computation

MDPI

Article

Some Remarks on Malicious and Negligent Data Breach Distribution Estimates

Maria Francesca Carfora *,† **and Albina Orlando** *,†

Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni del Calcolo "Mauro Picone", Via P. Castellino, 111, 80131 Naples, Italy

* Correspondence: f.carfora@na.iac.cnr.it (M.F.C.); a.orlando@na.iac.cnr.it (A.O.)

† These authors contributed equally to this work.

Abstract: Digitization offers great opportunities as well as new challenges. Indeed, these opportunities entail increased cyber risks, both from deliberate cyberattacks and from incidents caused by inadvertent human error. Cyber risk must be mastered, and to this aim, its quantification is an urgent challenge. There is a lot of interest in this topic from the insurance community in order to price adequate coverage to their customers. A key first step is to investigate the frequency and severity of cyber incidents. On the grounds that data breaches seem to be the main cause of cyber incidents, the aim of this paper is to give further insights about the frequency and severity statistical distributions of malicious and negligent data breaches. For this purpose, we refer to a publicly available dataset: the Chronology of Data Breaches provided by the Privacy Rights Clearinghouse.

Keywords: cyber risk; frequency and severity modeling; data breaches

check for updates

Citation: Carfora, M.F.; Orlando, A. Some Remarks on Malicious and Negligent Data Breach Distribution Estimates. *Computation* **2022**, *10*, 208. https://doi.org/10.3390/computation10120208

Academic Editor: Demos T. Tsahalis

Received: 18 October 2022
Accepted: 25 November 2022
Published: 30 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

According to [1], cyber threats continue to be an ongoing big concern for companies nowadays. The peril of data breaches, ransomware attacks or major IT outages worries companies even more than natural disasters, supply chain and business disruptions or the COVID-19 pandemic. For a very long time, cyber risk was classified as operational risk, which stems from external occurrences or bad and unfruitful internal systems, people and processes. However, it shows specific features that make it a class of risk of its own which requires specific approaches to study it [2]. Indeed, it is well known that the environment of cyber risk constantly evolves as a consequence of new technologies and the rapid development of computer information systems. Many researchers consider cyber risk to be natural catastrophes, at least in terms of scale. They stress that each catastrophe type has a different set of phenomena that need to be understood in detail and that cyber risk is of a somewhat novel nature [3]. A useful classification of the different types of cyber incidents as well as the types of losses that may occur is given in [4]. It must be pointed out that many employees, intentionally or not, are often the weakness of a successful cyberattack. Regarding the losses arising from a cyber incident, they include injury to tangible and intangible assets, losses connected to theft, and business disruption. Moreover, they can bring multiple liabilities to third parties such as employees, customers, suppliers, and shareholders. Reputation damage is another significant cost component, as loss of customers and stakeholders' confidence may result directly in a proper liability.

Based on the above, it is strictly necessary that companies consider cyber risk as a component of the overall risk management routine. They need to identify, assess, and treat risks connected to availability, integrity, and confidentiality of their assets, being fully aware that a residual risk must be accepted. The implementation of the best cyber security procedures and the setting up of several countermeasures in particular do not ensure avoiding cyber incidents, regardless of their huge cost [5].

Nevertheless, there is the option of transferring part of the risk to a third party by signing a cyber insurance contract. The purpose of this is to mitigate losses from cyber

incidents, which include data breaches, data theft, business interruptions, and network damage. Insuring cyber risks is not an easy task for insurance companies due to several reasons. Among others, we have the continuous evolution of information systems, increasingly sophisticated and challenging cyberattacks, interdependence of security levels, hard impact determination, information asymmetry, and a lack of data [6]. Regarding the shortage of data, this may be attributed principally to information sharing barriers as a result of the unwillingness of companies to reveal cyber incidents. The justification lies in the concern regarding reputation damages consequent to the release of a cyber incident. In many US states, reporting requirements have been introduced since 2002 in many US states. Therefore, the literature on this topic is mainly focused on US data. The situation is rapidly changing in Europe too with the introduction in May 2018 of the General Data Protection Regulation (GDPR), setting stricter procedures such as the requirement to notify the regulator and data owner of a data breach. Historical data are a key topic for actuarial valuations; insurance companies need to have information about the full loss distribution of their clients, both to set premiums and to determine the right amount of safety capital.

Cyber risk assessment is performed by referring to the probability of an incident and its impact on a firm. Indeed, the probability of an incident multiplied by the impact gives a measure of risk. The frequency with which a negative cyber event occurs makes it possible to estimate its likelihood, while its severity helps to provide a measure of the impact. Therefore, modeling the frequency and severity of cyber incidents is a crucial step.

