italy, Japan, Singapore, Spain, the United Kingdom and the United States) – 2019

- There has been a 60% increase in the average ransom payment (US$178,254) from the 1st quarter to the 2nd quarter of 2020*

- Insurance claims and cost per industry do not coincide

*) Accenture, 2020 Cyber Threatscape Report

A Highly Connected World



- □ Interviews of companies show that more and more of them use *cloud services*. Amazon Web Services (AWS) is one of the largest cloud providers
- □ In the hypothetical case of an *attack against AWS*, *many companies* using AWS would be *affected*

Source Geneva Association, Cyber Insurance as a Risk Mitigation Strategy, April 2018

From IT-Protection towards Cyber-Resilience

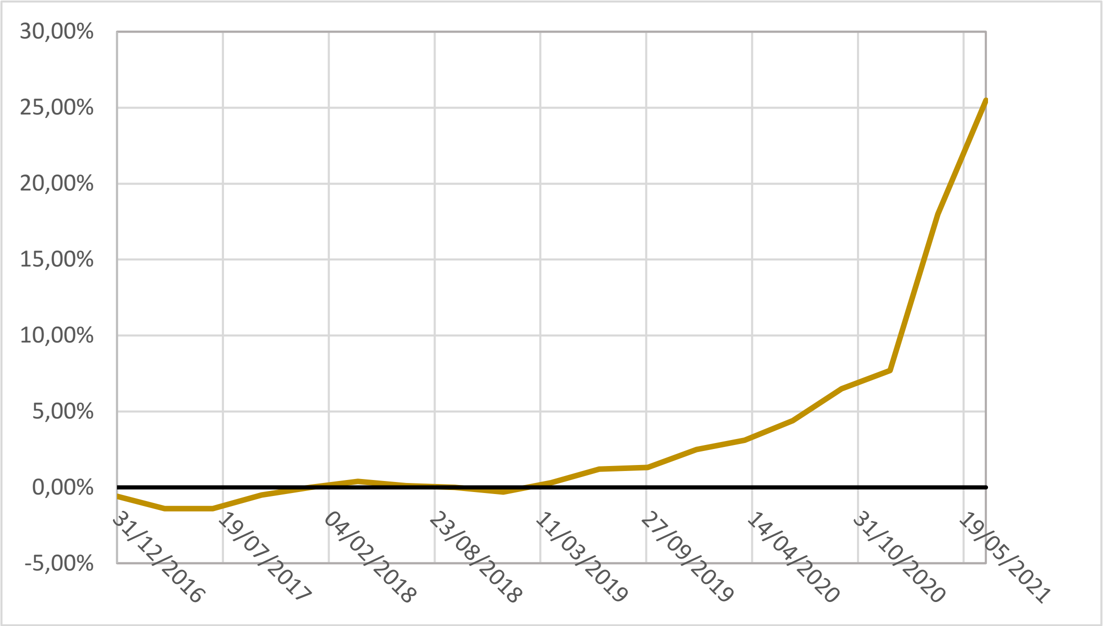
❑ Because there is no 100% security, businesses need to move towards *cyber-resilience* combining security measures, risk management and hedging instruments

❑ In this context, *insurance* will play an essential role to provide products that will help companies becoming more resilient

Cost of Insuring Against Cyber Risk is Growing

- ❑ In the US, following the increase in insured losses, insurance premiums are growing fast*
- ❑ *Demand* in insurance is *increasing*, while the *offer* is *retracting*; Allianz announced that they refuse ¾ of the demands because of lack of cyber security
- ❑ Insurers demand a high level of security before granting an insurance cover

Quarterly Changes of US Cyber Insurance Premium from Q4-2016



*) source: https://blog.alta.org/2021/09/cyber-coverage-premiums-increase-25-survey-shows.html
survey from the Council of Insurance Agents & Brokers (CIAB)

Cyber risk – characteristics*

❑ Increase of the frequency of attacks but also in their financial consequences (Netdiligence, 2022);

❑ High speed of changes in the risk landscape (stress scenarios);

❑ Targets of attacks often intangibles (data, reputation, political elections), so insurers limit payments for those;

❑ Potential of systemic failures due to attacks (highly connected world of IT systems); extreme risk

*) M. Dacorogna and M. Kratz (2023), Managing Cyber Risk, a Science in the Making. *Scandinavian Actuarial Journal*

Agenda

1 Cyber Risk Landscape

2 Statistical Analysis: The GN Cyber Complaints Database

3 Probabilistic Modelling (EVT)

4 Conclusion

Evaluating Quantitatively Cyber Risk:
The case of the GN database on cyber attacks

- ❑ Research Collaboration between:

  - the Center of Research in Econo-finance and Actuarial science on Risk – CREAR – of ESSEC Business School (Paris – Singapore)

  - and the SCRC (Service Central du Renseignement Criminel / Central Criminal Intelligence Service) of PJGN - Pôle Judiciaire de la Gendarmerie Nationale (Lieutenant Colonel Jérôme Barlatier)

