Insurance: Mathematics and Economics 82 (2018) 167–180

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

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

Copula approaches for modeling cross-sectional dependence of data breach losses

Martin Eling *, Kwangmin Jung

Institute of Insurance Economics, University of St. Gallen, Girtannerstrasse 6, 9010 St. Gallen, Switzerland

ARTICLE INFO

Article history:
Received December 2017
Received in revised form June 2018
Accepted 11 July 2018
Available online 26 July 2018

JEL classification:
C13
C15
C46
G22

Keywords:
Cyber risk
Data breach
Zero-inflated data
Pair copula construction
Vine copula
Risk measurement
Insurance pricing
Diversification effect

ABSTRACT

Many experts claim that cyber risks are correlated, but there is not much supporting empirical evidence. We consider 3327 data breach events from 2005 to 2016 and identify a significant asymmetric dependence of monthly losses in two cross-sectional settings: cross-industry losses in four categories by breach types (hacking, lost electronic device, unintended disclosure and insider breach) and cross-breach type losses in five categories by industries (banking and insurance, government, medical service, retail/other business and educational institution). To identify the method that best fits the dependence structure of the dataset, we implement copula modeling by separating the dependence into pairwise non-zero losses and zero loss arrivals. We model the former by pair copula construction (PCC) allowing for the flexible choice of copula functions, whereas the latter is modeled by Gaussian copula. We illustrate the usefulness of our results in two applications to risk measurement and pricing. Our findings are important for risk managers and actuaries who are designing cyber-insurance policies.

1. Introduction

Cyber risks are operational risks to information and technology assets that have consequences for confidentiality, availability, and integrity of information and information systems (Cebula and Young, 2010). Every day, media illustrate the growing economic and social importance of cyber risk (see, e.g., World Economic Forum, 2016). In addition, more businesses than ever are facing cyber risks and incurring considerable corporate losses (Allianz, 2015). Although numerous papers have researched cyber risks, there is a lack of understanding of how to model potential losses from cyber risks and how to price cyber-insurance (Böhme and Schwartz, 2010; Eling and Wirfs, 2018). One aspect often discussed in this context is what the dependence structure between cyber losses might look like. Many experts believe that cyber losses are correlated (see, e.g., Böhme and Schwartz, 2010; Ogut et al., 2011), for example, all companies are using the same software.

A few papers discuss the correlation of cyber risk, for example on information systems or infected computers (Böhme and Kataria, 2006; Herath and Herath, 2011; Mukhopadhyay et al., 2013; Shah, 2016; Xu and Hua, 2017; Peng et al., 2018)[1]; but there is still a need for a comprehensive empirical study that analyzes the dependence between actual cyber losses, and if such a dependence exists, what it looks like. The reason for the dearth of empirical research on this field might be a lack of data. The availability of data is, however, improving over time, especially with the emergence of the first database on data breaches.[2]

This paper aims at identifying the dependence structure between different cyber losses. For insurers, ex-post analysis on cyber

* Corresponding author.
E-mail addresses: martin.eling@unisg.ch (M. Eling), kwangmin.jung@unisg.ch (K. Jung).

https://doi.org/10.1016/j.insmatheco.2018.07.003
0167-6687/

[1] Online-Appendix A provides an overview of the literature and outlines the contribution of our paper. Only Böhme and Kataria (2006) consider a broader dataset (the number of potential attacks measured by honeypots), but they do not consider loss data and focus on the t-copula to capture potential tail dependencies. The other four papers rely either on simulations or on smaller datasets. Note that our focus is not on the global interconnection of different IT systems or computer networks, but on identifying the potential dependence structure of actual cyber losses, which is important for cyber risk management or to manage cyber-insurance portfolios.

[2] Since 2002, companies in many U.S. states have been legally required to report data breaches to their customers (NCSL, 2016) and with this data breach report databases are becoming increasingly available. Starting in 2018, companies in the European Union will be required to report data breach events (European Union, 2016); this will improve the availability of data. Private Rights Clearinghouse, a nonprofit organization in the U.S. is a good example of a database that has grown since 2005 (Privacy Rights Clearinghouse, 2016).

Table 1
Descriptive statistics of monthly loss data.

Panel A: Loss severity							***(in the number of breach records)***		
Variable	N	N of zeros	Mean	Std. dev	Skewness	Kurtosis	Min	Median	Max
HACK	144	4	5,378,936	24,606,777	9.141	95.518	0	52,316.7	270,131,250
ELET	144	12	185,391	868,203	9.378	97.310	0	24,300.3	9,550,998
DISC	144	12	353,149	2,736,163	10.925	125.022	0	7,336.0	31,835,867
INSD	144	32	305,371	2,457,333	10.643	118.941	0	1,700.0	28,200,000
BSF	144	51	1,990,783	10,801,531	6.548	42.735	0	1,255.5	80,000,000
GOV	144	22	640,856	3,104,316	6.470	44.048	0	14,330.2	25,333,655
MED	144	12	97,605	422,936	8.903	87.057	0	13,229.0	4,500,000
BSE	144	10	16,882,242	95,715,936	8.680	82.987	0	44,039.8	1,000,000,000
EDU	144	21	23,926	37,239	2.927	10.598	0	8,710.0	227,539
Panel B: Loss frequency									
HACK	144	4	5.764	4.591	1.862	4.968	0	4	27
ELET	144	12	7.160	5.752	1.244	1.913	0	6	30
DISC	144	12	4.757	3.343	0.654	−0.080	0	4	14
INSD	144	32	2.319	2.454	2.341	9.145	0	2	17
BSF	144	51	2.007	2.337	1.545	2.882	0	1	12
GOV	144	22	3.167	2.715	1.104	1.281	0	3	14
MED	144	12	7.125	6.671	1.350	1.581	0	5	31
BSE	144	10	3.757	3.299	2.019	5.667	0	3	18
EDU	144	21	3.944	3.148	0.848	0.749	0	3	15

Note (Privacy Rights Clearinghouse, 2016):
Breach type: HACK = Hacked by outside party or infected by malware; ELET = Lost, discarded or stolen portable devices or Stationary computer loss; DISC = Unintended disclosure, e.g., sensitive information posed in public or mishandled or sent to the wrong party; INSD = Insider breach by employee, contractor or customer
Entity(Industry) type: BSF = Financial and insurance services; GOV = Government and military; MED = Healthcare, medical providers and medical insurance services; BSE = Retail/Merchant and other business parties; EDU = Educational institutions.

losses is important because estimating the size of risk in a cyber-insurance risk pool is a key task in asset–liability management. The dependence structure in a cyber-insurance risk pool can provide diversification benefits; thus our modeling helps to correctly identify premiums and capital requirements. We construct a high-dimensional dependence model of cyber losses using different copula methods. For this purpose, we consider 3327 data breach events from 2005 to 2016 and apply the actuarial toolbox to identify the dependence structure between monthly loss events, in terms of both frequency and severity. We are interested in finding the dependence structure that most accurately describes the data, whether that structure is linear or non-linear. Since monthly losses include several zero values indicating no loss event in a certain month, we split the dependence structure into two parts: pairwise non-zero losses and zero loss arrivals. We then fit frequency and severity distributions using different parametric distributions and compound them by convolution. With the results of dependence modeling, we finally analyze the implications of the models in two applications to risk measurement and insurance pricing by aggregating cyber losses from different risk factors.

This paper contributes to cyber risk research in that we take two categorizations of cyber losses into account in an integrated structure and estimate a more accurate dependence structure of cyber losses in the risk pool. The two cross-sectional categorizations we consider are breach type (hacking, lost electronic device, unintended disclosure and insider attack) and industry (banking and insurance, governmental entity, medical service, retail/merchant and other business and educational institution); we call the former *cross-industry* structure and the latter *cross-breach type* structure.[3] Upon this cross-sectional setting, an up-to-date copula method, the pair copula construction (Aas et al., 2009), is used to build an empirical model for high-dimensional cyber risks. As a result, we find significant asymmetric tail dependence, providing evidence for non-linear dependence between different types of cyber risks. Our results are important for practitioners and regulators working on cyber risk management and for insurance underwriters working on the establishment of cyber-insurance policies.[4] The paper will motivate more academic research by outlining future research questions on the topic of cyber risk.

The rest of the paper is structured as follows. In Section 2, we describe the theoretical background on the high-dimensional copula method and the methodology of pair copula construction. Then in Section 3 the data is given. The results of the dependence modeling and applications to pricing and risk measurement are presented in Sections 4 and 5 respectively. Finally, the conclusion and possibilities for future research are shown in Section 6.

2. Theoretical background and methodology

2.1. Research background on cyber loss process

The loss process for cyber risk can be regarded as an operational loss process (Cebula and Young, 2010; Biener et al., 2015). Several methods have been developed to estimate an operational loss process. The loss distribution approach (LDA) has been widely used; it models the frequency and severity of operational risk losses separately (Panjer, 2006 Chapter 1.3). LDA is also frequently used to model underwriting claims in the collective risk model. It is based on the distributional fitting procedure for frequency and severity; the fitted distributions are then compounded by convolution (Wang, 1998; Frachot et al., 2001; McNeil et al., 2005, Chapter 10; Panjer, 2006, Chapter 6). A compound loss process

[3] Cross-industry setting consists of variables categorized by breach types, in which losses occurred across industry level, whereas cross-breach type setting contains variables categorized by industry, in which losses occurred across breach type level. This categorization is important because underwriting a cyber-insurance policy differentiates the sources of risk and industries, which have different risk exposures (Romanosky et al., 2017). The detail on the categorization is shown in the caption of Table 1.

[4] What risk managers are mainly concerned about is the likelihood of the tail risk from different risk factors (McNeil et al., 2005, p. 18). For an insurance company, loss aggregation is typically applied to the reserving process to measure the possible total loss amount from different lines of business (Kaas et al., 2008, Chapter 3). Similarly, each cyber risk factor can form an individual line of cyber insurance business with customized policies depending on a specific risk type or industry (Allianz, 2015; Eling and Wirfs, 2016). Thus, there is a need to analyze cyber risk in context of an insurance pool.

estimated for each risk factor j is a full predictive distribution to account for parameter uncertainty and can be described as:

$$\lambda_t^{(j)} = \sum_{i=1}^{N_t^{(j)}} X_{i,t}^{(j)}, \tag{1}$$

where $t = 1, 2, \ldots$ is discrete time in the monthly unit, $j = 1, \ldots, d$ is a breach type ($d = 4$) or an industry ($d = 5$), $N_t^{(j)}$ is the monthly count (frequency) process, $X_{i,t}^{(j)}$ is the monthly severity process and $\lambda_t^{(j)}$ is the monthly compound process. We assume that the count process and the severity process are independent (Wang, 1998; Shevchenko, 2010). In addition, we assume that the claim severity process, $X_{i,t}^{(j)}$, is independent and identically distributed (i.i.d.), hence we do not assume temporal dependency.[5] We also postulate that the aggregate claims process, $\lambda_t^{(j)}$, is a Markov process so that the development of claims at a certain time point does not rely on the development of the aggregate claims up to that time point (Bühlmann, 2007, p. 55).

