Insurance: Mathematics and Economics 106 (2022) 90–114

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

www.elsevier.com/locate/ime

Cyber risk frequency, severity and insurance viability

Matteo Malavasi [a], Gareth W. Peters [b,a], Pavel V. Shevchenko [a,*], Stefan Trück [a], Jiwook Jang [a], Georgy Sofronov [c]

[a] Department of Actuarial Studies and Business Analytics, Macquarie University, Australia
[b] Department of Statistics and Applied Probability, University of California Santa Barbara, USA
[c] Department of Mathematics and Statistics, Macquarie University, Australia

ARTICLE INFO

Article history:
Available online 24 May 2022

JEL classification:
G22
G32
C51
C52
L86
G28

Keywords:
Cyber risk
GAMLSS
Cyber risk insurance
Ordinal regression

ABSTRACT

In this study an exploration of insurance risk transfer is undertaken for the cyber insurance industry in the United States of America, based on the leading industry dataset of cyber events provided by Advisen. We seek to address two core unresolved questions. First, what factors are the most significant covariates that may explain the frequency and severity of cyber loss events and are they heterogeneous over cyber risk categories? Second, is cyber risk insurable in regards to the required premiums, risk pool sizes and how would this decision vary with the insured companies industry sector and size? We address these questions through a combination of regression models based on the class of Generalized Additive Models for Location Shape and Scale (GAMLSS) and a class of ordinal regressions. These models will then form the basis for our analysis of frequency and severity of cyber risk loss processes. We investigate the viability of insurance for cyber risk using a utility modeling framework with premiums calculated by classical certainty equivalence analysis utilizing the developed regression models. Our results provide several new key insights into the nature of insurability of cyber risk and rigorously address the two insurance questions posed in a real data driven case study analysis.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

The understanding, mitigation, reporting, risk management and insurance modeling related to cyber risk are still in the early stages of development. However, with the increasing focus on Information Technology (IT) and cyber related risk that many companies, governments and regulators are beginning to actively explore in all industry sectors, these aspects of cyber risk are coming to the forefront of the insurance industry and risk managers portfolios. In this work we seek to study the areas that are under explored in regards to insurance and loss modeling aspects of cyber risk.

In modern business practices the increasing intersection between automation and efficiency and the enhanced uptake of IT infrastructure, occurring in all areas of business, industry and service sector processes has resulted in a greater impact of cyber risk. This has driven an enhanced focus on cyber security and motivated a much stronger cyber risk awareness in both practice, governance, regulation and system design. This will continue to naturally progress due to the fact that organizations of all sizes in both the public and private sectors are increasingly reliant on IT systems in order to execute business processes that support the delivery of services. If there is a breakdown or failure in these systems, the organization will experience a direct negative impact on the processes it supports, resulting in reduction of service and disruptions that ultimately impact on the organizations ability to meet its objectives. This in turn can generate a variety of losses that are both directly and indirectly attributable to cyber events. In Eling and Schnell (2016), Peters et al. (2018a), Eling (2020) and book length discussions Refsdal et al. (2015), Bouveret (2018) several detailed perspectives of the current state of the cyber risk insurance market are presented. In this study we seek to extend the focus by utilizing a quantitative model based framework to assess critical questions regarding cyber risk modeling and risk transfer efficiency, effectiveness and availability. Such statistical modeling perspectives require a reliable data source, which is a challenge to obtain in the cyber risk context. We utilize the leading industry dataset from Advisen[1] (hereafter referred to as Advisen Cyber Loss Data) that maintains a detailed account of major cyber risk losses across all industry sectors around the world. Our analysis is

* Corresponding author.
E-mail addresses: Matteo.Malavasi@mq.edu.au (M. Malavasi), garethpeters@ucsb.edu (G.W. Peters), Pavel.Shevchenko@mq.edu.au (P.V. Shevchenko), Stefan.Trueck@mq.edu.au (S. Trück), Jiwook.Jang@mq.edu.au (J. Jang), Georgy.Sofronov@mq.edu.au (G. Sofronov).

https://doi.org/10.1016/j.insmatheco.2022.05.003
0167-6687/© 2022 Elsevier B.V. All rights reserved.

[1] https://www.advisenltd.com/data/cyber-loss-data/.

focused on the United States of America (USA) experience as this represents the most comprehensive and complete set of records in this dataset and we believe the findings we obtain naturally transfer to other developed economies.

In this work we seek to make novel contributions to the modeling and quantification of cyber risk along two key lines of questioning. First, what factors are the most significant explanatory variables that may explain the frequency or severity of cyber loss events, and are these factors heterogeneous over cyber risk categories? Second, is cyber risk insurable in regards to required premiums, risk pool sizes, and how would this decision vary with the insured company industry sector and size?

Regarding the modeling of cyber risk, there are studies that investigate basic statistical properties of cyber loss data in a non-regression framework, such as Biener et al. (2015), Eling and Wirfs (2015), Eling and Loperfido (2017), Eling and Jung (2018), Peters et al. (2018b), Bessy-Roland et al. (2021), Farkas et al. (2021), and Zeller and Scherer (2021). However, none of these studies has applied such a wide range of flexible regression models to an extensive dataset of losses from cyber events as we do. Therefore, our analysis will be one of the first to allow the thorough examination of key risk drivers of losses from cyber risk for different industry sectors and event types. In regards to questions of insurability, several authors have addressed aspects of the insurability of cyber risk losses, the efficiency and structuring of insurance for such risk types; see, Herath and Herath (2007), Peters et al. (2011), Mukhopadhyay et al. (2013), Biener et al. (2015), Camillo (2017), Fahrenwaldt et al. (2018), Eling and Wirfs (2019), Romanosky et al. (2019), Hillairet and Lopez (2021), Antonio et al. (2021), and Xu and Hua (2019). In this work we seek to explore insurability from a data driven quantitative perspective rather than an economic risk transfer perspective which has been undertaken in the aforementioned works. By this we aim to provide new perspectives and insights on the question of insurability of cyber risk.

Our analysis is uniquely based on a combination of carefully developed regression models that can adequately accommodate factors and key risk drivers that may influence loss frequency and loss severity in cyber risk loss modeling through generalizations of the well established Generalized Linear Modeling (GLM) regression framework to the Generalized Additive Models for Location Shape and Scale (GAMLSS) pioneered by Stasinopoulos and Rigby (2007). This class of models allows one to identify and associate the risk drivers or regression factors that influence the frequency and severity of cyber loss processes in the mean, variance, and unlike standard GLM regression models, we can also associate regression factors to higher order moments such as skewness and kurtosis. Effectively, the use of GAMLSS regression for modeling the severity and frequency of losses in cyber risk can prove to be a powerful framework in better understanding which factors are influential in the central moments of studied loss processes. Importantly, we can distinctly identify which factors are influential in the tail behavior of cyber losses. This is an advantage over GLM regression modeling that is particularly poignant when heavy tailed loss processes are under consideration. Then to ensure we have meaningful results not directly influenced by leverage effects arising from the sheer magnitude of the cyber losses studied in some risk lines, we also undertake an analysis of the dataset using an ordinal rank based regression, using a framework proposed in Giudici and Raffinetti (2020). This framework allows us to develop meaningful insights into concordance rankings for risk profiles in cyber risk loss categories not attainable readily from the standard GLM regression structures.

We conclude our study by focusing on the selected optimal regression structures to undertake an analysis of the insurability of cyber risk losses based on the developed frequency and severity regression models. We study the premium and risk pool size for insurance mitigation, using a utility based framework with certainty equivalence analysis for risk averse firms. The manner in which we have performed the analysis also allows us to provide insights with respect to the risk appetite and insurance cost on a firm by firm basis. This enhances the types of economic analysis undertaken in work such as Lis and Mendel (2019).

The paper is structured as follows. The model framework is defined in Section 2, while our dataset is described in Section 3. Sections 4 and 5 are devoted to GAMLSS and rank based regressions respectively. Applications for Value-at-Risk and insurance premium calculations are presented in Section 6. The paper is concluded in Section 7. Additional statistical results regarding various model fits are presented in Appendix A.

2. Model framework

Cyber risk related monetary losses (hereafter losses), as well as more general IT related Operational Risk (OpRisk) losses, exhibit extreme events. It is well documented in the literature that in the presence of a heavy tailed severity distribution, standard ordinary least squares (OLS) based techniques might not be appropriate (see, among others, Cope and Labbi, 2008; Dahen and Dionne, 2010; Ganegoda and Evans, 2013; Chavez-Demoulin et al., 2016). Moreover, cyber risk affects a variety of different actors, in many different ways such that flexible modeling approaches are needed in order to draw meaningful inference on the driving risk factors (see, e.g. Peters et al., 2018b).

A common statistical modeling framework used widely in OpRisk is the Loss Distribution Approach (LDA), in which risk related losses are a marked point process, occurring at random points in time. One can then model the frequency component with a counting process and the severity or size of the losses incurred over time at the times of events (the marks) with a positively supported heavy tailed loss distribution model; see e.g. Shevchenko (2011), Cruz et al. (2015). One popular approach adopted in cyber risk modeling has been to utilize a Generalized Pareto model that is obtained from the framework of Extreme Value Theory (EVT) via a classical estimation approach of the peaks-over-threshold (POT) method. This will accommodate the tail behavior to adequately explain and allow for the modeling of extreme events encountered in cyber risk loss modeling (see, among others, Ganegoda and Evans, 2013; Chavez-Demoulin et al., 2016; Eling and Wirfs, 2019).

To setup an LDA modeling framework with frequency and severity components, we establish the following notation. Let N be a discrete counting random variable for the number of loss events in a specified period of time (typically annual) and let $\widetilde{Y}_i$, $i = 1, \ldots, N$, be positive loss random variables, defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, obtained from a severity loss distribution. Then the aggregated loss or total loss over the specified period will be denoted by the compound random variable Z defined as:

$$Z = \sum_{i=1}^{N} \widetilde{Y}_i.$$

Typically, it is assumed that N and $\widetilde{Y}_i$, $i = 1, \ldots, N$ are all independent. As outlined in this set up, the LDA approach aims to find frequency and severity distributions supported by the data, in order to provide insights on the risk under investigation. In practice, one may consider several choices for the frequency and severity models. Popular examples include modeling the event frequency N by a Poisson distribution. If there is over- or under-dispersion in the observed frequency one may opt for a Negative Binomial or Binomial distribution, respectively; see discussion in the cyber risk modeling context in Edwards et al. (2016), Eling and Wirfs (2019).

In terms of the severity or loss size model, in order to accommodate extreme events, the severity distribution can be modeled with a positively supported distribution that can accommodate heavy tailed features.

Remark 1. Due to the lack of good datasets for cyber risk losses, some authors suggest to avoid the use of classical statistical methods for estimation of the LDA frequency and severity distributions using historical data (especially for low-frequency/high-impact risks) and rather to rely on scenario analysis approaches. For example, Rakes et al. (2012) argue that for sparse/high-impact IT security breaches one can rely on an expert's judgment defining worst-case scenarios and their likelihood. It is also clear that estimation based on historical losses is backward looking and it is challenging to account for a constantly changing environment. It has been a common practice in OpRisk modeling (in the banking industry) to use scenario analysis for estimation of frequency and severity distributions. Moreover, there was a Basel II regulatory requirement for bank's internal OpRisk capital models to include the use of internal data, external data, scenario analysis and factors reflecting the business environmental and internal control systems; see Basel Committee on Banking Supervision (2006). There are different methods to accomplish this task; see e.g. Shevchenko and Wüthrich (2006), Lambrigger et al. (2007) or for a book length treatment Cruz et al. (2015). In this paper we focus on the estimation of frequency and severity distributions using historical data because we have access to the leading industry dataset from Advisen that contains a substantial number of cyber loss events.

As mentioned, one popular choice in cyber risk modeling is to utilize the framework of EVT based on the POT method. Under the assumption that the true distribution of $\widetilde{Y}_i$, $i = 1, \ldots, N$ lies in the domain of attraction of the Generalized Extreme Value (GEV) distribution, loss exceedances over a high enough threshold u, i.e. $Y = \widetilde{Y} - u$ conditional on $\widetilde{Y} > u$, can be assumed to follow a Generalized Pareto Distribution (GPD) loss model with the following density:

$$g(y; \mu, \tau) = \frac{\tau}{\mu}\left(1 + \frac{y}{\mu}\right)^{-(1+\tau)}, \tag{1}$$

for $y > 0$ if $\tau > 0$ and $y \in [0, -\mu]$ if $\tau < 0$ (see, e.g. Ganegoda and Evans, 2013; Chavez-Demoulin et al., 2016). If $\tau > 0$, which is typically the case for financial and actuarial applications, in order to assume that losses can be described by the density given in Equation (1), the distribution function of $\widetilde{Y}$, $F_{\widetilde{Y}}$, must satisfy the following regular variation condition:

$$\overline{F}_{\widetilde{Y}}(x) = 1 - F_{\widetilde{Y}}(x) \sim x^{-\tau} L(x), \; x \to \infty \tag{2}$$

for some measurable, slowly varying function $L : (0, \infty) \to (0, \infty)$; see, Balkema and De Haan (1974), Pickands (1975). According to the asymptotic representation in Equation (2), it can be shown that if $\tau \in (0, 1)$ then $F_{\widetilde{Y}}$ does not have finite moments, while if $\tau \geq 1$ then $F_{\widetilde{Y}}$ has at least a finite first moment.

Whilst the LDA framework can provide insight into the modeling of the loss process in different classes of cyber risk, this model in its traditional form is incapable of providing a causal link between the leading drivers or risk factors that induce the loss process to occur. To achieve this further level of understanding of cyber risk loss process modeling one must introduce a regression structure into the LDA model components. Furthermore, in cyber risk loss modeling, the loss processes may vary significantly in their statistical attributes over time and over different companies, sectors or cyber risk event types. Therefore, while the LDA can be helpful in identifying the most appropriate distribution for the frequency and severity of cyber events, it might not be flexible enough to capture the nature of cyber risk to the full extent required to address the types of insurance questions posed at the onset of this manuscript.

Cyber risk events are heterogeneous in nature and can affect a wide variety of industries at varying degrees of penetration and with varying degrees of financial impact. Loss events or the process that leads to the loss can also arise from direct and indirect causes. This can result in many losses from a collection of connected events to the actual IT focused cyber attack. Therefore, to quantify and understand the risk drivers that affect cyber risk loss processes, a methodology allowing to identify the impact of explanatory variables on the frequency and severity distributions will be highly beneficial. The GAMLSS regression framework allows parameters in both the severity and frequency distributions to depend on covariates and, at the same time, to have a more flexible scaling structure than OLS based regression techniques (Stasinopoulos and Rigby, 2007; Rigby and Stasinopoulos, 2005; Ganegoda and Evans, 2013). Given that the risk from different types of cyber threats is likely to vary over time, we consider the extension to dynamic EVT in Chavez-Demoulin et al. (2016), allowing the parameters to depend also on time, and thus to investigate any non stationary behavior, with particular interest in the tail parameter τ. For any given set of covariates X and time t, we consider the following link functions for the intensity of the Poisson distribution of the loss exceedances $\lambda(X, t)$, the scale parameter of the loss distribution $\mu(X, t)$ and the tail of the loss distribution $\tau(X, t)$ that characterize the LDA loss process:

$$\begin{aligned} \log(\lambda(X,t)) &= f_\lambda(X) + h_\lambda(t), \\ \log(\mu(X,t)) &= f_\mu(X) + h_\mu(t), \\ \log(\tau(X,t)) &= f_\tau(X) + h_\tau(t), \end{aligned} \tag{3}$$

where h_λ, h_μ, h_τ are measurable, twice differentiable functions. The functional form of f_λ, f_μ and f_τ is assumed to be linear (linear predictors under a log-link).

In the estimation routine, we consider data at the company level, aggregated by year. Let n_i^c and y_i^c, be the number of cyber events and related loss in year i in company c, respectively, for $i = 1, \ldots, K$ and $c = 1, \ldots, C$. We aim to estimate the coefficients for the linear predictors that are connected to the parameters of the frequency and severity distribution via the link functions (3). Following Chavez-Demoulin et al. (2016), the estimation can be conducted via two penalized maximum likelihood optimization objectives as outlined below in Equation (4) for the frequency and Equation (5) for the severity:

$$\max \sum_{c=1}^{C} \sum_{i=1}^{K} \log(f_N(n_i^c; \lambda)) - \gamma_\lambda \int_0^K h_\lambda''(t)dt \tag{4}$$

and

$$\max \sum_{c=1}^{C} \sum_{i=1}^{K} \log(g(y_i^c; \mu, \tau)) - \gamma_\mu \int_0^K h_\mu''(t)dt - \gamma_\tau \int_0^K h_\tau''(t)dt, \tag{5}$$

where $f_N(\cdot; \lambda)$ is the probability density function of a Poisson random variable, and $\gamma_\lambda, \gamma_\mu, \gamma_\tau$ are the smoothing parameters. We adopt the algorithm proposed by Rigby and Stasinopoulos to solve the optimization problems and find the parameter estimates (see, Stasinopoulos and Rigby, 2007; Rigby and Stasinopoulos, 2005; Stasinopoulos et al., 2017). When h_λ, h_μ, h_τ are chosen to be cubic splines, the penalized maximum likelihood used in the *gamlss*

R-package coincides with the one in the *QRM* R-package (see, Stasinopoulos et al., 2008; Chavez-Demoulin et al., 2016).

3. Cyber losses data description

Our comprehensive data set is sourced from Advisen and contains more than 132,126 cyber events from 2008 to 2020, affecting 49,495 organizations across the world.[2] Advisen is a USA-based for-profit organization which collects and processes cyber reports from reliable and publicly verifiable sources such as news media, governmental and regulatory sources, state data breach notification sites, and third-party vendors. Given that the interest in cyber risk is on the rise, some recent studies on cyber risk have also made use of Advisen Cyber Loss Data (see, e.g. Romanosky, 2016; Aldasoro et al., 2020; Cyentia, 2020). More than 80% of the events recorded affect organizations residing in the USA and for each event accident timeline (i.e. first notice date, accident date, loss start date, and loss end date), a detailed explanation of the event is provided. One of the key advantages in comparison to other commonly used datasets – such as e.g. the "Chronology of Data Breaches" provided by the Privacy Rights Clearinghouse (PRC)[3] – is that the Advisen dataset offers direct information of monetary losses linked to each cyber risk event, providing an empirical measurement of financial losses that can be used for modeling purposes. Advisen also provides, when available, cost, economic losses, litigated loss amounts, fines, and penalties for cyber risk events from regulatory bodies. Although the dataset comprises more than 130,000 cyber risk related events, the nature of cyber risk itself determines that only a small percentage of the observation can be used for modeling. Following Edwards et al. (2016) and Eling and Loperfido (2017) we remove all observations that do not give information on the monetary losses, and restrict the analysis to the observation for which complete information on company specific characteristics, such as yearly revenue and number of employees are available. This leaves us with a total number of 3,792 observations, corresponding to roughly 2.6% of the total events.

