Insurance: Mathematics and Economics 75 (2017) 126–136

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Data breaches: Goodness of fit, pricing, and risk measurement

Martin Eling [a,*], Nicola Loperfido [b]

[a] Institute of Insurance Economics, University of St. Gallen, Kirchlistrasse 2, 9010 St. Gallen, Switzerland
[b] Dipartimento di Economia, Società e Politica – DESP, University of Urbino, Via Saffi 42, Urbino (PU) 61029, Italy

ARTICLE INFO

Article history:
Received December 2016
Received in revised form May 2017
Accepted 20 May 2017
Available online 26 May 2017

JEL classification:
G22
G31

Keywords:
Cyber risk
Risk measurement
Multidimensional scaling
Goodness of fit
Skew-normal distribution

ABSTRACT

Some research on cyber risk has been conducted in the field of information technology, but virtually no research exists in the actuarial domain. As a first step toward a more profound actuarial discussion, we use multidimensional scaling and goodness-of-fit tests to analyze the distribution of data breach information. Our results show that different types of data breaches need to be modeled as distinct risk categories. For severity modeling, the log-skew-normal distribution provides promising results. The findings add to the recent discussion on the use of skewed distributions in actuarial modeling (Vernic, 2006; Bolancé et al., 2008; Eling, 2012). Moreover, they provide useful insights for actuaries working on the implementation of cyber insurance policies. We illustrate the usefulness of our results in two applications on risk measurement and pricing.

1. Introduction

Cyber risks are operational risks to information and technology assets that have consequences for the confidentiality, availability, and integrity of information and information systems (see Cebula and Young, 2010). Although cyber risks, such as hacking attacks or unintended disclosures, are reported in the media every day and rank high in the business agenda of every Chief Financial Officer and Chief Risk Officer, to our knowledge no research on the topic has been done in the actuarial domain. This is surprising, given the high economic importance (global losses for cyber risk are estimated to surpass US$400 billion per year; see McAfee, 2014) and the increasing efforts of many insurance companies to further develop a market for cyber risk insurance (see Biener et al., 2015).

One reason for the lack of research in the actuarial domain is the lack of data. Recently, however, this situation has changed, especially with the establishment of first data breach databases. In the US, reporting requirements for data breaches have been introduced in many states since 2002 (National Conference of State Legislatures, 2016), and data breach databases are becoming increasingly available. This paper analyzes such data using both exploratory (multidimensional scaling, multiple factor analysis for contingency tables) and confirmatory approaches (goodness of fit).

The literature on cyber risk and information security is mainly limited to the field of information technology, but very little work has been done in business, finance, and economics. Our paper is closest to the data breach analyses of Maillart and Sornette (2010), Edwards et al. (2015), and Wheatley et al. (2016).[1] The intention of this paper is to link what has been done in those three papers with the current discussion on goodness of fit, pricing, and risk measurement in the actuarial domain (see Vernic, 2006; Bolancé et al., 2008; Eling, 2012; Miljkovic and Grün, 2016, among others).

Multidimensional scaling shows that different types of data breaches need to be modeled as distinct risk categories, given their different statistical nature—a result that has not been the focus of existing data breach analyses. For the severity model, it turns out that either the log-normal or the log-skew-normal distribution provides promising results. This is a relevant result, considering the

* Corresponding author.
E-mail addresses: martin.eling@unisg.ch (M. Eling), nicola.loperfido@uniurb.it (N. Loperfido).

[1] Maillart and Sornette's (2010) study of the statistical properties of data breaches between 2000 and 2008 reveals the existence of two distinct phases for the breach frequency (explosive growth up to about July 2006 and a stable rate thereafter). Breach size follows a heavy-tailed power-law distribution, remains stable over time and does not depend on the organization's type or size. Edwards et al. (2015) analyze time trends for the size and frequency of malicious and negligent data breaches and show that neither size nor frequency of breaches has increased in recent years; breach size is distributed log-normally and the frequency follows a negative binomial distribution. Wheatley et al. (2016) extend Maillart and Sornette's (2010) work by enlarging the dataset and focusing on the tails of the distribution (i.e., incidents with more than 50,000 records breached). They show that the frequency of large events is independent of time for the US and is increasing over time for non-US firms.

http://dx.doi.org/10.1016/j.insmatheco.2017.05.008
0167-6687/© 2017 Elsevier B.V. All rights reserved.

recent discussion on the use of skewed distributions in actuarial science (e.g., Vernic, 2006; Eling, 2012). Our results offer important information for insurance companies and regulators seeking to better understand the potential risk exposure when selling cyber insurance policies. Moreover, we also hope to encourage more research on the topic in the risk and insurance domain.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the methods employed in the paper. Section 3 presents the data and descriptive statistics. The results are given in Section 4, including two applications on risk measurement and pricing. Section 5 concludes the paper.

2. Methodology

We analyze data breach information using both exploratory and confirmatory approaches (Tukey, 1977). Exploratory data analyses aim to uncover previously unanticipated data features; we implement such analyses via multidimensional scaling (MDS) and multiple factor analysis for contingency tables (MFACT). Confirmatory data analyses aim to test statistical hypotheses; we implement such analyses by testing the goodness of fit for several well-known distributions, especially in the tails. For the sake of brevity, we only briefly describe the main methodological approaches used in the paper and refer to the literature for all details.

MDS is a multivariate statistical technique that is rarely used in the context of insurance (one exception is Brechmann et al., 2013), but is widely used in other fields. It aims to recover the data structure from the distances between data points. MDS approximates interpoint distances with Euclidean distances between numbers, pairs of numbers, or trios of numbers. When distances are Euclidean, MDS and principal component analysis lead to the same results; thus, the former generalizes the latter. A detailed description of MDS, together with its potential applications, is given by Mardia et al. (1979). In this paper, we use MDS to investigate differences between entities that suffer data breaches and differences between various types of attacks. In both cases, we first use the number of data breaches (to measure frequency) and then use the number of lost records (to measure severity). Differences between either entities or types will be evaluated via the chi-squared distance, which is the default choice for a distance when the data matrix is a contingency table.[2/3]

A dynamic approach taking the evolution of data breaches through time into account requires a joint analysis of several contingency tables. Bécue-Bertaut and Pagés (2004, 2008) introduced MFACT, a multivariate statistical method specifically developed for such situations. It has been applied to textual analysis (Bécue-Bertaut, 2014) and implemented in the R package FactoMineR (Husson et al., 2007), which has been thoroughly illustrated by Kostov et al. (2013).

In the confirmatory data analysis, we test several established actuarial models with respect to their goodness of fit. The data breach frequency is modeled by either a Poisson or a negative binomial distribution (see, e.g., Moscadelli, 2004). For the data breach severity, we fit the data to several distributions used in recent actuarial literature (see, e.g., Eling, 2012). Furthermore, we include a non-parametric transformation kernel estimation (see Bolancé et al., 2003, 2008)[4] and implement the peaks-over-threshold (POT) method from extreme value theory (EVT; see, e.g., Chapelle et al., 2008). In the latter approach, losses above a threshold (e.g. the 90% quantile) are modeled by a generalized Pareto distribution (GPD), while losses below the threshold are modeled with another common loss distribution, such as exponential, log-normal, or Weibull. To identify the best models, we apply various goodness-of-fit tests (log-likelihood value, the AIC, Kolmogorov–Smirnov (KS)-test, Anderson–Darling (AD)-test).[5]

All models are implemented in the R packages sn, ghyp, and MASS. We use all packages to derive the best-fitting parameters and compare these distributions. Some of the benchmark distributions are also involved in the risk measurement and pricing procedure, where we compare model results with the empirical results to evaluate the accuracy of different models. More details on skewed distributions can be found in Adcock et al. (2015) and Azzalini (2013); a description of the other benchmark models is given in actuarial textbooks, such as Mack (2002), Kaas et al. (2009), and Panjer (2007). It should also be mentioned that a better fit does not necessarily mean that a model is better, as actuaries need to keep in mind many other aspects, such as the risk of change of the underlying stochastic process.

