Annals of Actuarial Science (2025), 1–25
doi:10.1017/S1748499525100109

ORIGINAL RESEARCH PAPER

Cyber breach risk modeling for insurance: capturing temporal and cross-group dependence

Yijia Li[1], Xuanhe Wang[2], Peng Zhao[3] and Taizhong Hu[1]

[1]Department of Statistics and Finance, University of Science and Technology of China, Hefei, China; [2]School of Finance, Dongbei University of Finance and Economics, Dalian, China; and [3]School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, China
Corresponding author: Xuanhe Wang; Email: wxhmath@163.com

(Received 14 March 2025; revised 25 June 2025; accepted 28 July 2025)

Abstract

Cyber breaches pose a significant threat to both enterprises and society. Analyzing cyber breach data is essential for improving cyber risk management and developing effective cyber insurance policies. However, modeling cyber risk is challenging due to its inherent characteristics, including sparsity, heterogeneity, heavy tails, and dependence. This work introduces a cluster-based dependence model that captures both temporal and cross-group dependencies, providing a more accurate representation of multivariate cyber breach risks. The proposed framework employs a cluster-based kernel approach to model breach severity, effectively handling heterogeneity and extreme values, while a copula-based method is used to capture multivariate dependence. Our findings, validated through both empirical and synthetic studies, demonstrate that the proposed model effectively captures the statistical characteristics of multivariate cyber breach risks and outperforms commonly used models in predictive accuracy. Furthermore, we show that our approach can enhance cyber insurance pricing by generating more profitable insurance contracts.

Keywords: Copula; heavy-tail risks; heterogeneity; Rosenblatt transform; sparsity

1. Introduction

Data breaches have become a substantial risk to citizens and enterprises. According to the 18th Annual Survey of Emerging Risks (Key findings, 2025), cyber risk has remained the top 5 risk over the past 5 years. Based on the Identity Theft Resource Center's (ITRC) 2024 Annual Data Breach Report, there were 3,158 data compromises in the U.S. during 2024 (Identity Theft Resource Center, 2024). The number of victim notices issued surged dramatically. In 2024, over 1.7 billion notices were sent to individuals affected by data breaches, marking a 211% increase from the 419 million notices in 2023. The financial impact of these breaches is substantial. IBM's 2024 Cost of a Data Breach Report indicates that the global average cost of a data breach reached $4.88 million in 2024, a 10% increase from the previous year and the highest figure recorded to date (IBM Security, 2024).

The significant impact of cyber data breaches has driven extensive research into modeling cyber breach risks. For example, Maillart and Sornette (Maillart & Sornette, 2010) analyzed a dataset of cyber breach incidents collected between 2000 and 2008, demonstrating that heavy-tail distributions effectively model personal identity losses per incident. Sen & Borle (2015) explored the factors influencing the frequency of cyber breach incidents, applying theories such as the opportunity theory of crime, institutional anomie theory, and institutional theory. Wheatley et al. (2016) employed extreme value theory to examine a dataset from 2000 to 2015, finding that

breaches occur more frequently in large enterprises compared to smaller ones. Edwards et al. (2016) developed Bayesian models to analyze the temporal trends of data breach incidents, showing that breach sizes follow a log-normal distribution while frequencies adhere to a negative binomial distribution. Eling & Loperfido (2017) applied actuarial modeling to cyber breach incidents, demonstrating that breach sizes can be effectively modeled using a skew-normal distribution. Buckman et al. (2017) studied the intervals between cyber breaches for enterprises that experienced multiple incidents between 2010 and 2016, finding that these intervals vary based on several factors. Buckman et al. (2018) investigated the effects of data breach notification policies on enterprises, employing panel regressions with fixed effects models to examine their impact on repeat breaches. Xu et al. (2018) analyzed aggregate cyber hacking breach incidents, showing that both breach sizes and inter-breach arrival times could be modeled using stochastic processes. Eling & Jung (2018) identified cross-industry dependence in cyber breach-induced financial losses and proposed using copulas to model this dependence. Ikegami & Kikuchi (2020) examined a dataset of cyber breach incidents in Japan, demonstrating that inter-arrival times of breaches follow a negative binomial distribution. Woods et al. (2021) proposed a composite cyber loss model by aggregating multiple parametric distributions such as log-normal, Pareto, Burr, and Weibull, optimized through a particle swarm algorithm. Sun et al. (2021) developed a frequency-severity actuarial model for aggregate enterprise-level breach incident data to support insurance ratemaking and underwriting. Subsequently, Sun et al. (2023) focused on healthcare data breach incidents, formulating a multivariate frequency-severity framework that models frequency with a mixed-effects model and severity with a log-gamma distribution. Wu et al. (2023) developed an Recurrent Neural Network – Long Short-Term Memory (RNN-LSTM) framework to capture multidimensional dependencies in the number of cyber attacks and model the residual tail risks using a peaks-over-threshold approach with a generalized Pareto distribution. Malavasi et al. (2022) employed EVT in the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework to model frequency and severity with covariate effects and tail behavior. One may refer to He et al. (2024)] for a comprehensive review of modeling and management techniques for cyber risk across insurance, computer science, and finance.

However, studies focusing on modeling and predicting data breach risks from the insurer's perspective remain relatively scarce. This scarcity is largely due to the sparse, heterogeneous, and heavy-tailed nature of data breach incidents, which present substantial challenges for both modeling and prediction. These challenges are further compounded by the complex multivariate dependence among breach risks. While existing studies offer valuable insights into cyber risk modeling, they often treat entities independently by focusing solely on temporal dependence (Sun et al., 2023), or on the aggregate outcomes (Eling & Loperfido, 2017; Xu et al., 2018; He et al., 2024). Such approaches fail to capture the interconnected and evolving nature of cyber risks across sectors and time – critical considerations from the insurer's standpoint. Our work addresses this gap by jointly modeling the temporal and cross-sectional dependence of breach events across industries, with clustering incorporated to account for heterogeneity. This joint modeling framework enables insurers to better quantify systemic exposure, assess tail risks, and allocate capital more effectively. In particular, modeling how risks evolve over time and propagate across sectors is essential for dynamic risk scoring and premium pricing tasks that are inadequately addressed in much of the existing literature. From the insurer's perspective, the proposed framework is suitable for modeling multivariate data breach risks in the context of ratemaking. This framework integrates clustering and transformation techniques to manage heterogeneity and skewness, and employs S-vine copulas to simultaneously capture temporal and cross-group dependence. Both empirical analyses and synthetic data experiments confirm the effectiveness of our approach. In the insurance ratemaking context, the proposed model consistently outperforms existing methods in terms of predictive accuracy and interpretability.