To offer a reflection on this topic, we investigate the Chronology of Data Breaches provided by the Privacy Rights Clearinghouse (hereinafter, PRC data), by far the most complete and publicly free to access dataset. We chose this dataset on the grounds that data breaches, together with ransomware attacks, are the main cause of cyber incidents and the most concerning cyber exposure for companies [1]. Several contributions in the recent literature on cyber risk management analyzed this dataset [3,7,8].

In a previous study [9], we already considered PRC data and focused our analysis on a specific subset, constituted by breaches which occurred in healthcare, medical providers, and medical insurance services (MED data). These data represent a major part of the entire PRC dataset, but they are also affected by strong heterogeneity and a high frequency of missing information, so they deserved specific analysis. In the present study, we contribute to the existing literature by offering a further analysis of PRC data. We explore the current version of the dataset, which is much wider than the one considered in previous papers in the existing literature due to its constant updating. A more relevant difference is that we decided to exclude the MED data and focus on all the other organization types, with the main objective of investigating the differences in frequency and severity between two breach categories, which we denoted as malicious and negligent ones. The first are breaches due to malicious activities that actually aim for private information, while the second group originates from negligence by data owners, managers, or employees, resulting in records' accidental exposure. Our aim is to give further insights into the distribution of the frequency and severity of data breaches separately for the two categories of malicious and negligent ones and to assess the goodness of fit of the statistical models for frequency and severity of the breaches on the two considered subsets. Finally, we estimate the value at risk (VAR) in the two considered datasets. VAR is a well-known risk measure employed in the financial domain to assess the amount of money needed to front extreme negative events (worst cases). In the cyber risk domain, extreme events are characterized by a low probability of occurrence and a significant financial impact. For this reason, the scientific community started to use this risk measure in cyber risk management, identifying it as cyber value at risk (CVaR). VaR can help organizations to take into account the extreme events they are exposed to despite using particular risk control set-ups. The rest of the paper is structured as follows. Section 2 focuses on the main features of data breaches and relative studies in the recent literature. Section 3 describes the dataset and the statistical methodologies applied for its analysis, and Section 4 presents and discusses the main results. Section 5 concludes the paper.

2. Data Breach Risk

The authors of [4] gave an overview of both of the different types of cyber incidents and the kinds of losses that may occur. With regard to the incidents, four broad categories are included: data confidentiality breaches, system malfunctions or issues, data integrity or availability, and malicious activity. According to [1], data breaches together with ransomware attacks have dominated the cyber threat landscape in recent years. With the term "data breaches", we refer to incidents which involve the compromising of confidential data. The Chief Risk Officers Forum [10] classification divides data confidentiality incidents into two types: incidents involving one's own confidential data (e.g., financial data, trade secrets, and intellectual property) and incidents involving third party confidential data (e.g., customers' personal information). The release of confidential data through employee error has historically been the most common form of data confidentiality incidents. Many employees (intentionally or not) are often the weakest link that causes a successful cyber incident (e.g., accidental publication of confidential information or non-custody of laptop computers containing highly sensitive information) [4]. However, incidents caused by malicious attacks have accounted for an increasing share of data confidentiality incidents. Indeed, companies collect and use ever greater volumes of personal data, and breaches are becoming larger and more expensive. According to the Ponemon Institute [11], the average cost of a breach was USD 4.35 million in 2022, a new all-time high. The figure corresponds to an increase of 2.6% compared with the previous year, when the average cost of a violation amounted to USD 4.24 million. It also increased by 12.7% from USD 3.86 million in the 2020 report. Dealing with a mega breach (involving more than 1 million records) now costs USD 42 million on average. Moreover, data protection and privacy regulation, as well as subsequent penalties, are widening in scope and geographical reach. Organizations report cyber incidents to the public to be compliant with reporting requirements. Then, data brokers aggregate these reports, creating databases which are usually paid for, with a few exceptions. In general, large organizations are over-represented because their reports are more accessible. We can find several data breach studies in the existing literature. A very good review of the main results in this field is given in [12].

In the following, we analyze the data breach information, referred to as PRC data, obtained from the Chronology of Data Breaches compiled by the Privacy Rights Clearinghouse [13], a nonprofit organization. It is by far the most complete and publicly free to access dataset, as already stated by other researchers [7,8,14,15]. Several studies refer to PRC data to characterize data breach incidents. Among others, the authors of [16] investigated the duration between data breaches for enterprises having at least two incidents between 2010 and 2016. The authors of [14] develop Bayesian generalized linear models to investigate trends in data breaches and show that the size of data breaches is modeled well by the log-normal family of distributions and that the daily frequency of breaches is described by a negative binomial distribution. The authors of [7] used multidimensional scaling and goodness-of-fit tests to analyze the distribution of data breach information. As far as severity modeling is concerned, they showed that the log-skew-normal distribution provides promising results and provides useful insights for actuaries working on the implementation of cyber insurance policies. Based on PRC data, the authors of [9] proposed an actuarial approach to computing insurance premiums both from the insurer's and the insured's points of view. The authors of [8] proposed a novel frequency-severity model to analyze hacking breach risks at the individual company level. The authors of [15] analyzed a hacking breach dataset derived from PRC data, focusing on the incidents' inter-arrival times and the breach sizes, and they showed that both should be modeled by stochastic processes rather than distributions. Another issue concerns the dependability structure of losses, requiring a model able to deal with various (but dependent) classes of damages [17]. In order to model the dependency structure, copulas are commonly used. Referring to PRC data, the authors of [18] identified the significant asymmetric dependence of monthly losses. Models from extreme value theory (ETV) are popular too, because the data from operational risk show heavy tails. The peaks-over-threshold method (POT) is the most