3. Data and descriptive statistics

The data breach information we consider is taken from the "Chronology of Data Breaches" provided by the Privacy Rights Clearinghouse (PRC). This dataset has not yet been used in the context of actuarial science, but it has been applied in other fields (see Maillart and Sornette, 2010; Edwards et al., 2015; Wheatley et al., 2016). The PRC is a non-profit organization with the mission to engage, educate, and empower individuals to protect their privacy (Privacy Rights Clearinghouse, 2016); their data breach dataset is regularly updated and can be downloaded from the PRC website. The data sample we use here consists of data breaches in the US between January 10, 2005, and December 15, 2015. We follow Edwards et al. (2015) in erasing all observations that do not give information on the number of records; this yields a sample of 2266 observations. The data contain only the number of records affected by data breaches and do not include financial losses.[6]

[2] Information about data breaches is transformed into a contingency table, where each cell contains the number of data breaches of a given type and entity and each row (column) represents an entity (a type). A second contingency table contains the amount of data breached. The chi-squared distance (Izenman, 2008, p. 642) is the measure of discrepancy between rows (columns) and gives a weighted distance between rows (columns) of the table. The chi-squared distance is mostly used in correspondence analysis, a multivariate statistical technique used for exploring the association between cross-tabulated data (Izenman, 2008, Chapter 17).

[3] Chi-squared distance also plays a central role in correspondence analysis (CA), which is particularly apt for analyzing contingency tables. However, risk managers and insurance companies are primarily interested in whether different entities need different insurance contracts against cyber risks. Hence the need to focus on pairwise distances between entities, whose visualization is the primary goal of MDS. CA aims at approximating the total inertia of a contingency table that is a weighted average of squared distances between row profiles and the row centroid (see, for example, Izenman, 2008, p. 642). Proper approximation of pairwise distances between entities is not the primary goal of CA. For this reason, we focus on MDS when analyzing frequencies of differences among data breaches and use CA for validating MDS results.

[4] In the main body of the paper, the standard Silverman's rule smoothing parameter is implemented. In additional tests, available from the authors upon request, we also implement alternative estimation approaches following Alemany et al. (2013), which do not materially change our results; one explanation might be that the data are not too extreme in the tails.

[5] The bootstrap goodness-of-fit test by Villaseñor-Alva and González-Estrada (2009) is used to identify the optimal threshold value for the POT method.

[6] We apply our analyses to the original dataset (i.e., the number of records breached) and the natural logarithm of the number of records breached. An open research question is how to transform the number of records breached into actual loss data; one potential approach is the transformation described by Jacobs (2014); losses are estimated by $\ln(\text{loss}) = 7.68 + 0.76 \cdot \ln(\text{records breached})$. Jacobs (2014) generated this relationship between the number of records breached and the actual losses for the years 2013 and 2014 only, showing no significant differences in the two years. We use this formula to estimate insurance prices in Section 4.3. In additional tests, available upon request, we also present the results for alternative transformations presented by Jacobs (2014). The estimated prices vary substantially, depending on the type of transformation used, illustrating the need for future research on this topic.

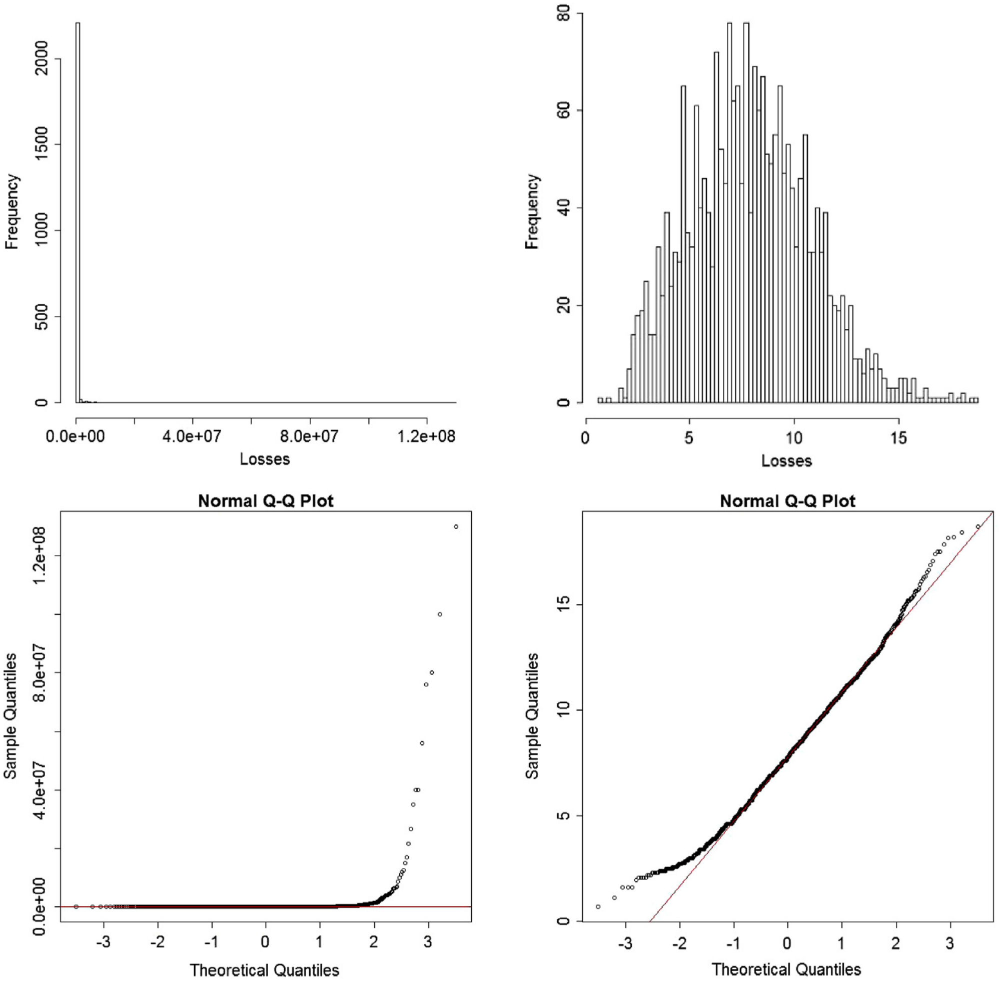


Fig. 1. Visualization of PRC data (left) and LN of PRC data (right).

We first provide a graphical illustration of the data in Fig. 1. Similar to insurance loss data (see, e.g., Eling, 2012), we observe that the distribution is skewed and exhibits a high kurtosis. The result from the quantile–quantile-plot (Q–Q-plot) for the original data shows that data breaches are not normally distributed. However, the Q–Q-plot for the logarithmic data shows a relatively good fit for the normal distribution. Edwards et al. (2015) already identified the log-normal distribution to be the optimal model for describing breach sizes.

The results from the Q–Q-plots and histograms can be underscored by the descriptive statistics in Table 1.[7] While the total sample exhibits high skewness and kurtosis parameters, these numbers are similar to the normal distribution for the log-transformed data. Summary statistics for the type of data breach and for the entity type in which the breach occurred are also presented. Here we also show the categorization in a malicious and negligent data breach, a differentiation used in Edwards et al. (2015) as an alternative measure for data breach type, which aggregates several breach categories. The numbers already illustrate substantial differences in frequency and severity of data breaches for different entity types. For example, the average amount of breached data in the category BSF (Businesses–Financial and Insurance Services, 1,284,040) is 47 times higher than is the average amount in the category EDU (Educational Institution, 27,372).

When looking at the development over time (Fig. 2), we observe the frequency (i.e., the number of incidents) decreasing over time, while the results for the severity (i.e., the mean/median number of records breached) show relatively small and constant values at the beginning of the observation period, with some more severe incidents at the end of 2009 and at the end of 2014. However, no clear upward or downward trend is observed.

The reliability of the dataset is important. Each breach event has been confirmed by at least one major media source and is, thus, easy to both trace and peer-review (see Maillart and Sornette, 2010). The dataset has already been used in numerous academic papers (e.g., Edwards et al., 2015) and is widely accepted in practice (e.g., Maillart and Sornette, 2010 mention that practitioners recognize the dataset as being the most complete dataset).

4. Results

4.1. Exploratory analysis

The results for the MDS are presented in Table 2. Here we analyze the frequency and severity of data breaches for both the different entities and the different types of data breaches. The first and second scores are the first and the second principal coordinates obtained from either frequency or severity data by classical MDS

[7] The summary statistics present the 2266 data entries after erasing all observations that gave no information on the number of lost records. When we apply MDS to frequencies, we also consider the cases where no information on the number of lost records was given.

Table 1
Variable description and summary statistics – PRC dataset.