- ❑ GN Database: data registered for the complaint:

1) Reporting date 2) Amount of damage 3) Date of birth of the victim 4) Victim gender

5) Category of the offence (GN) 6) Natinf (categorization by the Ministry of Justice)

Available in the database but not for this research (anonymization):

7) Location 8) Written description of the complaint (by the detective)

Disclaimer - Publications

Disclaimer: *The PJGN database we used for this study has been entrusted by the Gendarmerie under confidentiality agreement. Use and interpretation are the strict responsibility of the authors. As required by Gendarmerie Nationale, any communication on this study should mention that the source is from "Gendarmerie Nationale – PJGN – treated by ESSEC-CREAR".*

Publications:

- **Building up Cyber Resilience by Better Grasping Cyber Risk: A New Algorithm for Modelling Cyber Complaints Filed at the Gendarmerie Nationale***.* M. Dacorogna, N. Debbabi, M. Kratz. ***European Journal of Operational Research 2023*** (online)
- **Managing Cyber Risk, a Science in the Making**. M. Dacorogna and M. Kratz. ***Scandinavian Actuarial Journal 2023*** (Invited paper; online with open access until January 2024)
- **Moving from Uncertainty to Risk: The Case of Cyber Risk***.* M. Dacorogna and M. Kratz (2020). Chapter in "Cybersecurity in Humanities and Social Sciences" Ed. By H. Loiseau, D. Ventre and H. Aden, **ISTE SCIENCE PUBLISHING**, Montreal

Goals of the Study

- □ *Understanding of the data* (complaints by victims of cyber crimes – individual and companies)

- □ Statistical data exploration: another way to correct the database. Creation of a *benchmark reliable dataset*

- □ *Data Analytics* of selected variables in view of building predictive probabilistic & statistical models

- □ Insurability of cyber risk

Data Sampling and Description

- 208’037 complaints data from 07/2015 to 04/2019

Amount	Data Number	% of the total sample size
ND or 0	147’052	70.69%
< 500 €	29’074	13.97%
≥ 500 €	31’911	15.34%

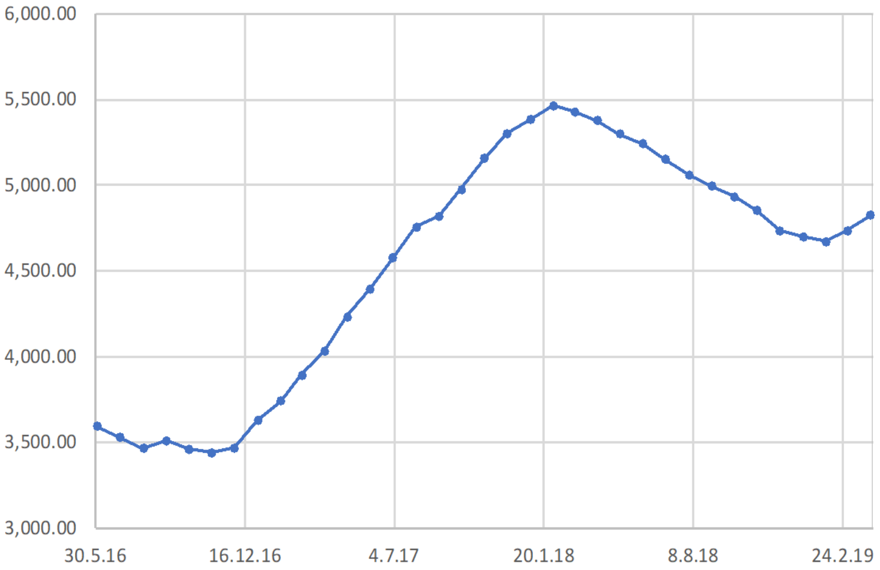
Gender	Data Number	% of the total sample size
F	91’599	44.03%
M	92’202	44.32%
ND	24’236	11.65%

- Damages classified **by type**: the first classes the most represented among the full sample