common EVT approach. This approach allows the modeling of losses above a threshold (e.g., the 90% quantile) by a generalized Pareto distribution (GPD), and losses below the threshold with another common loss distribution. Applications of these models in the cyber risk domain were proposed in [7]. It is pointed out that different kinds of cyber incidents often show different statistical natures, requiring separate modeling [9].

3. Materials and Methods

3.1. Data Description

Privacy Rights Clearinghouse (PRC) is a nonprofit organization which was founded in 1992 and focuses on data privacy rights and issues. Among the resources made available on their website (law overviews, reports, information on data brokers, etc.), they maintain the Chronology of Data Breaches in the US, sourced from official breach reports such as state Attorneys General and the U.S. Department of Health and Human Services. The chronology contains information on data breaches which occurred in the US between 10 January 2005 and 31 December 2019, always including the name, location, and type of the breached entity, description and type of breach, and often the number of breached records. All the reported events have been confirmed by major media sources.

Data are categorized by both business type and breach type (see Tables 1 and 2). Specifically, organizations are classified as businesses: financial and insurance services (BSF); businesses: retail or merchant, including online retail (BSR); businesses: other (BSO); educational institutions (EDU); government and military (GOV); healthcare, medical providers, and medical insurance services (MED); and nonprofit organizations (NGO). Information exposures (breaches) are reported as debit and credit cards frauds excluding hacking (CARD), data loss due to hacking or malware infection (HACK), data loss due to insiders, such as employees, contractors, or customers (INSD), physical data loss, such as lost, stolen, or discarded documents (PHYS), data loss due to lost, discarded, or stolen portable devices such as laptops, smartphones, memory sticks, or hard drives (PORT), stationary computer or server data loss (STAT), unintended disclosure of data not involving hacking, intentional breaches, or physical loss, such as sensitive information posted publicly, mishandled or sent to the wrong party via online publishing, emails, or faxes (DISC), and unknown causes or not enough information (UNKN).

Table 1. The organization type of the companies falling victim to the data breaches considered with their descriptions.

Acronym	Description
BSF	Businesses: Financial and Insurance Services
BSO	Businesses: Other
BSR	Businesses: Retail or Merchant, Including Online Retail
EDU	Educational Institutions
GOV	Government and Military
MED	Healthcare, Medical Providers, and Medical Insurance Services
NGO	Nonprofits

Since these data rely on publicly acknowledged breaches, they represent just a part of all the security incidents. The main issue in their use is a probable underestimation of the phenomenon. A further limitation is the lack of information on financial losses derived from the breaches. An older version of this dataset was thoroughly described and analyzed by Eling and Loperfido (2017). However, the current one is much wider than the one they considered due to the constant updating of the PRC dataset.

Table 2. The type of the considered breaches with their descriptions.

Acronym	Description
CARD	Payment card fraud (fraud involving debit and credit cards that is not accomplished via hacking)
HACK	Hacking or malicious software
INSD	Insider (someone with legitimate access, such as an employee, contractor, or customer, who intentionally breaches information)
PHYS	Physical loss (includes paper documents that are lost, discarded, or stolen (non-electronic))
PORT	Portable loss (lost, discarded, or stolen laptops, PDAs, smartphones, memory sticks, CDs, hard drives, data tapes, etc.)
STAT	Stationary loss (lost, inappropriately accessed, discarded, or stolen computers or servers not designed for mobility)
DISC	Unintended disclosure (not involving hacking, intentional breach, or physical loss): sensitive information posted publicly, mishandled, or sent to the wrong party via publishing online, sending in an email, etc.
UNKN	Unknown loss

As thoroughly discussed in [8], it can be assumed that the more recent data are more representative of the current cyber threat situation. In particular, breaches reported in the last 10 years, when cyberspace safety was clearly and widely recognized as a fundamental issue, should be analyzed. In the cited paper, the authors only considered breaches which occurred after 1 January 2010. This choice was motivated by the concurrency in that year of several interventions (legislative acts, the creation of specific task forces, etc.), confirming the increased attention of both government agencies and industries toward cyber risks. We decided to follow the same approach and focus our analysis on the breaches reported after 1 January 2010, whose number and severity (as measured by the number of breached records) are shown in Tables 3 and 4, respectively.