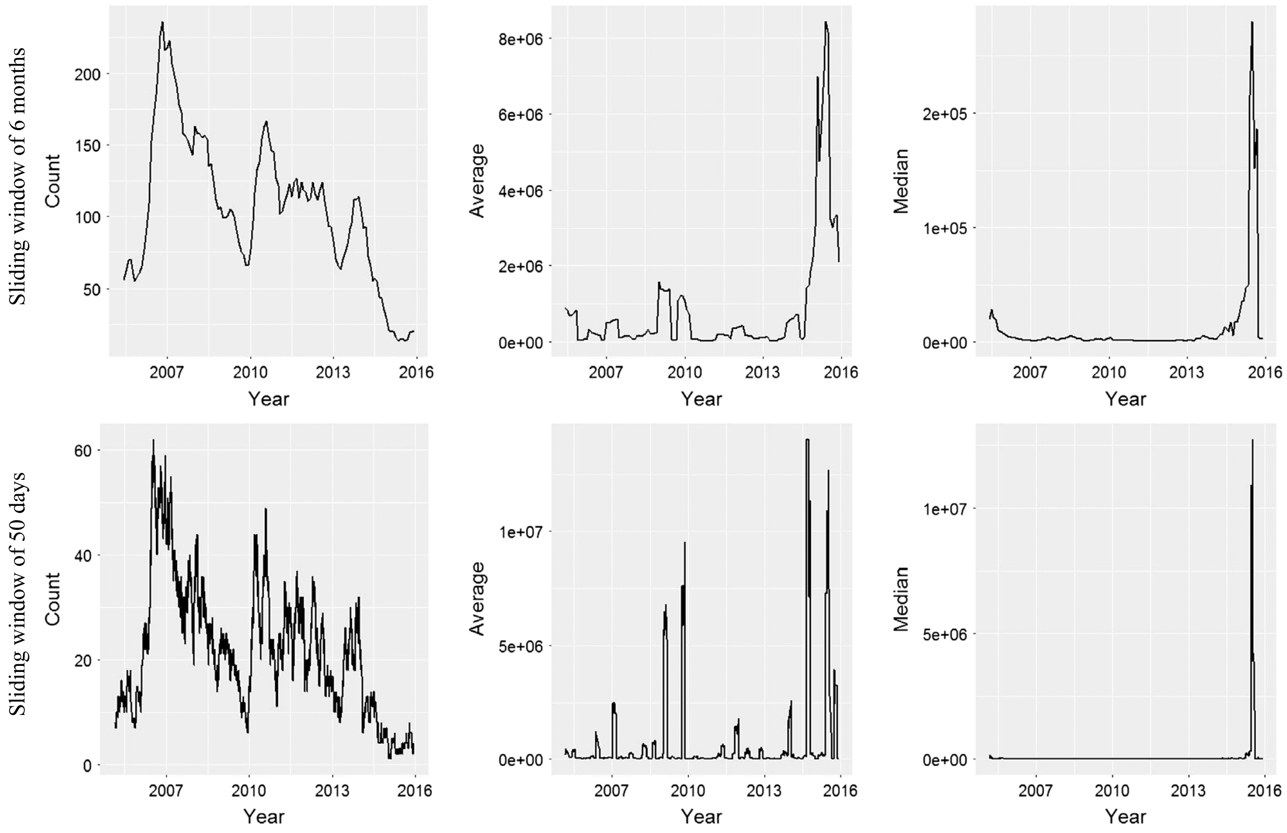
Panel A: Variable description	
Types of data breach	
CARD	Payment Card Fraud – fraud involving debit and credit cards that is not accomplished via hacking (e.g., skimming devices at point-of-service terminals)
DISC	Unintended disclosure – sensitive information either posted publicly on a website, mishandled, or sent to the wrong party via email, fax, or mail
HACK	Hacking or malware – electronic entry by an outside party, malware, and spyware
INSD	Insider – someone with legitimate access, such as an employee or contractor, intentionally breaches information
PHYS	Physical loss – lost, discarded, or stolen non-electronic records, such as paper documents
PORT	Portable device – lost, discarded, or stolen laptop, PDA, smartphone, portable memory device, CD, hard drive, data tape, etc.
STAT	Stationary device – lost, discarded, or stolen stationary electronic device, such as a computer or server not designed for mobility
UNKN	Unknown or other
Malicious vs. negligent data breaches	
MALB	Malicious data breaches; CARD + HACK + INSD
NEGB	Negligent data breaches; DISC + PHYS + PORT + STAT
Entity types	
BSF	Businesses – Financial and insurance services
BSO	Businesses – Other
BSR	Businesses – Retail/Merchant
EDU	Educational institution
GOV	Government and military
MED	Healthcare – Medical providers
NGO	Nonprofit organizations

Panel B: Summary Statistics	*N*	Mean	Standard Dev.	Median	Skewness	Kurtosis
Total sample	2266	395,203.80	4,655,598.00	2400.00	19.60	441.68
LN total sample	2266	7.91	2.91	7.78	0.33	2.99
Type of data breach						
CARD	30	240,101.20	1,276,882.00	300.00	4.94	26.19
DISC	459	69,803.75	480,911.80	1800.00	10.90	130.87
HACK	475	1,319,235.00	9,336,615.00	9000.00	10.11	116.99
INSD	283	127,795.50	1,146,738.00	431.00	12.70	175.80
PHYS	197	16,312.50	143,780.00	600.00	13.35	183.94
PORT	629	274,670.80	3,262,891.00	4400.00	20.79	468.18
STAT	135	85,694.39	382,233.50	5000.00	8.41	82.80
UNKN	58	102,344.40	400,299.60	1461.00	4.80	26.29
Malicious vs. negligent data breaches						
MALB	788	850,261.10	7,305,389.00	2750.00	12.93	191.47
NEGB	1420	154,641.00	2,194,282.00	2400.00	30.48	1020.68
Entity type						
BSF	295	1,284,040.00	9,326,960.00	2000.00	11.23	141.22
BSO	229	938,835.48	424,170.60	3800.00	9.07	97.65
BSR	198	1,300,592.00	8,953,309.00	665.50	8.60	84.41
EDU	538	27,371.90	126,891.70	2567.00	14.72	269.60
GOV	440	404,366.20	4,004,190.00	3000.00	16.28	294.60
MED	519	82,970.54	417,948.00	2995.00	8.35	79.36
NGO	47	43,101.40	153,700.20	1537.00	5.28	32.25

Table 2
MDS results.

Entity	Frequency		Severity		Type	Frequency		Severity	
	First	Second	First	Second		First	Second	First	Second
BSF	2.8629	−0.0782	−0.3657	0.9265	CARD	−13.6414	−4.3392	9.4172	−1.8346
BSO	−6.6298	1.0877	0.5721	0.7714	DISC	0.3416	5.1225	−2.0219	4.3558
BSR	−7.1158	0.6597	3.6734	2.5255	HACK	−5.7478	5.6719	1.5049	4.6809
EDU	−2.2754	−3.9755	3.4192	−3.0010	INSD	3.6849	−3.3348	−1.7610	−5.1351
GOV	4.3221	−3.3202	−1.8006	−3.0356	PHYS	5.7783	−1.2972	−4.2078	−1.8424
MED	8.5912	1.9238	−3.8808	0.7296	PORT	4.5149	−0.7872	−1.6448	−1.2392
NGO	0.2448	3.7028	−1.6176	1.0836	STAT	6.0842	−0.8530	−2.1703	0.2260
					UNKN	−1.0147	−0.1829	0.8837	0.7885
Reliability	94.13%		95.05%			92.47%		91.59%	

Note: See Table 1 for variable definitions.



(a) Number of incidents.

(b) Mean number of records breached.

(c) Median number of records breached.

Fig. 2. Development of number of incidents and the mean/median number of records breached over time.