The remainder of the article is structured as follows. Section 2 conducts an exploratory data analysis and reviews the statistical preliminaries used in this paper. Section 3 introduces the novel

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Table 1. The resulting groups from PRC dataset categories

Group	Industry	Breach type	Group	Industry	Breach type	Group	Industry	Breach type
G_1	BSF	HACK	G_2	BSF	DISC	G_3	BSF	INSD
G_4	BSF	PPS	G_5	BSF	OTHER	G_6	BSO	HACK
G_7	BSO	DISC	G_8	BSO	INSD	G_9	BSO	PPS
G_{10}	BSO	OTHER	G_{11}	BSR	HACK	G_{12}	BSR	DISC
G_{13}	BSR	INSD	G_{14}	BSR	PPS	G_{15}	BSR	OTHER
G_{16}	EDU	HACK	G_{17}	EDU	DISC	G_{18}	EDU	INSD
G_{19}	EDU	PPS	G_{20}	EDU	OTHER	G_{21}	MED	HACK
G_{22}	MED	DISC	G_{23}	MED	INSD	G_{24}	MED	PPS
G_{25}	MED	OTHER	G_{26}	OTHER	HACK	G_{27}	OTHER	DISC
G_{28}	OTHER	INSD	G_{29}	OTHER	PPS	G_{30}	OTHER	OTHER
G_{31}	UNKN	ALL						

multivariate dependence model. Section 4 discusses the estimation and prediction methodologies. Section 5 applies the proposed framework to a real-world dataset for a case study. Section 6 evaluates the model's performance in the context of insurance pricing. Finally, Section 7 concludes the study and presents discussions on the limitations and future directions. Additional analysis using empirical and synthetic data is provided in the supplementary material.

2. Exploratory data analysis and preliminaries

2.1 Exploratory data analysis

In our study, we utilize two publicly available datasets: the Privacy Rights Clearinghouse (PRC) (Privacy Rights Clearinghouse, 2025) and the ITRC. The PRC dataset spans from January 1, 2010, to December 31, 2018, and was publicly accessible. Similarly, the ITRC dataset covers the period from January 1, 2023, to December 31, 2024, and is also publicly available. In addition, a synthetic data study is conducted to further validate our approach. For demonstration purposes, we primarily use the PRC dataset due to its broader range of breach types, while the analysis of the ITRC and synthetic data is presented in the supplementary material.

The PRC dataset was categorized into 8 industries and 8 breach types. The 8 industries are businesses-financial and insurance services (BSF); businesses-retail/merchant including online retail (BSR); businesses-other (BSO); healthcare, medical providers and medical insurance services (MED); educational institutions (EDU); government, military (GOV); nonprofits (NGO); and Unknown (UNKN). Among these industries, we merge GOV and NGO into a new industry dubbed OTHER because they both are very sparse. The 8 breach types are: fraud involving debit and credit cards not via hacking (CARD); hacked by an outside party or infected by malware (HACK); insider-employee, contractor or customer (INSD); physical (PHYS); port (PORT); stationary computer loss (STAT); unintended disclosure not involving hacking (DISC); and Unknown (UNKN). Among these 8 types, we merge PHYS, PORT, and STAT as a new PPS type because they are related to physical activities; we merge CARD and UNKN into a new OTHER type because both contain few nonzero values. This leads to a total of 31 groups, as shown in Table 1.

For ratemaking purposes, we follow the standard insurance policy by using 6 months as the time period. This leads to 31 groups with 18 periods, namely $\{y_{i,t} | 1 \leq i \leq 31, 1 \leq t \leq 18\}$. For modeling and prediction purposes, we split the data into two parts: the first 15 time periods ($t \leq 15$) are used as the in-sample data for model fitting, and the remaining 3 periods ($16 \leq t \leq 18$) are used as the out-of-sample data for assessing the prediction performance. Table 2 presents the summary

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Table 2. Summary statistics of the various groups using the PRC data, where "SD" means standard deviation and "CV" means coefficient of variation

Group	n	Min	Q_{25}	Median	Q_{75}	Max	Mean	SD	CV
G_1	15	13	17,602	235,373	900,567	92,100,000	13,394,951	29,829,126	2.23
G_2	15	0	864	19,457	71,534	609,462	75,892	156,664	2.06
G_3	15	0	0	877	12,194	1,201,361	107,623	315,523	2.93
G_4	15	0	249	3,000	346,312	3,323,061	400,659	870,444	2.17
G_5	15	0	0	218	4,047	7,002,067	477,744	1,805,197	3.78
G_6	15	2.941×10^5	2.980×10^6	1.441×10^7	1.562×10^8	3.989×10^9	3.538×10^8	1,017,698,709	2.88
G_7	15	0	1,037	2,070	317,100	1.606×10^9	1.079×10^8	414,395,085	3.84
G_8	15	0	0	0	1,038	3,105,232	230,897	798,524	3.46
G_9	15	0	600	662	53,580	5,243,684	398,692	1,344,887	3.37
G_{10}	15	0	0	0	630	100,077,000	6,676,990	25,838,342	3.87
G_{11}	15	0	522,468	3,000,000	30,047,544	103,090,092	18,672,629	29,917,666	1.60
G_{12}	15	0	0	350	19,248	120,119,469	8,025,191	31,009,971	3.86
G_{13}	15	0	0	65	913	30,025	2,211	7,434	3.36
G_{14}	15	0	0	200	1,426	16,872	2,900	5,571	1.92
G_{15}	15	0	0	28	12,712	68,000,300	4,674,998	17,525,934	3.75
G_{16}	15	49	73,316	208,889	378,735	1,184,231	346,580	399,775	1.15
G_{17}	15	0	18,570	47,630	139,024	432,009	100,187	126,118	1.26
G_{18}	15	0	0	45	3,000	21,000	2,692	5,617	2.09
G_{19}	15	0	621	5,222	25,287	1,020,025	89,335	261,265	2.92
G_{20}	15	0	0	0	326	7,694,016	679,106	2,043,810	3.01
G_{21}	15	2,274	458,004	675,078	8,031,834	95,057,805	10,072,552	24,375,139	2.42
G_{22}	15	140,689	226,706	520,719	659,170	2,871,784	718,935	764,626	1.06
G_{23}	15	0	17,370	29,791	77,894	262,627	69,354	86,311	1.24
G_{24}	15	176,285	788,746	2,405,821	5,446,206	13,353,853	3,797,718	4,046,347	1.07
G_{25}	15	0	0	172	7,423	39,600	7,270	12,172	1.67
G_{26}	15	0	93,912	375,396	1,787,604	21,500,000	2,576,229	5,520,222	2.14
G_{27}	15	0	34,510	113,953	1,177,530	7,852,662	1,363,227	2,490,565	1.83
G_{28}	15	0	114	3,742	40,582	28,210,000	1,898,707	7,278,855	3.83
G_{29}	15	0	12,182	65,400	277,356	5,000,865	532,825	1,298,381	2.44
G_{30}	15	0	0	0	3,973	700,000	56,804	180,079	3.17
G_{31}	15	0	0	0	2,220	3,180,426	219,929	819,253	3.73