Remark 2. Cyber risk severity can be expressed as number of records or monetary losses. The type of relationship between the two is still the object of ongoing research and debate. According to a recent study by the Ponemon Institute in partnership with IBM security, the logarithm of number of records and the logarithms of cyber risk related monetary losses are linked via a linear relationship (Eling and Loperfido, 2017; Ponemon, 2019). Given that one of the objectives of our study is to quantify cyber risk in terms of monetary losses to provide insights to decision makers, practitioners, and insurance market participants, we focus on monetary losses.

In order to compare data characteristics with previous studies, Fig. 1 shows the time evolution of the number of events, as well as the mean and median losses from 31 January 2008 to 31 July 2020 in a sliding window of 6 months (see, for example Eling and Wirfs, 2019; Maillart and Sornette, 2010). One can readily observe that the number of cyber events increases between 2008 and 2014, and then it seems to follow a decreasing trend contrary to Maillart and Sornette (2010) and Eling and Wirfs (2019), who observe a decreasing trend in mean and median of cyber event related losses starting from 2008. We also find that both mean and median follow an increasing trend jointly with an increase in variability. It is also worth to notice that the median is significantly lower than the mean, since the majority of cyber events related losses are of small magnitude and relatively few, very extreme events are recorded. This is consistent with a heavy tailed loss generating process with rarely occurring but severe loss consequences for major cyber events.

The decreasing trend in the number of events observed after 2016 needs to be interpreted with caution. Other studies have analyzed cyber event magnitude by the means of number of records affected and observed increasing trends in cyber event frequency (see, for example Eling and Schnell, 2016; Eling and Loperfido, 2017; Xu et al., 2018; Woods and Böhme, 2021). In contrast to these studies, we measure cyber event severity directly on monetary losses reported in the Advisen dataset. As discussed above, only a small proportion of the observations reports a known monetary loss linked to the cyber event. In this context, Fig. 1, depicts the behavior of number of events, mean and median loss only for events that provide information on monetary losses. Furthermore, monetary losses are often caused by clusters of cyber events, making the relationship between monetary losses and number of records affected non proportional. The decreasing trend in panel (a) of Fig. 1 also needs to be interpreted in view of reporting delay that can be attributed to various factors. It is well known that businesses have a low tendency of reporting cyber related events to avoid repercussions on their reputation. Moreover, Advisen collects data under the USA Freedom of Information Act regulation which prescribes a 60 days window between discovering the data breach and reporting it to affected parties, while non-USA domiciled entities do not have such strict requirements.

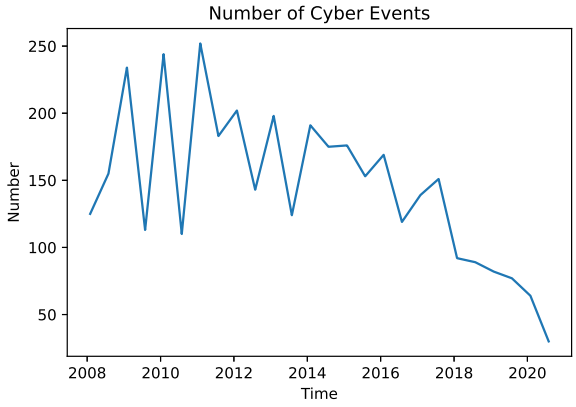
Typically, in OpRisk, loss events are classified according to business line and event type. On the one hand, Basel II defines seven official event types and eight business lines for Basel II banking OpRisk practices; see Basel Committee on Banking Supervision (2006). On the other hand, given the multidimensional heterogeneity that characterizes cyber risk, classifying cyber risk is a non-unique task to the point that a world wide accepted cyber risk classification does not exist to date as no consensus on standardization has been achieved yet; see further discussion in Peters et al. (2018a,b). Advisen provides its own classification based on the type of cyber threat[4]:

- **Privacy – Unauthorized Contact or Disclosure**: cases when personal information is used in an unauthorized manner to contact or publicize information regarding an individual or an organization without their explicit permission.
- **Privacy – Unauthorized Data Collection**: cases where information about the users of electronic services, such as social media, cellphones, websites, and similar is captured and stored without their knowledge or consent, or where prohibited information may have been collected with or without their consent.
- **Data – Physically Lost or Stolen**: situations where personal confidential information or digital assets have been stored on, or may have been stored on, computer, peripheral equipment, data storage, or printouts which has been lost, stolen, or improperly disposed of.
- **Data – Malicious Breach**: situations where personal confidential information or digital assets either have been or may have been exposed or stolen, by unauthorized internal or external actors whose intent appears to have been the acquisition of such information.

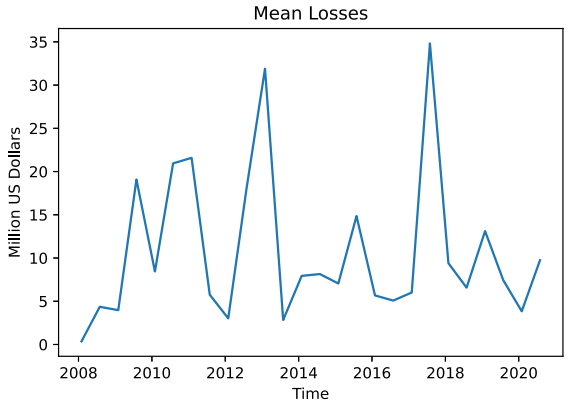
[2] Given the nature of cyber risk, it can be expected than a great number of events are not recorded in the dataset, since companies are reluctant to report cyber risk related events to the public in order to avoid, among other things, a loss of reputation and trust from their counterparties.

[3] https://privacyrights.org/data-breaches.

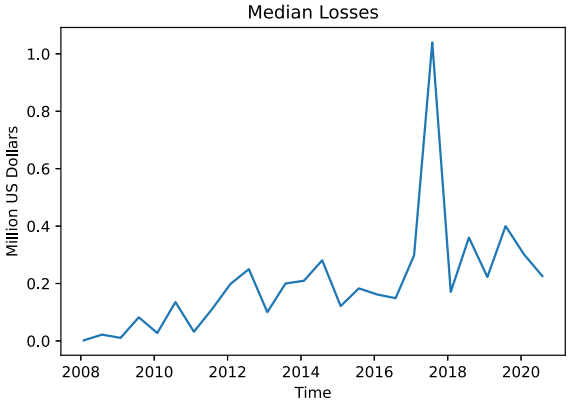
[4] For business line classification, given that the vast majority of the events affect companies residing in the USA, we adopt the North American Industry Classification System (NAICS), which comprises of 23 business sectors with corresponding codes available in the Advisen dataset.



(a) Number of losses resulting in disclosed losses during the period 2008-2020 in a sliding window of 6 months



(b) Mean loss during the period 2008-2020 in a sliding window of 6 months



(c) Median loss during the period 2008-2020 in a sliding window of 6 months

Fig. 1. These figures show the time evolution of the number of cyber events and resulting losses from 31 January 2008 to 31 July 2020 in a sliding window of 6 months.

- **Data – Unintentional Disclosure**: situations where personal confidential information or digital assets have either been exposed, or may have been exposed, to unauthorized viewers due to an unintentional or inadvertent accident or error.
- **Identity – Fraudulent Use/Account Access**: identity theft or the fraudulent use of confidential personal information or account access in order to steal money, establish credit, or access account information, either through electronic or other means.
- **Industrial Controls and Operations**: losses involving disruption or attempted disruption to "connected" physical assets such as factories, automobiles, power plants, electrical grids, and similar (including "the internet of things").
- **Network/Website Disruption**: unauthorized use of or access to a computer or network, or interference with the operation of same, including virus, worm, malware, digital denial of service (DDOS), system intrusions, and similar.
- **Phishing, Spoofing, Social Engineering**: attempts to get individuals to voluntarily provide information which could then be used illicitly, e.g. phishing or spoofing a legitimate website with a close replica to obtain account information, or sending fraudulent emails to initiate unauthorized activities (aka "spear phishing").
- **Skimming, Physical Tampering**: use of physical devices to illegally capture electronic information such as bank account or credit card numbers for individual transactions, or installing software on such point-of-sale devices to accomplish the same goal.
- **IT – Configuration/Implementation Errors**: losses resulting from errors or mistakes which are made in maintaining, upgrading, replacing, or operating the hardware and software IT infrastructure of an organization, typically resulting in system, network, or web outages or disruptions.
- **IT – Processing Errors**: losses resulting from internal errors in electronically processing orders, purchases, registrations, and similar, usually due to a security or authorization inadequacy, software bug, hardware malfunction, or user error.
- **Cyber Extortion**: threats to lock access to devices or files, fraudulently transfer funds, destroy data, interfere with the operation of a system/network/site, or disclose confidential digital information such as identities of customers/employees, unless payments are made.
- **Undetermined/Other**.

Other possible classifications of cyber risk events are available in the literature. For example, Eling and Wirfs (2019) suggests to divide cyber risk events into four categories, according to OpRisk classification: "Actions by People", "System and Technical Failure", "Failed Internal Process", "External Events" (see, Cebula and Young, 2010; Cebula et al., 2014). On the other hand, Romanosky (2016) provides cyber risk driven categories, such as "Data Breach", "Secu-

Table 1
This table reports descriptive statistics of cyber risk related losses aggregated by categories. All dollar values are reported in million dollars. The losses exhibit great variability in terms of median and first four moments across the considered risk types. "IT – Configuration/Implementation Error", "Privacy – Unauthorized Data Collection", and "Industrial Controls" have the highest average loss amongst all cyber risk categories.

Risk category	N	Mean	Median	StDev	Skew	Kurt
Privacy - Unauthorized Contact or Disclosure	1417	3.56	0.03	27.33	27.33	919.39
Privacy – Unauthorized Data Collection	113	54.69	0.45	472.56	10.15	103.14
Data - Physically Lost or Stolen	94	24.62	0.24	206.4	9.33	86.26
Identity – Fraudulent Use/Account Access	624	1.10	0.03	6.56	10.96	136.41
Data - Malicious Breach	719	24.94	0.53	187.02	15.94	303.05
Phishing, Spoofing, Social Engineering	179	8.91	0.55	54.28	12.11	153.21
IT - Configuration/Implementation Errors	57	18.76	0.82	45.58	2.95	8.61
Data - Unintentional Disclosure	175	1.52	0.12	9.74	11.59	141.28
Cyber Extortion	110	0.63	0.01	3.11	6.03	37.13
Network/Website Disruption	159	21.17	0.18	73.35	4.39	19.67
Skimming, Physical Tampering:	84	1.85	0.05	6.35	5.78	38.07
IT – Processing Errors	39	86.36	2.60	283.38	4.80	24.06
Industrial Controls	6	30.69	2.07	68.35	1.35	-0.1
Undetermined/Other	16	1.74	0.62	2.8	2.61	6.38

rity Incident", "Privacy Violation", "Phishing Skimming", and "Other".

Table 1 provides descriptive statistics of non-zero losses for each cyber risk category in the Advisen dataset. We find that the specified cyber risk classification exhibits great variability in terms of the number of recorded events in each risk category. This indicates that certain loss events are more prevalent, however this changes over time in an inhomogeneous manner. Furthermore, there is a heterogeneity in the severity distributions as evidenced by the first four moments of the loss distributions for each risk category. All risk categories exhibit a mean loss that is higher than the median, indicating skewness and potential for heavy tails and leptokurtic behavior. In some cases this effect is so pronounced that the mean is two orders of magnitude higher than the median and paired with high kurtosis. Additional goodness of fit analysis can be found in Section A.3 of Appendix A.

The descriptive statistics in Table 1 seem to be consistent with previous findings in the literature, that cyber risk related losses follow heavy tailed distributions; see, e.g. Maillart and Sornette (2010), Edwards et al. (2016), Eling and Loperfido (2017). One of the main objectives of the approach described in Section 2, is to estimate the tail parameter τ using a parametric approach to draw inference on how specific covariates can impact on the risk profile of cyber risk events. Here, we also present a non parametric estimate for the tail parameter, the Hill's estimator; see, e.g. Hill (1975), Grama and Spokoiny (2008), Durrieu et al. (2015), i.e. for a given threshold k, we consider the following estimator:

$$1/\widehat{\tau}_k(n) = \xi_k(n) := \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{y_{(n-i+1,n)}}{y_{(n-k,n)}}\right), \tag{6}$$

where $1 \le k \le n-1$ and $y_{(n,n)} \ge y_{(n-1,n)} \ge \cdots \ge y_{(1,n)}$ are the order statistics of the sample $y_i,\ i = 1, \ldots, n$.

Figs. 2 and 3 show the values for $\widehat{\tau}(k)$ broken down by cyber risk types and business sectors in the North American Industry Classification System (NAICS),[5] for different threshold values of the level in the Hill estimators order statistics used in estimator in Equation (6). As demonstrated in the box plots of the estimates for each risk category (Fig. 2) and NAICS business sector (Fig. 3), both decompositions demonstrate the typical behavior of the sample estimators for the population tail, i.e. both the median and the interquartile range (IQR) of the tail estimator for the companies involved in the aggregation suggest a tail index estimate lower than 0.5. This implies that the loss distributions exhibit extreme heavy tails and do not have finite first moments. There are some risk categories, where this behavior is less pronounced such as in "Privacy - Unauthorized Contact or Disclosure" where whilst the median and IQR of companies affected by this risk type have median and IQR consistent with infinite mean heavy tailed models, there is however a wide variety of variation in this case compared to other risk categories. We also suggest not to put too much emphasis on the results for the category "Undetermined/Other" due to its catch-all effect for many sources of loss events that cannot be allocated to any of the other 14 categories.

The key takeaway from this preliminary analysis is that, by and large, cyber related loss distributions are of the infinite mean type, regardless of the business sectors or risk category. The implications of these empirical findings are twofold. First, they justify the use of the POT method described in Section 2 for estimating the tail parameter, rather than focusing on OLS regression techniques, given that the data itself exhibits heavy tails. Second, the degree of heavy-tailedness of the cyber risk severity distribution poses a key aspects of our analysis. It could be argued that the parameter τ in Section 2 is one of the main factors discriminating the insurability of cyber risk. Having observed tail values of the Hill estimates lower than 1, this behavior of losses could jeopardize cyber risk insurability.

4. GAMLSS regression analysis of cyber risk losses

In this section, we present the empirical results of the applied GAMLSS regression to further examine the Advisen Cyber Loss Data. First, we discuss which covariates should be included in the analysis of the frequency and severity model parameters according to the literature. Second, we analyze whether cyber risk types should be modeled jointly or separately. Finally, we present the results of the POT method combined with the GAMLSS approach.

4.1. GAMLSS regression covariates

The GAMLSS approach allows us to identify relevant risk drivers and covariates as well as their impact on distributional parameters of the frequency and severity distribution in the LDA framework. The choice of covariates is dictated by problem specific instances, available information, as we all as previous findings in the literature on drivers of cyber risk. Several studies have observed changes in loss frequency and severity over time (Maillart and Sornette, 2010; Biener et al., 2015; Chavez-Demoulin et al., 2016; Eling and Wirfs, 2019). Observed changes in the number of events as well as in the magnitude of losses in Fig. 1 also seems to confirm the

[5] https://www.census.gov/naics/.

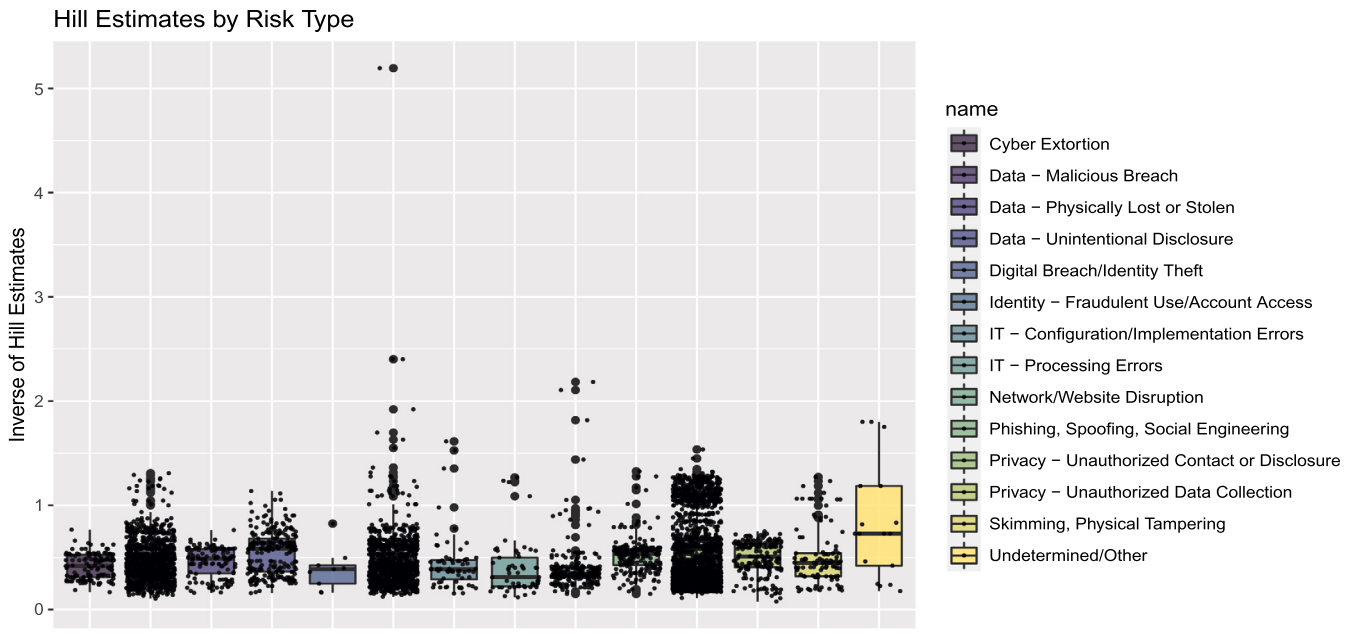


Fig. 2. Hill's estimates of monetary losses linked to cyber events from 2008 to 2020 by risk types.