As an example, a compound Poisson process with Poisson distribution for the frequency and lognormal or other continuous distribution for the severity is frequently used in operational risk modeling (Panjer, 2006, Chapter 5). Once a compound process has been estimated for each risk factor, we apply a dependence model for different risk factors, which we assume constitute the risk pool of a cyber-insurance provider. One challenge of the empirical study in this paper is that data breach risks on a monthly basis contain a number of zero values, which indicates that no loss occurred in a certain month (see column 3 in Table 1). This zero-inflation could generate a misspecification of a dependence model due to discontinuous probability function (Erhardt and Czado, 2012).[6] For this reason, following Erhardt and Czado (2012) and Brechmann et al. (2014), we model the parametric dependence structure in two separate approaches: dependence in positive loss pairs and dependence in zero value arrival. Hereafter, dependence in positive loss pairs is denoted by non-zero pair dependence and dependence in the zero value arrival is denoted by a zero loss dependence structure. The non-zero pair dependence is built upon Eq. (1), where monthly loss severity without zero loss is modeled by a parametric continuous distribution and monthly loss frequency is modeled by a parametric discrete distribution in Section 4.1. The zero loss dependence structure is based on a multivariate binary distribution, each margin of which gives 1 to zero value and 0 to non-zero value (Brechmann et al., 2014)[7]:

$$v_k := \begin{cases} 1 & \textit{zero loss} \\ 0 & \textit{otherwise}, \end{cases} \tag{2}$$

where v_k is a binary random variable.

When $v_k(\lambda) = 0$, the random variable has a positive loss and we can derive the cumulative distribution function of v_k, P_{v_k}, with the probability of a positive loss:

$$P_{v_k} := \begin{cases} p_{v_k}(0) & \textit{Zero loss} \\ 1 & \textit{otherwise}, \end{cases} \tag{3}$$

where p_{v_k} is a probability mass function of v_k and $p_{v_k}(0)$ is the probability of a positive loss. The methodology for dependence modeling in the two approaches is described in Section 2.3.

2.2. Theoretical background on copula modeling

The copula method is an effective and tractable way to identify complex, non-linear dependences inherent in multivariate distributions.[8] Copula functions can be classified into different classes. Elliptical copulas are the copula functions of elliptical distributions (e.g., Gaussian, student-t); thus if a bivariate copula function belongs to the elliptical class, margins will in general belong to the elliptical distribution (Embrechts et al., 2001). However, elliptical copulas are limited to symmetrical distributions and dependency (Embrechts et al., 2001). As shown in finance and insurance research, there might be a strong asymmetric dependence and tail dependence, for instance, between stock returns or insurance losses that cannot be captured by a symmetric and linear dependence measure.[9] As an alternative, Archimedean copulas incorporate different asymmetric tail dependence structures. However, since they explain the dependence structure by a single parameter only (via the generating function), using simple Archimedean copula to analyze a multivariate dependence structure is restricted in a multivariate case (Embrechts et al., 2001).[10]

To resolve these problems of multivariate dependence modeling, several advanced techniques have been developed (Aas and Berg, 2009). Among them, the pair copula construction (PCC) method, which is also called the vine copula model, is used to reduce the dimension by pairing the variable set (Bedford and Cooke, 2001). In this sense, the high-dimensional copula analysis can be transformed to a bivariate analysis to be more tractable. PCC is flexible in that any type of copula class can be applied to the construction and there is no mathematical complexity when one uses different copula functions in the modeling (Aas and Berg, 2009). These advantages result from the fact that a

[5] A criticism of this assumption is that there might be temporal dependency in claim amounts (see Araichi et al., 2016, for a general discussion in the context of auto insurance). Indeed, in our cyber context, a malicious attack by hacking (HACK) might trigger temporal dependency in losses on an hourly or perhaps a daily basis (e.g. Wannacry attack in 2017), but not on a monthly scale which is of interest for our modeling. Furthermore, it is not reasonable to assume temporal dependency for the other three risk sources considered in the paper (ELET, DISC and INSD), whose losses typically are firm-specific and occur independently over time; we could imagine temporal dependency within a firm, but not in the aggregate variables considered in this paper. We thus believe that for all four risks considered in this paper, the temporal structure of monthly data can be reasonably assumed away. Also empirically we do not observe non-stationarity and any serial dependence for our variables of interest (e.g. using the augmented Dickey–Fuller test; see Online-Appendix C).

[6] We test the dependence of original zero-inflated data by elliptical copulas, Archimedean copulas, joint Archimedean copulas and rotated Archimedean copulas (90°, 180° and 270°); however, we find that any type of parametric copula function does not fit the zero-inflated dataset.

[7] The mathematical description of a zero loss dependence structure is given in detail in Online-Appendix D.

[8] Copula modeling is widely used to examine dependence structures and helps to identify non-linear relations between different marginal distributions. The cyber risk literature that employs the copula method still has limitations. For example, Mukhopadhyay et al. (2013) make a normality assumption on each Bayesian network node, thereby using Gaussian copula to integrate all nodes. Furthermore, a simple copula method to identify the dependence of a high-dimensional structure is theoretically restricted due to lack of explanatory power (Embrechts and Hofert, 2013).

[9] For instance, multivariate asset returns or derivatives are not appropriately described by linear correlation measures (Chiou and Tsay, 2008). It also has been shown that the data breach information that we look at in this paper is non-normal and heavy tailed (Edwards et al., 2016) such that linear dependence might not fully illustrate the dependence structure of the data breach risk.

[10] Additionally, the single parameter for d-dimensional dependency can induce the permutation-symmetric property in multi-dimensional arguments, thereby resulting in exchangeability of margins (Savu and Trede, 2010). The exchangeability can be described as (in the three-dimensional case):

$$C(u_1, u_2, u_3) = C(u_1, u_3, u_2) = C(u_2, u_1, u_3) = \cdots$$

This property is called permutation-symmetric and the copula distribution is indifferent in d-exchangeable marginal variables; this exchangeability can become problematic when some variables come from the same sector and some from different sectors (Savu and Trede, 2010). Elliptical copulas also correspond to the exchangeability property in the bivariate case, but in a multi-dimensional case, it depends on the variance–covariance matrix of the marginal elliptical distributions (Harder, 2016).

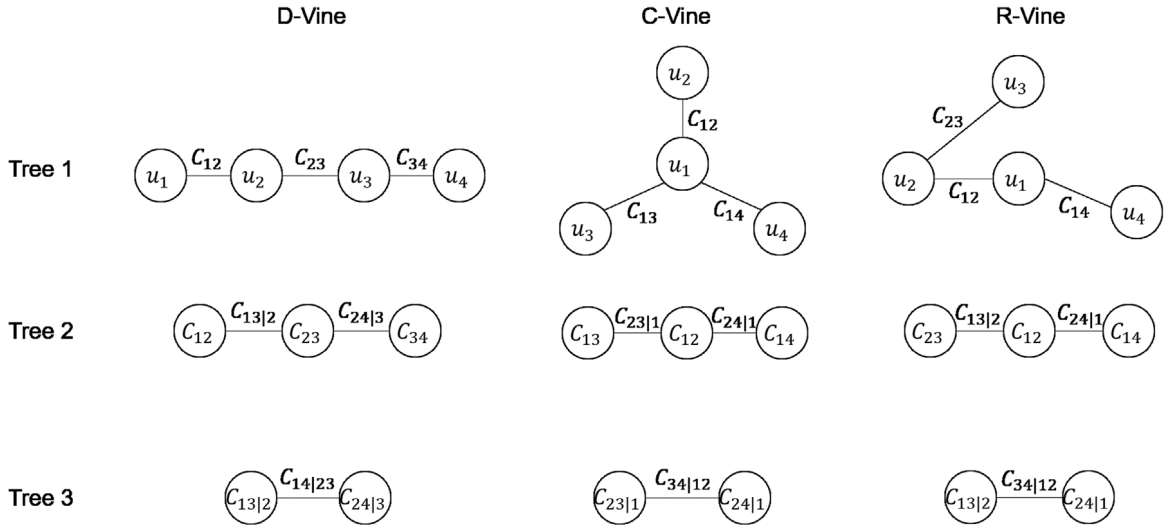

Fig. 1. The pair copula structure (four-dimensional case). The D-Vine is a structure showing the dependency in a row and forming a hierarchical tree, hence the variables are ordered by dependency. The C-Vine is a star-like structure, where a core variable placed in the center connects all other variables. The R-Vine flexibly links the variables by dependency without fixing a certain structure, thus the R-Vine can also project the D-Vine and the C-Vine structures (see Online-Appendix B for more details)

high-dimensional density function can be factorized by marginal density functions and conditional density functions.[11]

PCC is a tree-based model, where one builds the first tree with given random variables and continues to construct another tree with conditional variables estimated from the previous tree. The conditional variables are generated by copula densities as moving forward to the next trees (Aas and Berg, 2009). Three types of PCC have been developed (Kurowicka and Cooke, 2004; Czado, 2010): the drawable vine (D-Vine), the canonical vine (C-Vine) and the regular vine (R-Vine). The D-Vine is a hierarchical structure, the C-Vine is a dependence structure centered by a core risk factor and the R-Vine allows for more flexibility to structure dependency than the D-Vine and the C-Vine do. The D-Vine and the C-Vine can be represented by the R-Vine structure in accordance with the dependency in the first tree, thus we focus on the R-Vine structure in our empirical modeling. A four-dimensional case for three types is illustrated in Fig. 1; the mathematical definitions of the vine models are given in Online-Appendix B.

In Fig. 1, the marginal distribution in the first tree is transformed to the uniform distribution ($u_i \in [0,1], i = 1,\ldots,d$) and the estimated copula functions in the following trees are also accordingly transformed to the uniform conditional distributions. Due to high dimensional optimization for the parameter estimation, a two-step approach consisting of marginal estimation and copula estimation is customary (Czado et al., 2013). Joe and Xu (1996) introduce this two-step approach for the marginal estimation and the copula estimation using maximum likelihood method, known as the *inference function for margins (IFM)*.[12] The uniform distribution in our empirical study is estimated non-parametrically using the ranks of the observations, the distribution that is called *pseudo-observations* (Aas and Berg, 2009). The pseudo-observations are used for bivariate conditional copula functions in the sequential estimation (described in the next section) and the optimization of pair-wise likelihood function with pseudo-observations is conducted by maximizing the pseudo-likelihood (Aas et al., 2009).[13]

2.3. Methodology

In non-zero pair dependence modeling, we apply the estimated compound process for cyber risk into different dependence models. There have been several studies and textbooks on modeling dependence in the operational risk context with copula based on compound distribution method (e.g., Wang, 1998; Frachot et al., 2001; Panjer, 2006, Chapter 8; Embrechts and Puccetti, 2008; Giacometti et al., 2008; Shevchenko, 2010). We estimate the R-Vine, Gaussian, Student-t, Gumbel and Clayton and determine the best fit structure for the cyber loss processes. We consider different copula functions in the R-Vine model: independence, normal, student-t, Clayton, Gumbel, Frank, Joe, survival Archimedean copulas and rotated Archimedean copulas (90° and 270°). We take rotated Archimedean copulas into consideration to model negatively dependent variables (Dissmann et al., 2013). Gaussian and student-t copula models are used in many fields because of their tractability; the Gumbel and Clayton copulas are frequently used in dependence modeling due to the presence of asymmetric tail dependence (Genest and Rivest, 1993; Demanta and McNeil, 2005; Rosenberg and Schuermann, 2006).

[11] Aas et al. (2009) develop this methodology in the inferential way by decomposing a multivariate distribution into bivariate unconditional and conditional distributions based on the mathematical proofs by Joe (1996).