Type	% of the full sample
Fraud	59.4%
Identity theft	4.7%
Breach of trust	3.5%

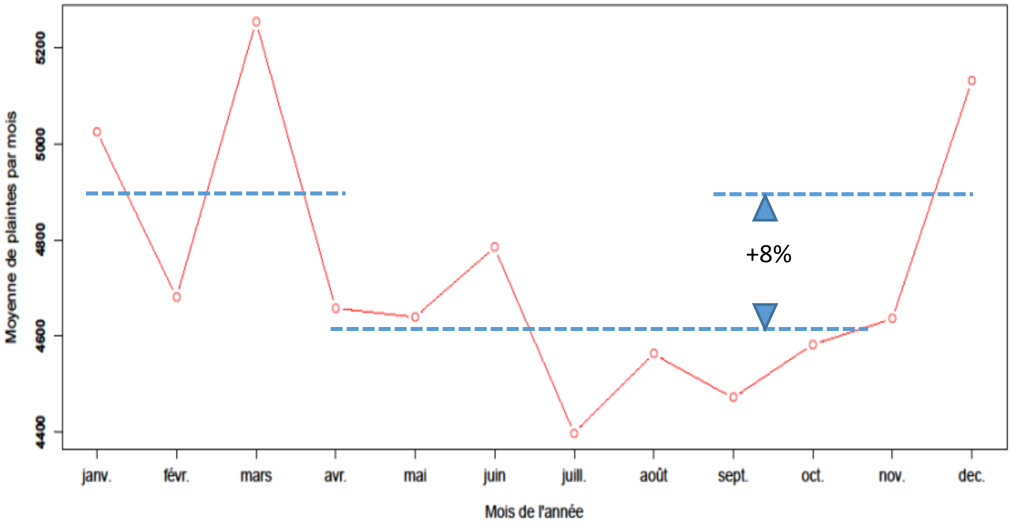
Frequency and Seasonality

Frequency: First Increasing then Levelling Off



Annual Moving Average of the monthly frequency of complaints

No strong *seasonality* for the monthly # of complaints



Severity for the sample of damages with amounts ≥ 500 €

Characteristics of the data: strong *asymmetry* and *kurtosis*

Sample	Mean	Median	Std deviation	CV	Skewness	Kurtosis
x ≥ 500	6’460	1500	61’892	10	91	10’513

Existence of **heavy tail**

Agenda

1 Cyber Risk Landscape

2 Statistical Analysis: The GN Cyber Complaints Database

3 Probabilistic Modelling

4 Conclusion

Probabilistic Modeling Through EVT

- ❑ As we have the Central Limit Theorem for the center of the distribution, we have the EVT for the rescaled maximum

- ❑ EVT Theorem:

  If $F \in MDA(G)$ then necessarily, G is of the same type as the *GEV cumulative distribution* H_γ (i.e. $G(x) = H_\xi(ax + b), a > 0$) defined as:

$$H_\gamma = \begin{cases} exp\left[-(1+\xi x)_+^{-\frac{1}{\xi}}\right] & \text{if } \xi \neq 0 \\ \exp(-e^{-x}) & \text{if } \xi = 0 \end{cases}$$

  where $y_+ = \max(0, y)$

- ❑ The tail index ξ ∈ ℝ determines *the nature of the tail distribution* and is called the extreme-value index: $\xi > 0$ (Fréchet), = 0 (Gumbel) or < 0 (Weibull)

Philosophy of Extreme Value Statistics*

- Extreme events are often *quite different* from ordinary everyday behavior and ordinary behavior often has little to say about extremes: then *only extreme events give useful information about future extreme events*

- *Theoretically motivated statistical models* give much better possibilities to learn from experience (and compare) than if everyone uses their own ad hoc method

- We apply EVT method to look at the extremes and try to *detect* certain *behaviors*

*) inspired by a talk by Prof. Holger Rootzén, European Statistics Day (ESSEC Paris La Défense), Oct. 2019

A general Hybrid Model as a basis for a Fitting Algorithm

- ❑ Frame: (right) heavy-tailed continuous data: Fit the tail using a GPD (Generalized Pareto Distribution) with a positive tail index (Fréchet domain of attraction)

- ❑ *For heavy tails:* **standard EVT graphical approaches** to determine the threshold to estimate the tail index (MEP, Hill, QQ, etc ...). *Main practical issue: supervised* methods

- ❑ *Main motivation*: to develop an **'unsupervised' method** to *determine the threshold* above which we fit the GPD, and to have a *good fit for the entire distribution*

A general Hybrid Model as a basis for a Fitting Algorithm

❑ Introduce a **simple** but **general hybrid model** with 3 components (LN-E-GPD):

1. A Lognormal distribution to model the mean behavior
2. A **GPD** for the **extreme behavior** (Pickands theorem): main component
3. An exponential distribution to bridge the mean and tail behaviors

Assumption: the distribution (that belongs to the **Fréchet** domain of attraction) has a C^1 density. *NO* assumption on the dependence

Remark. Main component in this hybrid model = the GPD one (for heavy tail). The mean behavior can be adapted to the context.