Table 3. Number of reported breaches.

	BSF	BSO	BSR	EDU	GOV	MED	NGO	UNKN	TOTAL
#N/A	0	0	0	0	0	87	0	0	87
CARD	8	2	15	1	0	0	0	0	26
DISC	39	33	34	83	85	991	5	0	1270
HACK	78	188	113	102	68	798	20	0	1367
INSD	29	15	32	8	36	160	5	0	285
PHYS	11	14	5	15	31	1281	5	0	1362
PORT	16	23	10	31	46	289	10	0	425
STAT	3	3	2	6	2	72	0	0	88
UNKN	41	11	7	34	14	32	2	465	606
TOTAL	225	289	218	280	282	3710	47	465	5516

Table 4. Number of breached records.

	BSF	BSO	BSR	EDU	GOV	MED	NGO	UNKN
#N/A	0	0	0	0	0	3,079,889	0	0
CARD	7,035,066	310	2,124,575	16	0	0	0	0
DISC	1,550,375	2,105,006,706	385,194,087	1,576,141	21,094,488	12,979,387	3,501,561	0
HACK	348,057,288	5,494,774,684	791,295,680	45,231,810	40,900,705	159,979,906	3,350,944	0
INSD	2,407,569	3,508,456	35,671	40,379	28,506,293	1,059,014	317	0
PHYS	58,909	64,007	4071	1,023,422	209,616	35,715,718	24,157	0
PORT	5,852,045	5,836,258	30,244	238,778	7,683,283	12,645,645	72,176	0
STAT	100,348	80,108	9189	78,177	3650	9,604,567	0	0
UNKN	421,366	100,155,387	68,000,391	10,352,675	849,587	109,731	2501	10,657,026

As can be observed in these tables, there are a certain amount of incomplete data. For example, in medical organizations, several breaches do not report the breach type. Additionally, there are breaches registered as happening in unknown organization types or with missing indication of the severity. In the following analysis, we decided to disregard the incomplete records with unknown, unreported, or missing hacking breach sizes or unknown causes. The resulting dataset contained 4823 breach incidents reported in the United States between 1 January 2010 and 25 October 2019.

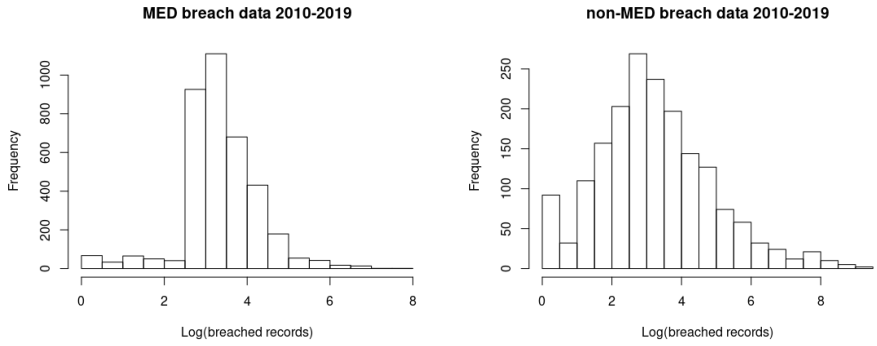
3.2. Modeling Methodology

Breach events can be characterized by their frequency (i.e., the number of daily, weekly, or monthly breaches) and their severity, generally estimated by the number s_i of compromised records in a breach i. With this specific dataset, previous studies [9,14] modeled the frequency of breach incidents by a negative binomial distribution, obtaining good results. It should be noticed, however, that this dataset includes data breaches of different types, as long as events occurred in different entities. As a consequence, the distributions of both the frequency and the severity of the breaches reflect this heterogeneity. This situation suggests a more adequate modeling that split the data into suitable subsets to consider separate risk categories. When the frequency was modeled on these subsets (again, by a negative binomial distribution), the same cited studies confirmed excellent performance of the estimate.

The distribution of data breach sizes, on the other hand, is generally modeled by a log-normal distribution, or better, by a skew-normal distribution [14,19]. A continuous random variable X has a skew-normal distribution with a shape parameter α if its probability density function has the form $f(x) = 2\phi(x)\Phi(\alpha x)$, where α is a real number, $\phi(\cdot)$ denotes the standard normal density function, and $\Phi(\cdot)$ is its distribution function [20]. The skew-normal distribution reduces to the standard normal distribution when $\alpha = 0$ and to the half-normal when α approaches infinity. The location (ξ) and scale (ω) parameters can be included via the linear transformation $Y = \xi + \omega X$. When $\alpha = 0$, the random variable Y is distributed as $\mathcal{N}(\xi, \omega^2)$.