[12] A simple case of IFM can be described in the following. According to Sklar's theorem (Sklar, 1959), we can describe the d-dimensional probability function for the random vector $\mathbf{X}$ as:

$$F(\mathbf{X};\tau_1,\ldots,\tau_d,\boldsymbol{\theta}) = C(F_1(x_1;\tau_1),\ldots,F_d(x_d;\tau_d);\boldsymbol{\theta}),$$

where $\tau_i, i = 1,\ldots,d$ is a parameter of a marginal function F_i, $\boldsymbol{\theta}$ is a set of dependence parameters by the copula function C. In our case, the random vector X is continuous and we can derive the joint density function of the random variables as:

$$f(\mathbf{X};\tau_1,\ldots,\tau_d,\boldsymbol{\theta}) = c(F_1(x_1;\tau_1),\ldots,F_d(x_d;\tau_d);\boldsymbol{\theta})\prod_{j=1}^{d} f_j(x_j;\tau_j),$$

where $c(\cdot)$ is a copula density function. Then, the log-likelihood function for the joint density function is:

$$L(\boldsymbol{\theta},\tau_1,\ldots,\tau_d) = \sum_{i=1}^{d}\log f(x_i;\tau_1,\ldots,\tau_d,\boldsymbol{\theta}).$$

The parameter estimation by IFM is comprised of separate optimizations for univariate margins and the optimization of the d-dimensional log-likelihood with the dependence parameter. The derivation of the parameters for vine models using IFM has been studied in Haff (2013) and Czado et al. (2013).

[13] The maximum pseudo-likelihood estimation (MPL) is introduced by Genest et al. (1995) and has been developed in Chen and Fan (2006) for time-series copula modeling and in Aas et al. (2009) for pair copula construction. The estimation using MPL is basically a semi-parametric approach, consisting of non-parametric marginal transformation and parametric estimation for dependence parameters (Genest et al., 1995).

The statistical process in the empirical study is designed to identify whether non-linear dependence modeling with the R-Vine is the best fit for the dependence structure of cyber losses or if it can be better described by linear dependence modeling (Gaussian or student-t) or by simple asymmetric tail dependence modeling (Gumbel and Clayton) using uniform margins. Furthermore, we check the statistical test results for the C-Vine and the D-Vine to see whether the R-Vine method formulates the structure identical to the C-Vine or the D-Vine.

We first determine the structure of the first tree and sequentially implement bivariate copula modeling for each adjacent pair and select the best fit copula with the following rules[14]:

Step 1: Select the most appropriate copula candidates by the Vuong–Clarke test (Vuong, 1989; Clarke, 2007; Brechmann and Schepsmeier, 2013)[15]:

Vuong test:

$$\tau = \frac{\frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{c_1\left(u_i|\hat{\theta}_1\right)}{c_2\left(u_i|\hat{\theta}_2\right)}\right]}{\sqrt{\sum_{i=1}^{n}\left(\log\left[\frac{c_1\left(u_i|\hat{\theta}_1\right)}{c_2\left(u_i|\hat{\theta}_2\right)}\right]-E\left(\log\left[\frac{c_1\left(u_i|\hat{\theta}_1\right)}{c_2\left(u_i|\hat{\theta}_2\right)}\right]\right)\right)^2}}, \tag{4}$$

Clarke test:

$$\nu = \sum_{i=1}^{n} 1_{(0,\infty)}\left(\log\left[\frac{c_1\left(u_i|\hat{\theta}_1\right)}{c_2\left(u_i|\hat{\theta}_2\right)}\right]\right), \tag{5}$$

where $u_i \in [0,1], i = 1,\ldots,n$, is a uniform margin, n is the number of margins (dimension), $c_j(\cdot), j = 1,2$ is a copula function to be compared, $\hat{\theta}_j, j = 1,2$ is the corresponding copula parameter.

Step 2: Check the statistical specification of those candidates by Cramer–von-Mises (CvM) goodness-of-fit test (Genest et al., 2009):

CvM test:

$$S_n = \int_{[0,1]^d}\left[\sqrt{n}\left(C_n(\mathbf{u}) - C_{\theta_n}(\mathbf{u})\right)\right]^2 dC_n(\mathbf{u}), \tag{6}$$

where $\mathbf{u}$ is a vector of uniform margins, $C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} 1_{(U_{i1}\le u_1,\ldots,U_{id}\le u_d)}$ is an empirical copula with uniform margins and $C_{\theta_n}(\mathbf{u})$ is a copula function of interest with a parameter θ_n. We exclude the candidates that fail to be accepted by the CvM test.

Step 3: Select the most appropriate copula function by AIC.[16] We choose the function with the minimum AIC for every single pair dependence.

For zero-loss dependence modeling, we use the Gaussian copula. According to Erhardt and Czado (2012) and Brechmann et al. (2014), this dependence modeling on binary margins is not standardized, hence some parametric copulas (Archimedean copulas) and the vine copula method are not appropriate due to the non-existence of a closed form and the non-heterogeneous pairwise dependence. Instead, we follow Brechmann et al. (2014) and model this binary dependence by Gaussian copula with rank correlation parameters as an efficient tool for dependence modeling.

With the estimated structures from non-zero and zero loss dependency, we aggregate monthly losses from both non-zero and zero loss dependence structures as (Brechmann et al., 2014):

$$\lambda_k = v_k \times \lambda_k^0 + (1 - v_k) \times \lambda_k^+ = (1 - v_k) \times \lambda_k^+, \tag{7}$$

where $\lambda_k \ge 0\,(k = 1,\ldots,d)$ is the kth loss out of the d-dimensional risk, $v_k \sim P_{v_k}$ is the occurrence of zero loss as a binary random variable. We denote zero loss by λ_k^0 and a positive loss by λ_k^+.

The aggregate loss distributions are applied to derive risk measures and insurance premiums in Section 5 and analyze the diversification effect of each dependence model. We compare the estimated dependence structures with the independence structure as a benchmark to see how our dependence modeling affects risk measures, insurance premiums, and diversification results. Fig. C1 in Online-Appendix C depicts the methodology of this study.

3. Data description

We consider data breaches occurring from January 1, 2005 to December 31, 2016, derived from the Privacy Rights Clearinghouse (PRC). The PRC dataset provides the entity, attack type, and the total number of records breached. PRC collects data breach information from government agencies and verifiable media sources.[17] In this dataset, most damages have been reported with positive breach records, whereas the rest of damages remain with zero record, because these cases either are not publicly acknowledged or are still being investigated (Edwards et al., 2016; Privacy Rights Clearinghouse, 2016). Such a data type with excess zeros is frequently presented in a variety of research areas, for example, insurance claim analyses and ecological studies, showing that a dataset consists of a substantial proportion of zero values and extremely skewed distribution of non-zero values (Fletcher et al., 2005; Erhardt and Czado, 2012). These zero values could give distorted information on the dependence structure of cyber losses, so we restrict breach data to those with counts (non-zero values). However, zero values appear again when we model the data on a monthly aggregate basis, because some months have no breach events. Thus, we consider dependence structures in the presence of zero values.

Descriptive statistics of the monthly loss data are shown in Table 1.[18] Variables are categorized in two cross-sectional settings. The underlying dataset contains 3327 data breach observations

[14] The choice of the fitted parametric copula function in each pair is still an open question (Erhardt and Czado, 2012). Furthermore, the pair copula method with parametric bivariate copula functions might have an inherent model risk arising from the misspecification of marginal copulas, thus the sequential estimation of pair copulas could be unstable due to the error-prone selection (Scheffer and Weiss, 2017).

[15] In this step, we implement the comparison test proposed by Belgorodski (2010) between different copulas in the bivariate setting, the test that is conducted by using both tests from Vuong (1989) and Clarke (2007). The comparison test allocates "1" to a copula model if it is preferred to another, otherwise "−1" is allocated. No point is assigned if there is no preference between copulas. The final decision on the most appropriate candidates is made by the highest score after all possible comparisons.

[16] The information criterion developed by Akaike (1973) is defined as:

$$AIC := -2Loglik_i(\Theta|\mathbf{u}) + 2k,$$

where $Loglik_i(\cdot)$ is the log-likelihood of i th model, $\Theta = (\theta_1,\ldots,\theta_k)$ is a set of parameters, $\mathbf{u} = (u_1,\ldots,u_d)$ is a d-dimensional set of uniform margins and k is the number of parameters.

[17] Given the absence of individual cyber-insurance loss data, we use the aggregate data for data breaches available from the PRC dataset. To use this dataset, the reliability of the data needs to be confirmed. Regarding reliability, each loss event has been confirmed at least by one major media source and is thus easily traceable and peer-reviewable. The dataset has already been used in numerous academic papers (e.g., Edwards et al., 2016; Eling and Wirfs, 2018; Eling and Loperfido, 2017; Rasoulian et al., 2017) and is widely accepted in practice. In terms of completeness, one limitation is that the data provider only includes losses that were publicly recognized (Edwards et al., 2016). However, PRC continuously updates the dataset to ensure the best possible completeness and the dataset is the largest public database about breached data information (Edwards et al., 2016).

[18] We choose monthly average data as the standard timeframe that can provide less excessive zeros than, for example, weekly data or bi-weekly data that are not sufficient for meaningful aggregation.

Table 2
Rank correlation matrices of cyber losses.

Cross-industry					***Cross-breach type***				
	HACK	ELET	DISC	INSD		BSF	GOV	MED	BSE	EDU
HACK	1				BSF	1				
ELET	0.119*	1			GOV	0.108*	1			
DISC	0.038	0.070	1		MED	0.091	0.071	1		
INSD	0.268***	−0.006	0.008	1	BSE	−0.067	−0.038	−0.066	1	
					EDU	−0.008	0.028	0.080	0.074	1

Note: The table shows the rank correlation matrices in both cross-sectional settings, which is comparable with the correlation classification in Table 1 of Böhme and Kataria (2006). The correlation measures in the table are calculated upon the monthly average losses, which we regard as the representative of individual loss processes. *, **, *** indicate that the *p*-value is less than the significance levels, 10%, 5% and 1% respectively.

Table 3
Goodness-of-fit and model comparison for loss frequency.

Distribution	Log-likelihood	AIC	Chisq-Test	
Panel A: Hacking (HACK)				
Poisson	−466.368	934.747	192.713	***
Zero-inflated Poisson	−460.888	925.777	>10,000	***
Negative Binomial	**−390.279**	**784.558**	**8.609**	
Zero-inflated Neg. Binomial	−390.279	786.558	53.347	***
Geometric	−408.058	818.116	52.927	***
Panel B: Electronic device (ELET)				
Poisson	−571.296	1,144.591	1,614.03	***
Zero-inflated Poisson	−522.838	1,049.675	>10,000	***
Negative Binomial	−430.241	864.482	8.347	
Zero-inflated Neg. Binomial	**−428.399**	**862.799**	**34.408**	
Geometric	−437.078	876.156	45.201	**
Panel C: Disclosure (DISC)				
Poisson	−403.684	809.368	148.609	***
Zero-inflated Poisson	−384.636	773.273	101.345	***
Negative Binomial	−366.322	736.644	4.266	
Zero-inflated Neg. Binomial	**−364.762**	**735.523**	**7.815**	
Geometric	−382.758	767.515	38.043	***
Panel D: Insider (INSD)				
Poisson	−321.857	645.713	108.172	***
Zero-inflated Poisson	−307.818	619.637	>10,000	***
Negative Binomial	**−288.792**	**581.583**	**3.093**	
Zero-inflated Neg. Binomial	−288.792	583.583	52.156	***
Geometric	−292.499	586.999	23.679	

Note: *, **, *** indicate that the *p*-value is less than the significance levels, 10%, 5% and 1% respectively. The bold indicates the best fit distribution for each loss distribution based on AIC and goodness-of-fit test result.

that we group into 144 monthly observations.[19] The monthly dataset is zero-inflated, because in some months no loss occurs. It is observed in panel A of Table 1 that all severity distributions of risk factors are highly skewed and leptokurtic. Particularly, hacking risk (HACK) and retail/other business risk (BSE) categories have severer and more frequent losses than other variables do.[20]

Böhme and Kataria (2006) define the correlations of different cyber security attacks in two categories: internal and global correlation, which are consistent, respectively, with cross-industry type and cross-breach type in our case.[21] Based on Table 2 showing rank correlation matrices for the size of the severity per event with statistical significance, we can empirically classify different types of data breach risks with respect to rank dependency.[22] The classification is determined by the test statistics for rank correlations at the 10% critical level.