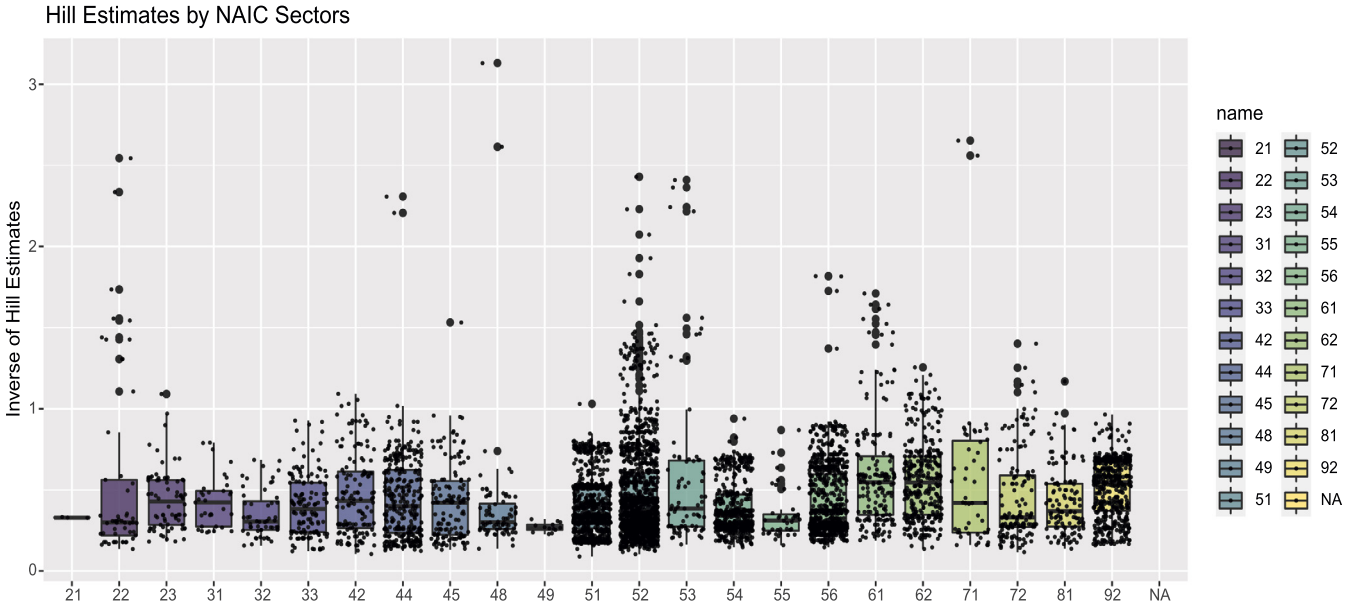


Fig. 3. Hill's estimates of monetary losses linked to cyber events from 2008 to 2020 by business sectors.

time-varying nature of cyber risk that has been observed by these authors for other datasets. Consequently, time is included as a covariate in the GAMLSS regression framework for both frequency and severity.

It is well documented in the literature that company size positively correlates with loss severity (see, among others, Cope and Labbi, 2008; Ganegoda and Evans, 2013; Eling and Wirfs, 2019). Thus, we develop a proxy for company size using revenue and the number of employees at the time of the cyber event. Note that we categorize revenue and size, i.e. the number of employees, into three equally sized groups. We acknowledge that discretizing continuous into categorical variables leads to a loss of information. However, as pointed out by Tsai and Chen (2019), a variety of data analytics techniques such as decision trees, naïve Bayes techniques, or many data mining algorithms take advantage of data discretization to construct more effective and efficient models. Categorization of continuous variables may also be used, if there is a concern that the regression results could be affected by extreme values, skewness or nonlinearities (Farrington and Loeber, 2000; Iacobucci et al., 2015). In our case, creating a balanced sample partition may reduce the effect of skewness in the proxies' distribution, at the same time allowing for nonlinear dynamics between company size and cyber event frequency and severity.[6] Controlling for the number of employees should also address the fact that human behavior is one of the major drivers for cyber risk as it has often been reported in the IT literature (Evans et al., 2016).

Another typical covariate, having a relationship with loss severity and frequency is the industry sector; see Dahen and Dionne

[6] Note that we also undertook a number of robustness checks that considered other possible splits with more categories and obtained qualitatively very similar results. Results for alternative splits of the revenue and number of employees variables are available upon request to the authors.

(2010), Ganegoda and Evans (2013), Chavez-Demoulin et al. (2016), Eling and Wirfs (2019). Since the majority of cyber events reported by Advisen are located in the USA and therefore regulated by the Freedom of Information Act (FOIA) and Health Information Technology for Economic and Clinical Health Act (HITECH), we distinguish between Financial industry, Healthcare industry, and other.[7] Following Eling and Wirfs (2019), we consider two additional covariates: *geographical location* and *contagion*. Location accounts for regulatory differences between different regions, i.e. in the USA-focused setting of this study for regulatory differences between the USA and the rest of the world.

Given that the majority of cyber risk events are in the USA, we distinguish between companies residing in the USA and the rest of the world. Recent findings in the literature have shown how the topology of a computer network can have an impact on how cyber threats spread into multiple entities (see, for example Antonio et al., 2021; Farkas et al., 2021). However, the Advisen dataset does not contain information regarding the network structure of companies affected by cyber events. Therefore, we adopt the definition of contagion provided by Advisen where events are related in case they share a common known source. For alternative definitions of contagion in the cyber risk context inspired by epidemiology, we refer the readers to Fahrenwaldt et al. (2018), Farkas et al. (2021), Antonio et al. (2021). We decided to classify contagion as three mutually exclusive realizations: events related to multiple losses in a single company, events related to multiple losses in different companies, and one shot events. Similarly, it is also common practice to include event categories as covariates to control for differences in the relationship for NAICS industry sectors or cyber risk loss categories.

Given that many covariates are company or institution specific we look at loss frequency and severity at the firm level. Therefore for each event we consider the following series of categorical variables:

- RT_s, $s = 1, \dots, S$ indicating the cyber risk type;
- L_l, $l \in \mathcal{L} = \{\text{USA, non-USA}\}$, indicating if the company that was affected by the cyber event resides in the USA or not;
- B_b, $b \in \mathcal{B} = \{\text{Finance, Healthcare, Other}\}$, indicating the company business sector;
- R_r, $r \in \mathcal{R} = \{\text{small,medium,big}\}$ indicating three classes of company revenue (small, if the revenue is lower than the 33rd percentile, medium if the company revenue is between the 33rd and 66th percentiles, and big if the company revenue is higher than the 66th percentile);
- E_e, $e \in \mathcal{E} = \{\text{small,medium,big}\}$ indicating three classes of company size with respect to the number of employees (small, if the number of employees is lower than the 33rd percentile, medium if the number of employees is between the 33rd and 66th percentiles, and big if the number of employees is higher than the 66th percentile).

Furthermore, we also include two additional dummy variables, ML and MC, indicating whether the event causes multiple losses in a single company or in various companies, respectively.

4.2. GAMLSS model selection

In this section we explore the ongoing debate regarding the effect of modeling cyber risk frequency and severity jointly across risk types or as individual models for each risk type within the LDA model framework (see, e.g. discussion in Edwards et al., 2016; Eling and Loperfido, 2017; Eling and Jung, 2018; Eling and Wirfs, 2019; Jung, 2021). Our findings suggest that due to the presence of heavy tailed features in the cyber event severity distribution, it's not possible to distinguish between separate or joint modeling. To explore this issue, we combine the POT method of EVT for severity within the regression framework of the GAMLSS approach. We compare a joint estimation of the GAMLSS regression structure across all risk types of the Advisen dataset against a separate estimation of a GAMLSS for the frequency and severity components of each risk type. Thus, we will be able to utilize formal hypothesis testing based on likelihood inference procedures to seek evidence for or against these approaches.

We distinguish the coupled model estimation approach as follows (termed the joint model), where we consider the following likelihood and link functions:

$$\log\left(L_1(Y;\mu,\tau)\right) = \sum_{i=1}^{m} \log\left(g(y_i;\mu,\tau)\right), \tag{7}$$

$$\log\left(\mu(X,t)\right) = \beta_0 + \sum_{s=1}^{S}\beta_s RT_s + \sum_{l\in\mathcal{L}}\beta_l L_l + \sum_{b\in\mathcal{B}}\beta_b B_b + \sum_{r\in\mathcal{R}}\beta_r R_r + \sum_{e\in\mathcal{E}}\beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \beta_t t,$$

$$\log\left(\tau(X,t)\right) = \beta_0 + \sum_{s=1}^{S}\beta_s RT_s + \sum_{l\in\mathcal{L}}\beta_l L_l + \sum_{b\in\mathcal{B}}\beta_b B_b + \sum_{r\in\mathcal{R}}\beta_r R_r + \sum_{e\in\mathcal{E}}\beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \beta_t t,$$

and the decoupled model (termed the separate risk category model) as follows:

$$\log\left(L_2(Y;(\mu_s)_{s=1,\dots,S},(\tau_s)_{s=1,\dots,S})\right) = \sum_{s=1}^{S}\sum_{i=1}^{n_s}\log(g(y_i^{(s)};\mu_s,\tau_s)),$$

$$\log\left(\mu_s(X,t)\right) = \beta_0 + \sum_{l\in\mathcal{L}}\beta_l L_l + \sum_{b\in\mathcal{B}}\beta_b B_b + \sum_{r\in\mathcal{R}}\beta_r R_r + \sum_{e\in\mathcal{E}}\beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \beta_t t,$$

$$\log\left(\tau_s(X,t)\right) = \beta_0 + \sum_{l\in\mathcal{L}}\beta_l L_l + \sum_{b\in\mathcal{B}}\beta_b B_b + \sum_{r\in\mathcal{R}}\beta_r R_r + \sum_{e\in\mathcal{E}}\beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \beta_t t,$$

where each coefficient β refers only to the equation it appears in and additional subscripts referring to parameter or risk type are omitted for notational convenience. In the joint model, the monetary loss due to cyber event is used as response variable, while risk types, along the other covariates, are used as explanatory variables in the link functions. By doing so, for specific values of covariates identifying a given company, it's then possible to obtain fitted parameters of the cyber event severity distribution at the company level. In the case of the separate model, for each risk type a separate model for cyber event severity is fitted. In each model, the response variable is the monetary loss due to a cyber event of the corresponding risk type. In this case, for any given company, with known characteristics, it's possible to obtain fitted parameters of the cyber event severity distribution at the company level for each

[7] Under the FOIA and HITECH, the Secretary for Health and Human Services has to publicly disclose breaches of unsecured protected health information affecting 500 or more individuals. For more information, see https://www.hhs.gov/.

Table 2
This table reports descriptive statistics of cyber risk related losses that exceed the corresponding threshold values for each risk category. All dollar values are reported in million dollars. The threshold for each risk category is chosen as the lowest possible value ensuring that the null hypothesis of GPD cannot be rejected for the exceeding losses.

Risk Type	u	Min	N	Mean	Median	StDev	Skewness	Kurtosis
Privacy - Unauthorized Contact or Disclosure	0.63	0.2	530	9.49	2.92	44.08	17.58	351.81
Privacy – Unauthorized Data Collection	-	0.00005	113	54.69	0.45	472.56	10.15	103.14
Data - Physically Lost or Stolen	-	0.0003	94	24.62	0.24	206.4	9.33	86.26
Identity – Fraudulent Use/Account Access	0.4	0.0156	375	1.83	0.16	8.38	8.47	80.82
Data - Malicious Breach	0.14	0.0263	619	28.96	0.94	201.29	14.78	261.09
Phishing, Spoofing, Social Engineering	-	0.0008	179	8.91	0.55	54.28	12.11	153.21
IT - Configuration/Implementation Errors	-	0.0012	57	18.76	0.82	45.58	2.95	8.61
Data - Unintentional Disclosure	0.04	0.0035	169	1.57	0.13	9.91	11.38	136.27
Cyber Extortion	0.02	0.0002	109	0.64	0.01	3.12	6	36.75
Network/Website Disruption	-	0.0006	159	21.17	0.18	73.35	4.39	19.67
Skimming, Physical Tampering	-	0.0003	84	1.85	0.05	6.35	5.78	38.07
IT - Processing Errors	-	0.0003	39	86.36	2.6	283.38	4.8	24.06
Industrial Controls and Operation	0.51	1.1348	4	46.03	6.5	82.73	0.75	-1.69
Undetermined/Other	-	0.0079	16	1.74	0.62	2.8	2.61	6.38

risk type. We apply Vuong's closeness test (Vuong, 1989) to determine if one can statistically distinguish between the joint and decoupled modeling approach. The test is essentially a likelihood-ratio test for model selection based on the Kullback–Leibler information criterion. To conduct the test, in a first step we formulate the null and alternative hypothesis as follows:

$$H_0: \text{var}\left(\log\left(\frac{L_1(Y;.,.)}{L_2(Y;.,.)}\right)\right) = 0 \qquad \text{vs}$$
$$H_1: \text{var}\left(\log\left(\frac{L_1(Y;.,.)}{L_2(Y;.,.)}\right)\right) \neq 0.$$

Based on the null hypothesis, we then aim to assess whether the two models are equally close to the true data generating process, against the alternative that one model is closer. Thus, in our setup we aim to verify if the following condition is satisfied based on the estimates for the joint and decoupled GAMLSS model structure:

$$g(y, \widehat{\mu}, \widehat{\tau}) = g(y, \widehat{\mu}_s, \widehat{\tau}_s), \qquad s = 1, \ldots, S,$$

where $\widehat{\mu}$ and $\widehat{\tau}$ are the fitted parameters of the joint model, and $\widehat{\mu}_s$ and $\widehat{\tau}_s$ are the fitted parameters under the decoupled model in which the model is fitted separately for each risk type.

The test statistic under the null hypothesis is given by the variance of the likelihood ratio and, under suitable regularity conditions, follows a χ^2 distribution with degrees of freedom equal to the number of parameters.

Applying the test to the proposed GAMLSS structure, our results suggest that we do not have enough evidence to reject the null hypothesis and therefore, the two alternatives appear to be indistinguishable. This last point further confirms the findings of the previous section and supports the statement that cyber event related losses are so heavy tailed that it is not possible to distinguish between the two nested models. In other words, across all Advisen risk type categories, one cannot statistically distinguish between the loss types, at least in their important attributes from an insurance perspective, relating to the tail behavior of cyber loss processes for each risk category. As a consequence, we are confident to therefore proceed the analysis by looking at separate models for each risk category given that the estimation routines require less computing time.

Table 2 shows threshold values and descriptive statistics for each category; "Privacy - Unauthorized Contact or Disclosure" has the highest threshold, both in percentage and in millions, while for the rest of the risk types the required threshold is almost negligible.

4.3. Dynamic extreme value theory estimates

In this section we present the results for cyber event frequency and severity regression models, which were estimated based on the assumptions of independence between frequency and severity and as such the estimation can be performed separately. The estimation results show how the POT method combined with the GAMLSS regression approach can capture the complex nature of cyber risk, allowing for heterogeneity among cyber risk types, and study the impact of time trend, company size, business sector, and contagion on cyber event frequency and severity distributions. For the frequency of cyber events we consider a Poisson distribution for each risk type, where the intensity parameter depends on covariates according to the following link function, where subscripts referring to the parameter λ and risk type are omitted for notational convenience:

$$\log(\lambda_s(X,t)) = \beta_0 + \sum_{l\in\mathcal{L}} \beta_l L_l + \sum_{b\in\mathcal{B}} \beta_b B_b + \sum_{r\in\mathcal{R}} \beta_r R_r + \sum_{e\in\mathcal{E}} \beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \beta_t t. \tag{8}$$

Here we assume h_λ to be linear in time; a detailed implementation of the spline smoothing approach in Chavez-Demoulin et al. (2016) to determine the functional form of h_λ can be found in Section A.4 of Appendix A. Table 3 reports the score ratios of each covariate for each risk type. Fitting the cyber risk types separately proves to be a flexible approach that allows us to capture the heterogeneity in the nature of losses and their driving factors for each risk category. By looking at the estimates, the impact of time on cyber event frequency clearly depends on the risk types. The frequency of cyber events related to privacy, identity, data breaches, and skimming and tampering is negatively related with time. These risk types compose the vast majority of cyber events in the dataset, which also explains the decreasing trend in Fig. 1. Even though we find consistent evidence of a decreasing trend in the number of cyber events, confirming the results in Eling and Wirfs (2019), we cannot rule out the fact that the frequency of events in the categories "Cyber Extortion" and "Phishing and Spoofing" is increasing over time. Moreover, many recent industry and technical reports have indicated that cyber extortion events are becoming more frequent, even though companies tend to report less cyber extortion events than other cyber event types (see, e.g. De Bolle, 2020; Falk, 2019). Recall that we measure the impact of company size on the frequency of events by considering the number of employees and revenue of the company. Also for these variables, we find substantial differences between the individual risk categories: iden-

Table 3
This table reports the regression score ratios for parameters in the link function log(λ). The significance of each coefficient is reported with the usual convention of 1,2, and 3 stars corresponding to a p-value of 10%, 5%, and 1%, respectively.

RiskType	β_0	Time	R_{medium}	R_{big}	E_{medium}	E_{big}	L_{USA}	$B_{financial}$	B_{health}	ML	MC
Privacy - Unauthorized Contact or Disclosure	17.8754***	-18.6083***	6.6067***	4.7541***	-4.2399***	-6.7822***	6.5116***	-4.7865***	-6.1643***	-0.0046	7.2530***
Privacy – Unauthorized Data Collection	0.4137	-2.5524**	0.4517	3.1365***	-1.3059	0.0869	1.4602	-3.8447***	-2.4056***	0.0514	0.0655
Data - Physically Lost or Stolen	4.6797***	-4.8496***	-2.5965***	-1.3731	0.3511	1.4979	-2.9452***	1.6965*	8.3679***	-0.0547	4.5084***
Identity – Fraudulent Use/Account Access	0.2037	-15.9455***	-8.4701***	-6.3861***	6.0657***	8.0248***	7.5698***	16.0172***	-1.4613	0.0087	0.0123
Data - Malicious Breach	4.6599***	-5.1782***	-1.03616	3.0066***	2.9143***	6.4114***	-0.8918	-0.3428	0.3386	0	8.3027***
Phishing, Spoofing, Social Engineering	-0.4375	7.7871***	-3.9149***	-1.9394*	5.1092***	5.3348***	-1.1372	-2.5014***	-1.9951*	0	0.0226
IT – Configuration/ Implementation Errors	-0.2305	0.6214	1.6194*	2.6201***	-0.6968	0.1246	-2.8849***	0.2772	5.1172***	0.0430	0.0571
Data - Unintentional Disclosure	0.2291	-2.1061**	0.68469	1.7449*	-1.3631	-1.2887	-8.4466***	-1.7927*	5.3561***	0	0.0627
Cyber Extortion	-1.5525	7.8595***	-3.9842***	-3.4462***	3.9356***	2.8025***	0.3568	-3.3763***	0.3533	0	0.0513
Network/Website Disruption	-0.3285	1.1902	-1.6372*	-1.5218	1.3012	4.8584***	-0.5805	-3.5876***	-1.3433	0	0.0767
Skimming, Physical Tampering	1.2323	-4.9652***	-1.4915	0.5161	-0.0864	1.2158	2.9849***	6.2142***	-0.1249	0.0540	0.0699
IT - Processing Errors	0.0843	-0.8353	-0.6775	0.1934	-1.0146	0.7012	-2.5285***	3.8826***	-0.0149	0	0.0305
Industrial Controls and Operations	-0.1066	1.1032	-0.0091	-0.193	0.0091	0.0091	-0.5626	-0.0069	-0.0043	0	0.0084
Undetermined/Other	-0.0940	0.4228	0.1462	1.3643	-0.6335	-0.725	1.1165	3.1339***	-0.0093	0	0.0196

tity, phishing, and cyber extortion are more frequent in companies with low average labor productivity (medium and high number of employees and low revenue), indicating that for these risk categories the frequency is higher for companies lacking efficient resource allocation, technological know-how, or returns of scale. On the other hand, companies with medium and high revenue suffer more frequently from privacy related events. The frequency of cyber events also depends on business sectors. Financial and health sectors have similar effects in privacy, "Data - Physically Lost or Stolen", and "Phishing, Spoofing, Social Engineering". "Cyber Extortion", and "Data - Unintentional Disclosure" happen to be less frequent in the financial sector than in healthcare, due to the strict regulation around data policy in the financial sector. Looking at the effect of contagion, only the frequency of "Privacy - Unauthorized Contact or Disclosure", "Data - Physically Lost or Stolen" and "Data - Malicious Breach" results appear to be impacted by events spilling over from company-to-company.