However, due to the strongly different statistical nature of the data related to different types of breaches and different organizations, none of the considered models for the severity of breaches could give completely satisfactory results on the entire dataset. Further investigation detected medical type organizations as a peculiar subset of the observations that deserved specific attention. This subset suffered the largest number of data breaches (about 75% of the total events), but most of these events can be described as *negligent* breaches, caused by accidental exposure or inadequate vigilance. In addition, the MED companies in the PRC database include both big institutions, such as hospitals, and small medical practices so that the empirical distributions of the data breach sizes reflect the coexistence of these two subpopulations. The number of breached records per single event can even be very low, as shown in the left panel of Figure 1. Such an anomalous distribution is very difficult to model with satisfactory results. On the other side, the severity of breaches in all the other (non-MED) organization categories, shown in the right panel of the same Figure, can be more easily modeled by a log-normal or a skew-normal distribution.

These evident differences led us in [9] to focus our interest on the specific analysis of the MED/negl subset, comprised of the breaches in MED entities related to negligence causes. Then, in the present study, we excluded the MED subset and focused on the other organization types to investigate the differences in frequency or severity between the negligent breaches (i.e., DISC, PHYS, PORT, and STAT) and the *malicious* ones (i.e., CARD, HACK, and INSD) originating from activities that actively targeted private information. After excluding the MED category, the remaining 1232 observations were then split into negligent (512 observations) and malicious (720 observations) ones.



(**a**) (**b**)

Figure 1. Empirical distribution of the (logarithm of the) number of breached records, assumed to be a measure of the breach severity for (**a**) MED type organizations and (**b**) all but MED type organizations. Data reported in the PRC dataset in the time period 2010–2019.

3.3. Cyber Risk Measures

Measures and methods to quantify risk in financial terms from the business perspective are widely used in the financial service industry. Risk managers are deeply interested in quantifying the probability of adverse outcomes in order to determine the amount of capital required to withstand them, either by investing in security or by transferring the residual risk to an insurance company. Among these measures, value at risk (VaR) is intensively used as a risk management tool. Financial institutions use VaR as a benchmark to measure their exposure to market risk. It is a quantile risk measure defined as the "predicted worst-case loss at a specific confidence level over a certain period of time"; in other words, it provides an estimate of the maximum probable losses for a given confidence interval. This measure can be applied to cyber risk management, where we identify it as cyber value at risk (CVaR).

For a given confidence level $\alpha \in [0, 1]$, the VaR at the level α is defined as the smallest number l such that the probability that the loss L exceeds l is not greater than $(1 - \alpha)$:

$$\mathrm{VaR}_\alpha(L) = \inf[l \in \mathbb{R} | P(L > l) \leq 1 - \alpha].$$

Then, VaR_α(L) corresponds to the α-quantile q_α of the distribution function of the losses. Typical values for α in the context of market risk management are α = 0.95 , α = 0.99, and α = 0.995.

VaR estimation requires modeling the aggregate losses, or the total amount of all losses occurring in a fixed time period. Now, assessing the cost of data breaches is still an open issue. Some regression models between breached records and losses have been suggested in the literature, but the use of a standard per record cost to estimate the total breach cost would be an excessive simplification. Even if some costs directly correlate with the number of compromised records, for some other costs, there is only an indirect correlation or no discernable correlation with them. This consideration, along with the lack of information on the real cost of recorded data breaches, forced our study to a slight misuse of the VaR definition, where the estimation of the cyber risk is performed by considering the severity of the breach events rather than their financial cost.

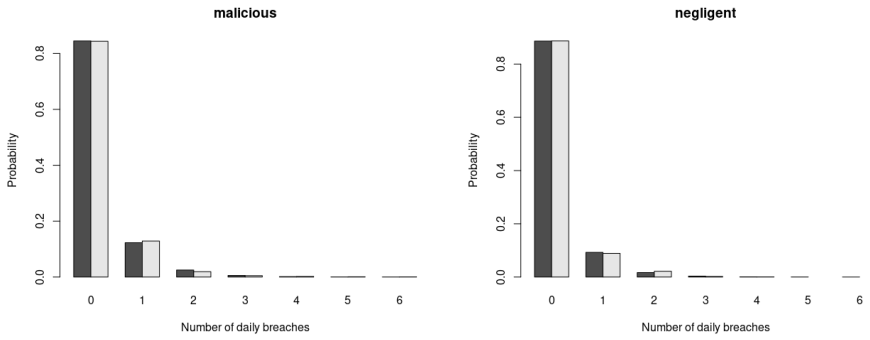
VaR can be estimated [21] by historical data as an empirical quantile of the severity distribution or also by Monte Carlo simulations of scenarios based on the modeled distribution of the historical data.