statistics of the in-sample data. We observe that most groups exhibit *sparsity* (e.g., 13 out of 31 groups have a 25th percentile equal to 0; and 5 of them have a median equal to 0). We observe that most groups exhibit *skewness* because their mean and median are significantly different. We observe *variability* because the standard deviation is always much larger than the mean, which is also reflected in the fact that the coefficient of variation ranges from 1.06 to 3.87. We observe *extreme values*, for example, there are some extremely large values in groups G_6 and G_7, which means that a heavy-tailed distribution is needed for accommodating these extreme values.

We examine the temporal and cross-group dependence by using Spearman's ρ and Kendall's τ. For the temporal dependence, 5.71% of the $|\rho|$'s are greater than 0.7, 41.90% are greater than 0.5, and 84.76% are greater than 0.3. In terms of the $|\tau|$'s, 13.33% are greater than 0.5 and 69.52% are greater than 0.3. Therefore, there is a strong temporal dependence in the data.

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

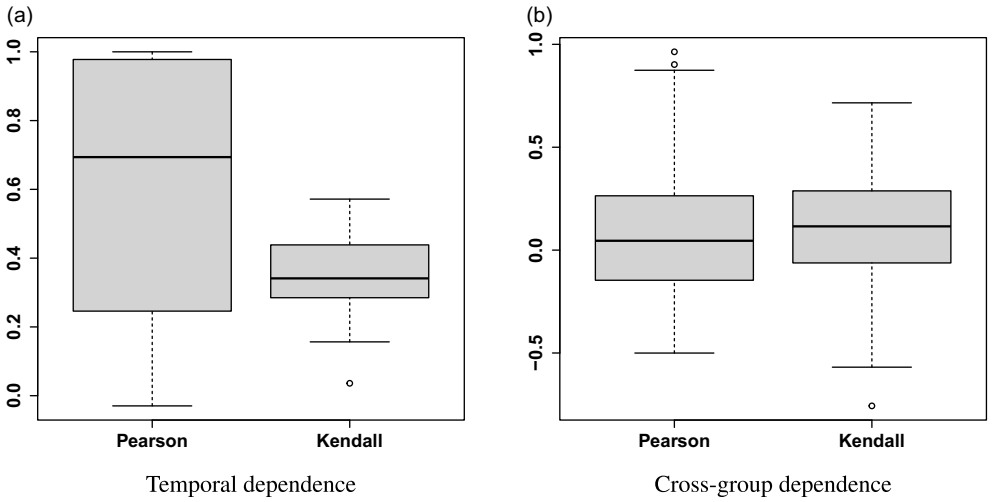

Figure 1. Boxplots of Pearson's ρ and Kendall's τ for temporal and cross-group dependence.

For the cross-group dependence, 2.36% of the $|\rho|$'s are greater than 0.7, 14.19% are greater than 0.5, 42.15% are greater than 0.3, and 80.43% are greater than 0.1. In term of the $|\tau|$'s, 4.52% are greater than 0.5 and 29.03% are greater than 0.3. Therefore, the cross-group dependence is non-negligible. In summary, there exists both *temporal and cross-group dependence* in the data. To further visualize these two correlations, Figure 1a and b show the boxplots of Pearson and Kendall correlation coefficients in temporal and cross-group dependence of breach sizes, respectively. We observe the presence of temporal and cross-group correlations, with the temporal dependence exhibiting strong positive correlations. These statistical characteristics are accounted for in the following modeling process.

2.2 Preliminaries

In order to model the temporal and cross-group dependence of group-level multivariate data breach incidents time series, we propose using copulas because they are widely used in modeling complex multivariate dependence (Joe, 2014). A theoretical foundation of copulas is Sklar's Theorem (Sklar, 1959), which says that a multivariate distribution can be represented as a certain composition of multiple univariate margins. Specifically, a d-variate distribution F with univariate marginal distributions $F_1, \ldots, F_d$ can be decomposed as

$$F(x) = C(F_1(x_1), \ldots, F_d(x_d)), \; x \in \mathbb{R}^d,$$

where the copula is a distribution function $C{:}[0,1]^d \to [0,1]$ with $U(0,1)$ margins. That is, a copula characterizes the dependence among the d random variables. If the marginal distributions, namely the F_i's, are continuous, the copula C is unique.