❑ A self-calibrating iterative algorithm, built on the solving of a set of non-linear least squares problems by the *Levenberg-Marquardt* technique, which combines Gauss–Newton and gradient descent methods to reach the desired minimum

Determining 4 Parameters (illustration on the G-E-GPD)

$$h(x;\theta) = \begin{cases} \gamma_1 \, f(x;\mu,\sigma), & \text{if} \quad x \leq u_1, \\ \gamma_2 \, e(x;\lambda), & \text{if} \quad u_1 \leq x \leq u_2, \\ \gamma_3 \, g(x-u_2;\xi,\beta), & \text{if} \quad x \geq u_2, \end{cases}$$

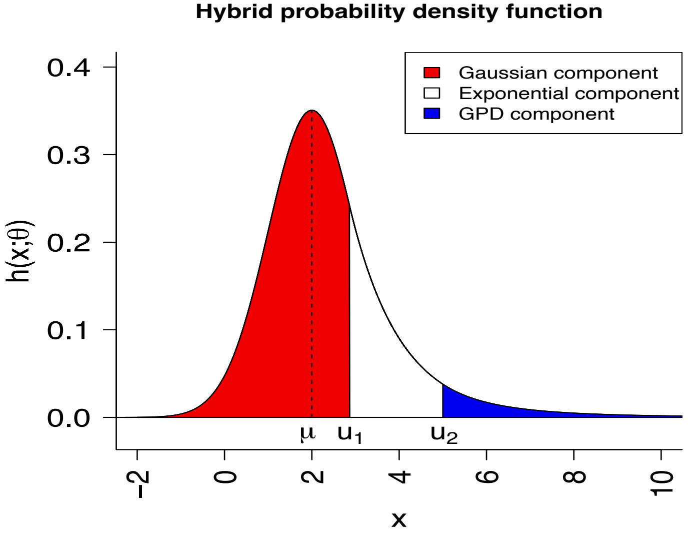
f: Gaussian pdf (μ, σ^2).

e: Exponential pdf with intensity λ.

g: GPD pdf with tail index ξ and scale parameter β.

$\theta = [\mu, \sigma, u_2, \xi]$: the parameters vector.

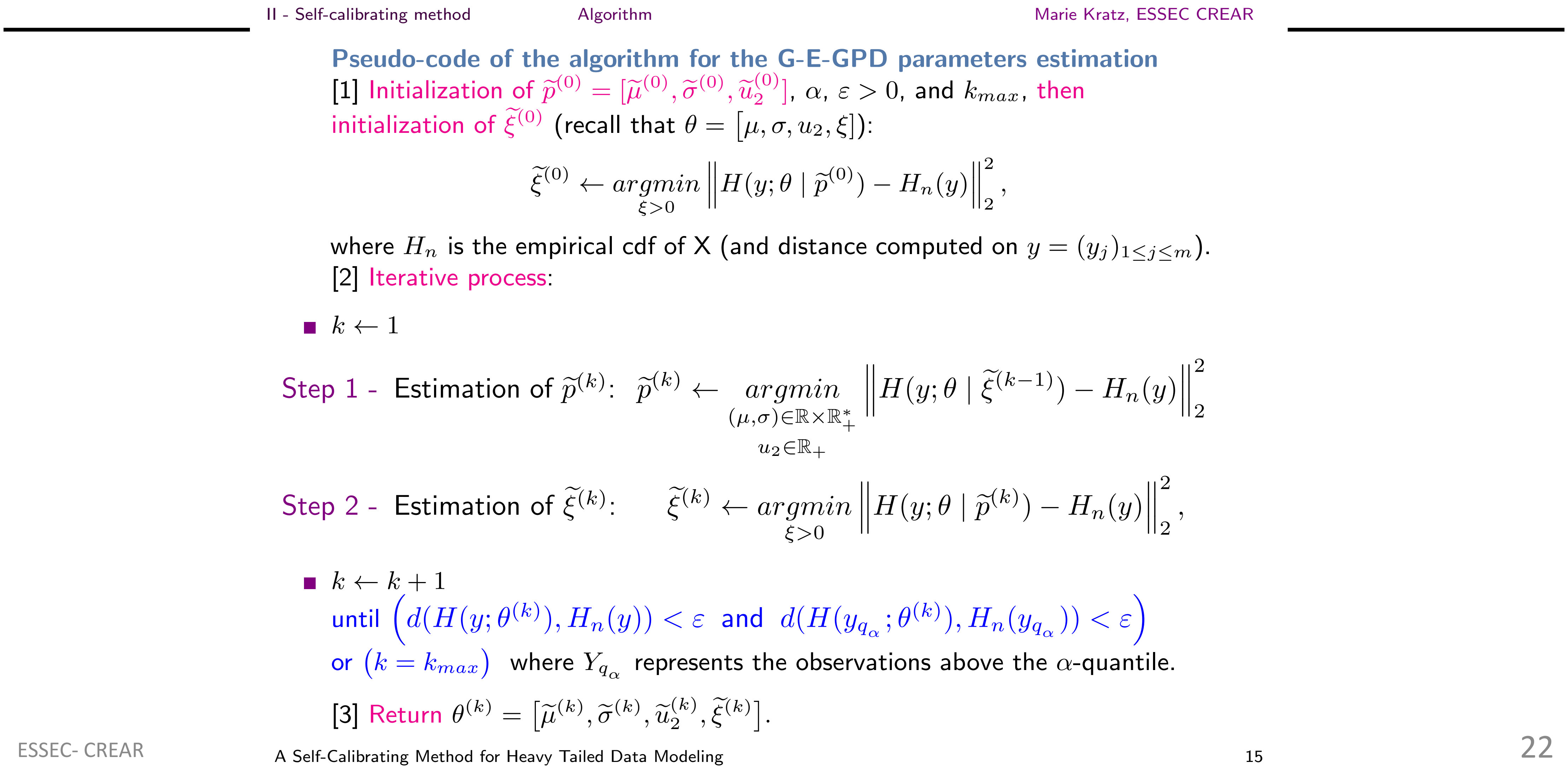
γ_1, γ_2 and γ_3: the weights (evaluated from the assumptions (in part. C^1))