4. Results

Data analysis was performed in the R software environment. All the models mentioned in this section were implemented in the R packages MASS and sn. In the following, we describe the estimates obtained for both the frequency and severity of the breaches

and discuss their reliability. Once we obtained the maximum likelihood estimate of the distribution parameters, we also performed a Kolomogorov–Smirnov test (a nonparametric test comparing a sample with a reference probability distribution) to determine if the parameterized distribution and the data were statistically significantly different.

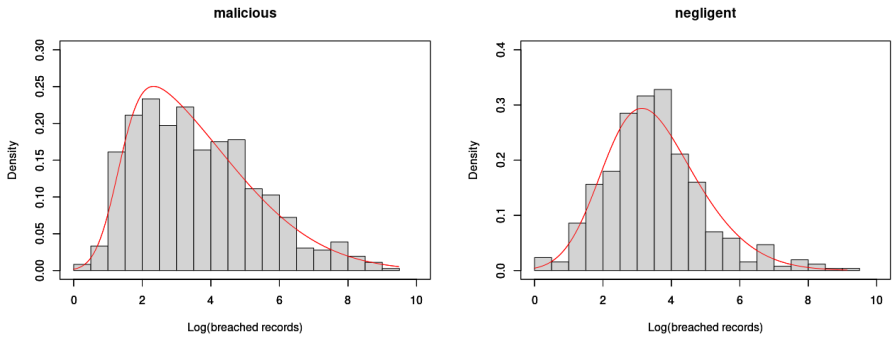
The breach frequency was modeled separately for the two groups (negligent and malicious). Specifically, we tried to fit both a Poisson and a negative binomial distribution to the daily frequencies of the breaches and found that an excellent fit was accomplished by the negative binomial distribution. The results of the Kolmogorov–Smirnov test showed a p-value = 1 in both the negligent and malicious subsets. Figure 2 compares the empirical daily frequencies to the estimated ones for the two populations.



(**a**) (**b**)

Figure 2. Empirical (light gray) vs. estimated (dark gray) distributions of the daily frequency of breaches for (**a**) malicious and (**b**) negligent breach types, respectively. Data for all but MED type companies as reported in the PRC dataset in the time period 2010–2019.

To model the severity of both malicious and negligent breaches, we explored several distributions among the ones proposed in the literature and found that the best fit was obtained by the skew-normal distribution (the Kolmogorov–Smirnov test provided p-values above 0.20 for both subsets). Specifically, we fit a skew-normal distribution to the logarithms of the number of breached records in each event. Figure 3 shows the superposition of the estimated severity (red line) compared to the empirical one in the two considered subsets.

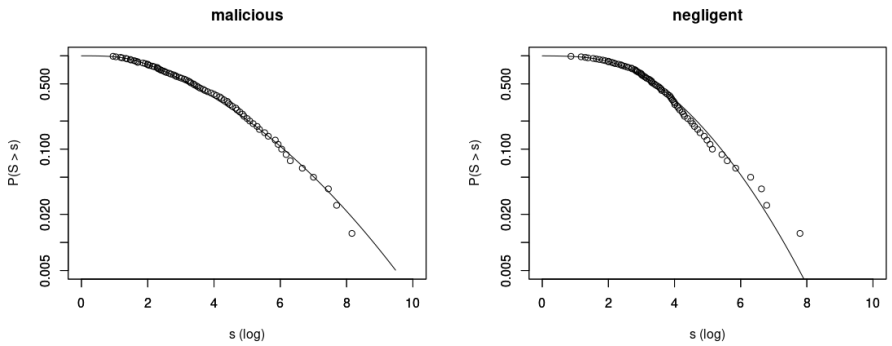


(**a**) (**b**)

Figure 3. Empirical (histograms) vs. estimated (red line) distributions of the severity of breaches, measured by the number of breached records for (**a**) malicious and (**b**) negligent breach types, respectively. Data for all but MED type companies as reported in the PRC dataset in the time period 2010–2019.

We also analyzed the tails of the severity distributions for the two groups. Because the distribution of breach sizes was heavy-tailed, large events occurred more frequently than expected, and it is worth investigating such large breaches to explore the possibility of different modeling (based on extreme value theory) for these tails. Figure 4 compares

the empirical distribution of the breach sizes with the modeled skew-normal distribution functions. In both subsets, there was just a single very large event not exactly fitted by the skew-normal distribution, so in the considered dataset, there was not enough experimental evidence to justify different modeling for the larger breach sizes. Our findings are in good agreement with the ones reported in [7], where the authors applied the peaks-over-threshold (POT) method to the same dataset and found just a relatively small improvement in the goodness of fit with respect to the log-normal and skew-normal models.