Our empirical classification in Table 2 is consistent with Böhme and Kataria (2006) in that hacking attacks (including worms and viruses) and insider attacks fall into the high dependence category (internal correlation). This classification could result from the fact that hacking and insider breaches are malicious attacks which are expected to be more correlated than negligent risks such as

[19] Although there is no clear standard on the sample size in the copula estimation, we do not consider all types of attacks included in the PRC dataset due to lack of data. Specifically, while variables in our model mainly include more than 100 data points, the subcategories of CARD (Payment card fraud), UNKN (Unknown attack) and NGO contain fewer than 20 data points. An analysis with small samples could give rise to a higher probability of assuming a false premise as true (Hogg et al., 2005). This can be also applied to the case of our dataset that the small sample size might lead to a distorted dependence structure between risks or between industries. Moreover, the industries BSO (Business others) and BSR (Retail/Merchant business) are combined to BSE (Business entities apart from finance and insurance) and the attack types PORT (Portable device) and STAT (Stationary device) combined to ELET (Electronic devices). Once enough data are accumulated, these industries and types of attacks might be analyzed in greater detail.

[20] Plots in Online-Appendix C offer graphical descriptions on the monthly data in both cross-sectional settings. Panels A and B of Figs. C2 and C3 display the histograms of frequency, severity and log-severity for cross-sectional variables. Both frequency and severity are right skewed, whereas the distributions of log-severity data seem to be closer to the normal distribution. Fig. C4 shows pairwise scatterplots with original monthly losses in panel A and with transformed uniform margins in panel B. Clustering in small losses is observed in panel A, but simultaneous extreme losses are scarce. Zero inflated pairs are more clearly identified in panel B, which are treated separately in the dependence modeling in Section 4.

[21] We determine cross-industry dependency (correlation) as equivalent to internal correlation in Böhme and Kataria (2006) in that we look into the dependence structure between losses from different breach types. Cross-breach type dependency is applied in the identical way.

[22] In this analysis, we use monthly average loss per risk factor, which can describe an expected association between individual losses from different risk factors.

disclosure risk. For example, if an insider who intends to breach some customer or financial information can access the company's security system to plant a malicious code into the system, an outsider might find it much easier to hack the system as well. Thus, the connection between a malicious insider and a hacker outside of the entity might facilitate a breach event. The correlation between HACK and ELET can be explained in a similar context, since ELET can be regarded as an internal risk. There might be also a case in which a driver by malware or ransomware could trigger an extreme loss event due to highly correlated information systems, but such a case could not be identified in this example. Note that other risks in Böhme and Kataria (2006), for example spyware/phishing and hardware failure, do not fully overlap with the risks in this study apart from insider/hacking attack, as our dataset is based on numerical values of data breach records.

4. Results

4.1. Marginal modeling

In this section, we conduct the distribution fitting for the frequency and severity data. Several studies provide evidence that the distribution of cyber risk frequencies follows a negative binomial distribution (see Edwards et al., 2016; Eling and Wirfs, 2018; Eling and Loperfido, 2017). These studies usually take into account Poisson and negative binomial as candidates, both of which are widely used in the insurance claim analysis. We consider three additional candidates: zero-inflated Poisson, zero-inflated negative binomial and Geometric, all of which could provide a better fit for our right-skewed and zero-inflated dataset.[23] We evaluate the distributions with the best fit based on the AIC and chi-squared goodness-of-fit test result. In Tables 3 and 4, we display fitting results for frequency and severity in the cross-industry setting, whereas the results in the cross-breach type setting are illustrated in Online-Appendix E. Monthly frequencies in Table 3 are best described by the negative binomial distribution, but some are better fitted by the zero-inflated negative binomial. However, in line with the literature we can conclude that the negative binomial distribution well describes the count process of data breach risks.

In the severity fitting (Table 4), we do not consider zero values because, otherwise, zero values are double considered both in frequency and severity. We test several continuous distributions known for right skewed distributions to fit severity: lognormal, skew normal, skew student-t, weibull, gamma, inverse Gaussian, Cauchy, burr, generalized Pareto and Peaks-over-Threshold (POT) with lognormal in the body and Pareto above 90% threshold.[24]

Table 4
Goodness-of-fit and model comparison for loss severity (non-zero values).

Distribution	Log-likelihood	AIC	K-S Test	
Panel A: Hacking (HACK)				
Lognormal	**−1,953.437**	**3,910.873**	**0.105**	
Skew normal	−2,582.646	5,169.292	0.907	***
Skew t	−2,507.013	5,022.027	0.986	***
Weibull	−1,972.419	3,948.838	0.160	***
Gamma	−2,002.671	4,009.343	0.284	***
Inverse Gaussian	−2,004.296	4,012.593	0.334	***
Cauchy	−2,080.374	4,164.748	0.321	***
Burr	−1,995.684	3,999.367	0.979	***
GPD	−2,103.127	4,212.254	0.417	***
POT (lognormal-GPD)	−1,952.288	3,912.576	0.121	
Panel B: Electronic Device (ELET)				
Lognormal	**−1,623.860**	**3,251.719**	**0.049**	
Skew normal	−1,997.298	3,998.597	0.962	***
Skew t	−1,877.755	3,763.510	0.985	***
Weibull	−1,639.617	3,283.235	0.111	
Gamma	−1,663.419	3,330.838	0.213	***
Inverse Gaussian	−1,638.764	3,281.527	0.179	**
Cauchy	−1,694.562	3,393.124	0.277	***
Burr	−1,624.796	3,257.591	0.059	
GPD	−1,680.591	3,367.182	0.351	***
POT (lognormal-GPD)	−1,623.369	3,254.737	0.138	
Panel C: Disclosure (DISC)				
Lognormal	−1,521.159	3,046.317	0.073	
Skew normal	−2,148.995	4,301.990	0.977	***
Skew t	−2,009.115	4,026.229	0.985	***
Weibull	−1,547.675	3,099.350	0.149	**
Gamma	−1,596.578	3,197.156	0.290	***
Inverse Gaussian	−1,566.356	3,136.713	0.279	***
Cauchy	−1,592.378	3,188.756	0.282	***
Burr	**−1,515.401**	**3,038.802**	**0.034**	
GPD	−1,546.286	3,098.573	0.226	***
POT (lognormal-GPD)	−1,516.954	3,041.908	0.083	
Panel D: Insider (INSD)				
Lognormal	**−1,216.649**	**2,437.298**	**0.079**	
Skew normal	−1,820.389	3,644.778	0.973	***
Skew t	−1,697.613	3,403.226	0.982	***
Weibull	−1,233.527	2,471.054	0.128	**
Gamma	−1,273.039	2,550.078	0.266	***
Inverse Gaussian	−1,253.774	2,511.548	0.314	***
Cauchy	−1,310.905	2,625.809	0.327	***
Burr	−1,215.820	2,439.641	0.045	
GPD	−1,317.948	2,641.896	0.464	***
POT (lognormal-GPD)	−1,215.768	2,439.537	0.143	

Note: *, **, *** indicate that the p-value is less than the significance levels, 10%, 5% and 1% respectively. The bold indicates the best fit distribution for each loss distribution based on AIC and goodness-of-fit test result.

These distributions are widely used in the operational risk modeling and the insurance context[25] and we choose them in accordance with the graphical description on the severity shape of body and tail (see panel B of Figs. B2 and B3). The best-fitting distribution is assessed by minimizing the AIC and the Kolmogorov–Smirnov test (K–S test).

Table 4 shows that the lognormal distribution is the best fit for most severity distributions (HACK, ELET and INSD) and DISC is better fitted by burr,[26] which implies that cyber loss severity is long-tailed in general, but extreme losses on the right tail are

[23] The zero-inflated Poisson distribution takes the distributional property from Poisson distribution, but more zero values are contained than expected (Zuur et al., 2009). The geometric distribution is a special case of the negative binomial, but different from the negative binomial in that it models the number of trials until the first success (Hogg et al., 2005). Hence it focuses on the number of failures (the number of no breach events in our dataset). We compare the five discrete distributions by Chi-square goodness-of-fit test results.

[24] We test different thresholds for POT from 50% to 99% and determine 90% threshold with the minimum AIC as the optimal value for the dataset. AIC is derived from the minimum negative log-likelihood from the body model and the tail model with the number of parameters for each variable. Scarrott and MacDonald (2012) argue that the tradeoff between bias and variance of the parameter estimates needs to be considered when the optimal threshold is estimated. In particular, they state that a sufficiently high threshold is required to confirm that the asymptotic estimates are reliable and unbiased, thereby making 90% threshold a stable level for the parameter estimates in this case. We use the R package, `evmix`, to carry out POT fitting with continuity constraints (see Scarrott and MacDonald (2012), for more detail on the continuity constraint at a threshold). The goodness-of-fit test in this distribution is implemented by two sample Smirnov test (Conover, 1971). The estimated shape parameters of the tail distribution (GPD) are for HACK: 0.1763, ELET: 0.2288, DISC: 0.3193, INSD: 0.3586, BSF: 0.1971, GOV: 0.1997, MED: 0.2357, BSE: 0.1977, EDU: 0.1533. We thus find that the distributions of DISC and INSD are more heavy-tailed than those of other variables.

[25] See Frachot et al. (2001), Moscadelli (2004), Fu and Moncher (2004), Shevchenko (2011) and Frees et al. (2016).

[26] Burr distribution is a continuous probability distribution allowing for only non-negative values. This distribution is statistically connected to Pareto distribution and consists of 12 different types as a distribution family (Kleiber and Kotz, 2003, Section 2.3). Among them, Burr type XII is most widely used and known, thus here we use Burr type XII for our severity analysis (see, e.g., Frees and Valdez, 2008; Frees et al., 2016).

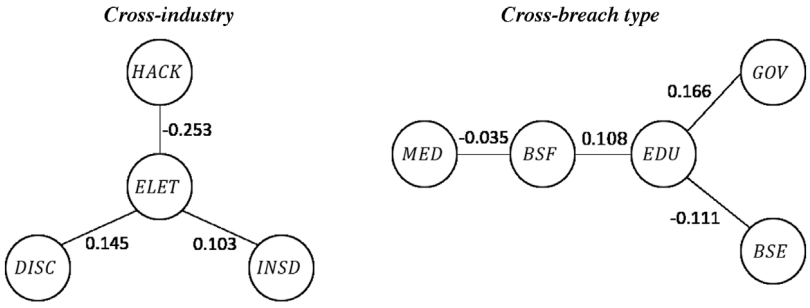

Fig. 2. The ordering of variables in the first tree (R-Vine).