For cyber event severity we consider a GPD for each risk type, with the following link functions:

$$\log(\mu_s(X,t)) = \beta_0 + \sum_{l\in\mathcal{L}} \beta_l L_l + \sum_{b\in\mathcal{B}} \beta_b B_b + \sum_{r\in\mathcal{R}} \beta_r R_r + \sum_{e\in\mathcal{E}} \beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \beta_t t, \tag{9}$$

$$\log(\tau_s(X,t)) = \beta_0 + \sum_{l\in\mathcal{L}} \beta_l L_l + \sum_{b\in\mathcal{B}} \beta_b B_b + \sum_{r\in\mathcal{R}} \beta_r R_r + \sum_{e\in\mathcal{E}} \beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \beta_t t. \tag{10}$$

Here we assume h_μ, and h_τ to be linear in time; a detailed implementation of the spline smoothing approach in Chavez-Demoulin et al. (2016) to determine the functional form of h_μ, and h_τ can be found in Section A.4 of Appendix A. Tables 4 and 5 show the results for cyber event severity. From an insurance and risk management perspective, one is primarily interested in the tail index relationships. In this section we therefore focus the discussion on the tail parameter τ_s. A tail parameter lower than 1 implies that the distribution has no finite moments. In such a case, the expected loss of the corresponding cyber risk type is infinite. Therefore, in Table 5, covariates with a negative estimated coefficient increase cyber risk type riskiness. Similarly to cyber event frequency, severity of privacy and identity related events follows a decreasing time trend, while "Data - Malicious Breach", "Phishing, Spoofing, Social Engineering", and "Cyber Extortion" severity has increased over time.

Looking at the impact of company size, the picture is less clear for the severity regression models. Losses in the category "Privacy – Unauthorized Data Collection" are more severe for labor intensive companies, suggesting the relevance of human behavior for these kinds of events. Nevertheless, no other privacy type of events supports the same findings. Losses in the category "Data - Physically Lost or Stolen" are less severe in companies with medium and high number of employees, while "Cyber Extortion" events are more severe in capital intensive companies (companies with high revenue and low number of employees). Data physically stolen and unintentionally disclosed are less severe in the USA, what might be a result of the more restrictive data protection policies in the USA in comparison to the rest of the world. The financial and healthcare sectors seem to suffer less severe losses in almost all cyber risk categories, except for "Data - Unintentional Disclosure", where the financial sector seems to be more exposed than other business sectors. The last two columns of Table 5 show the impact of contagion on cyber event severity. Data malicious breaches, caused by or causing other cyber events in the same company, have the potential to trigger heavier financial losses. While for the risk types "Privacy - Unauthorized Contact or Disclosure" and "Identity – Fraudulent Use/Account Access" spillover events in the same company are generally less severe. This may indicate that for these particular risk categories companies have cyber risk management practices in place that reduce losses from repeated breaches. Events linked to external enterprises increase the severity for the categories "Privacy – Unauthorized Data Collection" and "Data - Physically Lost or Stolen".

Additional analysis on the interaction effect among dummy covariates and Q-Q plot can be found in Section A.4 of Appendix A.

5. Rank-based regression of cyber risk loss processes

Results from both the non-parametric tail estimation and the GAMLSS regression analysis suggest that cyber event severity exhibits extreme heavy tails, with no finite first moments and no easily identifiable structure. Given the data structure, it is also possible that our results for the identified drivers of cyber risk are at least partially driven by few very extreme observations in the individual risk categories. We therefore seek to extend our analysis by applying an additional rank-based approach that removes the scale

Table 4
This table reports the regression score ratios for parameters in the link function $\log(\mu)$. The significance of each coefficient is reported with the usual convention of 1, 2, and 3 stars corresponding to a p-value of 10%, 5%, and 1%, respectively.

RiskType	β_0	Time	R_{medium}	R_{big}	E_{medium}	E_{big}	L_{USA}	$B_{financial}$	B_{health}	ML	MC
Privacy - Unauthorized Contact or Disclosure	6.3037***	-6.2111***	-1.6977*	2.6485**	0.3417	1.9169**	3.859***	8.413***	-0.0879		-4.1663***
Privacy – Unauthorized Data Collection	4.2443***	-4.2744***	5.7174***	7.9349***	-4.256	-8.1955***	3.4149***	13.7316***	4.4548***	12.3974***	
Data - Physically Lost or Stolen	1.1337	-1.1599	-0.5047	-0.7503	5.1968	5.2282***	-1.5124*	5.5719***	3.9608***		-1.591
Identity – Fraudulent Use/Account Access	1.2702	-1.2843	2.7723***	-2.6023*	0.6804	3.8916***	-3.083***	-2.8866***	-1.4372	1.0646	
Data - Malicious Breach	-7.1916***	7.3144***	0.4154	3.099***	-1.1811	0.1218	5.3967***	3.0396***	4.5634***	-6.6742***	-11.0582***
Phishing, Spoofing, Social Engineering	-6.3192***	6.3203***	4.6131***	0.1463	-0.9365	-1.5422*	-1.8377**	1.2422	7.8563***		
IT – Configuration/ Implementation Errors	6.0355***	-6.0597***	7.821***	3.6946***	3.6344	-4.1544***	-0.2476	3.4372***	5.2348***	6.6952***	
Data - Unintentional Disclosure	4.354***	-4.3879***	0.4246	-1.6519**	3.3095	3.9499***	-1.404*	-2.3172**	4.3793***		
Cyber Extortion	-16.203***	16.1703***	-4.9287***	0.8184	2.1502	-0.9929	0.867	3.7901***	10.8866***		
Network/Website Disruption	-4.3018***	4.2737***	2.5043**	1.7096*	-2.2092	-0.6527	0.6316	0.3401	6.0102***		
Skimming, Physical Tampering	1.7656*	-1.7583*	-2.642**	-1.0112	-0.5355	0.9971	-5.8438***	-1.5343*	3.4987***	-4.201***	
IT - Processing Errors	-1.0354	1.0261	0.4021	0.3733	0.6079	-1.3304	4.3801***	3.2282***			
Industrial Controls and Operation	-1.9461*	1.9537*			-2.8584						
Undetermined/Other	0.8327	-0.8327	1.8804	2.0735**	-1.6645	-1.8697*	1.1559	1.479*			

Table 5
This table reports the regression score ratios for parameters in the link function $\log(\tau)$. The significance of each coefficient is reported with the usual convention of 1,2, and 3 stars corresponding to a p-value of 10%, 5%, and 1%, respectively.

RiskType	β_0	Time	R_{medium}	R_{big}	E_{medium}	E_{big}	L_{USA}	$B_{financial}$	B_{health}	ML	MC
Privacy - Unauthorized Contact or Disclosure	6.6067***	-6.5183***	-0.1413	3.0814***	-1.0599	0.0451	-0.5747	6.635***	1.7084**		-2.6275***
Privacy – Unauthorized Data Collection	4.3034***	-4.3115***	6.0298***	4.7691***	-4.3361***	-6.3732***	0.1952	9.8741***	5.0142***	8.1338***	
Data - Physically Lost or Stolen	0.8894	-0.8623	-0.2714	-0.8924	2.6821***	3.139***	-2.2156**	1.6157**	2.1862**		-3.2789***
Identity – Fraudulent Use/ Account Access	2.7085***	-2.7197***	1.3889	-1.0612	1.2192	1.2382	0.9968	-1.273	-1.7623*	2.9839***	
Data - Malicious Breach	-2.3488**	2.4646**	-0.5338	-0.1638	-0.0974	-0.3404	2.3161	1.0256	7.1538***	-4.072***	-5.1573
Phishing, Spoofing, Social Engineering	-5.2824***	5.2772***	4.9515	-1.6083	0.9595	-0.8269	1.6028	2.1316**	7.6076***		
IT – Configuration/ Implementation Errors	2.7334***	-2.7344***	7.8107***	1.3124	6.3982***	-3.8609***	-0.8266	0.4187	3.4246***	3.3014***	
Data - Unintentional Disclosure	3.3412***	-3.3319***	1.0442	-1.5094	1.9885**	1.572	-7.3631***	-1.6987**	3.7304***		
Cyber Extortion	-5.3688***	5.3695***	-3.6261***	-2.5504**	2.8859***	-0.9208	0.2467***	1.797***	12.2652***		
Network/Website Disruption	0.6384	-0.657	3.2336***	-0.8623	-1.2996	-1.3532	2.0436**	0.5777	4.2778***		
Skimming, Physical Tampering	2.9529***	-2.9405***	-2.3306**	-0.0658	1.0398	-0.971	-3.2267***	0.6312	8.7332***	-1.0688	
IT - Processing Errors	-1.683*	1.6756*	3.8958***	3.1338***	-0.3692	-3.5478***	2.2917**	1.0939			
Industrial Controls and Operation	0.0042	-0.0004			0.0003						
Undetermined/Other	-0.1794	0.1891	-0.0265	-0.024	0.1261	-0.0299	-0.0431	-0.0355			

of the effect of extreme loss events. Thus, we apply an ordinal regression framework in order to control for the leverage effects that may be present in the regression estimators that could arise from the extreme magnitude of cyber risk losses. Ordinal regression can also be seen as an additional robustness check on the statistical significance of the covariates in the combined POT and GAMLSS approach. The main findings of the ordinal regression indicate that the heavy tails of cyber event severity are the major drivers behind the results of previous sections.

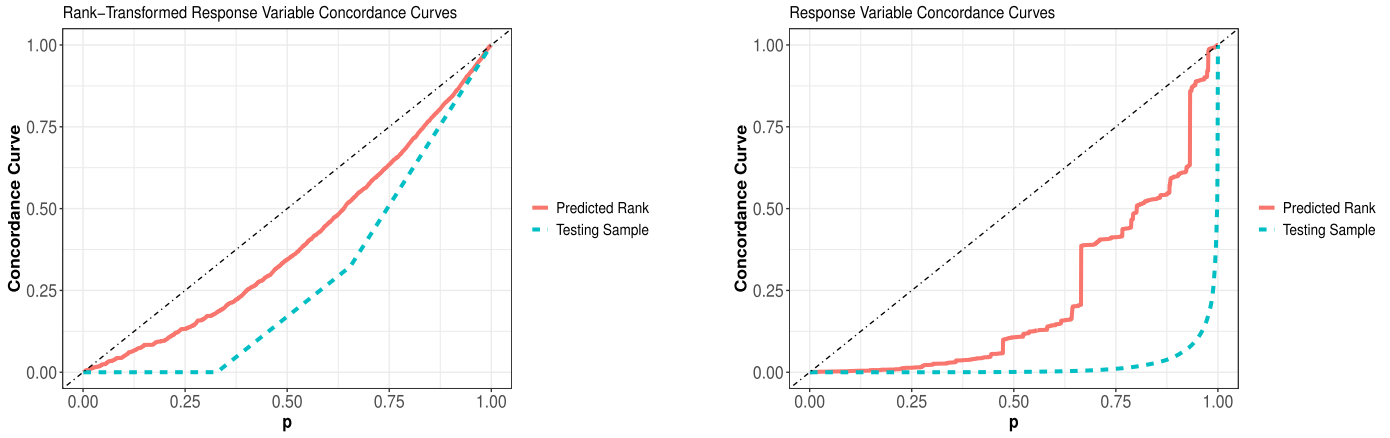
Ordinal regression analysis has been implemented in the context of cyber risk in various settings, where often data is expressed in terms of ordered level of severity (see, among others Hubbard and Evans, 2010; Raffinetti and Romeo, 2015; Sexton et al., 2015; Giudici and Raffinetti, 2020). Typically, cyber event severity Y can be transformed into its rank R, and then explained by a linear regression model. By construction, the rank transformed response variable does not present heavy tails anymore. Let Y to be expressed in k-levels, then cyber severity can be transformed into ranks according to the following steps: $r_1 = 1$ corresponds to the rank of the smallest category, $r_z = n_{z-1} + r_{z-1}$, with n_z being the absolute frequency of rank r_z, and $z = 1, \ldots, k$; see Iman and Conover (1979). Then the impact of covariates on the rank transformed cyber event severity can be addressed with the following linear regression:

$$R = \beta_0 + \sum_{s=1}^{S} \beta_s RT_s + \beta_t t + \sum_{l \in \mathcal{L}} \beta_l L_l + \sum_{b \in \mathcal{B}} \beta_b B_b + \sum_{r \in \mathcal{R}} \beta_r R_r + \sum_{e \in \mathcal{E}} \beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \epsilon, \tag{11}$$

with $\mathbb{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$. To evaluate the goodness of fit of the regression in (11), we employ the cross-validation method based on the Rank Graduation Accuracy (see, Giudici and Raffinetti, 2020):

$$RGA_R = \sum_{i=1}^{n} \frac{n}{i} \left(\frac{1}{n\bar{r}} \sum_{j=1} r_{ord(\hat{r}_j)} - \frac{i}{n} \right)^2,$$

where $\bar{r}$ is the average rank, $r_{ord(r_j)}$ is the rank transformed response variable reordered according to the predicted rank $\hat{r}_j$. RGA_R is based on the concordance curve given by the pairs



(a) Concordance curve of rank-transformed response variable

(b) Concordance curve of response variable

Fig. 4. This figure shows the concordance curves of cyber event severity, in terms of rank-transformed response variable (panel (a)) and response variable (panel (b)). Rank-transformed response variable and response variable are reordered according to the predicted rank in the training sample.

$\left(i/n, \frac{\sum_{j=1}^{i} y_{r_j}}{n\bar{y}}\right)$, for $i = 1, \ldots, n$. One interprets RGA_R as follows: a large value implies a better fit with respect to a random model with no predictive power and a lower value implies the presence of a model that may not be readily distinguished from an arbitrary random model with no predictive power.

Fig. 4 shows the concordance curve in terms of rank-transformed response variable (panel (a)) and in terms of the response variable reorder according to $\widehat{r}_j$ (panel (b)) for the linear model in (11). The bisector corresponds to the concordance curve of a model with no explanatory power (random model), where all observations share the same predicted rank. In other words, the further away a concordance curve is from the bisector, the better the performance of the corresponding linear model. The rank transformed concordance curves (in red in both panel (a) and (b)) lie below the bisector, implying the linear model in (11) has a better predictive power than a random model.

To test for the significance of the regression covariates in the ordinal regression we adopt the testing methodology based on the RGA_R in Raffinetti and Romeo (2015), Giudici and Raffinetti (2020). Here, we report only the main findings, a detailed description of the procedure and results can be found in Section A.5 of Appendix A. We first proceed in testing the significance of each coefficient of the linear model in (11). The outcome of this analysis indicates that time, medium number of employees and USA are the only statistical significant dummy covariates. Furthermore, we found that cyber risk types are not statistically significant, neither jointly nor separately. In terms of RGA, all the tested models present similar values of RGA, with the model where the covariate controlling for high revenue companies (R_{big}) is omitted, scoring the highest. These findings seem to suggest that the results on the significance of covariates in the previous section involving the study of the GAMLSS regression are mainly driven by extreme cyber events, an outcome we suspected could be the case and which led us to explore the ordinal regression as a means of confirmation.

Finally, as a robustness check on the result of the Voung's variance test, we test whether the rank transformed response variable should be fitted jointly on the risk categories, or separately. In other words, we perform the following test of hypothesis:

$$H_0 : \text{Joint Model} \qquad \text{vs} \qquad H_1 : \text{Separate Model}$$

where in the separate model, the rank-based regression is as follows:

$$R_s = \beta_0 + \beta_t t + \sum_{l \in \mathcal{L}} \beta_l L_l + \sum_{b \in \mathcal{B}} \beta_b B_b + \sum_{r \in \mathcal{R}} \beta_r R_r + \sum_{e \in \mathcal{E}} \beta_e E_e + \beta_{ml} ML + \beta_{mc} MC + \epsilon,$$

where R_s is the rank transformed response variable of risk type s. We find a significance value above 80%, implying that the models are almost always significantly different. This confirms once again the evidence to further support our conjecture regarding the findings in the previous sections, that one must be careful regarding the indifference one can take to the joint versus decoupled approaches, due to the sensitivity that this analysis may have to the extreme losses in the sample. In Section 4.2 it was determined that cyber event severity is so heavy tailed that it's not possible to distinguish between a joint and a separate model. However, once the effects of extreme events are removed through a rank based ordinal regression analysis, the conclusion changes and it is no longer the case that one can say that the joint and decoupled modeling approaches are indistinguishable. Moreover, it implies that the Voung's variance test result is more likely to be driven by extreme events in the loss data, rather than the independence assumption made in Section 2. Consequently, we have included further studies on the joint model for completeness of the analysis in Sections A.5 and A.6 of Appendix A.

As a conclusion to this debate, it comes down to what types of data one wishes to make inference about. If the raw monetary loss amounts are under consideration then this debate regarding joint versus decoupled is shown to be inconsequential in the aforementioned analysis. However, in the case where the relative severity effects are under investigation, which might be useful for management purposes and comparative analysis of effectiveness of certain management and business decisions relating to cyber mitigation for example, one may resort to model ranks. Then in this case, one will need to consider the joint modeling framework and decoupled framework as distinguishable and then perform analysis of each case.

Hence, in analyzing cyber event severity by the means of monetary losses, one can either proceed with the joint or the decoupled approach, the latter having the advantage of superior flexibility that can capture cyber risk heterogeneity. When instead one wishes to model rank transformed cyber severity, either for filtering out the effect of extreme events or due to other specific data characteristics, one needs to take a decision on which approach is more statistically sound. Such a decision can then be made, using the RGA_R and concordance curve framework, or any other sensible accuracy measure.

6. Addressing cyber risk capital and insurability via GAMLSS regression models

In this section we present two case studies that will utilize the GAMLSS regression models in the context of both risk management and insurance. The first case study is a Value-at-Risk calculation based on the estimates obtained in Section 4. The second application is related to insurance premium calculations, based on Advisen data and our estimates.