(Izenman, 2008, p. 481). We notice that the reliability of the results is, in general, very high, ranging from 91.59% to 95.05%.[8]

We first analyze the frequency for entities. When applying MDS to entities, we seek to uncover the differences between entities with respect to the types of data breaches they faced. The first MDS score is highly correlated with the HACK frequency (correlation of −99.5%), while the second score represents DISC (correlation of −94.2%).[9] Thus, although HACK and DISC represent less than 50% of the data, they account for most of the distances between units. In the left part of Fig. 3, BSO and BSR are very close to each other and are very different from MED. This is a meaningful finding, given that businesses are typically more prone to hacking attacks than are hospitals. NGOV, BSF, and GOV are close to each other regarding HACK, but they vary substantially in the second dimension, which is correlated with DISC. Thus, unintended disclosure seems to be a significant source of data breach problems, especially in the government and military. This is also the case in educational institutions. We conclude that BSO and BSR face very similar cyber problems, which are quite different from the cyber problems faced by NGOV, BSF, and GOV. The problems faced by EDU and MED are also very different from those faced by other entities.

By applying MDS to types, we aim to uncover the differences between types of data breaches, with respect to the entities in which they occurred. Again, the accuracy of the proposed approximation is very good (92.5%). The first score in Table 2 is highly correlated with the frequency of MED and BSR (correlation of 95.2% and −91.7%), while the second score is highly correlated with EDU (correlation of 89.3%). The MDS output is displayed in the right part of Fig. 3. Here we see that PORT, STAT, and PHYS are very close to each other; meanwhile, CARD is very different from all others, and UNKN is very close to the origin. We conclude that PORT, STAT, and PHYS appear with similar frequencies in all entities, whereas CARD is relatively rare. We also note that the classification into a malicious and negligent data breach given in Edwards et al. (2015)[10] is a good way for describing the source of the data breach.

[8] For reliability, we measure the discrepancy between the $n \times n$ distance matrix $D = \{d_{ij}\}$ to be approximated and the approximating $n \times n$ Euclidean distance matrix $\Delta = \{\delta_{ij}\}$ using the ratio $\Sigma_{ij}(d_{ij} - \delta_{ij})^2 / \Sigma_{ij}(d_{ij})^2$, where the indices i and j range from 1 to n. The ratio is bounded between zero (when the approximating matrix equals the approximated matrix) and one (when the approximating matrix is a null matrix). Multidimensional scaling is associated with each unit of the k scores, which minimizes the discrepancy ratio (Mardia et al., 1979, pp. 406–409). The smaller the ratio, the smaller the discrepancy between D and Δ and the greater the reliability of the scatter plots obtained through multidimensional scaling. The ratio, whose numerator is commonly known as raw stress in the MDS literature (Izenman, 2008, p. 497), is connected with the R^2 statistic in regression analyses. The latter is the ratio $1 - \Sigma_i(y_i - \hat{y}_i)^2 / \Sigma_i(y_i)^2$, where y_i and $\hat{y}_i$ are the ith observed and fitted value, respectively, and whose averages are taken to be zero, without the loss of generality, to better show the connection with the stress statistic. Kruskal (1964) provides some commonly accepted guidelines for the use of the stress function when assessing the fit of the MDS solution to the data at hand. For example, a value between 0.707 and 0.987 suggests a good fit. All ratios in Table 2 lie in that interval and are close to 0.987, suggesting a good fit (see, for example, Izenman, 2008, p. 501).

[9] The interpretation of the MDS result should focus on the (1) axes, (2) scores, and (3) graph. A good rule of thumb for interpreting the axis is to look at the correlation between each score and the respective variables. If variables with a similar meaning are highly correlated with a score, then the latter formalizes the concept behind the variables. For example, the first score is highly correlated with HACK (−99.5%), while the second score is highly correlated with DISC (−94.2%). Hence, we can say that the first (second) score measures the relevance of HACK (DISC) in explaining the difference between entities.

[10] This measure is used in Edwards et al. (2015) as an alternative measure for data breach type, which aggregates the previous measure into malicious data breaches (MALB = CARD + HACK + INSD) and negligent data breaches (NEGB = DISC + PHYS + PORT + STAT).

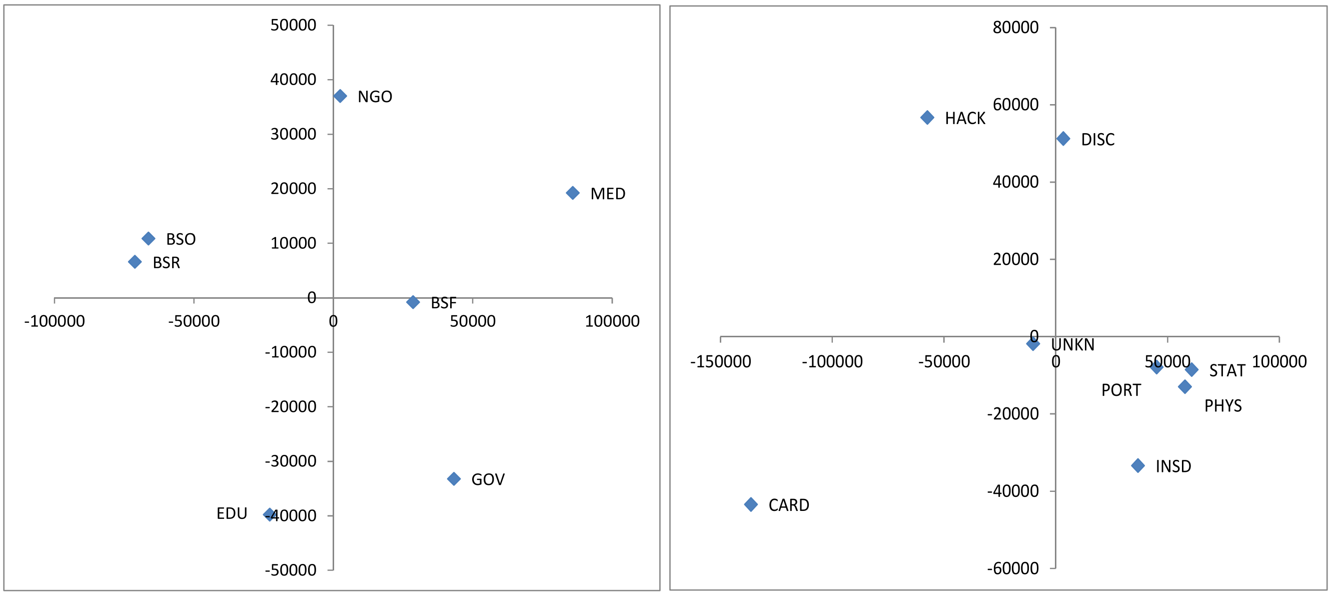


Fig. 3. MDS for frequency and different entities (left) and types of data breach (right). Notes: The horizontal axis represents the first score, and the vertical axis represents the second score. For illustration purposes, the scores from Table 2 are multiplied by 10,000. See Table 1 for variable definitions.

However, it cannot be further used for classifying from a statistical point of view. For example, CARD is very different from HACK and INSD. Moreover, DISC is very different from PHYS, PORT, and STAT. These differences are also meaningful, given for example that PHSY, PORT, and STAT are all physical events related to hardware, whereas DISC is typically electronic (e.g., unintended email).

In the second step, we analyze the severity for entities. We investigate the differences between entities with respect to the severity of data breaches originating from several sources. In the left part of Fig. 4, we observe that NGO, BSF, and BSO are very comparable in terms of severity, whereas MED, GOV, EDU, and BSR differ substantially in both the first and second score.

In the right part of Fig. 4, we again see that PORT, STAT, and PHYS are relatively close to each other, CARD is very different, and UNKN is very close to the origin. In this context, it is also worth mentioning that, although Figs. 3 and 4 look quite different, the correlation between the scores for frequency and severity is relatively high. This holds for the different entities (correlation of 83.08% for the two first scores for frequency and severity, 83.68% for the second score), but especially for the type of data breach (94.75% for the first score, 92.21% for the second score). Thus, for a particular type of attack, frequency and severity are highly correlated. For example, HACK and PORT are the two types with both the highest frequency and the highest severity.[11]

We also apply MFACT to the eleven contingency tables containing the frequencies of cyber problems recorded in the same year. The scatter plot in the left part of Fig. 5 shows that contingency tables from years 2005 to 2009 are close to each other and well separated from those of years 2010 to 2014, which are quite close to each other, too. The contingency table of year 2015 appears to be quite isolated, but still closer to the latter group than to the former one. In the right part of Fig. 5 we apply MFACT to eleven contingency tables containing the number of records lost in the same year, showing three well separated clusters (2009, 2012, 2015; 2008, 2011, 2013, 2014; and 2005, 2006, 2007, 2010). The first two clusters are nearly spherical and are less scattered than is the third one, which is more elongated. A comprehensive dynamic analysis of data breaches falls outside the scope of the present paper, but the cluster structures found with MFACT show that time matters and might merit further investigation in a future paper.