(**a**) (**b**)

Figure 4. The empirical distribution function of the (logarithm of the) number of breached records (circles) compared to the estimated distribution function (black line) for (**a**) malicious and (**b**) negligent breach types, respectively. Data for all but MED type companies as reported in the PRC dataset in the time period 2010–2019.

Finally, we estimated the VaR on the two data subsets. In Table 5, four confidence levels are reported (90%, 95%, 99%, and 99.5%), which are the most relevant ones for risk-based capital modeling. VaR is estimated on a daily basis. To better understand its significance, let us consider its estimated value for malicious breaches at a confidence level of 95% in Table 5. This means that there was a confidence level of 95% that, over the next day, the number of breached records due to malicious activity would not be higher than about 10^7. In an equivalent way, one can assert that there was a confidence level of 5% that, over the next day, the number of breached records would be higher than 10^7. The information given by VaR can be useful to a company in order to improve in the future its capability to reduce the likelihood of an extreme cyber threat by taking informed decisions on risk control adoption. It is worth remembering that the currently available data only allow for an aggregated analysis, where malicious (or negligent) breaches which occurred in all company types are considered together. More accurate indicators could be obtained by disaggregating information by company type, but such an analysis would require a larger dataset to be statistically robust. In the same table, the p value for testing the difference between the estimated and empirical values is also reported for all confidence levels. These p values confirm the evidence from Figure 4. While the fit for the severity of malicious breaches gave good results up to the 95th percentile and degraded after, the one for negligent breaches poorly approximated the tail of the empirical distribution. Indeed, this was already evident from the 90th percentile, and the worse fit concentrated in the first part of the tail, where more data were available.

Table 5. Risk measurement results: estimated (es) and empirical (em) values of the log of breached records for the two considered categories, malicious (mal) and negligent (negl), along with the p values (pv) of their agreement.

Type	**90 es**	**90 em**	**pv**	**95 es**	**95 em**	**pv**	**99 es**	**99 em**	**pv**	**99.5 es**	**99.5 em**	**pv**
mal	6.10	6.04	0.71	7.02	7.00	0.94	8.81	8.17	0.02	9.49	8.53	0.02
negl	5.41	5.14	0.02	6.05	6.30	0.19	7.30	7.95	0.16	7.78	8.18	0.45

To sum up our findings, in this study, we proceeded through the following steps:

- After observing the different statistical nature (in terms of distribution parameters) of the negligent breaches and the malicious ones, we separated the breach data into these two categories.
- We verified that the best fit for the daily frequency of both categories was given by a negative binomial distribution, and this was completely confirmed by a Kolmogorov–Smirnov (KS) test.
- Regarding severity, we observed that the best fit for both malicious and negligent breaches, among all the distributions proposed in the literature, was given by the skew-normal distribution. The fits were not as accurate as the frequency ones, but the KS test still confirmed the quite good performance of the model.
- Indeed, very few large breaches did not fit well with the skew-normal distribution, where negligent breaches were slightly underestimated and malicious ones were overestimated. All the conclusions were confirmed by the VaR estimates in Table 5.

5. Conclusions

In this final section, we relate the motivation and findings of this study along with some open issues and perspectives for future work. With the awareness that the first step toward modeling cyber risk is to collect data reflecting accurately this risk [2], we decided to provide further insight on the PRC dataset. The main reason for this choice was due to the fact that this dataset gives information about a particular type of cyber incident: the data breach. This point is of great interest because the data breach is the category of cyber incidents that causes the biggest losses nowadays. Moreover, PRC data are recognized as the most complete and reliable publicly available dataset, and it has been widely analyzed in the literature, as discussed in Section 2. Following [8], we decided to analyze the recent data (breaches reported after 1 January 2010), assuming that they reflected the more recent cyber threat situation. It is evident, though, that this dataset shows some relevant flaws. First of all, the data do not include information on financial losses. This reason forced our study into a slight misuse of the VaR definition by estimating cyber risk through considering the severity of the breach events rather than their financial impact. According to the findings of our previously published study, the empirical distribution of the subset of breaches which occurred in medical and health services showed several anomalies. Then, in the present study, we excluded the MED subset and focused on the other organization types, obtaining a quite accurate fit for the distributions of the frequency and severity of breaches when malicious and negligent events were separately modeled. A very positive feature of the considered PRC dataset is its structure. Indeed, it provides data on the type of company, the type of breach, and the dates of breach disclosure and relates these dates to the company's fiscal year. Our next step could be to repeat our analysis following [22] (i.e., by merging PRC data with existing accounting and finance datasets), which could allow for cross-sectional and longitudinal analyses. Moreover, other datasets should be analyzed in the future. In particular, it would be desirable to analyze data relating to European companies, and the availability of such data should be facilitated by the entry into force of the GDPR.