6.1. Value-at-risk for cyber risk capital reserves under GAMLSS regression model

The three sets of analysis undertaken in previous sections have provided strong statistical evidence to assert that cyber risk related losses are heavy tailed. It is well known that heavy tails have serious implications for capital requirement calculations (see, e.g. Nešlehová et al., 2006). This subsection deals with capital requirement calculations for cyber risk. It shows how the inclusion of covariates in cyber risk modeling can help to better represent the heterogeneous nature of cyber risk and how it affects capital requirement calculations. Estimates in Tables 3, 4 and 5 show that cyber event frequency and severity depend on cyber threat types and company characteristics. To illustrate the impact of covariates on capital requirements we present the Value-at-Risk calculation using the Single Loss Approximation (SLA); see Degen (2010), Peters et al. (2013) and book length reviews in Peters and Shevchenko (2015), Cruz et al. (2015):

$$\mathrm{VaR}_\alpha(Z) \approx \begin{cases} u + GPD^{-1}\left(1-\frac{1-\alpha}{\widehat{\lambda}};\widehat{\mu},\widehat{\tau}\right)\left(1-\frac{(1-\alpha)c(\widehat{\tau})}{1-\widehat{\tau}}\right) \\ \quad \text{for } 0<\widehat{\tau}<1 \\ u + GPD^{-1}\left(1-\frac{1-\alpha}{\widehat{\lambda}};\widehat{\mu},\widehat{\tau}\right)+\widehat{\lambda}\frac{\widehat{\mu}}{\widehat{\tau}-1} \\ \quad \text{for } \widehat{\tau}>1, \end{cases} \tag{12}$$

where u is the threshold, $c(\widehat{\tau}) = \frac{1}{2}(1-\widehat{\tau})\frac{\Gamma^2(1-\widehat{\tau})}{\Gamma(1-2\widehat{\tau})}$, $GPD^{-1}(\cdot;\widehat{\mu},\widehat{\tau})$ is the inverse distribution function of a GPD random variable with parameters $\widehat{\mu}$ and $\widehat{\tau}$, and $\Gamma(\cdot)$ is the gamma function. Using the fitted values of λ, μ and τ from the GAMLSS approach in (12), it is possible to relate specific business models and structures to capital requirements for any give cyber risk type. Thanks to the covariates formulation and model specification, for a given company with known characteristics, the capital requirement calculation can be carried out at a company level, for each cyber risk type. In line with the discussion of the previous section, we present two hypothetical examples to illustrate how the GAMLSS based SLA for this regression based LDA model is instructive for risk management and capital reserving. The first example shows the impact on capital requirements of a fitted value of τ lower than 1. The second one focuses on the effects of contagion, allowing cyber related events of the type "Data - Malicious Breach" to be linked to multiple losses in various companies. Fig. 5 depicts the fitted values of λ, τ, and the corresponding Value-at-Risk of two risk types ("Network/Website Disruption" in red and "Privacy - Unauthorized Contact or Disclosure" in blue) for one company, residing outside the USA, operating neither in financial nor in healthcare sector, and having high revenue and high number of employees.

Panel (a) of Fig. 5 shows the fitted values of λ on the log scale. "Network/Website Disruption" frequency was higher than "Privacy - Unauthorized Contact or Disclosure" frequency during the period 2008-2020. Moreover, "Network/Website Disruption" frequency is following an increasing trend, while "Privacy - Unauthorized Contact or Disclosure" frequency is decreasing over time, in agreement with the score ratios in Table 3. Looking at panel (b), both the risk types show decreasing trends in the values of $\log(\widehat{\tau})$, meaning that their severity distributions have become more heavy tailed over time, with an increasingly higher probability of extreme events. What really sets the difference between these two risk types is that "Network/Website Disruption" has $\log(\widehat{\tau})$ lower than 0, implying that the corresponding distribution has no finite first moment, and therefore, events from such type can generate an infinite expected loss. This last aspect is further shown in panel (c) with the Value-at-Risk (on the log scale) computed using formula (12). The Value-at-Risk for "Network/Website Disruption" is increasing over time following what it seems to be an exponential growth path, and indicating that the capital required in the case of infinite mean loss distribution is much higher than in the case of "Privacy - Unauthorized Contact or Disclosure".

The effect of contagion on "Data - Malicious Breach" is shown in Fig. 6. Fig. 6 shows the fitted values of λ, τ, and the corresponding Value-at-Risk for two companies, both residing outside the USA, having medium revenue and high number of employees, and suffering from multiple events in red (henceforth company A) and suffering only from one shot events in blue (henceforth company B). Data malicious breaches are more frequent and severe in company A than B, having higher fitted values of λ and lower fitted values of τ. Considering contagion type of events, increases the riskyness of data malicious breaches considerably, since the corresponding values of $\log(\widehat{\tau})$ are shifted downwards below the threshold level of one, turning a cyber risk types with finite first moments into the infinite mean loss distribution case. These two hypothetical examples show the importance of considering covariates into capital requirement calculation. The heterogeneous nature of cyber risk is then reflected into the variations in capital requirements for different cyber risk types. Similar results can be obtained with other combination of covariates and risk types.

6.2. Cyber risk insurance

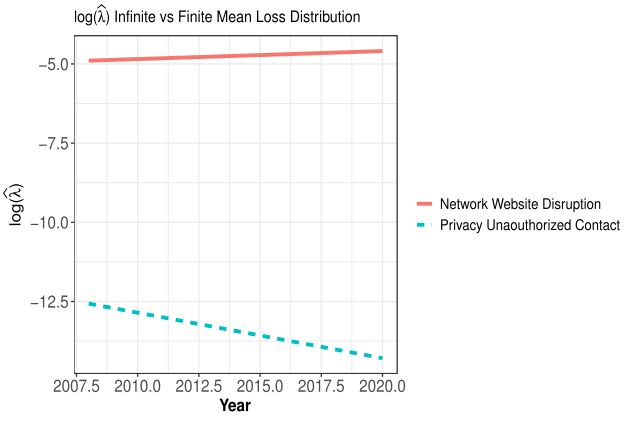
The statistical features of cyber risk have important implications on modeling and on prudential capital requirements. In this section we address the implication of our findings on the insurabiltiy of cyber risk. By using a simple insurance premium calculation, we show that cyber risk reveals to be a formidable foe from the insurance perspective, having many aspects that undermine its insurability.

Consider a non satiable risk averse decision maker with total wealth w, facing the possibility to suffer from $\widetilde{Y}_1, \ldots, \widetilde{Y}_N$ random losses, where N follows a poisson distribution. The decision maker is offered 1 year insurance policy to protect against each random loss up to a top cover limit equivalent to a percentage of the company wealth. According to the zero utility principle, the maximum premium he is willing to pay will be the solution P^+:

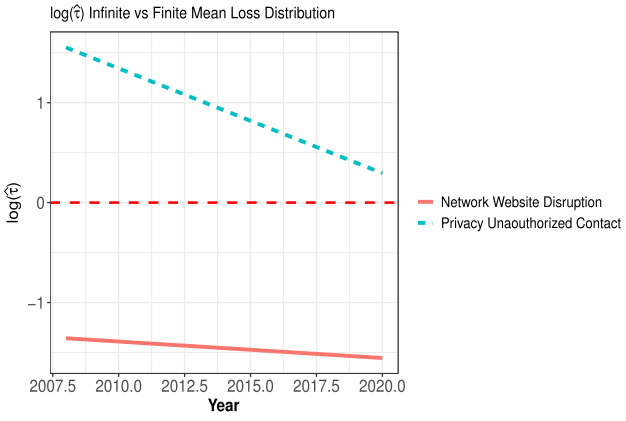
$$\begin{aligned} &\mathbb{E}\left[u\left(w-\sum_{i=1}^{N}\widetilde{Y}_i\right)\right] \\ &= \mathbb{E}\left[u\left(w-P^+-\sum_{i=1}^{N}\widetilde{Y}_i+\sum_{i=1}^{N}\min(\widetilde{Y}_i,kw)\right)\right], \end{aligned} \tag{13}$$

where $u(x) = \widetilde{u}(\max(x,1))$, $\widetilde{u}$ is a concave, non decreasing utility function, and k is the percentage corresponding to the top cover limit. The minimum premium an insurer with utility function v and wealth W is willing to accept to insure the decision maker is given by P^- solution of the following nonlinear equation:

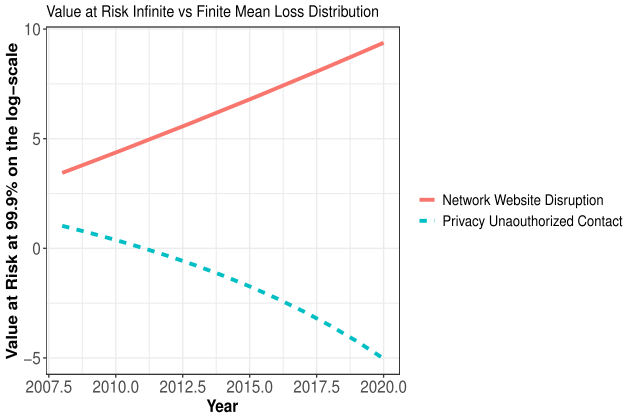
$$v(W) = \mathbb{E}\left[v\left(W+P^- - \sum_{i=1}^{N}\min(\widetilde{Y}_i,kw)\right)\right].$$



(a) Fitted values of $\log(\widehat{\lambda})$



(b) Fitted Values of $\log(\widehat{\tau})$



(c) Value-at-Risk

Fig. 5. This figure shows the Value-at-Risk and fitted values for λ, τ of "Network/Website Disruption" in red and "Privacy - Unauthorized Contact or Disclosure" in blue for the hypothetical company. The values of τ under the red dashed line in panel (b) correspond to the infinite mean case. Capital requirements for "Network/Website Disruption" are much higher in the case of "Privacy - Unauthorized Contact or Disclosure". (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

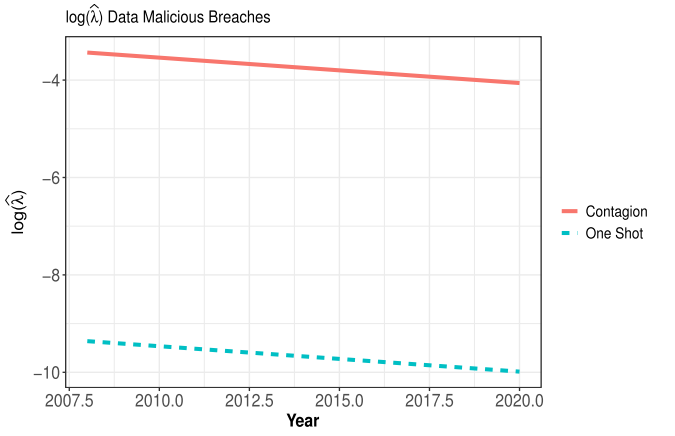
The decision then to insure and to be insured depends, among other things on the decision maker and insurance company total wealth. To get rid of one dimension, we consider the hypothetical case where P^+ is the equilibrium premium (i.e. $P^+ = P^-$) and then we focus on *how big* the insurance company should be in order to insure the decision maker. In other words, assume that $W = mP^+$, where m is the size of the insurance pool. Then, the problem for the insurer becomes, finding the optimal pool size, in order to be indifferent between insuring or not. The optimal pool size is given by m, solution of the following equation:

$$v(mP^+) = \mathbb{E}\left[v\left(mP^+ + P^+ - \sum_{i=1}^{N} \min(\widetilde{Y}_i, kw)\right)\right]. \tag{14}$$

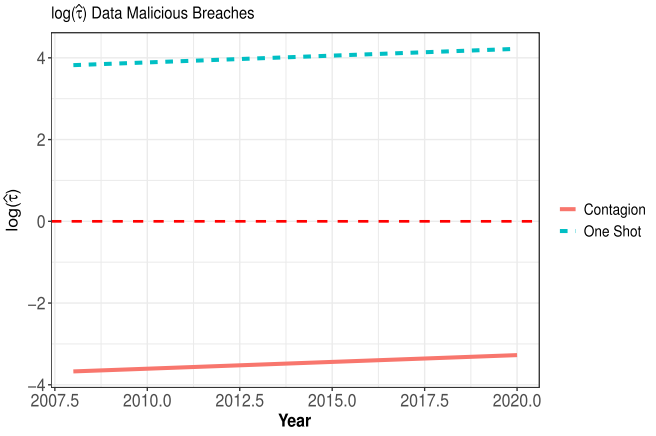
As a comparison tool between insurer and insurance, we also compute the insurer relative wealth $w_i = mP^+/w$, which indicates how much bigger in terms of size, the insurer should be in order to issue the insurance contract to the company. Risk types with high relative wealth could be considered hard to insure, since the insurer would be required to have a much higher capital than the case of risk types with lower relative wealth.

The results of the combined POT and GAMLSS approach can be used to compute insurance premiums specific for a given company. This should help to address the insurabiltiy question at a company specific level, allowing to consider the heterogeneous nature of cyber risk and at the same time reducing uncertainty. Similar to the Value-at-Risk case, the calculated insurance premium would reflect the riskiness of a given risk type for predetermined company characteristics. We consider a financial company, residing in the USA, with a high number of employees and high revenue, including also contagion type of events, and compute the insurance premiums for each risk types using a simulation approach. Figs. 7, 8, and 9 show the premiums and the corresponding insurer relative wealth in million dollars and on the log scale for "Privacy - Unauthorized Contact or Disclosure", "Data Malicious Breach", and "Cyber Extortion" respectively, from 2008 to 2020, using logarithmic utility function, CRRA utility function with $\gamma = 0.2$, and $\gamma = 0.7$. Company capital w is equal to 1 billion dollars and the top cover limit k is set to 10% of company capital.

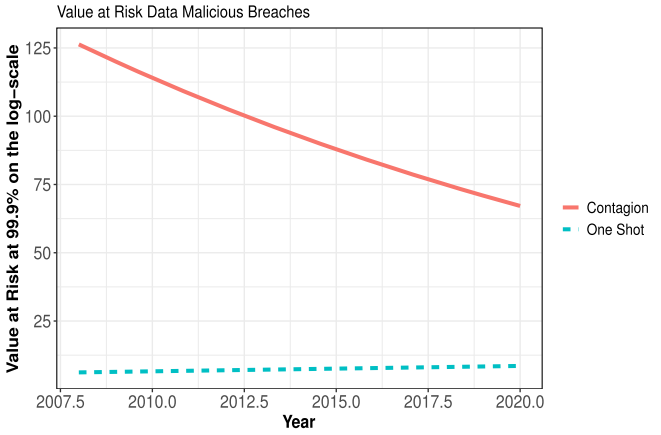
"Privacy - Unauthorized Contact or Disclosure" premiums and relative wealth in Fig. 7 show the case when cyber risk is insurable, since both premiums and relative wealth are low. Nonetheless, this situation of low premiums is only attainable in the case of finite mean loss distribution, as shown in Fig. 8. Fig. 8 shows the premiums and relative wealth of "Data Malicious Breach", to which corresponds a case of infinite mean loss distribution. For the considered company, the computed insurance premiums are too costly, making it very hard, if not impossible to buy the insurance. The same applies to the offer side. In this case, the insurer requires to have a wealth more than e^{20} times greater than the company total wealth. In cases where the underling cyber event severity distribution is of the infinite mean type, insurability is compromised.



(a) Fitted values of $\log(\widehat{\lambda})$



(b) Fitted Values of $\log(\widehat{\tau})$



(c) Value-at-Risk

Fig. 6. This figure shows the Value-at-Risk and fitted values for λ, τ for "Data - Malicious Breach". The values of τ under the red dashed line in panel (b) correspond to the infinite mean case. Capital requirements for "Data - Malicious Breach" due to contagion type of events are much higher than those for one shot events.

Finally, Fig. 9 depicts the case of "Cyber Extortion" as an example of the effect of non stationary in the tail index on insurance premiums and relative wealth. As shown in Tables 3, 4, and 5, cyber event frequency and severity depend also on time. This time dependence also affects the insurability of cyber risk types, since the necessary wealth for the insurance company would need to be adjusted accordingly. The "Cyber Extortion" depicted in Fig. 9 shows that, while the premiums and relative wealth drop from non feasible to more feasible values, it also requires a rapid adjustment in the insurance company wealth, decreasing from being e^7 times higher the company wealth, to being a fraction of the company wealth in only 3 years.

7. Conclusions

The expanding reliance of businesses and enterprises on information technology has resulted into an increase of the importance of cyber risk. Decision and policy makers have started to investigate the matter recently, and so does the actuarial community, with academic research, insurance industry, and risk managers. Nevertheless, the scarcity of good quality datasets is a common limitation among the many areas of study on cyber risk. We used the industry leading dataset provided by Advisen, to study cyber risk modeling and insurance aspects. In particular, we focused on two main unresolved questions on cyber risk: which factors are important explanatory variables for cyber event frequency and severity, and what can be inferred on cyber risk insurability.

In our analysis, we found that cyber event severity distributions often are of the infinite mean type, having important implications both from a modeling and insurance perspective. Regarding the modeling of cyber risk, we found that no standard OLS based techniques would be able to estimate correctly relevant parameters. We then used the POT method combined with the GAMLSS approach, allowing cyber event frequency and severity distributional parameters to depend on covariates. Using Voung's closeness test (Vuong, 1989), we found that a joint GAMLSS regression structure across all cyber risk types is not statistically distinguishable from a separate estimation of a GAMLSS for each risk type. This implies that distinguishing statistical attributes of cyber risk types, important from an insurance perspective such as the tail behavior, turns out to be a very challenging task. Then, combining the POT method with the GAMLSS regression approach, we captured the complex and heterogeneous nature of cyber risk, showing that time trends, and the impact of company size, business sector, and contagion on cyber event frequency and severity varies with cyber risk type. A further investigation using a ordinal rank regression framework confirmed that the results are likely to be mainly affected by extreme event, regardless of the cyber risk type under consideration.

Our extensive empirical analysis allowed us to translate statistical features of cyber event frequency and severity to capital requirements and cyber risk insurability. We showed how the inclusion of covariates in cyber risk modeling can improve capital

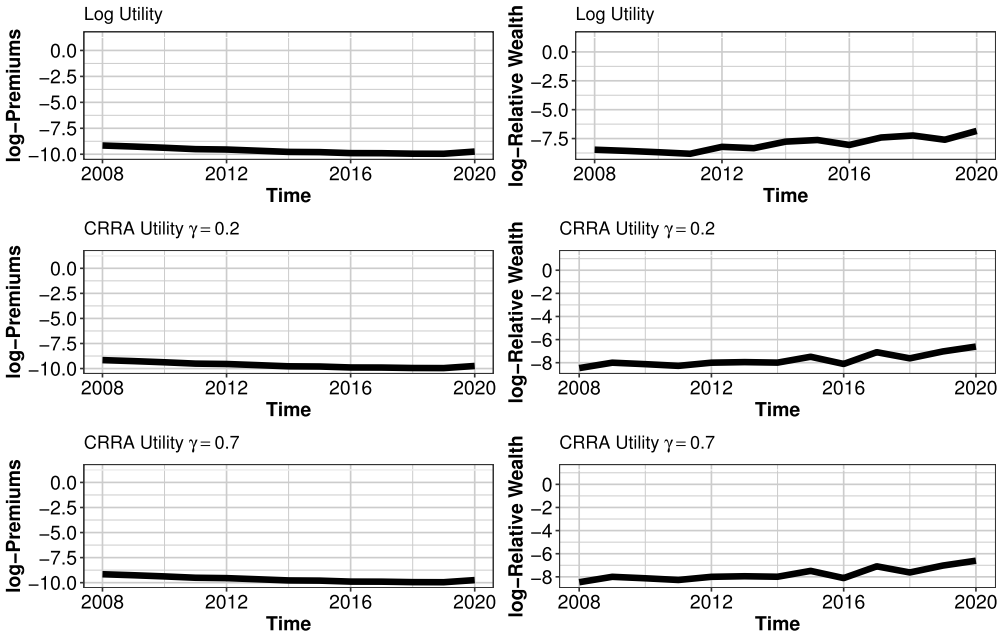


Fig. 7. This figure shows premiums and the corresponding insurer relative wealth in million dollars, on the log scale, estimated using company specific characteristic, using logarithmic utility function, CRRA utility with $\gamma = 0.2$, and $\gamma = 0.7$, of "Privacy - Unauthorized Contact or Disclosure". Company capital w is equal to 1 billion dollars and the top cover limit k is set to 10% of company capital. The case depicted corresponds to a situation where the estimated tail parameter τ is greater than 1 and therefore the risk type appears to be insurable.