4.2. *Goodness of fit*

The results in Section 4.1 illustrate the different statistical nature of data breaches, when both the different types of attacks and the different entities are considered. These results suggest that the different entities and types of data breaches should not be put in one basket. For this reason, in the following goodness-of-fit tests, we model the different types of data breaches and entities as separate risk categories. We present the results for the biggest breach category, which is PORT, with 629 observations. The results for all other categories are available from the authors upon request.[12]

We first analyze the data breach frequency. Table 3 presents the results for the negative binomial and the Poisson distributions. As indicated by Edwards et al. (2015), the negative binomial distribution provides the best fit for the daily frequencies of data breaches. The analysis shows that this distribution cannot be rejected by the K–S test. This finding also holds for all other types of data breaches and entities.[13]

In Table 4, we analyze the goodness of fit for the severity distribution for different single parametric distributions, the transformation kernel, and the POT method. As indicated in Section 3, the data might be well described by a log-normal distribution. Regarding Panel A the log-normal distribution is the only distribution for which the null hypothesis of the K–S and A–D tests is not rejected. For all other distributions, the fit is not as good as

[11] We also applied correspondence analysis (CA) to the contingency table containing the frequencies of data breaches, displaying the same pattern as those in Fig. 3. The only noticeable difference is NGO, which is quite close to the origin in the MDS scatter plot and is far away in CA. When applying CA to the contingency table containing the number of records lost due to data breaches, we again see the same pattern as those seen in Fig. 4. The only noticeable difference between patterns in the two graphs are the relative positions of EDU, MED, and GOV. We conclude that, apart from few minor discrepancies, the results from CA and MDS are consistent. The results of CA are available upon request.

[12] For comparison purposes, we also calculated the results for the full dataset, which confirm the findings from previous studies (Edwards et al., 2015, and others).

[13] In additional tests, available upon request, we also consider the truncated negative binomial and the truncated Poisson distribution, which, however, do not materially improve the estimation results.

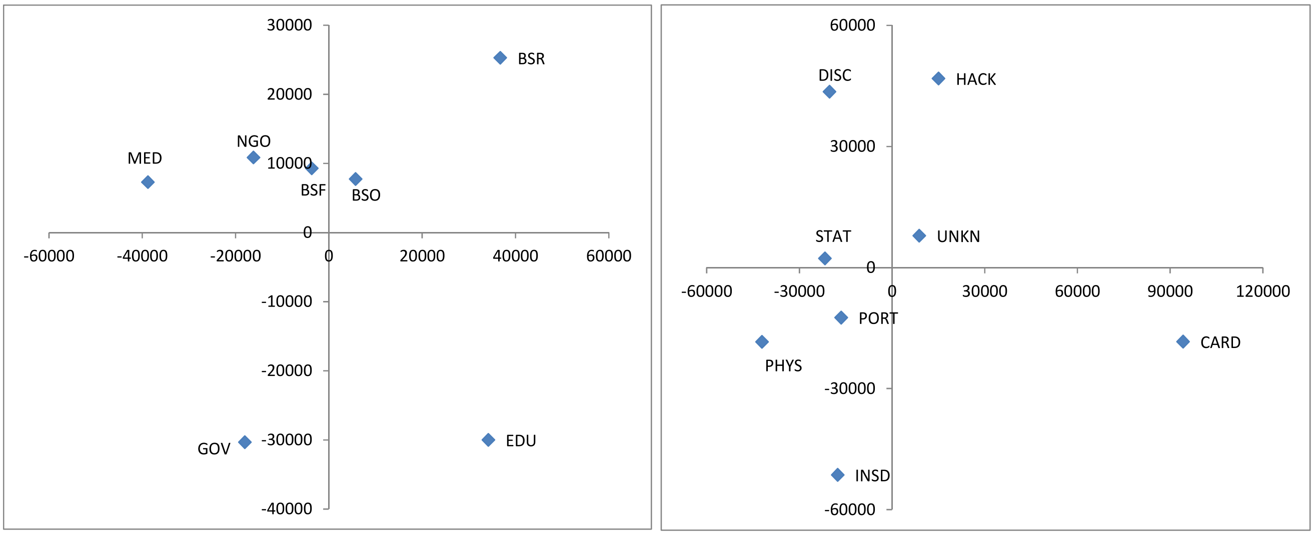


Fig. 4. MDS for severity and different entities (left) and types of data breach (right). Notes: The horizontal axis represents the first score, and the vertical axis represents the second score. For illustration purposes, the scores from Table 2 are multiplied by 10,000. See Table 1 for variable definitions.

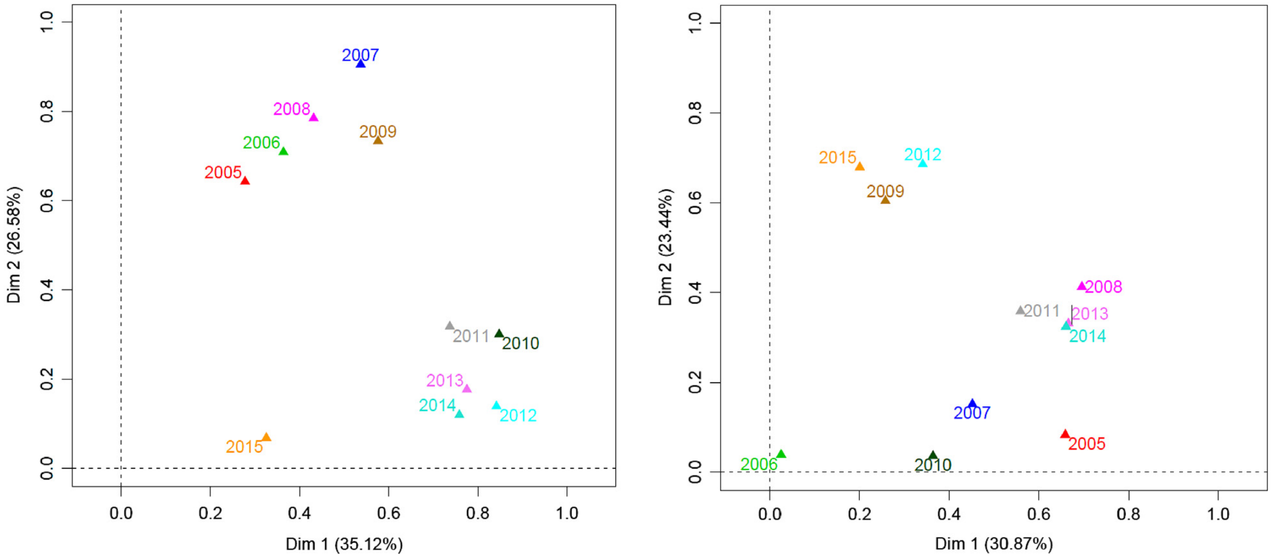


Fig. 5. MFACT for frequency (left) and severity (right). Notes: The horizontal (vertical) axis reports the scores of the first (second) maximum inertia dimension that MFACT assigns to each group. The number in brackets is the percentage of total inertia explained by the axis itself. The sum of the two percentages represents the reliability of the graph in summarizing total inertia.

Table 3
Goodness-of-fit analysis – frequency – PORT.

Model	Log-likelihood	AIC	Chi-square-test		K–S test	
Poisson	−4733.25	9468.50	>10,000.00	***	0.80	***
Negative binomial	−33.83	71.67	51.57	***	0.22	

Notes: AIC = Akaike information criterion; for the Chi-square and the Kolmogorov–Smirnov tests (for discrete distributions, see Arnold and Emerson, 2011), we present the value of the test statistic and the significance level of rejecting the null hypothesis (H_0: the given distribution is equal to the sample distribution).
* Indicate the 10% significance level.
** Indicate the 5% significance level.
*** Indicate the 1% significance level.

for the log-normal distribution. In addition, the results for the POT approach and the Transformation Kernel are very good, but they do not show a significant improvement in the distribution fit (see log-likelihood and AIC values). This might indicate that the distribution of data breach sizes is not as severely tailed as are other data from the area of operational risk (see e.g. Chavez-Demoulin et al., 2016). Regarding Panel B, for the log data the normal, skew normal and skew student distribution are not rejected by the K–S and A–D tests.[14] The POT only provides a relatively small improvement in

[14] Note that the skew-normal in log-transformed data is the log-skew-normal in the original data; similarly, the normal in log-transformed data is the log-normal in the original data.