2.2.1 R-vine

A *regular vine* (R-vine) is a graphical structure for decomposing a multivariate distribution to a sequence of nested trees (Bedford & Cooke, 2001). Specifically, an R-vine with d elements is an acyclic graph which consists of $d-1$ trees, denoted by $T_1, \ldots, T_{d-1}$, where each tree has a set N_i of nodes and a set E_i of edges with $i = 1, \ldots, d-1$. Such a structure $\mathcal{V} = (N_i, E_i)_{i=1}^{d-1}$ is an R-vine if the following conditions hold:

- T_1 is the first tree with d nodes in node set N_1 and $d-1$ edges in edge set E_1;

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

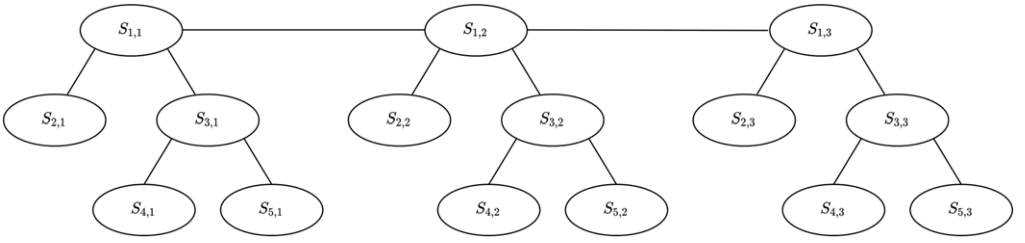

Figure 2. The first tree of a 5-dimensional S-vine with the first three time points.

- for $i = 2, \ldots, d-1$, nodes in T_i are edges in T_{i-1}, namely $N_i = E_{i-1}$;
- (proximity condition) for $i = 2, \ldots, d-1$, if two nodes $m, n \in N_i$ are connected by an edge in E_i, then the corresponding edges m, n in T_{i-1} share a common node.

In an R-vine copula, each edge is associated with a bivariate copula which accommodates the dependence or conditional dependence between a pair of variables. The density of an R-vine copula is the product of the densities of the linking pair copulas and the marginal densities, that is,

$$c(x) = \prod_{k=1}^{d-1} \prod_{e \in E_k} c_{a_e, b_e | D_e}(u_{a_e|D_e}, u_{b_e|D_e} | \boldsymbol{u}_{D_e}) \prod_{j=1}^{d} f_j(x_j),$$

where f_j represents the marginal density, D_e is the conditioned set of edge e, $c_{a_e,b_e|D_e}$ represents the conditional bivariate copula density, $u_{a_e|D_e} := C_{a_e|D_e}(F_{a_e}(x_{a_e})|\boldsymbol{F}_{D_e})$, $\boldsymbol{F}_{D_e} := (F_l(x_l))_{l \in D_e}$ is a subvector of $(F_1(x_1), \ldots, F_d(x_d))$, and

$$C_{a_e|D_e}(F_{a_e}(x_{a_e})|\boldsymbol{F}_{D_e}) = \frac{\partial C_{a_e,b_e|D_e}(F_{a_e|D_e}, F_{b_e|D_e})}{\partial F_{b_e|D_e}},$$

$$C_{b_e|D_e}(F_{b_e}(x_{b_e})|\boldsymbol{F}_{D_e}) = \frac{\partial C_{a_e,b_e|D_e}(F_{a_e|D_e}, F_{b_e|D_e})}{\partial F_{a_e|D_e}}.$$

2.2.2 S-vine

A stationary vine (S-vine) copula, as discussed by Nagler et al. (2022), allows for arbitrary R-vines in the cross-sectional structure and connects two cross-sectional trees at arbitrary variables. It extends the other two well-known stationary vine copula models for modeling multivariate time series: D-vine (Smith, 2015) and M-vine of Beare & Seo (2015).

The S-vine selects the edge with the highest correlation between the time points. Specifically, let $\mathcal{V} = (N_k, E_k)_{k=1}^{mn-1}$ be a regular vine on $\{1, \ldots, m\} \times \{1, \ldots, n\}$, and the regular vine $\mathcal{V}$ is called a *S-vine* if its vine structure does not change with time. More specifically, let

$$\mathcal{V}_{t,t+w} = (N_{k,t}, E_{k,t})_{k=1}^{(w+1)m-1}$$

be a vine on $\{1, \ldots, m\} \times \{t, \ldots, t+w\}$,

$$\mathcal{V}_{s,s+w} = (N_{k,s}, E_{k,s})_{k=1}^{(w+1)m-1}$$

be a vine on $\{1, \ldots, m\} \times \{s, \ldots, s+w\}$, and $\mathcal{V}_{t,t+w}$ and $\mathcal{V}_{s,s+w}$ are the restriction of vine $\mathcal{V}$ (Beare & Seo, 2015). $\mathcal{V}$ on $\{1, \ldots, m\} \times \{1, \ldots, n\}$ is an S-vine if for all $w = 0, \ldots, n-1$, $1 \le t \le n-w$, $k = 1, \ldots, (w+1)m-1$, and edges $e \in E_{k,t}$, there is an edge $e' \in E_{k,s}$ such that $e = e' + (0, t-s)$.

Figure 2 depicts the first tree of a 5-dimensional S-vine at the first three time points. We observe that the edges among $(S_{1,1}, S_{1,2}, S_{1,3})$ model the temporal dependence. Specifically, the

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

edges between nodes $(S_{1,1}, S_{1,2})$ and $(S_{1,2}, S_{1,3})$ specify the temporal dependence between two adjacent time points. For describing the cross-group dependence at time 1, the structure consists of variables $S_{1,1}, S_{2,1}, S_{3,1}, S_{4,1}$ and $S_{5,1}$. The same structure applies to the other time points, while noting that the same copula is used for the edges with the same structure. For example, for modeling temporal dependence, the edge between $(S_{1,1}, S_{1,2})$ and the edge between $(S_{1,2}, S_{1,3})$ are modeled by the same bivariate copula with the same parameter(s). Similarly, for modeling cross-group dependence, the edge between $(S_{1,1}, S_{3,1})$, the edge between $(S_{1,2}, S_{3,2})$, and the edge between $(S_{1,3}, S_{3,3})$ are all modeled with the same copula with the same parameter(s). The edge between $(S_{1,1}, S_{2,1})$, the edge between $(S_{1,2}, S_{2,2})$, and the edge between $(S_{1,3}, S_{2,3})$ are all assumed to have the same copula.