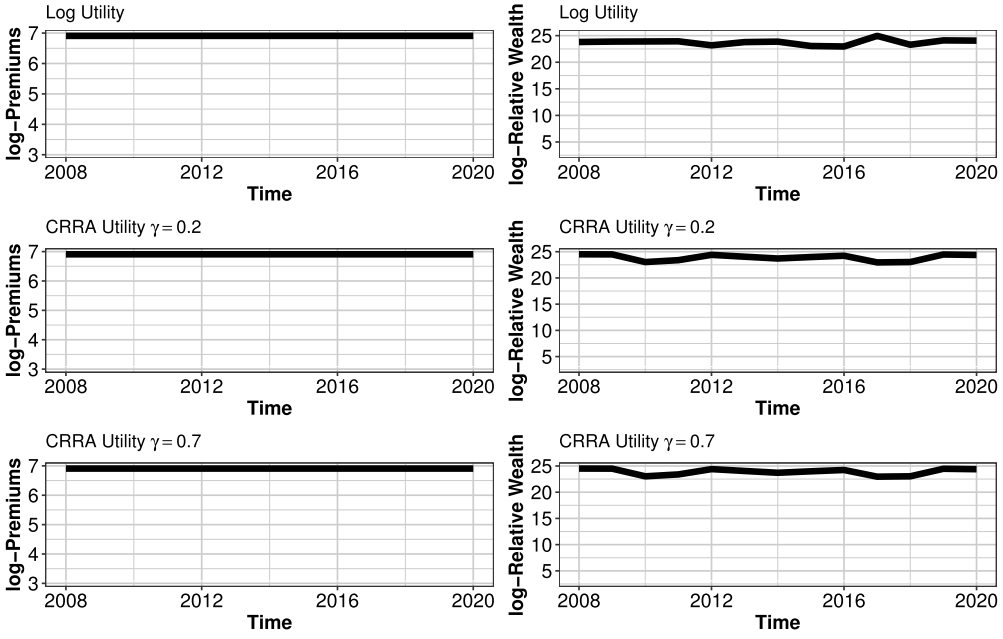


Fig. 8. This figure shows premiums and the corresponding insurer relative wealth in million dollars, on the log scale, estimated using company specific characteristic, using logarithmic utility function, CRRA utility with $\gamma = 0.2$, and $\gamma = 0.7$, of "Data Malicious Breach". Company capital w is equal to 1 billion dollars and the top cover limit k is set to 10% of company capital. The case depicted corresponds to a situation where the estimated tail parameter τ is lower than 1 and therefore the risk type appears to be not insurable.

requirement calculations, allowing for company and risk management tailored approaches for cyber risk types. Finally, we discussed the implications of our statistical findings on insurance premium calculation and more broadly, cyber risk insurabiltiy. Using a utility based argument, we show that consistently with the findings of the empirical analysis, insurance premiums vary with cyber risk types. According to our calculations, based on the combined POT method and GAMLSS approach, tail behavior and non stationarity in the tail parameter could jeopardize both demand and offer for cyber risk insurance, to the point where insurance premiums and insurance pools would be too high to be realistically adopted, or require a excessively high level of active management that insurance companies have very little incentive to undertake.

Our study is one of first to thoroughly investigate statistical features of monetary losses related to cyber risk. Our findings provide useful insights for industry and market participants, as well as policy makers and regulators. There are many important aspects of cyber risk that still require thorough investigation by the academic community and are left for future research. Investigating cyber risk using Bayesian non parametric approaches, such as

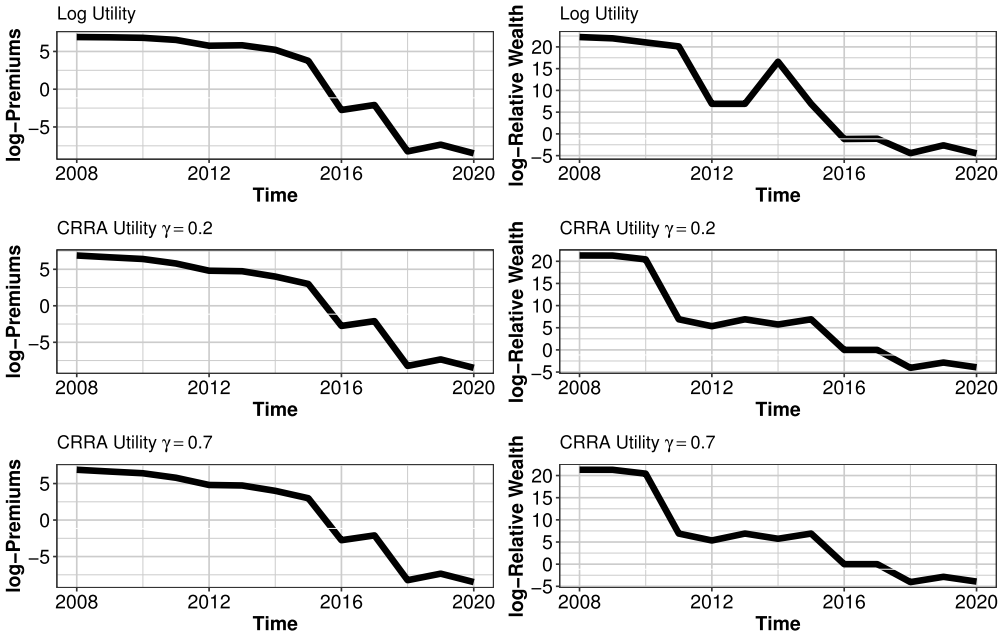


Fig. 9. This figure shows premiums and the corresponding insurer relative wealth in million dollars, on the log scale, estimated using company specific characteristic, using logarithmic utility function, CRRA utility with $\gamma = 0.2$, and $\gamma = 0.7$, of "Cyber Extortion". Company capital w is equal to 1 billion dollars and the top cover limit k is set to 10% of company capital. The case depicted corresponds to a situation where the estimated tail parameter τ increases over time becoming greater than 1. The rapid change in the required wealth makes the risk type not insurable.

the Approximate Bayesian Computation proposed in Goffard and Laub (2021), allowing for less restrictive assumptions in the LDA formulation may provide additional useful insights. As more data becomes available and the insurance market further develops, the presence of reverse causality between insurance protection and cyber risk should also be addressed.

Declaration of competing interest

There is no competing interest.

Acknowledgement

This research has been conducted within the Optus Macquarie University Cyber Security Hub and has been funded by its Risk Management, Governance and Control Program. We would like to acknowledge valuable comments from participants at the international congress "Insurance: Mathematics and Economics" 2021.

Appendix A

A.1. Hill's estimates by company size

Figs. 10 and 11 show the Hill's estimates by risk types and business sectors occurring in the top 10% companies by size measured in number of employees (panels a) and revenue (panels b). As it can be seen, the overall structures resemble the ones in Fig. 2 and 3. Even at a more granular level, the tail index estimates for the biggest companies by number of employees and revenue consistently indicate the presence of heavy tails in the cyber event severity distribution.

A.2. Dependence between covariates

Fig. 12 shows the linear correlation coefficient estimates between the covariates used in the GAMLSS regression. All the mutually exclusive event dummies present a negative linear correlation coefficient. Interestingly enough, the dummies Health and Multiple Losses positively correlate, indicating that companies in the healthcare sector could suffer more often from multiple internal breaches, than companies in other business sectors. Moreover, Big Revenue and Big Employee negatively correlate with Multiple Companies, suggesting that big corporations tend to be less affected by cyber events generated in external entities.

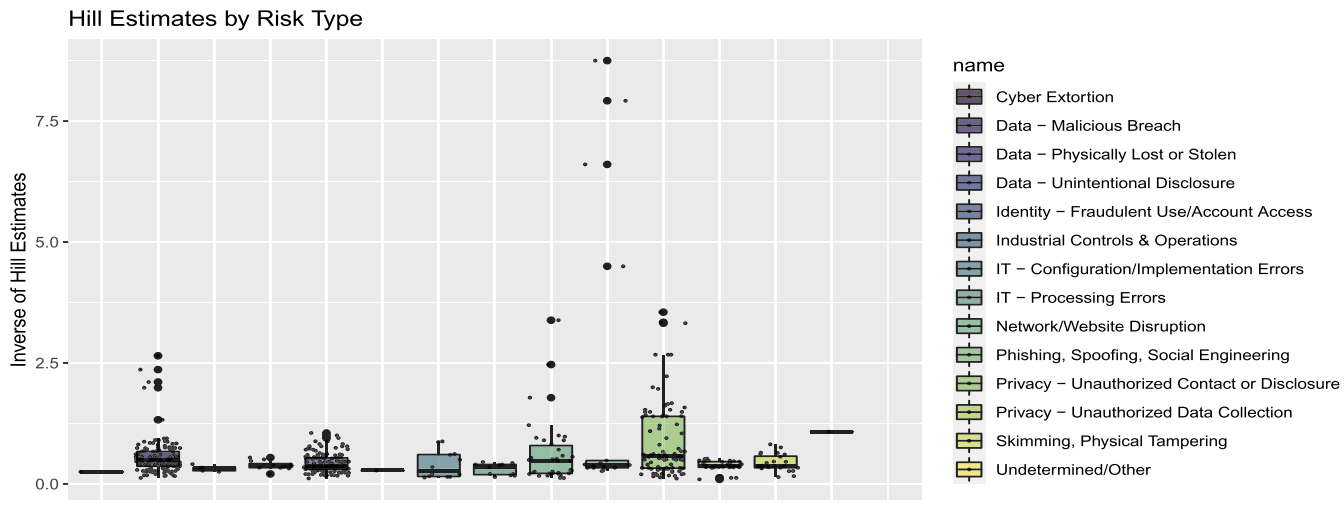
Table 6
This table reports the log-likelihood values, AIC and KS p-values for different distributional choices for cyber event severity. The null hypothesis is rejected for all the considered distributions.

Distribution	Log-Likelihood	AIC	KS test
Exponential	-30,211.2855	60,426	0
Gamma	0	6	0
Generalized Pareto	-8,091.2171	16,188	0
Log-logistic	-8,186.3720	16,378	0
Lognormal	-7,825.6678	15,657	0
Weibull	-10,350.5367	20,707	0
skew-Normal	-40,442.5800	80,891	0

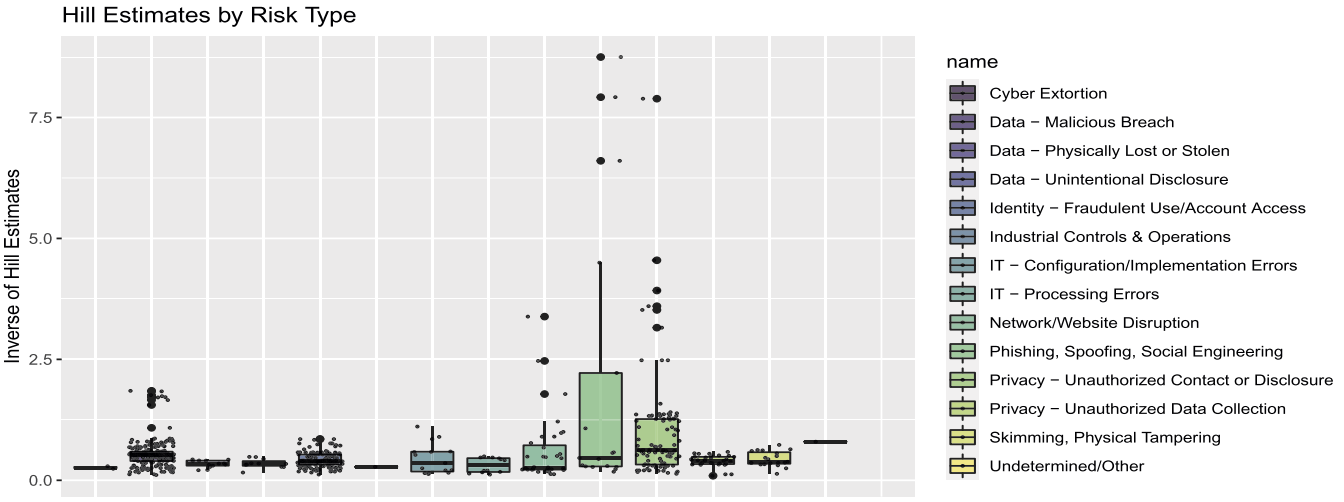
A.3. Goodness of fit

Table 6 shows log-likelihood, Akaike information criterion (AIC) and p-values for the Kolmogorov-Smirnov (KS) test for some commonly used distributions for cyber event severity. The null hypothesis is rejected for every considered distribution (the null hypothesis being no difference between cyber related loss distribution and one of the considered distributions), suggesting that more complex estimation procedures, such as the combined POT and GAMLSS approaches, may be required. Among the chosen distributions, lognormal, generalized Pareto, and Log-logistic preform better in terms of AIC and log-likelihood. To allow for heterogeneity in cyber event severity, we also perform the KS test on cyber event severity by cyber risk types.

Table 7 shows the p-values for the KS test for some commonly used distributions for cyber event severity, broken down by cyber risk types. With the exception of cyber related events of the type "Privacy - Unauthorized Contact or Disclosure", for losses belong-



(a) Hill's estimates of monetary losses linked to cyber events by risk types occurring in the top 10% companies by number of employees.

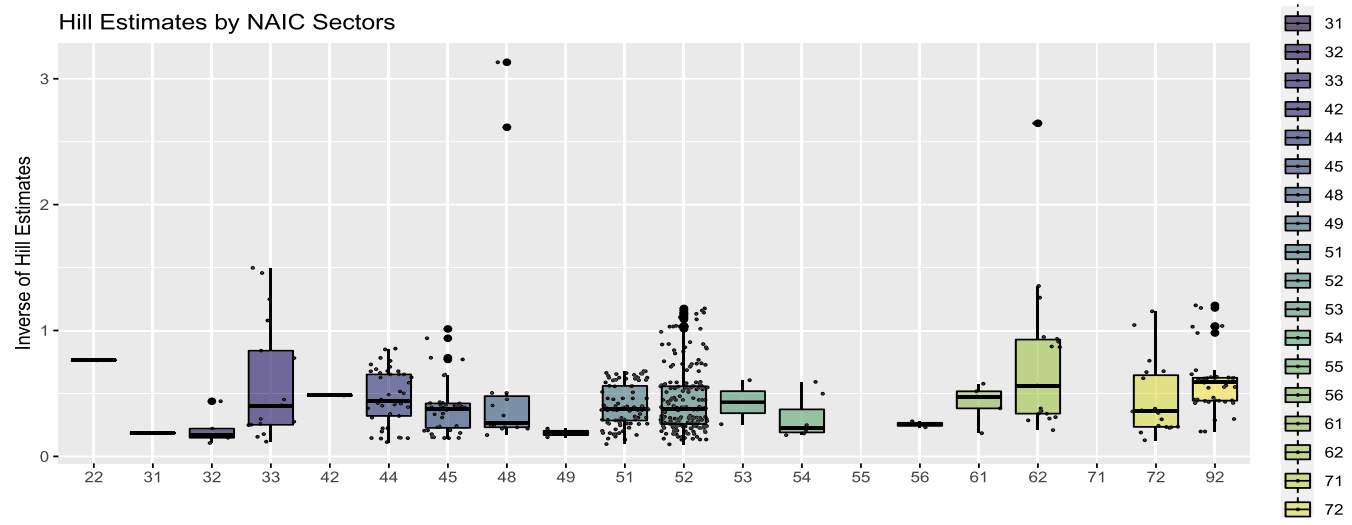


(b) Hill's estimates of monetary losses linked to cyber events by risk types occurring in the top 10% companies by revenue.

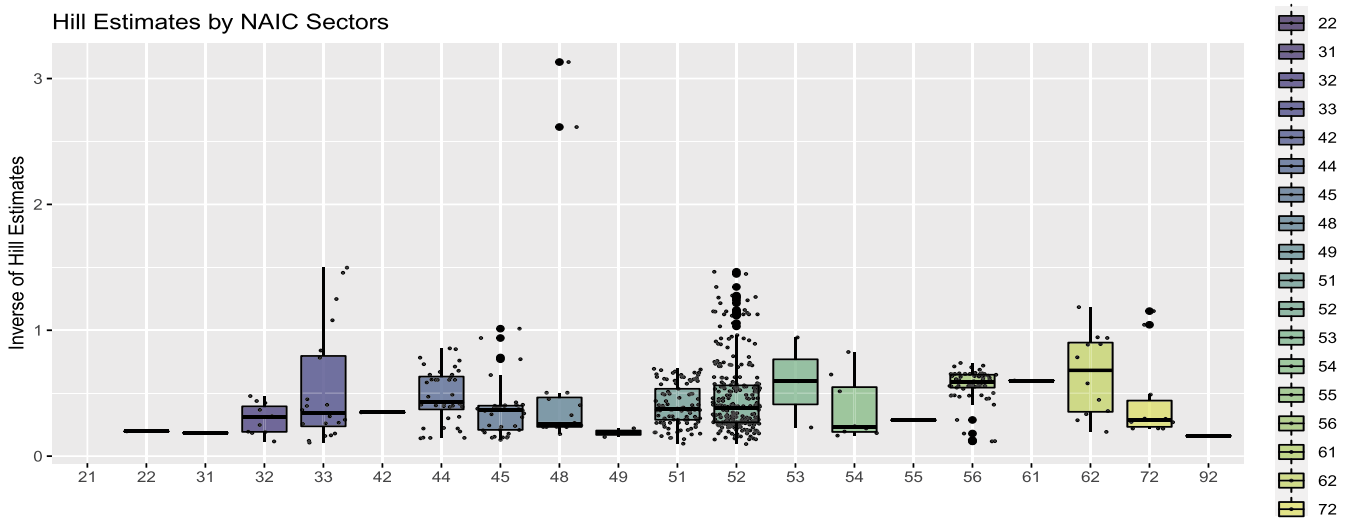
Fig. 10. This figure shows the Hill's estimates of monetary losses linked to cyber events by risk types, occurring in the top 10% companies by size. Panel (a) considers the number of employees, while Panel (b) looks at the revenue.

Table 7
This table reports the p-values for a KS test for different distributions broken down by risk types. Except for "Privacy - Unauthorized Contact or Disclosure", every other cyber risk type can be fitted on a generalized Pareto, Lognormal, or Log-logistic distribution.

Risk Type	Exponential	Gamma	GPD	Log-logistic	Lognormal	Weibull	skew-Normal
Privacy - Unauthorized Contact or Disclosure	0	0	0	0	0	0	0
Privacy – Unauthorized Data Collection	0	0	0.41	0.69	0.16	0	0
Data - Physically Lost or Stolen	0	0	0.31	0.16	0.24	0	0
Identity – Fraudulent Use/Account Access	0	0	0	0.54	0.85	0	0
Data - Malicious Breach	0	0	0	0	0.92	0	0
Phishing, Spoofing, Social Engineering	0	0	0.60	0.16	0.06	0	0
IT - Configuration/Implementation Errors	0	0.02	0.28	0.24	0	0	0
Data - Unintentional Disclosure	0	0	0.14	0.55	0.43	0	0
Cyber Extortion	0	0	0.06	0.001	0.83	0	0
Network/Website Disruption	0	0	0.79	0.12	0.49	0	0
Skimming, Physical Tampering	0	0	0.60	0.27	0.58	0	0
IT - Processing Errors	0	0	0.15	0.08	0	0.01	0
Industrial Controls	0.02	0.89	0	0.70	0.22	0.44	0
Undetermined/Other	0.22	0.75	0.94	0.45	0.45	0.20	0.01



(a) Hill's estimates of monetary losses linked to cyber events by business sectors occurring in the top 10% companies by number of employees.