The 3 other parameters (β, λ, u_1) also deduced from the C^1 assumption.

f will be replaced by a *Lognormal distribution* (our case) and the parameters evaluated with the new relations

Pseudo-code of the algorithm for the parameters estimation



Pseudo-code of the algorithm for the G-E-GPD parameters estimation

[1] Initialization of $\widetilde{p}^{(0)} = [\widetilde{\mu}^{(0)}, \widetilde{\sigma}^{(0)}, \widetilde{u}_2^{(0)}]$, α, $\varepsilon > 0$, and k_{max}, then initialization of $\widetilde{\xi}^{(0)}$ (recall that $\theta = [\mu, \sigma, u_2, \xi]$):

$$\widetilde{\xi}^{(0)} \leftarrow \underset{\xi>0}{argmin} \left\| H(y;\theta \mid \widetilde{p}^{(0)}) - H_n(y) \right\|_2^2 ,$$

where H_n is the empirical cdf of X (and distance computed on $y = (y_j)_{1 \le j \le m}$).

[2] Iterative process:

- $k \leftarrow 1$

Step 1 - Estimation of $\widetilde{p}^{(k)}$: $\widetilde{p}^{(k)} \leftarrow \underset{\substack{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+^* \\ u_2\in\mathbb{R}_+}}{argmin} \left\| H(y;\theta \mid \widetilde{\xi}^{(k-1)}) - H_n(y) \right\|_2^2$

Step 2 - Estimation of $\widetilde{\xi}^{(k)}$: $\widetilde{\xi}^{(k)} \leftarrow \underset{\xi>0}{argmin} \left\| H(y;\theta \mid \widetilde{p}^{(k)}) - H_n(y) \right\|_2^2 ,$

- $k \leftarrow k+1$

until $\left(d(H(y;\theta^{(k)}), H_n(y)) < \varepsilon \text{ and } d(H(y_{q_\alpha};\theta^{(k)}), H_n(y_{q_\alpha})) < \varepsilon \right)$

or $(k = k_{max})$ where Y_{q_α} represents the observations above the α-quantile.

[3] Return $\theta^{(k)} = [\widetilde{\mu}^{(k)}, \widetilde{\sigma}^{(k)}, \widetilde{u}_2^{(k)}, \widetilde{\xi}^{(k)}]$.

Performance and convergence of the iterative algorithm

- ❑ *Performance*: tests via MC simulation with training and tests sets

- ❑ Proof of the convergence in 2 steps:

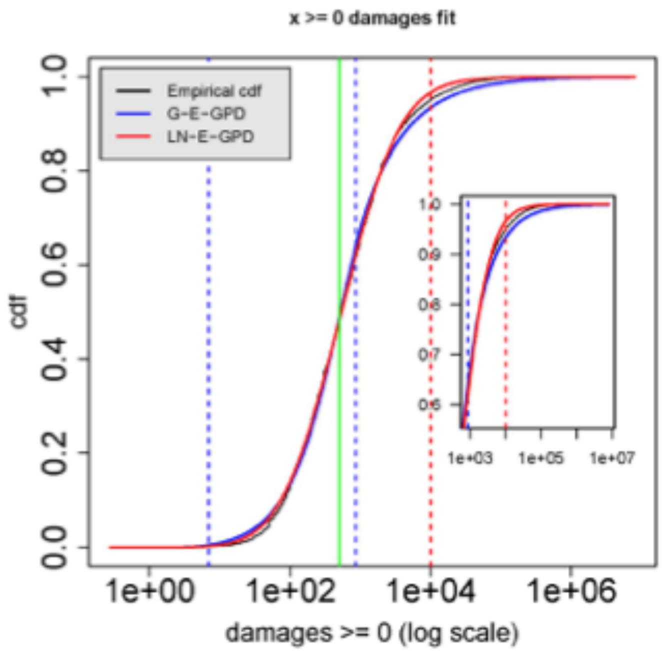
1. **Analytical** proof of **stationary points existence**, supplemented by numerical simulations.
   Algorithm: sequence of minimization, does not rely on the optimization of a cost function by seeking a trajectory to reach an extremum of an error surface.
   As a cq, existence of a stationary point not guaranteed, neither the convergence towards it: It has to be proven