Table 4
Goodness-of-fit analysis – severity – PORT.

Model	Log-likelihood	AIC	Kolmogorov–Smirnov test	Anderson–Darling test
Panel A: Original data				
Exponential	−8506.17	17,014.35	0.66***	1083.10***
Gamma	No convergence			
GPD	−6870.14	13,744.28	0.06**	4.12***
Log-logistic	−6853.29	13,710.58	1.00***	0.55
Log-normal	**−6848.40**	**13,700.81**	**0.03**	**0.27**
Normal	−10,325.83	20,655.66	0.47***	218.74***
Weibull	−6916.13	13,836.25	0.08***	9.51***
Skew-normal	−9893.24	19,792.49	0.86***	1532.70***
Skew-student	−9919.29	19,846.58	0.94***	1141.60***
POT (90%, log-normal)	−6945.30	13,900.60	/	/
Transformation Kernel	−6850.99	/	/	/
Panel B: Log-transformed data (LN)				
Exponential	−1841.34	3684.69	0.30***	103.75***
Gamma	−1679.30	3362.60	0.14***	22.55***
GPD	−1690.78	3385.55	0.24***	63.81***
Log-logistic	−1680.87	3365.75	0.99***	11.57***
Log-normal	−2244.12	4492.23	0.28***	110.02***
Normal	−1513.75	3031.50	0.03	0.28
Weibull	−1585.81	3175.63	0.08***	4.90***
Skew-normal	−**1512.75**	**3031.51**	**0.02**	**0.26**
Skew-student	−1512.75	3033.51	0.02	0.25
POT (90%, Weibull)	−**1511.71**	**3033.43**	/	/
Transformation kernel	−1513.42	/	/	/

Note:
* indicate the 10% significance level.
** Indicate the 5% significance level.
*** Indicate the 1% significance level.

goodness of fit, while the Transformation Kernel is very good again but does not further improve the fit.

In Table 5, we summarize the results of the goodness-of-fit tests across the 15 different categories, i.e. the seven different entities and eight different types of breach. Again we distinguish between the original data and the log data. Based on the results in Table 5, the log-normal and the skew-normal in particular turn out to be promising for severity modeling. The result for the log-normal is in line with Edwards et al. (2015). However, the skew-normal seems to do an even better job in describing the data; noticeably, the skew-normal is not rejected in any cases while the log-normal is rejected in one of the seven cases for the different entities and one of the eight cases for the different types of data breaches.[15] The result that the skew-normal is a promising model for the log data confirms the results from other actuarial analyses (e.g. Eling, 2012).

The question arising from Table 5 is whether the log-normal on the original data or, alternatively, the skew-normal distribution on the log data should be used for actuarial modeling.[16] Answering this question also requires considering whether the effort to estimate one additional parameter for the skew-normal distribution justifies the gain from a slightly better fit (the above goodness-of-fit tests already incorporate the different number of parameters to be estimated, e.g., in the AIC criterion). To further answer the question on whether the log-normal or the skew-normal is the preferred model, Table 6 shows the estimated parameters of the skew-normal distribution. The column "Normality" shows the p-values of the skewness-based tests for normality, thereby illustrating the importance of the additional parameters in the skew-normal distribution (if the p-value is larger than 0.1, then the normality is given).[17] The columns "Skewness" and "Kurtosis" contain the third and fourth sample standardized moments—namely, $\Sigma_i(z_i)^3/n$ and $\Sigma_i(z_i)^4/n$, respectively, where z_i is the ith standardized score.

The results from Panel A of Table 6 can be summarized as follows. First, some distributions of records' losses are normal (BSO, EDU, MED, and NGO), given that the p-value in the row normality is higher than 0.1. Second, the parameters do not change much from one entity to the other. Third, skewness is always positive, although only slightly so. Overall, we can, thus, conclude that the log-normal is the preferred model for BSO, EDU, MED, and NGO, whereas the skew-normal should be used for BSF, BSR, and GOV. For Panel B of Table, we can conclude that the log-normal is the preferred model for PORT, STAT, and UNKN, whereas the skew-normal should be used for CARD, DISC, HACK, INSD, and PHYS.

4.3. Applications

In Table 7 we use the model results to derive estimators for value at risk and compare them with the empirical data. Four confidence levels are considered (90%, 95%, 99%, and 99.5%), with the 99% and 99.5% confidence intervals being the most important applications, especially for risk-based capital modeling (e.g., Solvency II relies upon value at risk at the 99.5% confidence level). The table shows the estimated value (est.), the empirical value (emp.), and the p-value (p.-val.) for testing the difference between

[15] We mention the existence of the skew-log-normal distribution (Azzalini et al., 2003), which could be an attractive complement for the original data.

[16] The log-skew-normal is identical to the skew-normal applied to log data, which is why we use them synonymously in the following.

[17] Sample skewness is the test statistic of the locally most powerful (location and scale) invariant test (UMPI) for normality within the class on skew-normal distributions (Salvan, 1986). More formally, let $X \sim SN(\xi, \psi, \lambda)$ be our sampled distribution and assume we wish to test the null hypothesis $H_0 : \lambda = 0$ (X is normally distributed) against the alternative hypothesis $H_1 : \lambda > 0$ (X is positively skewed). Then, the test rejects H_0 if $g = \Sigma_i(z_i)^3/n$ is larger than a fixed value, where z_i is the ith standardized score. Under the null hypothesis, the asymptotic distribution of $n^{1/2}g$ is $N(0, 1/6)$.

Table 5
Summary of goodness-of-fit analysis for severity (KS and AD test).

	Original data	LN data
Best fit	Log-normal not rejected in – 6 out of 7 cases for the different entities – 7 out of 8 cases for the different types of data breaches	Skew-normal not rejected in – 7 out of 7 cases for the different entities – 8 out of 8 cases for the different types of data breaches
2nd best fit	GPD not rejected – 3 out of 7 cases for the different entities – 6 out of 8 cases for the different types of data breaches	Normal not rejected in – 6 out of 7 cases for the different entities – 7 out of 8 cases for the different types of data breaches

Table 6
Estimated parameters for the log-skew-normal distribution.

Entity	Location	Scale	Shape	Normality (p-value)	N	Mean	Standard deviation	Skewness	Kurtosis
Panel A: Entity type									
BSF	4.0734	5.3162	2.3769	0.0000	295	7.9833	3.6022	0.5488	2.7750
BSO	6.2779	3.4087	0.9726	0.4263	229	8.1742	2.8325	0.1288	2.2546
BSR	2.7606	5.6171	7.9938	0.0000	198	7.2077	3.4314	0.9343	3.4952
EDU	9.2780	2.7499	−0.8356	0.3909	538	7.8712	2.3627	0.0906	2.5087
GOV	5.3093	4.0418	1.7018	0.0018	440	8.0897	2.9335	0.3654	2.9837
MED	7.2663	2.8355	0.3255	0.9473	519	7.9666	2.7476	0.0071	2.8121
NGO	5.5129	3.4898	0.9914	0.7068	47	7.4733	2.8871	0.1344	2.0008
Panel B: Type of data breach									
CARD	2.6033	4.7580	20.222	0.0039	30	6.3996	2.8682	1.2919	4.8611
DISC	5.3153	3.4047	1.5024	0.0085	459	7.5767	2.5452	0.3011	2.9216
HACK	6.5600	3.9755	1.1599	0.0957	475	8.9623	3.1675	0.1873	3.1434
INSD	3.0708	4.5134	3.6400	0.0000	283	6.5433	2.8831	0.7499	3. 2049
PHYS	4.1479	3.1749	1.8310	0.0203	197	6.3711	2.2665	0.4051	3.1986
PORT	6.6179	3.2681	1.0214	0.4111	629	8.4812	2.6849	0.1434	3.0493
STAT	9.6411	2.9960	−0.6651	0.8078	135	8.3172	2.6877	−0.0513	2.8185
UNKN	4.7979	4.0953	1.7573	0.2341	58	7.6378	2.9506	0.3827	2.4714

Note: See Table 1 for variable definitions.

the estimated and theoretical values.[18] If the p-value is lower than 10%, then the values are not identical. We focus the presentation of the results on the skew-normal and the log-normal distribution. The cases where the p-value is lower than 10% are printed in bold.