(b) Hill's estimates of monetary losses linked to cyber events by business sectors occurring in the top 10% companies by revenue.

Fig. 11. This figure shows the Hill's estimates of monetary losses linked to cyber events by business sectors, occurring in the top 10% companies by size. Panel (a) considers the number of employees, while Panel (b) looks at the revenue.

ing to any other cyber risk type it's not possible to reject the null hypothesis of being distributed as, generalized Pareto, Lognormal or Log-logistic.

These results are in line with the findings in the literature (see, e.g. Edwards et al., 2016; Eling and Loperfido, 2017; Eling and Wirfs, 2019), where Lognormal, generalized Pareto, and Log-logistic distributions are often considered valid alternatives for cyber event severity. The implication of these results on the goodness of fit of the GAMLSS approach is investigated in Section A.6.

A.4. Non-linearity and interactions

Dynamic EVT allows for a flexible structure in the link functions, including parametric, semi-parametric and non parametric relationship. We follow the spline smoothing approach in Chavez-Demoulin et al. (2016) and select the values of $\gamma_\lambda, \gamma_\mu$ and γ_τ according to the AIC (see, also Ganegoda and Evans, 2013; Eling and Wirfs, 2019). Fig. 13 shows the AIC curve for different values of degrees of freedom in the smoothing spline in Equation (3). The minimum AIC is reached at γ_λ equals to 8, suggesting that cyber event frequency depends nonlinearly on time.

Fig. 14 shows the AIC surface for different values of degrees of freedom in the smoothing splines in the case of μ and τ. The minimum values of the AIC suggest that the tail parameter τ should not depend on time, while $\log(\mu)$ appears to depend nonlinearly on time.

We also use the AIC to investigate if any interaction effect among covariates is present, with particular interest in the effects on severity. For any possible combination, we include the interaction term as covariate in both $\log(\mu)$ and $\log(\tau)$ and estimate the corresponding model. Fig. 15 shows changes in the AIC surface when interaction terms are included. The values on the main diagonal correspond to the no interaction term case. As it can be seen from Fig. 15, some interaction terms show improvements in the AIC values with respect to the restricted model. Nonetheless, the impact of including any interaction term on the goodness of fit is minimal. To illustrate this, we compare the residual Q-Q plot of the model without interaction term and two interaction cases: interac-

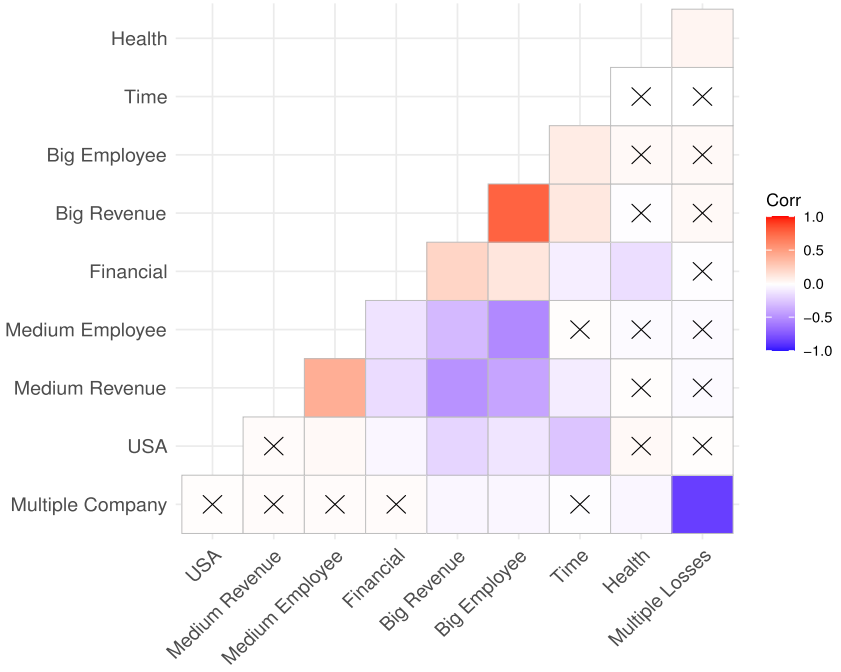


Fig. 12. This figure shows the correlation between covariates. "Xs" correspond to non statistically significant correlation at 5%.

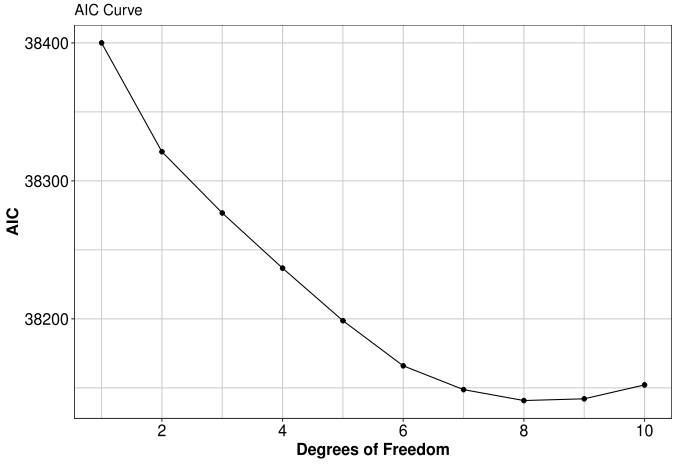


Fig. 13. AIC curve for cyber event frequency for different values of γ_λ. $\gamma_\lambda = 8$ corresponds to the minimum value of AIC.

tion term between USA and Financial (with an AIC equal to 7759), and interaction term between time and company size (with an AIC equal to 7767). Fig. 16 shows the residual Q-Q plot of cyber event severity fitted in three different configurations: no interaction term in the link functions (panel (a)), interaction between time and size (panel (b)), and interaction between Finance and USA dummy covariates (panel (c)). No particular difference arises from a graphical inspection of the three selected alternatives, implying that while adding an interaction term might improve a information criterion, it does not necessarily increase the goodness of fit.

A.5. RGA_R based testing methodology

In the context of ordinal regression, a formal hypothesis test to evaluate two competing models can be constructed using the following test statistic (see, Giudici and Raffinetti, 2020):

$$T = n\bar{r}\left(RGA_{R_{full}} - RGA_{R_{rest}}\right),$$

where $\bar{r}$ is the average rank, $RGA_{R_{full}}$ and $RGA_{R_{rest}}$ are the RGA_R of the full and the restricted model, respectively. The test statistic T is then distributed as variance gamma distribution, with parameters $\lambda = n/2$, $\alpha = 1/2$, $\beta = 0$, and $\mu = 0$, with $\lambda > 0$, $\alpha \in \mathbb{R}$, β is the asymmetry parameter, and μ is the location parameter. Given the fact that, for large values of λ the distribution of the test statistic does not well behave, Raffinetti and Romeo (2015) suggests to utilize d sub-samples to robustify the test statistics and compute the significance values as follows:

$$s\text{-value} = \mathbb{P}\left[T \geq |t_{\alpha/2}|\right] = \frac{1}{d}\sum_{i=1}^{d} \mathbb{I}_{T \geq |t_{\alpha/2}|},$$

where $\mathbb{I}_{T \geq |t_{\alpha/2}|}$ is equal to 1 whenever $T \geq |t_{\alpha/2}|$ and 0 otherwise.[8] Following Raffinetti and Romeo (2015), the quantity $\sum_{i=1}^{d} \mathbb{I}_{T \geq |t_{\alpha/2}|}$ follows a binomial distribution with parameters d and p, with the following probability mass function:

$$f(n; d, p) = \binom{d}{n} p^n (1-p)^n \tag{15}$$

with $n = 1, \ldots, d$. In other words, s-value is an estimator for the quantity p to which corresponds the probability of success in the binomial distribution. To test whether the estimates for s-value are consistent with every possible subsampling combination, a three way testing procedure can be used.
Call $z = (\widehat{s\text{-value}} - p_0)/\sqrt{p_0(1-p_0)/d}$

- If $\widehat{s\text{-value}} \in (0.7, 1]$, perform $H_0 : p \leq 0.7$ vs $H_1 : p > 0.7$. Reject the null if $\mathbb{P}[Z > z] \leq \alpha_s$
- If $\widehat{s\text{-value}} \in (0.0, 0.3]$, perform $H_0 : p > 0.3$ vs $H_1 : p \leq 0.3$. Reject the null if $\mathbb{P}[Z \leq z] \leq \alpha_s$

[8] It is important to notice that the s-value is not to be interpreted as a p-value, but rather as a test statistics for the proportion of times the value of T is greater than a given threshold in the subsamples. The authors themselves suggest to perform a further hypothesis test on the s-value based on the binomial distribution (see, Raffinetti and Romeo, 2015).

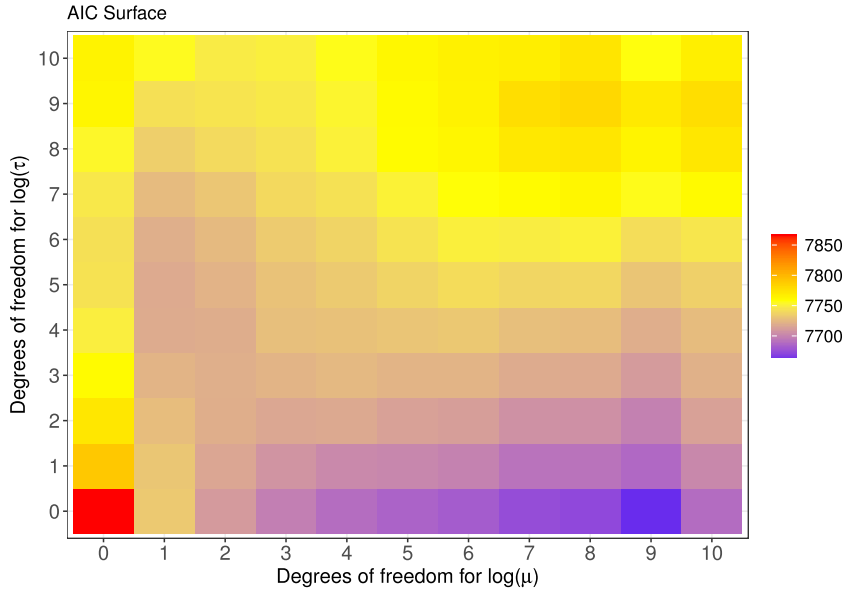


Fig. 14. AIC surface for cyber event severity for varying values of γ_μ and γ_τ. The minimum value of the AIC suggests the tail parameter τ should not depend on time and a non-linear relationship between $\log(\mu)$ and time.

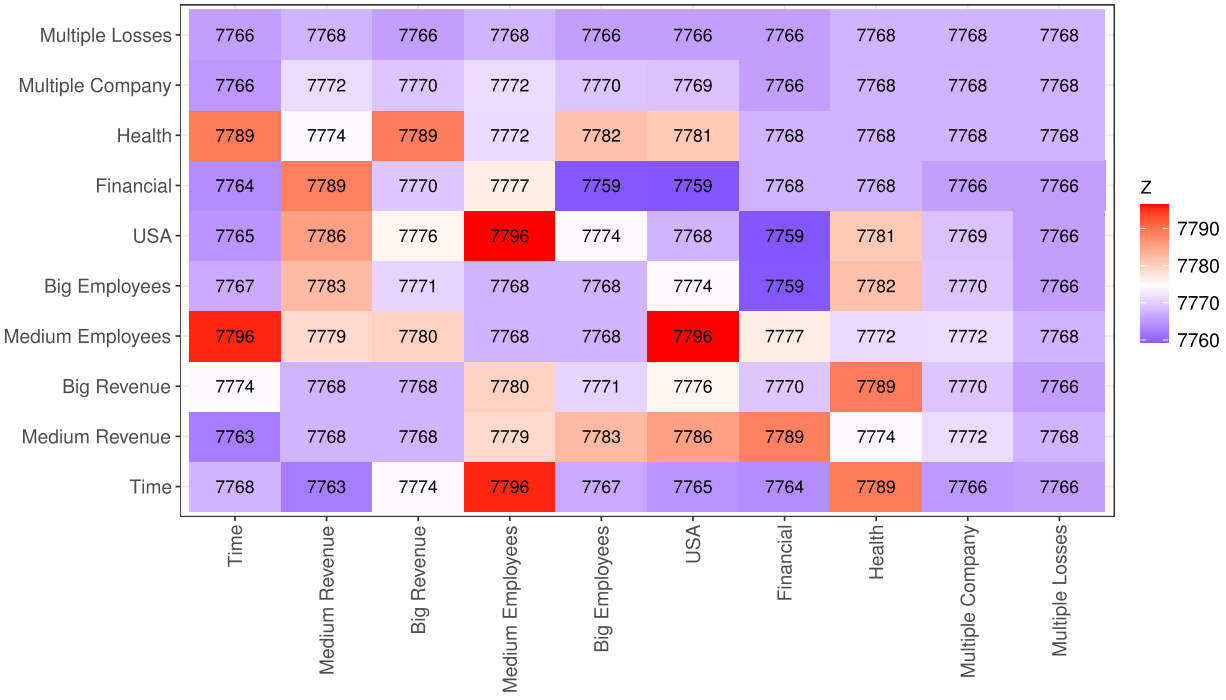


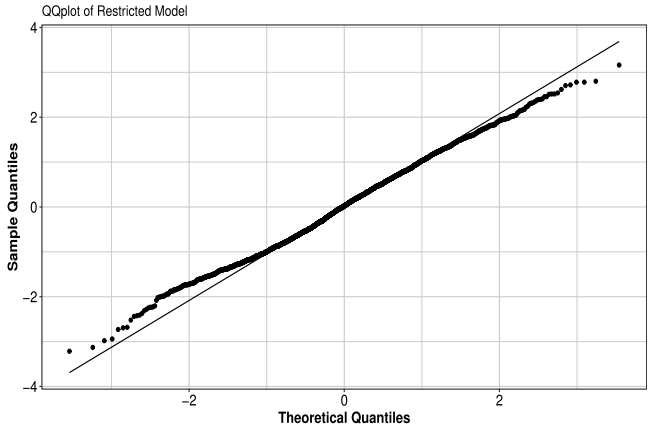
Fig. 15. AIC values for models including different interaction terms between covariates.

- $\widehat{s\text{-value}} \in (0.3, 0.5]$, two tests need to be performed: $H_0 : p \leq 0.3$ vs $H_1 : p > 0.3$ and, $H_0 : p > 0.5$ vs $H_1 : p \leq 0.5$. Then the p-values of both test are combined using Holm's two steps approach Holm (1979). The case of $\widehat{s\text{-value}} \in (0.5, 0.7]$ follows analogously.

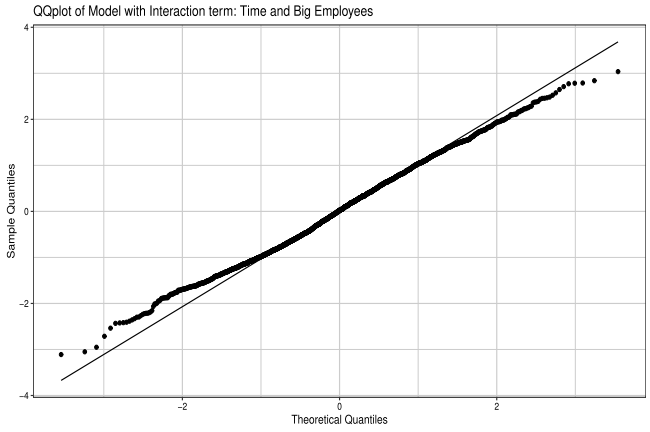
Table 8 shows the result on the significance of each individual covariates in the linear model in Equation (11), using the Advisen classification. For each covariate Table 8 shows the RGA, s-value, and the s-class of the restricted model. The s-class is computed accordingly to Raffinetti and Romeo (2015) and Giudici and Raffinetti (2020). As it can be seen from the table, only Time, E_{big} and L_{USA} belong to the significance class "Almost Always" significant, while none of the Advisen Risk types are statistically significant. To further investigate this we also have run the test for the joint significance of the risk types getting a s-value of 0.4982.

A.6. Coupled model estimates

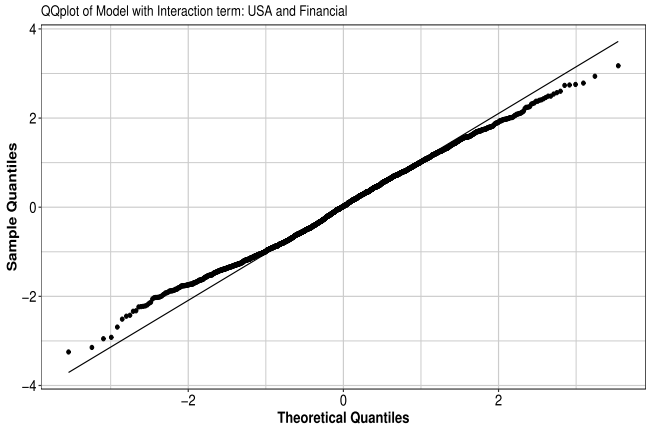
In this section we present the dynamic EVT estimates for the coupled model for cyber severity. We proceed with the approach discussed in Section 2, finding that the threshold value for the coupled model is 59%, similarly the results discovered in Eling and Wirfs (2019). Then, for cyber related losses exceeding such threshold we fit the model given by likelihood and link functions in Equation (7). Given the results in Section A.3 of the KS test,



(a) Q-Q plot of residual of model without interaction term



(b) Q-Q plot of residual of model with interaction term between time and company size



(c) Q-Q plot of residual of model with interaction term between USA and Finance

Fig. 16. This figure shows the Q-Q plots in three different configurations: no interaction term in the link functions (panel (a)), interaction between time and size (panel (b)), and interaction between Finance and USA (panel (c)).

Table 8
This table shows the results of the rank based regression analysis on model 0. The s-values corresponds to the test of hypothesis for individual significance of each coefficient. The size of the sub-sample is set equal to 10, and the number of sub-samples drawn is 5,000. None of the Advisen risk types are statistically significant.