Table 5
Comparison of dependence models by different pair copula structures.

		Cross-industry				Cross-breach type			
Model	LogLik	AIC	GoF test	LogLik	AIC	GoF test			
PCC	D-Vine	22.154	−32.308	12.842	**	19.952	−19.904	15.842	**
	C-Vine	27.840	−43.680	10.003		25.654	−37.307	16.365	*
	R-Vine	**27.840**	**−43.680**	**10.003**		**26.274**	**−38.548**	**32.822**	
Elliptical	Gaussian	0.0001	1.999	0.057	**	0.0005	1.999	0.032	**
	Student-t	16.472	−18.943	0.977		15.388	−8.776	1.862	*
Archime-dean	Gumbel	0.048	1.904	0.055	**	0.013	1.974	0.033	**
	Clayton	1.183	−0.365	0.734	**	4.740	−7.481	0.083	**

Note: The table illustrates the results of statistical tests in both cross-sectional settings: Log-likelihood, Akaike Information Criteria and goodness-of-fit test. These measures help compare the model fits and determine which model is the best fit for each dependence structure. The parametric PCCs (C-Vine and R-Vine) are sequentially determined by R-package VineCopula and the D-Vine structure is estimated via `CDVine`.[29] *, **, *** indicate that the p-value of GoF test is less than the significance levels, 10%, 5% and 1% respectively. The bold indicates the best fit distribution for each loss distribution based on AIC and goodness-of-fit test result.

not necessarily modeled by an infinite-mean model from extreme value theory. The estimation for cross-breach type frequency and severity distributions in Online-Appendix E yields similar results; for instance, for severity the lognormal shows the best fit in three out of five cases, while the burr is the best fit for BSE and Weibull for EDU (see Tables E1, E2). In Online-Appendix E, we provide the graphical diagnosis in the fitting outcomes using QQ plots and CDF plots for both cross-sectional cases, where the plots illustrate good fits of the estimated distributions for the risk factors by showing an almost exact match of the theoretical line and the fitting line (see Figs. E1, E2).

4.2. Modeling non-zero pair dependence

We now model the dependence structures for pairwise non-zero losses in different dependence settings. We estimate the R-Vine, Gaussian, Student-t, Gumbel and Clayton and determine the best fit structure for the cyber loss processes. Kendall's tau is used to order random variables in the first tree since Kendall's tau can estimate non-parametric correlation independently of the hypothetic distribution and provide a global measure for non-Gaussian families (Dissmann et al., 2013).

Fig. 2 illustrates the orders of uniform margins with Kendall's tau at each edge in the R-Vine structure.[27] In the cross-industry setting, ELET connects other risk factors under strong dependency, which draws the identical structure to the C-Vine structure. Similarly, EDU plays a central role in the cross-breach type setting except for indirect connection to MED, providing a more flexible structure than that of D-Vine or C-Vine. It can be assumed that the dependences in the first tree are stronger than those in the other trees because the conditional correlation between variables given a certain variable is smaller than the unconditional correlation in the first tree (see Dissmann et al., 2013, Section 3.1). This property reduces the number of the model parameters by using independence copula in the later trees when the dependence parameters in these trees are close to independence.[28] For this reason, we consider the independence copula in the selection process and find its validity in the later trees in the estimation (see Table 6).

Table 5 shows the goodness-of-fit results and the information criteria of pair-wise dependence structures. Overall, the pair copula structures (D-Vine, C-Vine and R-Vine) turn out to be the better fit to describe the dependency in both cross-industry risks and cross-breach type risks according to AIC and goodness-of-fit results than elliptical copulas and Archimedean copulas. Among vine models, the R-Vine structure is the most appropriate model for data breach losses and is identical to the C-Vine structure in the cross-industry setting (see Fig. 2).

Based on the results from Table 5, the estimation parameters for the best fit pair copula structure (R-Vine) are illustrated in Table 6. In addition, the tree structures in both cross-sectional settings are graphically illustrated in Online-Appendix F (Fig. F1). Parameters are obtained by the maximum likelihood estimation based on the copula function fitted for each pair. As above-mentioned, ELET

[27] The measures of Kendall's tau in Fig. 2 are based on the monthly loss sum per risk factor, which better describe the estimated compounding process. For this reason, the rank correlations are different from those in Table 2.

[28] Typically, modeling R-Vine structure is computationally intensive due to a substantial number of possible R-Vine structures (Dissmann et al., 2013).

[29] Note that the C-Vine and the D-Vine in Table 5 are unique structures by determining the shape of the first tree in accordance with the structure of each vine (see Fig. 1). In the implementation of the vine models the package `VineCopula` might convert the C-Vine and the D-Vine to the optimal R-Vine model because the R-Vine can represent the two other models depending on the dependence structure of variables. The C-Vine estimation can be implemented via this R-package using the default option for this model in the sequential estimation, whereas the D-Vine estimation needs to be conducted via the R-package, `CDVine`, to obtain a unique D-Vine structure. `CDVine` is still available in R, but not actively developed anymore so that the package authors ask users to transfer to `VineCopula` package. To model a unique D-Vine structure in the `VineCopula` package, one can employ a function facilitating to convert the D-Vine structure to an R-Vine matrix representation.

Table 6
Parameter estimations of flexible pair copula structure.

Copula	Parameter	Lower-tail dependency	Upper-tail dependency
Panel A: Cross-industry			
θ_{21}; $C_{ELET,HACK}$ (270° Rotated Clayton)	−0.60	–	–
θ_{23}; $C_{ELET,DISC}$ (Survival Joe)	1.46	0.391	–
θ_{42}; $C_{ELET,INSD}$ (Survival Joe)	1.27	0.275	–
$\theta_{13\|2}$; $C_{HACK,DISC\|ELET}$ (Frank)	0.92	–	–
$\theta_{41\|2}$; $C_{INSD,HACK\|ELET}$ (Student-t)	0.08 ($df = 4.04$)	0.093	0.093
$\theta_{43\|1,2}$; $C_{INSD,DISC\|HACK,ELET}$ (Independence)	–	–	–
Panel B: Cross-breach type			
θ_{52}; $C_{EDU;GOV}$ (Survival Joe)	1.52	0.424	–
θ_{51}; $C_{EDU,BSF}$ (Survival Joe)	1.44	0.382	–
θ_{54}; $C_{EDU,BSE}$ (90° Rotated Joe)	−1.29	–	–
θ_{13}; $C_{BSF,MED}$ (Student-t)	−0.05 ($df = 2.69$)	0.121	0.121
$\theta_{12\|5}$; $C_{BSF,GOV\|EDU}$ (Independence)	–	–	–
$\theta_{41\|5}$; $C_{BSE,BSF\|EDU}$ (Independence)	–	–	–
$\theta_{53\|1}$; $C_{EDU,MED\|BSF}$ (Independence)	–	–	–
$\theta_{42\|15}$; $C_{BSE,GOV\|BSF,EDU}$ (Independence)	–	–	–
$\theta_{23\|51}$; $C_{GOV,MED\|EDU,BSF}$ (Survival Joe)	1.12	0.146	–
$\theta_{43\|251}$; $C_{BSE,MED\|GOV,EDU,BSF}$ (Survival Joe)	1.12	0.141	–

Note: The parametric PCCs are sequentially determined by R-package VineCopula. The numbers as subscripts of the copula parameters indicate in the following: *Cross-industry*: 1 = HACK (Hacking), 2 = ELET (Lost electronic device), 3 = DISC (Disclosure), 4 = INSD (Insider attack); *Cross-breach type*: 1 = BSF (Banking and insurance), 2 = GOV (Governmental entity), 3 = MED (Medical service), 4 = BSE (Retail/Merchant and other business), 5 = EDU (Educational institution).

and EDU play key roles in the relationship of the risk factors. In the cross-industry setting, ELET is placed in the center of the dependence being connected to all other risk factors in the first tree; similarly in the cross-breach type setting EDU is directly connected to the other factors except for MED. We observe lower tail dependence between DISC/INSD and ELET (θ_{23} and θ_{42} in panel A of Table 6) and between GOV/BSF and EDU (θ_{52} and θ_{51} in panel B of Table 6). Smaller losses from DISC/INSD and ELET as well as GOV/BSF and EDU are thus dominant in the dependence structure. Moreover, ELET and HACK are negatively dependent described by 270° rotated Clayton copula, indicating a tendency to have a small loss by ELET and a larger loss by HACK at the same time. This also applies to the relationship between EDU and BSE.[30]

This common relationship can be explained by the fact that the risk by losing an electronic device containing personal/corporate data can lead to a small loss due to the limit of the device, whereas the risk by hacking can be significantly large due to the interconnected system in the corporation. In the similar context, it can be implied that a business entity operating small sub-units such as a university or school is more likely to be exposed to less risk than other entities possessing a centralized and complex operation in the retail/franchise industry are (Wheatley et al., 2016).

The asymmetric tail dependency on the left tail has implications for an insurance company managing a cyber-insurance portfolio in that data breach losses with high frequency and low severity could mainly constitute the risk pool. That is, small losses of data breach can more frequently occur, whereas the frequency of large losses tends to be rather rare and stable (Wheatley et al., 2016). If extremely large losses frequently occur simultaneously, an insurance company might need more reserves to prepare against those simultaneous claims and these losses might not be insured. In line with the analysis by Biener et al. (2015) we conclude that cyber risk can be insured in terms of the criteria on maximum possible loss and loss exposure. Such an asymmetric co-movement in losses could not be captured by a linear dependence modeling, which might lead to an inaccurate estimation on the risk level for the insurance company.

[30] Estimated 270° Clayton copula indicates the correlation between small loss of ELET and large loss of HACK. Similarly, estimated 90° Joe demonstrates the correlation between small loss of EDU and large loss of BSE.

4.3. Modeling zero loss dependence

If we want to derive risk measures from a portfolio loss distribution with excessive zeros, zero value dependence needs to be separately considered (Brechmann et al., 2014). As mentioned in Section 2.2, each variable is binary distributed, where 1 is given to a zero value and 0 is given to a non-zero value. We implement Gaussian copula modeling to identify the dependence of binary distributions, displaying how dependent zero value arrival processes are. For the Gaussian copula model, we generate parameter matrices for both cross-industry and cross-breach type (see Table 7).

We observe very weak dependences in zero loss arrivals where many of the parameters are close to 0. Here, a positive dependence means that risks are more likely to simultaneously arrive than arrive not simultaneously, whereas a negative dependence indicates a higher likelihood of counter-monotonicity in zero loss arrivals. It can be inferred that there is no significant dependency between zero losses. We generate the probability of zero losses for each loss distribution from the Gaussian dependence structure and use it to measure risk levels and price insurance premiums by Eq. (7) in the next section.

4.4. Statistical difference in loss aggregate distributions

Prior to the applications, we want to evaluate whether there is a difference in the estimated loss aggregate distributions. To achieve this, we conduct two statistical tests: the Wilcoxon signed-rank test, which is a non-parametric statistical hypothesis test to compare two independent samples by using population mean ranks, and the Kolmogorov–Smirnov test (Wilcoxon, 1945; Smirnov, 1948). We thus test whether two loss aggregate distributions from different dependence structures are significantly different. Each aggregate distribution of breach records is generated based on Eq. (7) with 500,000 simulations. Four comparisons are tested among the R-Vine structure, the independence structure, the Gaussian structure and the empirical structure, which are of main interest in the applications. The null hypothesis is[31]:

$$H_0 : L_i - L_j = 0 \quad (i \neq j),$$

where L_i is a loss vector from a dependence structure considered ($i = 1, \ldots, 3$). The results in Table 8 show that the differences

[31] All tests are two-sided paired difference tests as related to the null hypothesis.