As shown in Table 7, both the skew-normal and the log-normal distributions provide an excellent fit to the data. The estimated value from the log-normal distribution is not identical in 5 of the 64 cases presented in Table 7. For the skew-normal distribution, the estimator is not identical in even 2 of the 64 cases. Again, it seems that the skew-normal performs slightly better than does the log-normal distribution.

A second possible application is pricing of insurance contracts. Applying the results of this paper to pricing is especially interesting given the infinite model problem often observed with risk categories where data are sparse (see, e.g., Eling and Wirfs, 2016; Chavez-Demoulin et al., 2016).[19] To yield a price for a typical cyber insurance policy, we define two sample companies and predict the loss frequency and loss severity distribution.[20] Afterwards, we simulate (1) the actuarial fair premium (P = E[X]) and (2) three net premium principles (expectation principle, P = E[X] +$\delta\cdot$ E[X] and the standard deviation principle, P = E[X] + $\delta\cdot$ (Var[X])$^{1/2}$; and (3) the semi-variance principle, E[X] +$\delta\cdot$ E[(X − E[X])$^{+}$]^{2}). We refer to Embrechts (2000) for a more detailed summary of those actuarial pricing principles. The results are given in Table 8.

The log-skew-normal produces insurance prices that are much closer to the empirical distribution than the log-normal distribution does. This might be another argument why the log-skew-normal is a useful model for actuarial work. We also see that the results very much depend on the pricing principle. In particular, when the standard deviation is included in the estimation, premiums are extremely high. Furthermore, we observe that company #2 has much higher premiums than does company #1, which can be explained by the differences in industries and types of attacks.[21]

5. Conclusion

This paper has two main findings: The first is that different types of data breaches and different entities need to be analyzed as separate risk categories when modeling data breaches. The second main finding is that the skew-normal is a good distribution in describing the amount of the data breach. For the pricing of cyber risk insurance and the estimation of risk measures, understanding the properties and behaviors of the data breach information is of fundamental importance. Thus, the results of this paper might offer important insights for both risk management and insurability. The findings are relevant not only for insurance managers but also for

[18] Let q_k be the $100\cdot k$ percentile of the empirical distribution of the data. Also, let $F(\cdot)$ be the cumulative distribution function of the skew-normal model estimated from the same data. Hence $p_k = F(q_k)$ is the probability that the estimated model generates a value not greater than q_k. Small values of $|p_k - k|$ suggest good fit of the estimated distribution to the upper tail of the empirical distribution, and we formalize the concept as follows. Under the null hypothesis that data are skew-normal, and the statistic $z_k = n1/2(p_k-k)/[k\cdot(1-k)]1/2$ is approximately standard normal. We assess the closeness of corresponding percentiles of the empirical and fitted distribution by the p-value $1 - 2\Phi(|z_k|)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. See Bartoletti and Loperfido (2010) for a description of the test.

[19] Those distributions often have a shape parameter that is greater than one, indicating infinite higher moments and, by that, an infinite expected value, so that random sampling from the distribution is not meaningful.

[20] For company #1 we consider a university that faces a HACK attack ($N = 174$). For company #2 we consider a hospital and determine the premium for an incident from the PORT category ($N = 180$). To transfer the amount of breached data to a dollar loss, we apply the Jacobs (2014) transformation, i.e., the log-losses are described by ln(loss) = 7.68 + 0.76 · ln(records breached).

[21] Note that we assume no deductibles or cover limits in this application; in practice the insurance prices would thus be substantially lower. Also, the loss probability, which we assume to be 10%, is critical for the result and might be substantially lower in practice.

Table 7
Risk measurement results.

	Num	90 Est.	90 Emp.	90 *p*-val.	95 Est.	95 Emp.	95 *p*-val.	99 Est.	99 Emp.	99 *p*-val.	99.5 Est.	99.5 Emp.	99.5 *p*-val.
Panel A	Entities	Log-skew-normal											
BSF	295	13.05	12.82	0.96	14.93	15.06	0.32	18.61	16.70	0.48	19.99	17.87	0.49
BSO	229	11.84	11.89	0.78	12.95	12.88	0.95	15.10	14.07	0.26	15.90	14.85	0.34
BSR	198	12.02	11.72	0.66	13.81	13.71	0.95	17.31	17.51	0.77	18.62	17.85	0.81
EDU	538	10.87	10.88	0.97	11.69	11.53	0.41	13.22	12.68	0.06	13.78	13.43	0.48
GOV	440	11.97	11.93	0.98	13.27	13.06	0.83	15.80	15.65	0.94	16.75	16.65	0.76
MED	519	11.49	11.45	0.92	12.49	12.49	0.99	14.36	14.44	0.81	15.05	15.19	0.83
NGO	47	11.22	10.98	0.83	12.37	12.03	0.96	14.59	13.31	0.62	15.42	13.56	0.23
ALL	2266	11.75	11.56	0.14	13.00	12.71	**0.05**	15.45	15.43	0.92	16.37	16.46	0.59
Panel B	Entities	Log-normal											
BSF	295	12.60	12.82	0.53	13.91	15.06	**0.04**	16.36	16.70	0.45	17.26	17.87	0.51
BSO	229	11.80	11.89	0.72	12.83	12.88	0.87	14.76	14.07	0.49	15.47	14.85	0.64
BSR	198	11.61	11.72	0.75	12.85	13.71	0.14	15.19	17.51	0.21	16.05	17.85	0.40
EDU	538	10.90	10.88	0.89	11.76	11.53	0.26	13.37	12.68	**0.02**	13.96	13.43	0.23
GOV	440	11.85	11.93	0.69	12.92	13.06	0.46	14.91	15.65	0.28	15.65	16.65	0.27
MED	519	11.49	11.45	0.91	12.49	12.49	0.98	14.36	14.44	0.79	15.04	15.19	0.81
NGO	47	11.17	10.98	0.81	12.22	12.03	0.99	14.19	13.31	0.74	14.91	13.56	0.34
ALL	2266	11.65	11.56	0.44	12.71	12.71	0.99	14.69	15.43	**0.01**	15.42	16.46	**0.02**
Panel C	Types	Log-skew-normal											
CARD	30	10.30	10.25	0.24	11.68	11.07	0.46	14.38	14.51	0.63	15.39	15.14	0.77
DISC	459	10.92	10.82	0.73	12.01	11.77	0.58	14.14	14.07	0.97	14.94	15.06	0.71
HACK	475	13.07	12.61	**0.06**	14.32	14.04	0.48	16.73	17.50	0.37	17.63	18.07	0.61
INSD	283	10.68	10.50	0.95	12.25	11.89	0.95	15.33	14.17	0.94	16.48	15.54	0.84
PHYS	197	9.37	9.32	1.00	10.37	10.31	0.88	12.33	11.93	0.68	13.07	12.20	0.79
PORT	629	11.95	11.99	0.81	13.00	12.85	0.52	15.02	14.85	0.77	15.77	15.41	0.97
STAT	135	11.75	11.52	0.55	12.71	12.64	0.98	14.50	14.01	0.83	15.15	14.50	0.92
UNKN	58	11.75	11.34	0.93	13.26	12.82	0.76	16.21	14.51	0.67	17.32	14.62	0.27
Panel D	Types	Log-normal											
CARD	30	10.08	10.25	0.78	11.12	11.07	0.83	13.07	14.51	0.61	13.79	15.14	0.74
DISC	459	10.84	10.82	0.96	11.76	11.77	0.80	13.50	14.07	0.30	14.13	15.06	0.28
HACK	475	13.02	12.61	**0.07**	14.17	14.04	0.83	16.33	17.50	0.16	17.12	18.07	0.31
INSD	283	10.24	10.50	0.38	11.29	11.89	0.14	13.25	14.17	0.21	13.97	15.54	0.29
PHYS	197	9.28	9.32	0.77	10.10	10.31	0.59	11.64	11.93	0.60	12.21	12.20	0.43
PORT	629	11.92	11.99	0.67	12.90	12.85	0.84	14.73	14.85	0.68	15.40	15.41	0.56
STAT	135	11.76	11.52	0.50	12.74	12.64	0.89	14.57	14.01	0.71	15.24	14.50	0.81
UNKN	58	11.42	11.34	0.99	12.49	12.82	0.53	14.50	14.51	0.93	15.24	14.62	0.70

Note: See Table 1 for variable definitions.