Covariates	RGA	s-value	p_s-value	Result	s-class
Privacy - Unauthorized Contact or Disclosure	0.2168	0	-	-	Never significant
Data - Unintentional Disclosure	0.2168	0	-	-	Never significant
Privacy – Unauthorized Data Collection	0.2168	0	-	-	Never significant
Data - Malicious Breach	0.2168	0	-	-	Never significant
Identity – Fraudulent Use/Account Access	0.2168	0	-	-	Never significant
Data - Physically Lost or Stolen	0.2168	0	-	-	Never significant
Skimming, Physical Tampering	0.2168	0	-	-	Never significant
IT - Processing Errors	0.2168	0	-	-	Never significant
Phishing, Spoofing, Social Engineering	0.2168	0	-	-	Never significant
IT - Configuration/Implementation Errors	0.2168	0	-	-	Never significant
Network/Website Disruption	0.2168	0	-	-	Never significant
Cyber Extortion	0.2168	0	-	-	Never significant
Undetermined/Other	0.2168	0	-	-	Never significant
Industrial Controls/Operations	0.2168	0	-	-	Never significant
Time	0.1861	0.9002	< 0.0001	$p > 0.7$	Almost Always Significant
R_{medium}	0.2094	0.5344	0	$p > 0.3$	Frequently significant
R_{big}	0.2190	0.2112	0.0633	$p > 0.3$	Rarely significant
E_{medium}	0.2051	0.5852	0.0022	$p \in [0.5, 0.7]$	Frequently significant
E_{big}	0.1881	0.8130	0	$p > 0.7$	Almost Always Significant
L_{USA}	0.2176	0.7322	0.2563	$p \leq 0.7$	Almost Always Significant
$B_{financial}$	0.2154	0.3042	0.4136	$p > 0.3$	Sometimes significant
B_{health}	0.2164	0.0722	< 0.0001	$p \leq 0.3$	Rarely significant
MC	0.2165	0.0386	< 0.0001	$p \leq 0.3$	Rarely significant
ML	0.2167	0.0172	< 0.0001	$p \leq 0.3$	Rarely significant

Table 9
Coefficient Estimates for three different severity loss distribution regression fits under the GAMLSS setting. Note: the threshold for the GAMLSS GPD regression model was selected at 59%.

Covariates	Lognormal		GPD		Log-logistic	
	μ	$\log(\sigma)$	$\log(\mu)$	$\log(\tau)$	μ	$\log(\sigma)$
β_0	-316.6092***	0.5983	-63.8505*	-17.9724	-340.8670***	0.5983***
Time	0.1647***	0.0001	0.0407**	0.0095	0.1769***	0.0001***
R_{medium}	-0.2223	-0.0750*	-0.3536**	-0.1263	-0.2655*	-0.0750*
R_{big}	0.5367**	0.0055	0.0435	-0.2623**	0.5353***	0.0055
E_{medium}	0.2635 *	0.0387	0.1971	0.1166	0.2894*	0.0387
E_{big}	1.0606***	0.0315	0.4278**	0.1440	1.1461***	0.0315
L_{USA}	0.7166***	-0.0465	0.6562***	0.3093***	0.7710***	-0.0465
$B_{financial}$	0.2708*	0.0144***	0.6492***	0.3318***	0.3368*	0.0144
B_{health}	0.0235	-0.2480	0.4185**	0.7531***	0.0625	-0.2480***
ML	-2.0848	-0.1139	-0.5483**	0.9510	-2.3437	-0.1139
MC	-3.2110	-0.1313	-4.7323***	-1.2791*	-3.8585*	-0.1313
Privacy - Unauthorized Contact or Disclosure	-1.6194**	0.3844*	1.2107**	0.0272	-1.6418**	0.3844
Privacy – Unauthorized Data Collection	-0.1683	0.2438	-0.1922	-0.5236	0.0014	0.2438
Data - Physically Lost or Stolen	-0.7374	0.3264	-0.6194	-0.7809*	-0.6587	0.3264
Identity – Fraudulent Use/Account Access	-2.5693***	0.2877	-0.3174	-0.2325	-2.4116***	0.2877
Data - Malicious Breach	-0.1813	0.2428	-0.0340	-0.5986	-0.0989	0.2428
Phishing, Spoofing, Social Engineering	-0.3590	0.0393	-0.3976	-0.4555	-0.2996	0.0393
IT - Configuration/Implementation Errors	0.1785	0.2843	0.0494	-0.7946*	0.2943	0.2843
Data Unintentional Disclosure	-1.2169*	0.0559	0.2962	0.1148	-1.0761	0.0559
Cyber Extortion	-4.0761***	0.0429	-0.7456	-0.2864	-3.9993***	0.0429
Network/Website Disruption	-0.8125	0.3320	-0.2308	-0.8477**	-0.8070	0.3320
Skimming, Physical Tampering	-2.1929	0.2582	-0.1169	-0.1578	-2.1180**	0.2582
IT – Processing Errors	0.6990***	0.4725*	1.1321	-0.8106*	0.8234	0.4725
Industrial Controls	0.4209	0.2258	0.8873	-0.5504	0.3442	0.2258

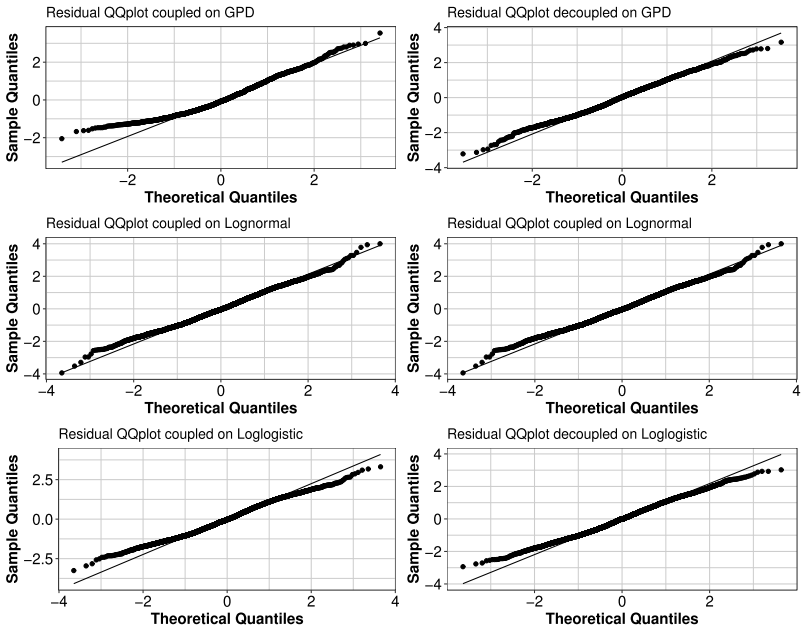


Fig. 17. This figure shows the residual Q-Q plots of cyber event severity fitted on generalized Pareto, Lognormal, and Log-logistic distributions coupled (on the left) and decoupled (on the right).

we consider also two alternative distributions, commonly used in Operational Risk applications, Log-logistic and Lognormal. Table 9 shows the estimates for each distribution. Comparing the results for the Generalized Pareto distribution with the results of the decoupled model it can be seen that fewer coefficients appear to be statistically significant. Looking at the tail parameter τ, it shows that a company with a high level of revenue tends to experience heavier cyber related losses, while for a company residing in the USA the severity seems to be lower than that for companies residing outside the USA. Indicating that perhaps cyber awareness, mitigation or risk management processes are more complete in companies located in the USA versus equivalent entities outside the USA. Similar consideration can be made for companies operating in the healthcare or financial business sectors: cyber event losses appear to be less severe in financial or healthcare companies with respect to enterprises operating in other business sectors. This can once again be explained by the higher degree of cyber awareness present in these industries. Looking at the temporal dependence, the evidence appears to be mixed, with Lognormal and generalized Pareto distributions returning conflicting, although not statistically significant, results. Looking at the effect of cyber risk types, very few coefficients are statistically significant, in line with the previous results obtained in Section 5. Overall, when the coupled model is considered, cyber event severity appears to be mainly driven by company specific characteristics, while cyber risk types, time, and contagion are only rarely significant variables.

Finally, Fig. 17 compares Q-Q plots of cyber event severity fitted on generalized Pareto, Lognormal, and Log-logistic distributions coupled (on the left) and decoupled (on the right). Both in the case of coupled and decoupled models, the Q-Q plots look very similar for all the distributions. This fact has a twofold implication. First, it further confirms that, coupled and decoupled models for cyber event severity are statistically indistinguishable. Second, while the combined POT and GAMLSS approach is a powerful but sophisticated framework, other commonly used distribution such as Log-logistic and Lognormal performs adequately well in fitting cyber event severity.

References

Aldasoro, I., Gambacorta, L., Giudici, P., Leach, T., 2020. The drivers of cyber risk. Technical Report. Bank of International Settlements.

Antonio, Y., Indratno, S.W., Saputro, S.W., 2021. Pricing of cyber insurance premiums using a Markov-based dynamic model with clustering structure. PLoS ONE 16, e0258867.

Balkema, A.A., De Haan, L., 1974. Residual life time at great age. Annals of Probability 2, 792–804.

Basel Committee on Banking Supervision, 2006. International convergence of capital measurement and capital standards: a revised framework. Technical Report. Bank for International Settlements, Basel.

Bessy-Roland, Y., Boumezoued, A., Hillairet, C., 2021. Multivariate Hawkes process for cyber insurance. Annals of Actuarial Science 15, 14–39.

Biener, C., Eling, M., Wirfs, J.H., 2015. Insurability of cyber risk: an empirical analysis. The Geneva Papers on Risk and Insurance. Issues and Practice 40, 131–158.

Bouveret, A., 2018. Cyber Risk for the Financial Sector: A Framework for Quantitative Assessment. International Monetary Fund.

Camillo, M., 2017. Cyber risk and the changing role of insurance. Journal of Cyber Policy 2, 53–63.

Cebula, J.J., Popeck, M.E., Young, L.R., 2014. A taxonomy of operational cyber security risks version 2. Technical Report. Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst.

Cebula, J.L., Young, L.R., 2010. A taxonomy of operational cyber security risks. Technical Report. Carnegie-Mellon University, Pittsburgh Software Engineering Inst.

Chavez-Demoulin, V., Embrechts, P., Hofert, M., 2016. An extreme value approach for modeling operational risk losses depending on covariates. Journal of Risk and Insurance 83, 735–776.

Cope, E., Labbi, A., 2008. Operational loss scaling by exposure indicators: evidence from the orx database. Journal of Operational Risk 3, 25–46.

Cruz, M.G., Peters, G.W., Shevchenko, P.V., 2015. Fundamental Aspects of Operational Risk and Insurance Analytics: A Handbook of Operational Risk. John Wiley & Sons.

Cyentia, 2020. A Clearer Vision for Assessing the Risk of Cyber Incidents. Technical Report. Cyentia Institute.

Dahen, H., Dionne, G., 2010. Scaling models for the severity and frequency of external operational loss data. Journal of Banking & Finance 34, 1484–1496.

De Bolle, C., 2020. Internet Organized Crime Threat Assessment 2020. Technical Report. Europol, European Union Agency for Law Enforcement Cooperation.

Degen, M., 2010. The calculation of minimum regulatory capital using single-loss approximations. Journal of Operational Risk 5, 3–17.

Durrieu, G., Grama, I., Pham, Q.K., Tricot, J.M., 2015. Nonparametric adaptive estimation of conditional probabilities of rare events and extreme quantiles. Extremes 18, 437–478.

Edwards, B., Hofmeyr, S., Forrest, S., 2016. Hype and heavy tails: a closer look at data breaches. Journal of Cybersecurity 2, 3–14.

Eling, M., 2020. Cyber risk research in business and actuarial science. European Actuarial Journal 10, 303–333.

Eling, M., Jung, K., 2018. Copula approaches for modeling cross-sectional dependence of data breach losses. Insurance. Mathematics & Economics 82, 167–180.

Eling, M., Loperfido, N., 2017. Data breaches: goodness of fit, pricing, and risk measurement. Insurance. Mathematics & Economics 75, 126–136.

Eling, M., Schnell, W., 2016. What do we know about cyber risk and cyber risk insurance? The Journal of Risk Finance 17, 474–491.

Eling, M., Wirfs, J., 2019. What are the actual costs of cyber risk events? European Journal of Operational Research 272, 1109–1119.

Eling, M., Wirfs, J.H., 2015. Modelling and management of cyber risk. In: Proceedings of Colloquium of International Actuarial Association Life Section Colloquium. Oslo, pp. 1–24. https://www.actuaries.org/oslo2015/papers/IAALS-Wirfs&Eling.pdf.

Evans, M., Maglaras, L.A., He, Y., Janicke, H., 2016. Human behaviour as an aspect of cybersecurity assurance. Security and Communication Networks 9, 4667–4679.

Fahrenwaldt, M.A., Weber, S., Weske, K., 2018. Pricing of cyber insurance contracts in a network model. ASTIN Bulletin: The Journal of the IAA 48, 1175–1218.

Falk, A., 2019. Notifiable Data Breaches Report. Technical Report. Office of Australian Information Commissioner.

Farkas, S., Lopez, O., Thomas, M., 2021. Cyber claim analysis using generalized Pareto regression trees with applications to insurance. Insurance. Mathematics & Economics 98, 92–105.

Farrington, D.P., Loeber, R., 2000. Some benefits of dichotomization in psychiatric and criminological research. Criminal Behaviour and Mental Health 10, 100–122.

Ganegoda, A., Evans, J., 2013. A scaling model for severity of operational losses using generalized additive models for location scale and shape (gamlss). Annals of Actuarial Science 7, 61–100.

Giudici, P., Raffinetti, E., 2020. Cyber risk ordering with rank-based statistical models. AStA Advances in Statistical Analysis, 1–16.

Goffard, P.O., Laub, P.J., 2021. Approximate bayesian computations to fit and compare insurance loss models. Insurance. Mathematics & Economics 100, 350–371.

Grama, I., Spokoiny, V., 2008. Statistics of extremes by oracle estimation. The Annals of Statistics 36, 1619–1648.

Herath, H.S., Herath, T.C., 2007. Cyber-insurance: copula pricing framework and implication for risk management. In: WEIS.

Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. The Annals of Statistics, 1163–1174.

Hillairet, C., Lopez, O., 2021. Propagation of cyber incidents in an insurance portfolio: counting processes combined with compartmental epidemiological models. Scandinavian Actuarial Journal 2021, 671–694.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Hubbard, D., Evans, D., 2010. Problems with scoring methods and ordinal scales in risk assessment. IBM Journal of Research and Development 54. 2–1.

Iacobucci, D., Posavac, S.S., Kardes, F.R., Schneider, M.J., Popovich, D.L., 2015. Toward a more nuanced understanding of the statistical properties of a median split. Journal of Consumer Psychology 25, 652–665. https://doi.org/10.1016/j.jcps.2014.12.002.

Iman, R.L., Conover, W.J., 1979. The use of the rank transform in regression. Technometrics 21, 499–509.

Jung, K., 2021. Extreme data breach losses: an alternative approach to estimating probable maximum loss for data breach risk. North American Actuarial Journal, 1–24.

Lambrigger, D.D., Shevchenko, P.V., Wüthrich, M.V., 2007. The quantification of operational risk using internal data, relevant external data and expert opinions. Journal of Operational Risk 2, 3–27.

Lis, P., Mendel, J., 2019. Cyberattacks on critical infrastructure: an economic perspective. Economics and Business Review 5, 24–47.

Maillart, T., Sornette, D., 2010. Heavy-tailed distribution of cyber-risks. The European Physical Journal B 75, 357–364.

Mukhopadhyay, A., Chatterjee, S., Saha, D., Mahanti, A., Sadhukhan, S.K., 2013. Cyber-risk decision models: to insure it or not? Decision Support Systems 56, 11–26.

Nešlehová, J., Embrechts, P., Chavez-Demoulin, V., 2006. Infinite mean models and the lda for operational risk. Journal of Operational Risk 1, 3–25.

Peters, G., Shevchenko, P.V., Cohen, R., 2018a. Statistical machine learning analysis of cyber risk data: event case studies. In: Maurice, D., Fairman, D., Freund, J. (Eds.), Fintech: Growth and Deregulation. Risk Books, United Kingdom, pp. 303–330. chapter 12.

Peters, G., Shevchenko, P.V., Cohen, R., Maurice, D., 2018b. Statistical machine learning analysis of cyber risk data: event case studies. In: Maurice, D., Fairman, D., Freund, J. (Eds.), Fintech: Growth and Deregulation. Risk Books, United Kingdom, pp. 75–99. chapter 3.

Peters, G., Targino, R., Shevchenko, P.V., 2013. Understanding operational risk capital approximations: first and second orders. Journal of Governance and Regulation 2, 58–78. https://doi.org/10.22495/jgr_v2_i3_p6.

Peters, G.W., Byrnes, A.D., Shevchenko, P.V., 2011. Impact of insurance for operational risk: is it worthwhile to insure or be insured for severe losses? Insurance. Mathematics & Economics 48, 287–303.

Peters, G.W., Shevchenko, P.V., 2015. Advances in Heavy Tailed Risk Modeling: A Handbook of Operational Risk. John Wiley & Sons.

Pickands, J., 1975. Statistical inference using extreme order statistics. The Annals of Statistics 3, 119–131.

Ponemon, 2019. Cost of Data Breach Report 2019. Technical Report. Ponemon Institute.

Raffinetti, E., Romeo, I., 2015. Dealing with the biased effects issue when handling huge datasets: the case of INVALSI data. Journal of Applied Statistics 42, 2554–2570.

Rakes, T.R., Deane, J.K., Rees, L.P., 2012. It security planning under uncertainty for high-impact events. Omega 40, 79–88.

Refsdal, A., Solhaug, B., Stølen, K., 2015. Cyber-risk management. In: Cyber-Risk Management. Springer, pp. 33–47.

Rigby, R., Stasinopoulos, D., 2005. Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54, 507–554.

Romanosky, S., 2016. Examining the costs and causes of cyber incidents. Journal of Cybersecurity 2, 121–135.

Romanosky, S., Ablon, L., Kuehn, A., Jones, T., 2019. Content analysis of cyber insurance policies: how do carriers price cyber risk? Journal of Cybersecurity 5, 1–19.

Sexton, J., Storlie, C., Neil, J., 2015. Attack chain detection. Statistical Analysis and Data Mining: The ASA Data Science Journal 8, 353–363.
Shevchenko, P.V., 2011. Modelling Operational Risk Using Bayesian Inference. Springer Science & Business Media.
Shevchenko, P.V., Wüthrich, M.V., 2006. The structural modelling of operational risk via bayesian inference: combining loss data with expert opinions. Journal of Operational Risk 1, 3–26.
Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software 23, 1–46.
Stasinopoulos, M., Rigby, B., Akantziliotou, C., 2008. Instructions on How to Use the Gamlss Package in R, second edition.
Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., De Bastiani, F., 2017. Flexible Regression and Smoothing: Using GAMLSS in R. CRC Press.
Tsai, C.F., Chen, Y.C., 2019. The optimal combination of feature selection and data discretization: an empirical study. Information Sciences 505, 282–293. https://doi.org/10.1016/j.ins.2019.07.091.
Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.
Woods, D.W., Böhme, R., 2021. Systematization of knowledge: quantifying cyber risk. In: 2021 IEEE Symposium on Security and Privacy (SP), pp. 211–228.
Xu, M., Hua, L., 2019. Cybersecurity insurance: modeling and pricing. North American Actuarial Journal 23, 220–249.
Xu, M., Schweitzer, K.M., Bateman, R.M., Xu, S., 2018. Modeling and predicting cyber hacking breaches. IEEE Transactions on Information Forensics and Security 13, 2856–2871.
Zeller, G., Scherer, M., 2021. A comprehensive model for cyber risk based on marked point processes and its application to insurance. European Actuarial Journal, 1–53.