Table 7
Dependence parameters of Gaussian copula for zero loss arrival.

Cross-industry					*Cross-breach type*					
	HACK	ELET	DISC	INSD		BSF	GOV	MED	BSE	EDU
HACK	1				BSF	1				
ELET	−0.017	1			GOV	0.004	1			
DISC	−0.009	0.0002	1		MED	−0.008	−0.003	1		
INSD	−0.006	0.0005	−0.007	1	BSE	0.010	0.018	0.004	1	
					EDU	−0.009	0.001	−0.006	−0.012	1

Note: The table displays the correlation matrices of zero-loss processes estimated by Gaussian copula model. The acronyms of variables are described as follows:
Cross-industry: HACK = Hacking; ELET = Lost electronic device; DISC = Disclosure; INSD = Insider attack
Cross-breach type: BSF = Banking and insurance; GOV = Governmental entity; MED = Medical service; BSE = Retail/Merchant and other business; EDU = Educational institution.

Table 8
The result of statistical difference test.

		Wilcoxon test	K-S test
Cross-industry	$L_{Rvine} - L_{emp}$	242,297	0.024
	$L_{Rvine} - L_{ind}$	268,598**	0.069***
	$L_{Rvine} - L_{Gauss}$	231,743*	0.059*
	$L_{Ind} - L_{emp}$	229,505**	0.069***
Cross-breach type	$L_{Rvine} - L_{emp}$	259,209	0.035
	$L_{Rvine} - L_{ind}$	279,855***	0.078***
	$L_{Rvine} - L_{Gauss}$	265,948*	0.068**
	$L_{Ind} - L_{emp}$	229,635**	0.076***

Note: The table shows how statistically close different aggregate distributions are based on two statistical tests. Wilcoxon test is a non-parametric statistical test using ranks of two distributions of interest and K-S test is a goodness-of-fit test using an empirical distribution to measure the distance. *, **, *** indicate p-value less than a significant level at 10%, 5% and 1% respectively.

in aggregate distributions between the R-Vine structure and the independence structure, between the independence structure and the empirical structure and between the R-Vine structure and the Gaussian structure are statistically significant for both cross-industry and cross-breach type settings; no significant difference between the R-Vine structure and the empirical structure is identified. Tests in both cross-sectional settings arrive at the same conclusion and the testing results are confirmed at the 10% critical level.[32] The R-Vine model is thus close to the empirical model, but different from the independence and Gaussian models.

5. Applications to risk measurement and pricing

We now apply the estimated PCC dependence structure from Section 4 to risk measurement and insurance pricing. The applications are based on the aggregate distributions from different dependence models with 500,000 copula-simulated values. The risk measures and prices using PCC are then compared with the measures under the (a) independence assumption, (b) linear dependence assumption (Gaussian copula), (c) linear and symmetric tail dependence assumption (Student-t copula) and (d) the dependence structure from the empirical copula.[33]

The applications are again carried out in two cross-sectional settings: cross-industry and cross-breach type. With regard to insurance pricing, since the values of the aggregated distribution are the number of breach records, it is necessary to convert them to dollars to derive insurance prices. Several cyber security companies offer data breach cost calculators (e.g. Imperva, 2016; eRiskHub, 2016; FireEye, 2016). However, there is no established method on how to conduct these calculations.[34] Jacobs (2014) analyzes the Ponemon datasets[35] for 2013 and 2014 and proposes the following regression model to compute data breach cost that we use to calculate insurance prices:

$$\textit{Dollar loss} = \exp\left[7.68 + 0.76 * \ln\left(\textit{breach records}\right)\right]. \tag{8}$$

Two risk measures are derived from the aggregate distribution of breach records:

Value at Risk:

$$VaR_{(1-\alpha)}(X) = \inf\left\{X \in \mathbb{R} : X \geq F_X^{-1}(1-\alpha)\right\}, \tag{9}$$

Expected Shortfall:

$$ES_{(1-\alpha)}(X) = E\left[X \in \mathbb{R} : X | X \geq F_X^{-1}(1-\alpha)\right], \tag{10}$$

where X is a non-negative random variable with finite variance, α is a risk threshold and F is the aggregated loss distribution. We calculate risk measures at critical levels of 90%, 95%, 99% and 99.5% (i.e. α is 10%, 5%, 1% and 0.5%).[36] In addition, we derive insurance premiums using three pricing principles that incorporate different expected utility functions (see Embrechts, 2000):

Fair Premium:

$$P = EX, \tag{11}$$

Standard Deviation Principle:

$$P = EX + \delta \cdot \sqrt{Var\,(X)}, \tag{12}$$

Exponential Principle:

$$P = \frac{1}{\gamma} \ln\left(E\left(e^{\gamma X}\right)\right), \tag{13}$$

where δ is a cost-loading and γ is the risk-aversion parameter.[37]

[32] The test result between the R-Vine and Gaussian is determined at the 10% level, whereas other test results are confirmed at the 5% level. The statistical distance between the R-Vine and Gaussian is relatively closer than the distances between the R-Vine and independence and between independence and empirical structure.

[33] The independence assumption is a baseline benchmark without any dependence modeling. Gaussian and student-t copula models are used in many fields, hence the measures by these two models might be close to the values that are used in practice. Lastly, the empirical setting can serve as a benchmark for our estimated model on how close our model is to the historically, empirically observed dependency.

[34] The cost per data breach might be different for company size, industry or other factors and insured losses by a cyber-insurance provider typically consist of data recovery/replacement of intellectual property, third-party liability and forensics (Allianz, 2015). Thus, the estimated risk measures in this section do not perfectly reflect the actually insured losses in reality.

[35] Ponemon Institute LLC. (2016) provides parameters about cost of data breach by different years, industries and type of attacks in its annual reports. The approximation of Jacobs (2014) to transfer the number of breach records into actual loss data is useful to carry out the first applications, but is clearly only a crude and rough approximation of the real loss.

[36] We choose 99.5% as the most extreme critical level estimated since Solvency II requires VaR at 99.5% as the equity capital. Swiss Solvency Test (SST) requires tail VaR (Expected shortfall) at 99% level (FINMA, 2016).

[37] Eq. (11) is based on the *zero-utility* model where the expected utility before paying the insurance premium is the same as that after paying the insurance premium,

In addition to risk measurement and insurance pricing, we investigate the effect of each dependence model on portfolio diversification. The effect demonstrates the extent to which risk measures of the aggregate distribution under the estimated model are reduced compared to the sum of the risk measures from the individual loss distributions. That is the relative risk reduction compared to the comonotonicity assumption (the equally weighted summation of individual risks; Brechmann et al., 2014). We estimate the effects using the expected shortfall since this measure is coherent and satisfies sub-additivity (Artzner et al., 1999; Acerbi and Tasche, 2002).[38] The loss aggregation is based on the equally weighted aggregation that can be derived (Jorion, 2007):

$$\varphi_{j,1-\alpha} = \frac{ES_{j,1-\alpha}\left(\sum_{k=1}^{d}\lambda_{jk}\right) - \sum_{k=1}^{d} ES(\lambda_{jk})}{\sum_{k=1}^{d} ES(\lambda_{jk})}, \tag{14}$$

where $\varphi_{j,1-\alpha}$ is a diversification effect of j th model at $1-\alpha$ quantile ($j = 1, \ldots, 3$) and λ_{jk} is a loss vector of k th risk from j th model ($k = 1, \ldots, d$).

The risk measures in panel A of Table 9 indicate the potential monthly breach records at different critical levels for the entire U.S. market consisting of risks from five industries (BSF, GOV, MED, BSE and EDU) and exposed to four breach types (HACK, ELET, DISC and INSD). For instance, if we use the pair copula model for cyber risk aggregation with the cross-industry risk categorization, we expect 59.54 million monthly breach records at the 90% critical level and around 1.03 billion at the 99.5% level from the potential U.S. risk pool.

The independence structure produces a lower level of risk than other structures do in both cross-sectional settings, resulting from the fact that the structure does not consider correlated risk among the risk factors in the portfolio. The student-t provides higher values at a more extreme level (e.g. 99 or 99.5%), especially in case of the expected shortfall. This might be because the restrictive model using student-t copula with only one tail dependence parameter can hinder the accurate estimation of the true tail dependency and causes the risk measure to be overestimated (Brechmann et al., 2014). In line with the statistical tests from Section 4.4, the risk measures with the R-Vine structure are closer to the empirical structure than in other cases, demonstrating that the pairwise dependence structure with different copula functions serves as a useful tool for this modeling.[39]

The risk measures estimated in the cross-industry setting are larger than the measures in the cross-breach type setting, although the number of risk factors in the cross-industry setting is smaller.[40] There are three plausible reasons for this outcome. Firstly, it can result from the estimated dependence models and the dependence structure of the cross-industry risk pool incorporates a higher correlated risk (i.e. less number of pairs are modeled by independence copula in the cross-industry setting than in the cross-breach type setting; see Table 6). Secondly, the likelihood of zero loss occurrence is affected by different zero loss dependences when integrated by Eq. (7). Lastly, a smaller size in the cross-breach type setting could be addressed by the key factor of the dependence structure, EDU, which demonstrates the lowest severity among all considered risk factors. We thus suggest aggregating the cyber risks in the cross-industry setting, which could lead a cyber-insurer to be on the safe side against cyber-insurance claims.

Based on the results of risk measurement, cyber-insurance premiums are estimated in panel B of Table 9, again on a monthly and an aggregated industry level. Note that there are no cover limits or deductibles assumed in the pricing application and current cyber-insurance policies usually provide the protection against a certain type of risk with restricted coverage and consider internal factors of an insured.[41] Therefore, the estimated premium size will be different from that in practice. If we assume independence, $724.81 million would be needed as fair premium to cover the possible monthly loss in the cyber-insurance portfolio with four types of risk and five industries. The value estimated for the PCC model ($761.97 million) is 5.1% higher and very close to the empirical value ($758.67 million); it thus seems that not enough premium could be collected when independence is assumed, especially in the cross-industry setting. Gaussian and student-t models lead to an underestimation of the insurance premium in the cross-industry setting, illustrating the need to be accurate in describing the dependence structure.[42] As with the risk measurement, the premium size in the cross-industry setting is generally higher than in the cross-breach type setting due to the stronger dependence.

Diversification effects of the estimated models are presented in Table 10 and Figure F2 (Online-Appendix F). In line with the literature (see Brechmann et al., 2014) we observe diversification effects in all structures in both cross-industry and cross-breach type modeling and this benefit becomes larger at more extreme levels. Overall, the pair copula structure shows a bigger diversification effect across quantiles for both cross-sectional settings than other structures do, accounting for the strength of the pairwise dependence model.

6. Conclusion and further research

In this paper, we implement the pair copula construction (PCC) with a range of parametric copulas to investigate the dependence structure of data breach losses. Since monthly breach records

if the policyholder is risk neutral (Kaas et al., 2008). Eq. (12) includes a cost-loading, δ, representing the level of transaction cost and requiring some risk aversion to accept the insurance contract. Following Mukhopadhyay et al. (2013), we assume a cost loading of 0.1 for the standard deviation principle. Eq. (13) explicitly includes the exponential utility function $U(X) = -\gamma e^{-\gamma X}$, where increasing γ augments the premium, implying that a more risk averse insured is willing to pay a bigger premium to cover the risk. If γ converges to 0, the premium converges to the fair premium. We specify the risk aversion level as 1/1000, 1/10000, 1/100000 to see the difference in premiums with different risk aversion levels.