Author Contributions: Conceptualization, A.O.; methodology, M.F.C. and A.O.; software, M.F.C.; data curation, M.F.C. and A.O.; writing M.F.C. and A.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data from Privacy Right Clearinghouse are publicly available at https://privacyrights.org/databreaches, accessed on 15 June 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Allianz Global Corporate & Specialty. Allianz Risk Barometer: Top Business Risks for 2022. Report. 2022. Available online: https://www.agcs.allianz.com/news-and-insights/reports/allianz-risk-barometer.html (accessed on 10 February 2021).
2. Dacorogna, M.; Debbabi, N.; Kratz, M. Building up Cyber Resilience by better grasping cyber risk via a new algorithm for modelling heavy-tailed data. *arXiv* **2022**, arXiv:2209.02845.
3. Weathley, S.; Hofmann, H.; Sornette, D. Data breaches in the catastrophe framework and beyond. *arXiv* **2019**, arXiv:1901.00699v2.
4. OECD. Types of cyber incidents and losses. In *Enhancing the Role of Insurance in Cyber Risk Management*; OECD Publishing: Paris, France, 2017. [CrossRef]
5. Martinelli, F.; Orlando, A.; Uuganbayar, G.; Yautsiukhin, A. Preventing the Drop in Security Investments for Non-competitive Cyber-Insurance Market. In *Risks and Security of Internet and Systems*; Cuppens, N., Cuppens, F., Lanet, J.L., Legay, A., Garcia-Alfaro, J., Eds.; CRiSIS 2017. Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 10694, pp. 159–174.
6. Marotta, A.; Martinelli, F.; Nanni, S.; Orlando, A.; Yautsiukhin, A. Cyber-insurance survey. *Comput. Sci. Rev.* **2017**, *24*, 35–61. [CrossRef]
7. Eling, M.; Loperfido, N. Data breaches. Goodness of fit, pricing, and risk measurement. *Insur. Math. Econ.* **2017**, *75*, 126–136. [CrossRef]
8. Sun, H.; Xu, M.; Zhao, P. Modeling Malicious Hacking Data Breach Risks. *N. Am. Actuar. J.* **2021**, *25*, 484–502. [CrossRef]
9. Carfora, M.F.; Martinelli, F.; Mercaldo, F.; Orlando, A. Cyber Risk Management: An Actuarial Point of View. *J. Oper. Risk* **2019**,*14*, 77–103.
10. Chief Risk Officers Forum—CRO Forum. *Concept Paper on a Proposed Categorisation Methodology for Cyber Risk CRO Forum*; CRO Forum: Amsterdam, The Netherlands, 2016.
11. Ponemon Institute. *2022 Cost of Data Breach Study: Global Analysis*; Ponemon Institute LLC: Traverse City, MI, USA, 2022.
12. Woods, D.W.; Böhme, R. SoK: Quantifying Cyber Risk. In Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 24–27 May 2021.
13. Privacy Rights Clearinghouse Chronology of Data Breaches. 2022. Available online: https://privacyrights.org/data-breaches (accessed on 15 June 2022).
14. Edwards, B.; Hofmeyr, S.; Forrest, S. Hype and heavy tails: A closer look at data breaches. *J. Cybersecur.* **2016**, *2*, 3–14. [CrossRef]
15. Xu, M.; Schweitzer, K.M.; Bateman, R.M.; Xu, S. Modeling and Predicting Cyber Hacking Breaches. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2856–2871. [CrossRef]
16. Buckman, J.; Bockstedt, J.; Hashim, M.J.; Woutersen, T. Do organizations learn from a data breach. In Proceedings of the Workshop on the Economics of Information Security (WEIS), La Jolla, CA, USA, 26–27 June 2017; pp. 1–22.
17. Bentley, M.; Stephenson, A.; Toscas, P.; Zhu, Z. A multivariate model to quantify and mitigate cybersecuity risk. *Risks* **2020**, *8*, 61. [CrossRef]
18. Eling, M.; Jung, K. Copula approaches for modeling cross sectional dependence of data breach losses. *Insur. Math. Econ.* **2018**, *82*, 167–180. [CrossRef]
19. Eling, M. Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insur. Math. Econ.* **2012**, *51*, 239–248. [CrossRef]
20. Azzalini, A. A class of distributions which includes the normal ones. *Scand. J. Stat.* **1985**, *12*, 171–178.
21. Carfora, M.F.; Orlando, A. Quantile based risk measures in cyber security. In Proceedings of the 2019 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA), Oxford, UK, 3–4 June 2019; pp. 1–4.
22. Rosati, P.; Lynn, T. A dataset for accounting, finance and economics research on US data breaches. *Data Brief* **2021**, *35*, 106924. [CrossRef] [PubMed]