2. **Cv** to a **unique** stationary point. Done **numerically,** performing various simulations changing each time the initialization. Analytical proof of this 2nd step: still open pb

Application of the Method on Cyber Crimes Complaints

Ex: modelling of amounts > 0 (log scale) (raw data)

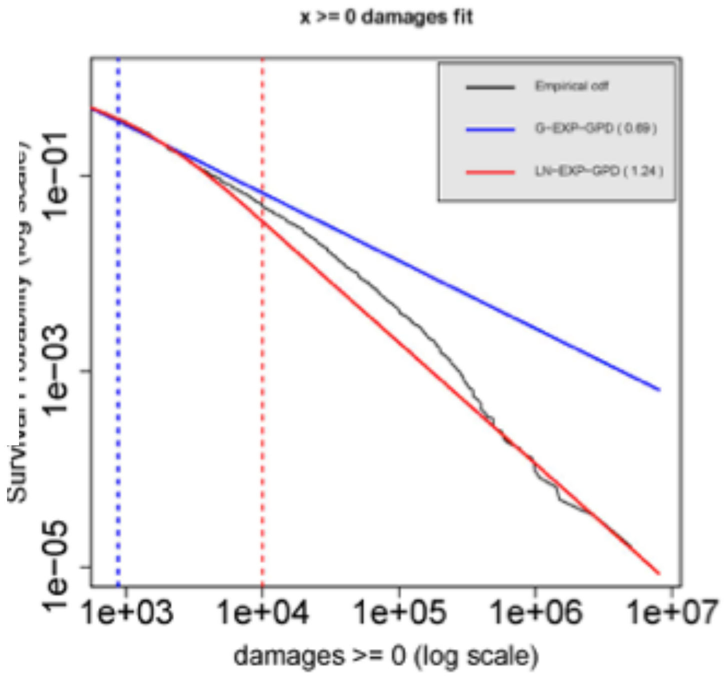
Cumulative Probability Distribution



Tail index = 0.81
<=> α=1.24
(no finite variance)

Threshold for Tail:
9’999 € (quantile: 96.6%)

Survival Probability



Exploration of the Statistical Significance of the Results

- ❑ Any statistical result must be given with a *significant band* (usually for the 95% confidence)
- ❑ This band can be computed *theoretically*, if the underlying *process* is *known*, or can be directly estimated with the data
- ❑ In our case, we do not know the data generating process, thus we have to resort to *direct methods* either *bootstrap* or *Jackknife*
- ❑ The bootstrap method consists in artificially *constructing many samples* by picking randomly (*with replacement*) values from the original data and redo on those samples the fitting procedure
- ❑ We then obtain a *distribution* of the *fitting results* and pick the 95% confidence range (between the 2.5% and 97.5% quantiles)

The Jackknife Method to Get the Significance Band

- ❑ With the *amount of data* we have, there is a straighter forward method to estimate the confidence range: The *Jackknife* method that measures the variability of the estimation across sub-samples

- ❑ To determine a *numerical confidence range*, we build randomly subsamples and run on each one the algorithm for calibrating the hybrid model

- ❑ We *omit* some randomly *selected data points* that amount to 10% of the original dataset of size n = 60985, making sure that each of those selected observations is omitted only once (*without replacement*), while used in the 9 other computations

- ❑ Note that it means that each observation in the whole sample will be removed in 1 of the 10 subsamples

Variability of Our Results

❑ Using the Jackknife method, we get:

	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{u_2}$
Estimation	1.236	8’087	9’999
95% Confidence Range	[1.213 ; 1.260]	[7’929 ; 8’245]	[9’980 ; 10’018]

where $\alpha = {}^{1}/_{\xi}$ and $\beta = u_2\xi$

❑ We can conclude that the *expectation exists* $(\alpha > 1)$ but *not the variance* $(\alpha < 2)$