[38] Sub-additivity can be defined in the following. Let X and Y be two risk factors and let ρ be a function of a risk measure. The risk measure, ρ, which can be defined in the real space of random variables is sub-additive if it satisfies the following property (Artzner et al., 1999):

$$\rho(X + Y) \leq \rho(X) + \rho(Y).$$

This property does not hold in case of Value-at-Risk so that there could exist the following case in Value-at-Risk:

$$\rho(X + Y) > \rho(X) + \rho(Y).$$

[39] Compared to these industry-level estimates, company-level estimates are derived in Online-Appendix G, informing a cyber-insurer of the potential loss amount per event in the cyber-insurance portfolio. The company-level estimation is more complicated than the industry-level estimation due to additional assumptions required. For example, we need to reduce the dimension from aggregate level to individual level by using an estimator how many companies are breached in the U.S. market. Furthermore, we also need to break down the aggregate level of premium size into the individual level by specifying a certain industry and a certain risk type.

[40] Brechmann et al. (2014) also compare the estimated risk measures from two settings of operational risk (business line: BL and event type: ET). The difference between the log-scaled estimates in their paper (see Fig. 6) is not significant and comparable with the difference between the estimates in panel A of Table 9 with log-transformation.

[41] We do not specify any details such as company size, revenue or type of security system in place. In addition, a range of risk types including malicious and accidental risks are considered in the price. However, risk classification in practice is based on the specific cyber risk profile of a customer, which relies on company size, industry, existing security systems and other factors (Allianz, 2015; KPMG, 2016).

[42] In the cross-breach type setting, it is observed that the Gaussian model generates a higher fair premium than others do. This can be explained by a higher density of the aggregate distribution by the Gaussian model at less extreme levels, which can influence the expectation value of the distribution.

Table 9
Applications to risk measurement and insurance pricing.

Panel A: Risk measurement									***(in million breached records, monthly time horizon)***
Data type	Dependence structure	Value-at-risk				Expected shortfall			
		90%	95%	99%	99.5%	90%	95%	99%	99.5%
Cross-industry	Independence	46.87	157.88	764.80	994.87	288.41	472.25	1,011.61	1,050.55
	PCC	**59.54**	**163.22**	**984.61**	**1,041.85**	**313.98**	**522.75**	**1,033.12**	**1,053.11**
	Empirical	59.77	163.03	986.18	1,045.82	310.11	515.85	1,035.70	1,055.18
	Gaussian	55.91	161.43	980.39	1,040.28	301.51	502.24	1,033.78	1,055.87
	Student-t	57.20	161.06	983.76	1,045.13	306.85	512.52	1,039.03	1,062.97
Cross-breach type	Independence	45.39	152.59	732.02	946.05	276.49	452.80	957.41	983.84
	PCC	**51.60**	**153.99**	**930.56**	**930.88**	**276.09**	**467.91**	**933.49**	**936.16**
	Empirical	51.74	154.35	930.57	931.04	274.47	463.43	933.93	937.09
	Gaussian	51.78	156.06	930.62	931.25	280.95	475.16	934.82	938.79
	Student-t	50.88	154.35	930.60	932.09	279.24	474.29	942.20	953.38

Panel B: Insurance pricing						***(monthly premium in million $)***
Data type	Dependence structure	Fair premium	Standard dev. principle	Exponential premium principle		
				$\gamma = 10^{-3}$	$\gamma = 10^{-4}$	$\gamma = 10^{-5}$
Cross-industry	Independence	724.81	938.58	10,924.60	1,065.05	748.35
	PCC	**761.97**	**979.94**	**10,911.78**	**1,114.47**	**786.59**
	Empirical	758.67	975.03	10,927.01	1,106.02	782.93
	Gaussian	739.42	952.01	10,906.24	1,074.83	762.84
	Student-t	749.61	964.88	11,024.94	1,094.51	773.63
Cross-breach type	Independence	688.44	894.99	10,257.92	999.85	710.38
	PCC	**689.86**	**889.76**	**9,763.18**	**978.17**	**710.52**
	Empirical	686.77	885.50	9,748.62	971.25	707.19
	Gaussian	695.86	897.84	9,812.90	990.53	716.96
	Student-t	685.96	888.05	9,929.22	982.63	707.09

Note: The risk measurements are specified by Value-at-risk (VaR) and Expected shortfall (ES) at three critical levels, 90%, 95%, 99% and 99.5%. For insurance pricing, we calculate the premium by three different pricing principles on an annual basis: fair premium principle, standard deviation principle and exponential premium principle. γ is the risk aversion parameter, where $\gamma \to 0$ indicates risk neutrality. The bold model is the preferred model from Section 4.

Table 10
Diversification effects on ES per quantile.

Model	90%	95%	99.5%
Panel A: Cross-industry			
PCC	**−5.6%**	**−9.6%**	**−19.9%**
Empirical	−5.1%	−9.5%	−19.0%
Gaussian	−5.5%	−9.5%	−19.6%
Student-t	−5.4%	−9.5%	−19.0%
Panel B: Cross-breach type			
PCC	**−7.8%**	**−9.7%**	**−13.6%**
Empirical	−7.6%	−9.6%	−13.3%
Gaussian	−7.7%	−9.6%	−13.4%
Student-t	−7.0%	−9.0%	−11.9%

Note: The diversification effects are derived using expected shortfalls. The bold model is the preferred model from Section 4.

include excessive zero losses, we model the dependence by using non-zero pair dependence and zero-loss dependence. We describe the modeling results in two cross-sectional settings: cross-industry by four breach types (HACK, ELET, DISC and INSD) and cross-breach type by five industries (BSF, GOV, MED, BSE and EDU). We find a significant asymmetric tail dependence among risk factors, especially lower tail dependency dominated by small losses. This asymmetric tail dependency is estimated pairwise, resulting in a more accurate tail dependence estimation than the widely used elliptical models or Archimedean copula models. The likelihood of simultaneous small losses might in general not be a big concern for a cyber-insurance provider, however, when aggregating the estimated loss distributions, the insurer must take the potential dependence into account to avoid an underestimation of the risk and necessary premiums. We also show that on an aggregate industry level the U.S. could be faced with approximately 1 billion monthly breach records at 99.5% critical level (one event with such amount likely to occur over 200 months), in our case considering four breach types and five industries.

The estimated pair copula structure produces a higher level of potential loss than under independence assumption. Hence there might be a possibility for a risk manager to underestimate the risk level when neglecting the potential correlated risk. In addition, the correlated risk in the cross-industry setting (between risk factors by breach types) turns out to be higher than the risk in the cross-breach type setting (between risk factors in different industries). This result illustrates the importance of determining the risk factors considered in underwriting and risk management of cyber risk. We propose considering the risk aggregation in the cross-industry setting, since this setting incorporates higher correlated risks in the portfolio, leads a cyber-insurer to be on the safe side with higher capital requirement against cyber risk claims and provides a higher diversification benefit at more extreme levels (see Fig. F2). Given that cyber insurance policies typically cover different types of risk and the policies are then aggregated across companies from different industries, risk managers in insurance companies might follow this approach. We also show that if an insurer offers different types of data breach risks in the coverage, simultaneous losses on the tails might have different impacts on the capital basis when different dependence structures are considered. This finding is relevant for recent equity capital standards such as US Risk-Based Capital (RBC), Solvency II or the Swiss Solvency Test, which typically do not model non-linear dependence among risk factors.

This study identifies a non-linear dependence in data breach losses, but there are several limitations which open directions for future research. For example, the cyber-insurance market development must clearly define damage by a single risk and identify how this damage can be accurately measured. Moreover, the dependence structures in data breach loss could be affected by different characteristics of industries or other causes, such as geographical variation or the degree of the development of security system. For instance, we might expect the costs of data breaches to vary by industry, e.g. when comparing banking and healthcare. If more data on breach events are available across the globe, we might compare the dependence among different regions, because there

could be geographically different appearance, size and frequency in cyber risk events. Moreover, the time variation in data breach risk should be studied in more detail, especially given the dynamic nature of cyber risk and the risk of change; it is thus not clear whether a dependence structure observed in historical data will also hold in the future. As indicated above, we could not include all types of data breaches in our analyses (payment card fraud, unknown attacks) and not all types of companies (NGO's), because the samples are too small. When enough data on those events and industries are accumulated, those should be also analyzed in detail. Additionally, the insurance pricing example presented here should not be interpreted as more than a first rough indication, because it is based on the number of breached data and not on actual loss information that needs to be verified when more and better information becomes available. Another interesting avenue for future research could be to dig deeper into the potential drivers of correlations.

Acknowledgments

We are grateful to two anonymous referees for valuable suggestions. We also thank Christian Biener, Omid Ghavibazoo, Felix Irresberger, Werner Schnell, Shaun Wang and Jan Wirfs for helpful comments and suggestions. Furthermore, we appreciate valuable comments by academics at four research seminars: the 2017 annual meeting of the American Risk and Insurance Association, the conference "Innovations in Insurance, Risk- and Asset Management" at Technical University of Munich, the workshop "Recent developments in dependence modeling with applications in finance and insurance" by Vrije Universiteit Brussel and the research seminar by Institute of Insurance Economics at the University of St. Gallen.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.insmatheco.2018.07.003.