To further clarify the construction of this the 5-dimensional S-vine, let $\mathbf{S} = (S_{1,1}, \ldots, S_{5,1}, \ldots, S_{1,3}, \ldots, S_{5,3})^\top$ represents the vector of variables at the three time points. For the S-vine copula representation, the joint distribution of $\mathbf{S}$ is decomposed into its univariate marginal densities $f_i(s_{i,t})$ and a set of copula densities that captures the dependence structure among them. Denoting $u_{i,t} = F_i(s_{i,t})$, the copula density of $\mathbf{S}$ can be represented as

$$c(s_{1,1}, \ldots, s_{5,1}, s_{1,2}, \ldots, s_{5,2}, s_{1,3}, \ldots, s_{5,3}) = \prod_{e \in \mathcal{T}_1} c_{a_e, b_e}(u_{a_e}, u_{b_e}) \prod_{k=2}^{14} \prod_{e \in \mathcal{T}_k} c_{a_e, b_e | D_e} \left(C_{a_e|D_e}(u_{a_e}|\mathbf{u}_{D_e}), C_{b_e|D_e}(u_{b_e}|\mathbf{u}_{D_e}) \right) \prod_{i=1}^{5} \prod_{t=1}^{3} f_i(s_{i,t}),$$

where $\mathcal{T}_k$ is the set of edges in tree k, and each $c_{a_e, b_e|D_e}$ is a conditional bivariate copula density associated with edge e, conditioned on the variable set D_e. For example, the pair $(S_{1,1}, S_{1,2})$ captures temporal dependence between two adjacent time points, while pairs like $(S_{1,t}, S_{j,t})$ for $j \neq 1$ capture cross-group dependence. For more discussions on the S-vine copula, please refer to Nagler et al. (2022).

3. A multivariate dependence model

Let $y_{i,t}$ denote the breach size of group i at time t with $y_{i,t} = 0$ meaning no beach incidents for group i at time t, m the number of groups, and n the length of time periods. The breach incidents can be represented as an m-variate time series

$$\{y_{i,t} | 1 \leq i \leq m, 1 \leq t \leq n\}.$$

As discussed in Section 2, cyber breach sizes exhibit skewness and heavy tails. We propose using a proper transformation $g(\cdot)$ to reduce the skewness and the high variability. We call the transformed data the *severity* time series, which is denoted by

$$\{s_{i,t} = g(y_{i,t}) | 1 \leq i \leq m, 1 \leq t \leq n\}.$$

To model the distribution of severity, for the PRC data, we observe that there are many small values along with extremely large ones. This prompts us to use a semi-parametric model to estimate the marginal distribution of cyber breach sizes. The key idea is to use a nonparametric distribution to fit small breach severity while using a parametric distribution to fit the large breach severity (Scarrott, 2016; McNeil et al., 2015). Specifically, we propose the following kernel density to model the nonparametric distribution of small breach severity, where $s_{i,t} \leq \mu_i$, and μ_i is an unknown parameter that needs to be estimated:

$$h_i(s; s_{i,t}, \lambda_i) = \frac{1}{nm\lambda_i} \sum_{t=1}^{n} \sum_{i=1}^{m} K\left(\frac{s - s_{i,t}}{\lambda_i} \right), \tag{1}$$

where λ_i is the bandwidth that can be determined by the cross-validation likelihood method (Wand & Jones, 1994), and $K(x) = \exp(-x^2)$ is the kernel function. For the large breach severity, namely $s > \mu$, we propose using the *Generalized Pareto distribution* (GPD) because it is

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

directly related to the extreme value theory (Scarrott, 2016). The GPD distribution function can be written as

$$G_i(s|\mu,\sigma_\mu,\xi_i)=\begin{cases}1-\left[1+\xi_i\left(\frac{s-\mu_i}{\sigma_{\mu_i}}\right)\right]_+^{-1/\xi_i}, & \xi_i\neq 0,\\ 1-\exp\left[-\left(\frac{s-\mu_i}{\sigma_{\mu_i}}\right)_+\right], & \xi_i=0,\end{cases} \tag{2}$$

where $s_+=\max(s,0)$, μ_i is the threshold, $\sigma_{\mu_i}>0$ and ξ_i are, respectively, the scale and shape parameters. If $\xi_i<0$, the support is $\mu_i<s<\mu_i-\sigma_{\mu_i}/\xi_i$; otherwise, the support is unbounded from above. Combining the nonparametric distribution and GPD, we have the following flexible mixed model:

$$F_i(s|\boldsymbol{\eta})=\begin{cases}(1-\phi_{\mu_i})H_i(s|\boldsymbol{\eta}_i), & s\le\mu_i,\\ (1-\phi_{\mu_i})+\phi_{\mu_i}G(s|\boldsymbol{\eta}_i), & s>\mu_i,\end{cases} \tag{3}$$

where $\boldsymbol{\eta}_i$ is the parameter vector, $H_i(\cdot)$ is the distribution with density given by Eq. (1), ϕ_{μ_i} is the proportion of breach sizes above the threshold μ_i, and $G_i(\cdot)$ is the GPD in Eq. (2). Please note that, given the relatively limited size of the dataset, we assume that the marginal distributions are time-invariant, that is, a stationarity assumption imposed in the modeling process. This assumption is important for effective modeling and prediction, as it ensures that the marginal distributions remain stable over time.

Let $\boldsymbol{S}_t=(s_{1,t},\ldots,s_{m,t})$, $t=1,\ldots,n$, and we assume sequence $\boldsymbol{S}_t$ is stationary with Markov order p. Consider a stationary sequence $\boldsymbol{S}_1,\ldots,\boldsymbol{S}_T$ which takes the values in $\mathbb{R}^m$ with Markov order p, then, the distribution of all finite dimensions can be uniquely determined by the joint distribution of $(\boldsymbol{S}_1,\ldots,\boldsymbol{S}_{p+1})$. The dependence of the variables can be captured by an S-vine model on the $m\times(p+1)$ array of nodes $\{1,\ldots,m\}\times\{1,\ldots,p+1\}$. Given the S-vine tree structure and copula families of the edges in the identified S-vine, we employ the likelihood-based method to estimate their parameters.