Some consequences for risk management

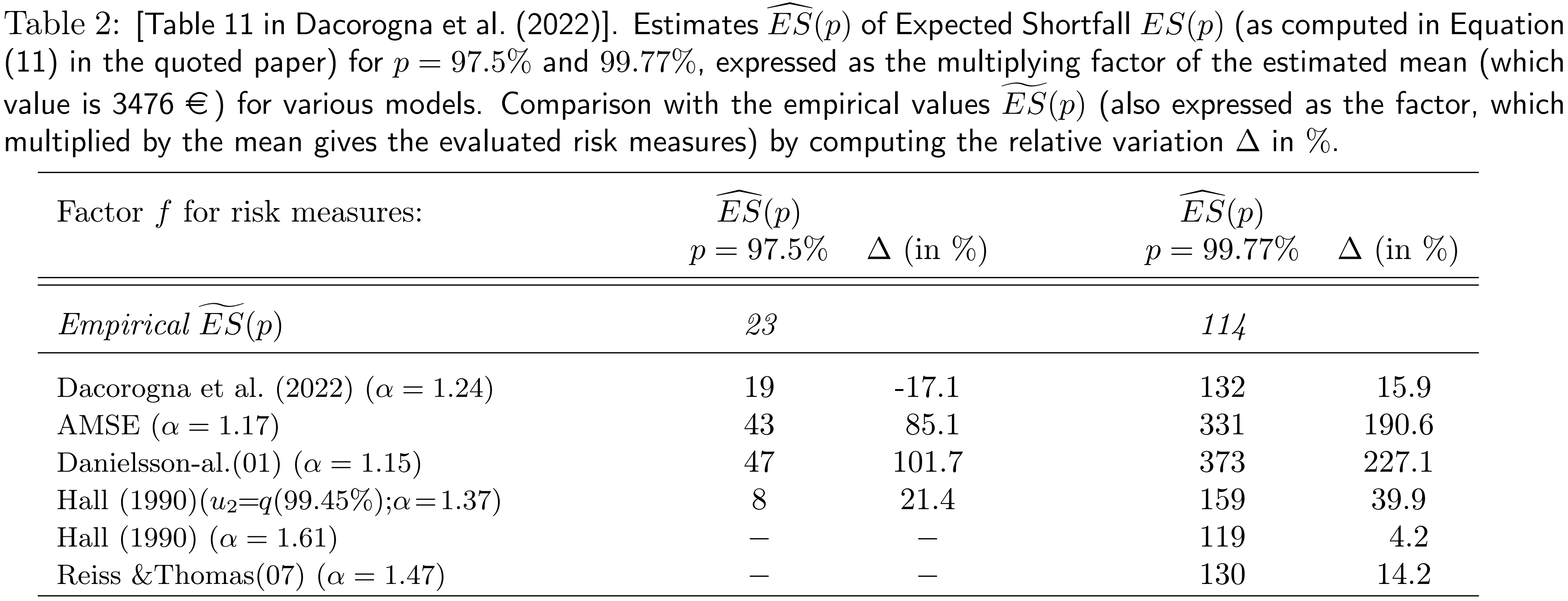


Table 2: [Table 11 in Dacorogna et al. (2022)]. Estimates $\widehat{ES}(p)$ of Expected Shortfall $ES(p)$ (as computed in Equation (11) in the quoted paper) for $p = 97.5\%$ and 99.77%, expressed as the multiplying factor of the estimated mean (which value is 3476 €) for various models. Comparison with the empirical values $\widetilde{ES}(p)$ (also expressed as the factor, which multiplied by the mean gives the evaluated risk measures) by computing the relative variation Δ in $\%$.

Factor f for risk measures:	$\widehat{ES}(p)$ $p = 97.5\%$	Δ (in $\%$)	$\widehat{ES}(p)$ $p = 99.77\%$	Δ (in $\%$)
Empirical $\widetilde{ES}(p)$	*23*		*114*	
Dacorogna et al. (2022) ($\alpha = 1.24$)	19	-17.1	132	15.9
AMSE ($\alpha = 1.17$)	43	85.1	331	190.6
Danielsson-al.(01) ($\alpha = 1.15$)	47	101.7	373	227.1
Hall (1990)(u_2=q(99.45%);α=1.37)	8	21.4	159	39.9
Hall (1990) ($\alpha = 1.61$)	–	–	119	4.2
Reiss &Thomas(07) ($\alpha = 1.47$)	–	–	130	14.2

Cyber risk presents clearly a *catastrophic nature; High capital intensity*: ES is about 20 times the mean!