Table 8
Pricing example (in US$).

	Actuarial fair premium	Expectation principle	Standard deviation principle	Semi-variance principle
	Log-normal			
Company #1	813,254.00	894,689.50	1,780,723.00	887,713.40
Company #2	1,345,070.00	1,479,577.00	3,972,762.00	1,469,653.00
	Log-skew-normal			
Company #1	616,872.70	678,560.00	1,015,452.00	673,092.40
Company #2	1,063,593.00	1,169,952.00	2,422,789.00	1,161,660.00
	Empirical			
Company #1	605,974.50	666,571.90	964,240.40	661,122.10
Company #2	991,353.80	1,090,489.00	1,763,130.00	1,082,535.00

Notes: The prices in this table are based on one million simulated losses for the appropriate distributional parameters of company #1 and #2. We assume a loss probability of 10%. For the pricing parameter, we follow Mukhopadhyay et al. (2006) and assume $\delta = 0.1$.

policymakers and regulators who need to develop sound policies for the treatment of this new, dynamic risk category.

There are manifold limitations of this research that might provide opportunities for future research. For example, the analysis focuses only on the number of data breaches and the amount of lost data, but not on real loss data. The approximation of Jacobs (2014) to transfer the number of records breached into actual loss data can help carry out the first applications, but this is, clearly, only a crude and rough approximation. For this reason, the analysis presented here should not be interpreted as more than a first step toward a more thorough analysis of data breaches and cyber risk. However, international organizations and reinsurance firms are now starting to set up cyber loss databases so that data availability will improve in the near future. It will then be interesting to apply our techniques and compare with our results when these data pools reach a sufficient size. Also, our analyses are static in nature, although the results for MFACT illustrate that time matters; the identified cluster structure might be explored in more detail by future research.

Acknowledgments

The authors are grateful to Jan H. Wirfs, Kwangmin Jung, Philipp Schaper, Werner Schnell and two anonymous referees for their helpful suggestions. Special thanks to Jan H. Wirfs for valuable research assistance.

References

Adcock, C., Eling, M., Loperfido, N., 2015. Skewed distributions in finance and actuarial science: a review. Eur. J. Finance 21 (13–14), 1253–1281.

Alemany, R., Bolancé, C., Guillén, M., 2013. A nonparametric approach to calculating value-at-risk. Insurance Math. Econom. 52, 255–262.

Arnold, T.B., Emerson, J.W., 2011. Nonparametric goodness-of-fit tests for discrete null distributions. The R Journal 3 (2), 34–39.

Azzalini, A., 2013. The Skew-Normal and Related Families. In: Institute of Mathematical Statistics Monographs, Book, vol. 3, Cambridge University Press.

Azzalini, A., Dal Cappello, T., Kotz, S., 2003. Log-skew-normal and log-skew-t distributions as model for family income data. J. Income Distrib. 11, 12–20.

Bartoletti, S., Loperfido, N., 2010. Modelling air pollution data by the skew-normal distribution. Stoch. Environ. Res. Risk Assess. 24 (4), 513–517.

Bécue-Bertaut, M., 2014. Tracking verbal-based methods beyond conventional descriptive analysis in food science bibliography. Stat. Approach. Food Qual. Pref. 32, 2–15.

Bécue-Bertaut, M., Pagés, J., 2004. A principal axes method for comparing contingency tables: MFACT. Comput. Statist. Data Anal. 45, 481–503.

Bécue-Bertaut, M., Pagés, J., 2008. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. Comput. Statist. Data Anal. 52, 3255– 3268.

Biener, C., Eling, M., Wirfs, J.H., 2015. Insurability of cyber risk–an empirical analysis. Geneva Pap. Risk Insur. 40 (1), 131–158.

Bolancé, C., Guillen, M., Nielsen, J.P., 2003. Kernel density estimation of actuarial loss functions. Insurance Math. Econom. 32, 19–36.

Bolancé, C., Guillen, M., Pelican, E., Vernic, R., 2008. Skewed bivariate models and nonparameteric estimation for the CTE risk measure. Insurance Math. Econom. 43, 386–393.

Brechmann, E.C., Hendrich, K., Czado, C., 2013. Conditional copula simulation for systemic risk stress testing. Insurance Math. Econom. 53, 722–732.

Cebula, J.J., Young, L.R., 2010. A Taxonomy of Operational Cyber Security Risks. Technical Note CMU/SEI-2010-TN-028, Software Engineering Institute. Carnegie Mellon University.

Chapelle, A., Crama, Y., Huebner, G., Peters, J.-P., 2008. Practical methods for measuring and managing operational risk in the financial sector: a clinical study. J. Bank. Finance 32 (6), 1049–1061.

Chavez-Demoulin, V., Embrechts, P., Hofert, M., 2016. An extreme value approach for modeling operational risk losses depending on covariates. J. Risk Insurance 83 (3), 735–776.

Edwards, B., Hofmeyr, S., Forrest, S., 2015. Hype and Heavy Tails: A Closer Look at Data Breaches. Working Paper, http://www.econinfosec.org/archive/weis2015/papers/WEIS_2015_edwards.pdf, (accessed 08.02.16).

Eling, M., 2012. Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? Insurance Math. Econom. 51, 239–248.

Eling, M., Wirfs, J.H., 2016. Cyber Risk Is Different. University of St. Gallen, Working Paper.

Embrechts, P., 2000. Actuarial versus financial pricing of insurance. J. Risk Finance 1 (4), 17–26.

Husson, F., Josse, J., Lê, S., Mazet, J., 2007. FactoMineR: Factor analysis and data mining with R. R package version 1.19. http://www.cran.R-project.org/package=FactoMineR.

Izenman, A.J., 2008. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer.

Jacobs, J., 2014. Analyzing Ponemon Cost of Data Breach. http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/, (accessed 20.06.16).

Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M., 2009. Modern Actuarial Risk Theory. Springer.

Kostov, B., Bécue-Bertaut, M., Husson, F., 2013. Multiple factor analysis for contingency tables in the FactoMineR package. R J. 5, 29–38.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 115–129.

Mack, T., 2002. Schadenversicherungsmathematik. VVW-Verlag.

Maillart, T., Sornette, D., 2010. Heavy-tailed distribution of cyber-risks. Eur. Phys. J. B 75 (3), 357–364.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press.

2014. Net Losses–Estimating the Global Cost of Cybercrime. http://www.mcafee.com/mx/resources/reports/rp-economic-impact-cybercrime2.pdf, (accessed 16.03.15).

Miljkovic, T., Grün, B., 2016. Modeling loss data using mixtures of distributions. Insurance Math. Econom. 70, 387–396.

Moscadelli, M., 2004. The modelling of operational risk: Experience with the analysis of the data collected by the Basel Committee, Technical Report 517, Banca d'Italia.

Mukhopadhyay, A., Chatterjee, S., Saha, D., Mahanti, A., Sadhukhan, S.K., 2006. E-Risk management with insurance: A framework using copula aided Bayesian belief networks. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences, HICSS'06, vol. 6, pp.1-6.

National Conference of State Legislatures 2016. Security Breach Notifications Laws. http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx.

Panjer, H.H., 2007. Operational Risk. John Wiley & Sons.

Privacy Rights Clearinghouse 2016. About the Privacy Rights Clearinghouse. https://www.privacyrights.org/node/1398, (accessed 08.02.16).

Salvan, A., 1986. Test localmente più potenti tra gli invarianti per la verifica dell'ipotesi di normalità. In: Atti Della XXXIII Riunione Scientifica Della Società Italiana Di Statistica, Volume II. Bari. Società Italiana di Statistica, Cacucci, pp. 173–179.

Tukey, J.W., 1977. Exploratory Data Analysis. Pearson.

Vernic, R., 2006. Multivariate skew-normal distributions with applications in insurance. Insurance Math. Econom. 38, 413–426.

Villaseñor-Alva, J.A., González-Estrada, E., 2009. A bootstrap goodness of fit test for the generalized Pareto distribution. Comput. Statist. Data Anal. 53 (11), 3835–3841.

Wheatley, S., Maillart, T., Sornette, D., 2016. The extreme risk of personal data breaches and the erosion of privacy. Eur. Phys. J. B 89 (7), 1–12.