The total log-likelihood function of $\boldsymbol{S}_t$, $t=1,\ldots,n$, can be represented as

$$l(\boldsymbol{\theta},\boldsymbol{\eta})=\sum_{i=1}^{m}\sum_{t=1}^{n}\log f_i(s_{i,t}|\boldsymbol{\eta}_i)+\sum_{k=1}^{m(p+1)-1}\sum_{e\in T_k}\log c_{a_e,b_e|D_e}(z_{a_e|D_e},z_{b_e|D_e}|\boldsymbol{\eta}_{[e]},\boldsymbol{\theta}_{[e]}), \tag{4}$$

where m denotes the number of considered industries, n denotes the length of the time series used to fit the model, and p denotes the order of the Markov chain. $f_i(\cdot\,|\boldsymbol{\eta}_i)$ represents the marginal density of group i with parameter vector $\boldsymbol{\eta}_i$. The set $\{a_e,b_e\}$ includes the two conditioning nodes associated with edge e, and D_e is the conditioned set corresponding to edge e. $\boldsymbol{\theta}_{[e]}$ are the copula parameters associated with edge e. $z_{a_e|D_e}$ is the conditional distribution of $F_{i_{a_e}}(s_{a_e};\boldsymbol{\eta}_{[e]})$ given $F_{i_{D_e}}(s_{D_e};\boldsymbol{\eta}_{[e]})$, that is, of $C_{a_e|D_e}(z_{a_e}|z_{D_e})$, where i_{a_e} and i_{D_e} denote the group indexes related to nodes a_e and D_e, respectively, $s_{D_e}:=(s_{i,t})_{(i,t)\in D_e}$, and $\boldsymbol{\eta}_{[e]}$ denotes the set of marginal parameters associated with all nodes involved in edge e, including those in the conditioned set D_e.

Note that, we have

$$c_{a_e,b_e|D_e}(z_{a_e|D_e},z_{b_e|D_e}|\boldsymbol{\eta}_{[e]},\boldsymbol{\theta}_{[e]})=\begin{cases}\Delta_{w_2}^{w_2'}\Delta_{w_1}^{w_1'}C_{a_e,b_e|D_e}(v_1,v_2), & s_{a_e}=s_{b_e}=0,\\ \Delta_{w_2}^{w_2'}\dfrac{\partial C_{a_e,b_e|D_e}(F_{a_e|D_e}(s_{a_e}|\boldsymbol{s}_{D_e}),v_2)}{\partial F_{a_e|D_e}(s_{a_e}|\boldsymbol{s}_{D_e})}, & s_{a_e}>0,s_{b_e}=0,\\ \Delta_{w_1}^{w_1'}\dfrac{\partial C_{a_e,b_e|D_e}(v_1,F_{b_e|D_e}(s_{b_e}|\boldsymbol{s}_{D_e}))}{\partial F_{b_e|D_e}(s_{a_e}|\boldsymbol{s}_{D_e})}, & s_{a_e}=0,s_{b_e}>0,\\ c_{a_e,b_e|D_e}(F_{a_e|D_e}(s_{a_e}|\boldsymbol{s}_{D_e}),F_{b_e|D_e}(s_{b_e}|\boldsymbol{s}_{D_e})), & s_{a_e}>0,s_{b_e}>0,\end{cases}$$

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

where $w_1' = F_{a_e|D_e}(0|\boldsymbol{s}_{D_e})$, $w_1 = F_{a_e|D_e}(0^-|\boldsymbol{s}_{D_e})$ is the left-hand limit of $F_{a_e|D_e}$ at 0, $w_2' = F_{b_e|D_e}(0|\boldsymbol{s}_{D_e})$, $w_2 = F_{b_e|D_e}(0^-|\boldsymbol{s}_{D_e})$, v_i is the index of difference, $C_{a_e,b_e|D_e}$ and $c_{a_e,b_e|D_e}$ are the distribution function and density function of a bivariate copula, respectively. Specifically, the conditional distribution can be recursively represented as

$$F_{a_e|D_e}(s_{a_e}|\mathbf{s}_{D_e}) = \begin{cases} \Delta_{w_3}^{w_3'} C_{a_e,b_e'|D_e'}(F_{a_e|D_e'}(s_{a_e}|\mathbf{s}_{D_e'}), v_3), & s_{b_e'} = 0, \\ \dfrac{\partial C_{a_e,b_e'|D_e'}(F_{a_e|D_e'}(s_{a_e}|\mathbf{s}_{D_e'}), F_{b_e'|D_e'}(s_{b_e'}|\,\mathbf{s}_{D_e'}))}{\partial F_{b_e'|D_e'}(s_{b_e'}|\mathbf{s}_{D_e'})}, & s_{b_e'} > 0, \end{cases}$$

where $b_e' \in D_e$, $D_{e'} = D_e \backslash b_{e'}$, $w_3' = F_{b_e'|D_e'}(0|\mathbf{s}_{D_e'})$, and $w_3 = F_{b_e'|D_e'}(0^-|\mathbf{s}_{D_e'})$. And

$$F_{b_e|D_e}(s_{b_e}|\mathbf{s}_{D_e}) = \begin{cases} \Delta_{w_4}^{w_4'} C_{a_e',b_e|D_e''}(v_4, F_{b_e|D_e''}(s_{b_e}|\mathbf{s}_{D_e''})), & s_{a_e'} = 0, \\ \dfrac{\partial C_{a_e',b_e|D_e''}(F_{a_e'|D_e''}(s_{a_e'}|\mathbf{s}_{D_e''}), F_{b_e|D_e''}(s_{b_e}|\,\mathbf{s}_{D_e''}))}{\partial F_{a_e'|D_e''}(s_{a_e'}|\mathbf{s}_{D_e''})}, & s_{a_e'} > 0, \end{cases}$$

where $a_e' \in D_e$, $D_{e''} = D_e \backslash a_{e'}$, $w_4' = F_{a_e'|D_e''}(0|\mathbf{s}_{D_e''})$, and $w_4 = F_{a_e'|D_e''}(0^-|\mathbf{s}_{D_e''})$. The conditional distribution of $z_{a_e|D_e}$ can be recursively written as

$$z_{a_e|D_e} = \frac{\partial C_{a_e,b_e'|D_e'}(z_{a_e|D_e'}, z_{b_e'|D_e'})}{\partial z_{b_e'|D_e'}},$$

where $b_e' \in D_e$ and $D_{e'} = D_e \backslash b_{e'}$.

4. Estimation and prediction

4.1 Stepwise maximum likelihood estimation

For parameter estimation, we propose using the *Inference Functions for Marginals (IFM)* method because it can efficiently reduce the computational time, which has been popularly used in the literature (Joe, 2014). The IFM method has two steps: (i) estimating parameters $\boldsymbol{\eta}$ of the marginal distribution, which equates to the severity distribution in this paper; and, (ii) estimating parameters $\boldsymbol{\theta}$ in the S-vine model via the fitted marginal distribution.