Towards a classification

- ❑ Classification by the GN

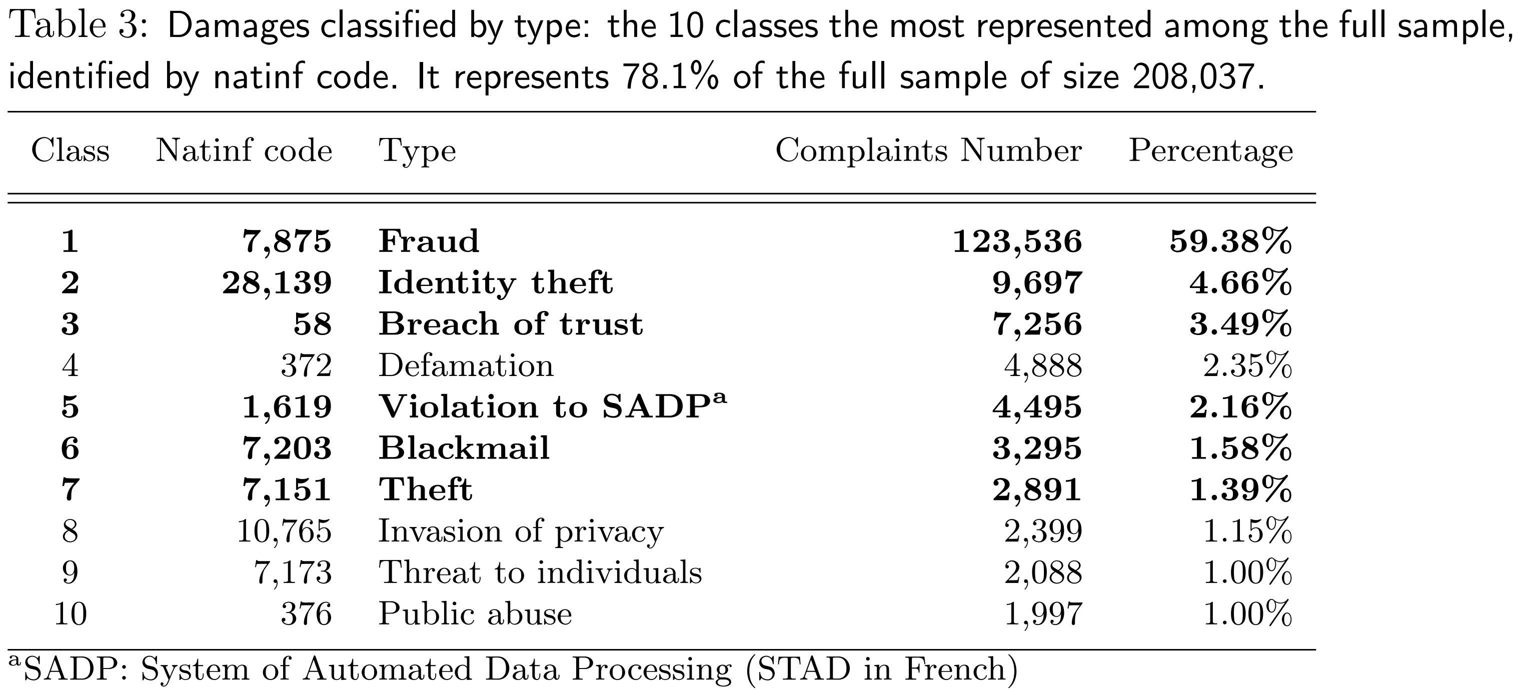


Table 3: Damages classified by type: the 10 classes the most represented among the full sample, identified by natinf code. It represents 78.1% of the full sample of size 208,037.

Class	Natinf code	Type	Complaints Number	Percentage
1	**7,875**	**Fraud**	**123,536**	**59.38%**
2	**28,139**	**Identity theft**	**9,697**	**4.66%**
3	**58**	**Breach of trust**	**7,256**	**3.49%**
4	372	Defamation	4,888	2.35%
5	**1,619**	**Violation to SADP**[a]	**4,495**	**2.16%**
6	**7,203**	**Blackmail**	**3,295**	**1.58%**
7	**7,151**	**Theft**	**2,891**	**1.39%**
8	10,765	Invasion of privacy	2,399	1.15%
9	7,173	Threat to individuals	2,088	1.00%
10	376	Public abuse	1,997	1.00%

[a] SADP: System of Automated Data Processing (STAD in French)

- ❑ Comparing the types of cyber attacks via their *tail index*

What do we learn from this study on cyber risk?

1. The GN database is a *precious source of data* for studying cyber risk (large database and different from usual ones, completing the panorama)

2. We confirm that *cyber risk* is *insurable* (existence of expectations; tail index <1)

3. Non-stationarity but not in the extremes (Poisson-GPD model for freq-severity)

4. *Cyber risk* presents clearly a *catastrophic nature* (extreme risk)

5. Further research to exploit the complaint descriptions via *semantic analysis* of the text

Comparing properties observed on *different cyber databases* will help find the main cyber characteristics

Agenda

1 Cyber Risk Landscape

2 Statistical Analysis: The GN Cyber Complaints Database

3 Probabilistic Modelling

4 Conclusion

Conclusion: On the method

❑ Proposition of a *general* and *simple* hybrid model for asymmetric non-negative *heavy-tailed* data;

A 3-components model: *bulk* + **tail** *with exponential* ***bridge***

❑ Development of a *fast*, *inexpensive* and *unsupervised algorithm* for calibrating the model on heavy-tailed data

❑ Can be used in many fields (OR, finance, …) and combined with other models (e.g. EV regression: see Hambuckers et al. (2023), Efficient estimation in extreme value regression models of hedge fund tail risks)

Conclusion: On Cyber Risk

❑ Presence of *extremes*, signature of systemic risk, but *finite loss expectation*, necessary condition for insurability

❑ Cyber, a *very high risk*, with tail heaviness in the same range as natural catastrophes

❑ Cyber risk creates a *new risk landscape*, but also *opportunities for insurance* companies to offer *hedging solutions* to companies. The *market* is going to *grow,* and the insurance industry cannot stay away of the social demand

❑ *Accumulation control* and modelling are *key* to develop a successful business

❑ Creating *links with cyber security firms* is a way to improve the risk profile of insured and to design products that incentivize customer to invest in cyber security