References

Aas, K., Berg, D., 2009. Models for construction of multivariate dependence - A comparison study. Eur. J. Finance 15 (7–8), 639–659.
Aas, K., Czado, C., Frigessi, A., Bakken, H., 2009. Pair-copula constructions of multiple dependence. Insurance Math. Econom. 44 (1), 182–198.
Acerbi, C., Tasche, D., 2002. Expected shortfall: A natural coherent alternative to value at risk. Econ. Notes 31 (2), 379–388.
Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Proceedings of the Second International Symposium on Information Theory Budapest. Akademiai Kiado, pp. 267–281.
Allianz, 2015. A Guide to Cyber Risk. Munich: Allianz Global Corporate & Specialty Communications. Retrieved January 5, 2017, from http://www.agcs.allianz.com/assets/PDFs/risk%20bulletins/CyberRiskGuide.pdf.
Araichi, S., Peretti, C., Belkacem, L., 2016. Solvency capital requirement for a temporal dependent losses in insurance. Econ. Model. 58, 588–598.
Artzner, P., Delbaen, F., Eber, J.M., Heath, D., 1999. Coherent measures of risk. Math. Finance 9 (3), 203–228.
Bedford, T., Cooke, R., 2001. Probability density decomposition for conditionally dependent random variables modeled by vines. Ann. Math. Artif. Intell. 32 (1–4), 245–268.
Belgorodski, N., 2010. Selecting Pair-Copula Families for Regular Vines with Application To the Multivariate Analysis of European Stock Market Indices (Diploma Thesis), Technische Universität München.
Biener, C., Eling, M., Wirfs, J., 2015. Insurability of cyber risk: An empirical analysis. Geneva Papers Risk Insurance Issues Pract. 40 (1), 131–158.
Böhme, R., Kataria, G., 2006. Models and measures for correlation in cyber-insurance. Workshop on the Economics of Information Security (WEIS). University of Cambridge, UK.
Böhme, R., Schwartz, G., 2010. Modeling cyber-insurance: Towards a unifying framework. Workshop on the Economics and Insurance Security (WEIS). Havard University, US.
Brechmann, E.C., Czado, C., Paterlini, S., 2014. Flexible dependence modeling of operational risk losses and its impact on total capital requirements. J. Banking & Finance 40, 271–285.
Brechmann, E.C., Schepsmeier, U., 2013. Modeling dependence with C- and D-vine copulas: The R-package CDVine. J. Stat. Softw. 52 (3), 1–27.
Bühlmann, H., 2007. Mathematical Methods in Risk Theory. Springer Science & Business Media, Heidelberg.
Cebula, J., Young, L., 2010. A Taxonomy of Operational Cyber Security Risks. Software Engineering Institute, Carnegie Mellon.
Chen, X., Fan, Y., 2006. Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. J. Econometrics 135 (1–2), 125–154.
Chiou, S., Tsay, R., 2008. A Copula-based approach to option pricing and risk assessment. J. Data Sci. 6 (3), 273–301.
Clarke, K., 2007. A simple distribution-free test for nonnested model selection. Political Anal. 15 (3), 347–363.
Conover, W., 1971. Practical Nonparametric Statistics. John Wiley & Sons, New York.
Czado, C., 2010. Pair-copula constructions of multivariate copulas. In: Copula Theory and Its Applications. Springer, Berlin Heidelberg, pp. 93–109.
Czado, C., Jeske, S., Hofmann, M., 2013. Selection strategies for regular vine copulae. J. SFDS 154 (1), 174–191.
Dissmann, J., Brechmann, E., Czado, C., Kurowicka, D., 2013. Selecting and estimating regular vine copulae and application to financial returns. Comput. Statist. Data Anal. 59, 52–69.
Edwards, B., Hofmeyr, S., Forrest, S., 2016. Hype and heavy tails: A closer look at data breaches. J. Cybersecurity 2 (1), 3–14.
Eling, M., Loperfido, N., 2017. Data breaches: Goodness of fit, pricing and risk measurement. Insurance Math. Econom. 75, 126–136.
Eling, M., Wirfs, J., 2018. What are the actual costs of cyber risk events?, Eur. J. Oper. Res., forthcoming.
Eling, M., Wirfs, J., 2016. Cyber risk: too big to insure? Risk transfer options for a mercurial risk class. Institute for Insurance Economics, University of St.Gallen. Retrieved June 28, 2017, from http://www.ivw.unisg.ch/~/media/internet/content/dateien/instituteundcenters/ivw/studien/cyberrisk2016.pdf.
Embrechts, P., 2000. Actuarial versus financial pricing of insurance. J. Risk Finance 1 (4), 17–26.
Embrechts, P., Hofert, M., 2013. Statistical inference for copulas in high dimensions: A simulation study. J. Internat. Actuar. Assoc. 43 (2), 81–95.
Embrechts, P., Lindskog, F., McNeil, A., 2001. Modelling dependence with copulas and applications to risk management. Rapport technique. Département de mathématiques, Institut Fédéral de Technologie de Zurich.
Embrechts, P., Puccetti, G., 2008. Aggregating risk across matrix structured loss data: The case of operational risk. J. Oper. Risk 3 (2), 29–44.
Erhardt, V., Czado, C., 2012. Modeling dependent yearly claim totals including zero claims in private health insurance. Scand. Actuar. J. 2, 106–129.
eRiskHub, 2016. NetDiligence® Mini data breach cost calculator. Retrieved December 15, 2016, from https://eriskhub.com/mini-dbcc.
European Union, 2016. The protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Regulation (EU) 2016/679 of the European Parliament and of the Council. Retrieved June 24, 2016, from http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679{&}from=EN.
FINMA, 2016. FINMA Circular 2017/3 SST. Retrieved December 7, 2016, from https://www.finma.ch/en/~/media/finma/dokumente/dokumentencenter/myfinma/rundschreiben/finma-rs-2017-03.pdf?la=en.
FireEye, 2016. Calculate your cyber security cost. Retrieved December 15, 2016, from https://www.fireeye.com/current-threats/tco/calculator.html.
Fletcher, D., Mackenzie, D., Villouta, E., 2005. Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. Environ. Ecol. Stat. 12 (1), 45–54.
Frachot, A., Georges, P., Roncalli, T., 2001. Loss distribution approach for operational risk. Available at SSRN. Retrieved April 17, 2017, from https://ssrn.com/abstract=1032523.
Frees, E., Lee, G., Yang, L., 2016. Multivariate frequency-severity regression models in insurance. Risks 4 (1), 4.
Frees, E.W., Valdez, E.A., 2008. Hierarchical insurance claims modeling. J. Amer. Statist. Assoc. 103 (484), 1457–1469.
Fu, L., Moncher, R., 2004. Severity distributions for GLMs: Gamma or lognormal? Evidence from Monte Carlo simulations. Casualty Actuar. Soc. Discuss. Paper Program 14, 9–230.
Genest, C., Ghoudi, K., Rivest, L., 1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82 (3), 543–552.
Genest, C., Remillard, B., Beaudoin, D., 2009. Goodness-of-fit tests for copulas: A review and a power study. Insur. Math. Econ. 44 (2), 199–213.
Genest, C., Rivest, L., 1993. Statistical inference procedures for bivariate Archimedean copulas. J. Amer. Statist. Assoc. 88 (423), 1034–1043.
Giacometti, R., Rachev, S.T., Chernobai, A., Bertocchi, M., 2008. Aggregation issues in operational risk. J. Oper. Risk 3 (3), 3–23.

Haff, I.H., 2013. Parameter estimation for pair-copula constructions. Bernoulli 19 (2), 462–491.
Harder, M., 2016. Exchangeability of Copulas (Doctoral dissertation), Universität Ulm.
Herath, H., Herath, T., 2011. Copula-based actuarial model for pricing cyber-insurance policies. Insur. Markets Companies: Anal. Actuar. Comput. 2 (1), 7–20.
Hogg, R., McKean, J., Craig, A., 2005. Introduction To Mathematical Statistics. Pearson Education International, New Jersey.
Imperva, 2016. DDoS downtime cost calculator. Retrieved December 15, 2016, from https://lp.incapsula.com/ddos-downtime-cost-calculator.html.
Jacobs, J., 2014. Analyzing Ponemon cost of data breach. Retrieved December 11, 2016, from http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/.
Joe, H., 1996. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. Distrib. Fixed Marginals Relat. Topics 28, 120–141.
Joe, H., Xu, J., 1996. The estimation method of inference functions for margins of multivariate models. Technical Report 166. Department of Statistics, University of British Columbia.
Jorion, P., 2007. Value-At-Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York.
Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M., 2008. Modern Actuarial Risk Theory. Springer-Verlag, Heidelberg.
Kleiber, C., Kotz, S., 2003. Statistical Size Distributions in Economics and Actuarial Sciences. John Wiley & Sons, New Jersey.
KPMG, 2016. Cyber Insurance: Are insurers finding growth or looking for trouble? Retrieved January 5, 2017, from https://assets.kpmg.com/content/dam/kpmg/us/pdf/cyber-insurance-whitepaper.pdf.
Kurowicka, D., Cooke, R., 2004. Distribution - free continuous Bayesian belief nets. In: Fourth International Conference on Mathematical Methods in Reliability Methodology and Practice.
McNeil, A., Frey, R., Embrechts, P., 2005. Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press, New Jersey.
Moscadelli, M., 2004. The modelling of operational risk: Experience with the analysis of the data collected by the Basel Committee. Available at SSRN. Retrieved April 17, 2017, from https://ssrn.com/abstract=557214.
Mukhopadhyay, A., Chatterjee, S., Saha, D., Mahanti, A., Sadhukhan, S., 2013. Cyber-risk decision models: To insure IT or not? Decis. Support Syst. 56 (1), 11–26.
National Conference of State Legislative (NCSL), 2016. Security breach notifications laws. Retrieved June 6, 2016, from http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx.
Ogut, H., Raghunathan, S., Nirup, Menon., 2011. Cyber security risk management: Public policy implications of correlated risk, imperfect ability to prove loss, and observability of self-protection. Risk Anal. 31 (3), 497–512.
Panjer, H., 2006. Operational Risk: Modeling Analytics. John Wiley & Sons, New Jersey.
Peng, C., Xu, M., Xu, S., Hu, T., 2018. Modeling multivariate cybersecurity risks. J. Appl. Stat. 1–23.
Ponemon Institute LLC., 2016. Cost of data breach study. Traverse City: Ponemon Institute LLC. Retrieved December 15, 2016, from http://www-03.ibm.com/security/data-breach/.
Privacy Rights Clearinghouse. (PRC), 2016. Data breaches. Retrieved January 5, 2017, from https://www.privacyrights.org/data-breaches.
Rasoulian, S., Gregorie, Y., Legoux, R., Senecal, S., 2017. Service crisis recovery and firm performance: insights from information breach announcements. J. Acad. Marketing Sci. 1–18.
Romanosky, S., Ablon, L., Kuehn, A., Jones, T., 2017. Content analysis of cyber insurance policies: How do carriers write policies and price cyber risk? Available at SSRN. Retrieved March 21, 2018, from https://ssrn.com/abstract=2929137.
Rosenberg, J., Schuermann, T., 2006. A general approach to integrated risk management with skewed, fat-tailed risks. J. Financ. Econom. 79 (3), 569–614.
Savu, C., Trede, M., 2010. Hierarchies of Archimedean copulas. Quant. Finance 10 (3), 295–304.
Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT 10 (1), 33–60.
Scheffer, M., Weiss, G.N., 2017. Smooth nonparametric Bernstein vine copulas. Quant. Finance 17 (1), 139–156.
Shah, A., 2016. Pricing and risk mitigation analysis of a cyber liability insurance using Gaussian, t and Gumbel copulas - A case for cyber risk index. In: Canadian Economic Association (CEA) Ottawa Meetings Paper.
Shevchenko, P.V., 2010. Implementing loss distribution approach for operational risk. Appl. Stoch. Models Bus. Ind. 26, 277–307.
Shevchenko, P.V., 2011. Modelling Operational Risk using Bayesian Inference. Springer, Heidelberg.
Sklar, A., 1959. Fonctions de repartition a n dimensions et leurs marges. Inst. Statist. Univ Paris 8, 229–231.
Smirnov, N., 1948. Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. 19 (2), 279–281.
Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 (2), 307–333.
Wang, S., 1998. Aggregation of correlated risk portfolios: models and algorithms. Proc. Casualty Actuar. Soc. 85 (163), 848–939.
Wheatley, S., Maillart, T., Sornette, D., 2016. The extreme risk of personal data breaches and the erosion of privacy. Eur. Phys. J. B 89 (1), 7.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1 (6), 80–83.
World Economic Forum, 2016. The Global Risks Report 11th Edition. Geneva: World Economic Forum. Retrieved January 5, 2017, from www3.weforum.org/docs/Media/TheGlobalRisksReport2016.pdf.
Xu, M., Hua, L., 2017. Cybersecurity Insurance: Modeling and Pricing. Society of Actuaries (SoA, Illinois.
Zuur, A., Leno, E., Walker, N., Saveliev, A., Smith, G., 2009. Zero-truncated and zero-inflated models for count data. In: Mixed Effects Models and Extensions in Ecology with R. Springer, New York, pp. 261–293.