We can represent the joint log-likelihood function of an S-vine copula model after estimating the marginal parameters as

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{m(p+1)-1} \sum_{e \in T_k} \log c_{a_e,b_e|D_e}(z_{a_e|D_e}, z_{b_e|D_e}|\hat{\boldsymbol{\eta}}_{[e]}, \boldsymbol{\theta}_{[e]}).$$

Since there are many parameters for the S-vine copula, we propose using the stepwise maximum likelihood estimation approach to estimate parameters of $\boldsymbol{\theta}$, namely by sequentially estimating the bivariate copula models corresponding to each edge of the vine structure from the first tree to the last tree. Specifically, corresponding to edge $e' \in T_k$, we can obtain the parameters of the copula from

$$\hat{\boldsymbol{\theta}}_{[e']} = \arg\max_{\boldsymbol{\theta}_{[e']}} \sum_{e \sim e'} \log c_{[e]}(\hat{z}_{a_e|D_e}, \hat{z}_{b_e|D_e}; \boldsymbol{\theta}_{[e']}),$$

where $e \sim e'$ means edges e and e' are equivalent. For copula selection, the AIC criterion can be used (Joe, 2014). The key insight into determining the S-vine structure is that we should model the strongest dependence as early as possible. This prompts us to use the following two steps to determine the structure of the first tree of the S-vine:

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Algorithm 1. Estimating an S-vine copula model to simultaneously accommodate the cross-group dependence and the temporal dependence in a multivariate time series.

INPUT: Breach severity $\{s_{i,t}|1 \leq i \leq m, 1 \leq t \leq n\}$; a set of candidate pair copulas Ω; Markov order p.

Fit the marginal distribution of severity $s_{i,t}$ via Eq. (3);
Determine the first tree T_1 of the S-vine copula;
for $j = 1, \ldots, m(p+1) - 1$ **do**
  if $j = 1$ **then**
    for The edges of different copulas in T_1 **do**
      For a given copula in Ω, estimate parameters of a set of equivalent edges in T_1, while using the AIC to select pairwise copulas;
    end for
  else
    for The edges of different copulas in T_j **do**
      Fix the estimated copula structures in $T_1, \ldots, T_{j-1}$, and determine the tree structure T_j under the constraint of T_{j-1};
      Estimate parameters of a set of equivalent edges in T_j for a copula in Ω, and select copulas by AIC;
    end for
  end if
end for
return $\{T_1, \ldots, T_{m(p+1)-1}\}$.

OUTPUT: Estimated S-vine copula.

(i) Determine the cross-group structure of the S-vine. The empirical Kendall's $\hat{\tau}_{i,j}, 1 \leq i < j \leq m$ between any two of the m cross-group variables are calculated. We propose using the maximum spanning tree algorithm (Dißmann et al., 2013) to select the cross-group R-vine tree that has the maximum sum of empirical absolute Kendall's τ, namely $\max \sum_{e=\{i,j\} \text{in spanning tree}} |\hat{\tau}_{i,j}|$.

(ii) Identify the temporal vine structure. For this, we propose computing Kendall's τ with lag 1 in between any two groups and selecting the largest $|\tau|$ to specify the temporal dependence for the first tree. That is, we compare all Kendall's τ correlation coefficients between an industry and its lag of time points. The variable with the largest absolute correlation coefficient is the optimal in-/out-vertices to specify temporal dependence. The same approach can be used to deal with the other trees.

After determining the structure and parameters of the first tree, these parameters are fixed and used as input for coping with the second tree. The rest of the tree structures and parameters are determined in a similar way as the first tree. Algorithm 1 presents the detailed estimation process of the S-vine copula model.

This stepwise approach is especially advantageous for high-dimensional copula models like the S-vine, where full joint maximum likelihood estimation becomes computationally prohibitive. We estimate the parameters edge-by-edge following the tree structure, which provides at least a local optimum for each bivariate copula term. While it does not guarantee a global optimum of the full likelihood, it is widely used in vine copula modeling due to its scalability and traceability, especially important given the large number of parameters involved, even in moderate dimensions (Shi & Yang, 2018; Nagler et al., 2022; Sun et al., 2023). A potential limitation of this sequential approach is that it does not guarantee a global optimum and may be sensitive to the ordering of the vine

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Algorithm 2. Simulating the predictive distribution from a fitted S-vine copula model.

INPUT: The fitted S-vine copula model $\mathcal{V}, \mathcal{C}(\mathcal{V})$; historical observations $s_t = (s_{1,t}, \ldots, s_{m,t})$; number of simulations L.

1: $z_{j,t} \leftarrow F_j(s_{j,t})$ for $j = 1, \ldots, m$;
2: $z_t \triangleq (z_{1,t}, z_{2,t}, \ldots, z_{m,t})$;
3: $\boldsymbol{u}_1 \leftarrow R_{\mathcal{C}(\mathcal{V})}(z_t)$;
4: **for** $l = 1, \ldots, L$ **do**
5: Simulate $\boldsymbol{u}_2^{(l)} = \left(u_{1,2}^{(l)}, \ldots, u_{m,2}^{(l)}\right)$ with $u_{j,2}^{(l)} \backsim u(0,1)$ which is the uniform distribution in $(0,1), j = 1, \ldots, m$;
6: $(z_t, z_{t+1}^{(l)}) \leftarrow R_{\mathcal{C}(\mathcal{V})}^{-1}(\boldsymbol{u}_1, \boldsymbol{u}_2^{(l)})$;
7: $s_{j,t+1}^{(l)} \leftarrow F_j^{-1}\left(z_{j,t+1}^{(l)}\right)$ for $j = 1, \ldots, m$;
8: **end for**
9: **return** $s_{t+1}^{(l)} = \left\{s_{1,t+1}^{(l)}, \ldots, s_{m,t+1}^{(l)}\right\}, l = 1, \ldots, L$.

OUTPUT: Simulated predictive samples $s_{t+1}^{(l)}, l = 1, \ldots, L$.

structure, particularly under the simplifying assumption. Nevertheless, it provides a reliable and scalable solution for complex dependence modeling in multivariate time series.

4.2 Predicting the distribution of breach sizes

To predict the distribution of cyber breach incidents in the next time period, we propose using the *Rosenblatt transform* to simulate the predictive distribution based on the developed S-vine copula (Czado & Nagler, 2022).

The purpose of the Rosenblatt transform is to transform a random vector to independent uniforms via certain conditional distributions; correspondingly, the inverse Rosenblatt transform can transform independent uniforms to a vector with a certain joint distribution. Specifically, the Rosenblatt transform of a random vector $Z = (Z_1, \ldots, Z_d)$ in a d-dimensional vine copula model $\mathcal{C}(Z)$ can be represented as

$$u_1 = F(z_1),\ u_2 = C(z_2|z_1),\ \ldots,\ u_d = C(z_d|z_1, \ldots, z_{d-1}),$$

namely $R_{\mathcal{C}(Z)}(z_1, z_2, \ldots, z_d) \triangleq (u_1, u_2, \ldots, u_d)$, where $C(z_i|z_1, .., z_{i-1})$, $i = 2, \ldots, d$ is the conditional distribution of Z_i given $Z_1, \ldots, Z_{i-1}$. The inverse transformation is

$$z_1 = F^{-1}(u_1), z_2 = C^{-1}(u_2|u_1), \ldots, z_d = C^{-1}(u_d|u_1, \ldots, u_{d-1}),$$

namely $R_{\mathcal{C}(Z)}^{-1}(u_1, u_2, \ldots, u_d) \triangleq (z_1, z_2, \ldots, z_d)$.

Algorithm 2 presents the prediction algorithm based on the fitted S-vine copula model.

5. Case study

In this section, we conduct an empirical study using the real dataset obtained from the PRC. The first 15 time periods are used as in-sample data for model fitting, while the remaining 3 periods serve as out-of-sample data to evaluate predictive performance. Additional analyses using the ITRC and synthetic data are provided in the supplementary material.

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Table 3. The AIC and BIC of the clustered severity when fitted as a mixed distribution with different numbers of clusters

Number of clusters	1	2	3	4	5
AIC	6,556.6	6,171.8	6,165.1	**5,994.6**	6,251.3
BIC	6,557.3	6,205.8	6,210.3	**6,049.6**	6,311.9

Table 4. K-means clustering of breach sizes

Cluster	Groups	Label
1	G_1 G_6 G_{11} G_{21} G_{22} G_{24} G_{26}	Extreme
2	G_2 G_4 G_9 G_{16} G_{17} G_{19} G_{23} G_{27} G_{29}	Large
3	G_5 G_7 G_{12} G_{15} G_{20} G_{30} G_{31}	Medium
4	G_3 G_8 G_{10} G_{13} G_{14} G_{18} G_{25} G_{28}	Small

5.1 Risk group determination and classification

As discussed in Section 2, we observe the variability and skewness existing in the breach sizes. To alleviate the impact of skewness and variability, we perform the square root transform with $s_{i,t} = g(y_{i,t}) = \sqrt{y_{i,t}}$ for $i = 1, \ldots, 31$ and $1 \le t \le 15$. The square root transformation is chosen for the following reasons: (i) *Variance Stabilization*. It reduces the range of values, which helps in stabilizing the variance across different levels of the data. This is particularly useful when the data exhibit heteroscedasticity, where the variance is not constant across observations. (ii) *Reduction of Skewness*. The square root transformation compresses the range of large values more than it does for smaller values, which helps in making the distribution more symmetrical. (iii) *Ease of Interpretation*. The square root transformation is less aggressive and retains more of the original data's characteristics, making it easier to interpret the transformed values in the context of the original data.

Note that there exist extremely large variations among the groups; that is, some groups have extremely large breach sizes (e.g., G_6, G_{11}, and G_{21}), while others have small breach sizes (e.g., G_{14}, G_{18}, and G_{25}). To accommodate the heterogeneity between the groups, we use the *K-means clustering algorithm* (Hartigan & Wong, 1979) to classify the 31 groups into 4 clusters. The K-means algorithm is well-suited for this task because it partitions the data into clusters by minimizing the within-cluster variance, thereby ensuring that each cluster contains groups with similar breach size characteristics. This method effectively reduces the complexity and variability within each cluster, allowing for more accurate modeling and analysis. The criteria for determining the number of clusters are AIC and BIC, which are obtained by fitting the training data in each cluster with the mixed distribution in Eq. (3).

Table 3 shows AICs and BICs for the number of clusters, ranging from 1 to 5. We observe that when clustering the 31 groups into 4 clusters, it has the smallest AIC and BIC. Restated, we use 4 clusters for two reasons: (i) The data are sparse and the size of the data is small. If we use too few clusters (e.g., 1), the result cannot accommodate the heterogeneity; if we use too many clusters (e.g., 5), the result leads to non-robust statistical estimates and inferences because of sparsity and small sample sizes. (ii) In practice, especially for insurance ratemaking purposes, groups that have similar characteristics are used to determine risk and premium for their members. Indeed, it is a common practice to cluster them into four categories: extreme, large, medium, and small.

Table 4 presents the clustering result based on the training data. We observe that cluster 1 is labeled as *extreme* risk, containing 7 groups and most of them have the HACK breach type (e.g., G_1, G_6, G_{11}, G_{21}, and G_{26}); cluster 2 is labeled as *large* risk, containing 9 groups; cluster 3 is labeled as *medium* risk, containing 7 groups; and cluster 4 is labeled as *small* risk, containing 8 groups. We fit the mixed distribution of Eq. (3) for each cluster, which

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Table 5. Estimated parameters and their standard errors for the marginal distribution of severity in each cluster, where "Est." represents the estimates and "SE" is the standard error

Cluster	Parameter	λ	μ	σ_μ	ξ	ϕ_μ
1	Est.	33.397	750.567	1,220.549	0.863	0.600
	SE	13.038	0.003	329.207	0.261	0.048
2	Est.	6.643	299.083	299.990	0.427	0.363
	SE	1.4065	0.011	135.299	0.320	0.041
3	Est.	2.839	142.582	204.662	1.702	0.257
	SE	0.333	7.538	95.566	0.548	0.043
4	Est.	1.314	117.781	95.644	1.344	0.183
	SE	0.124	2.876	43.590	0.481	0.035

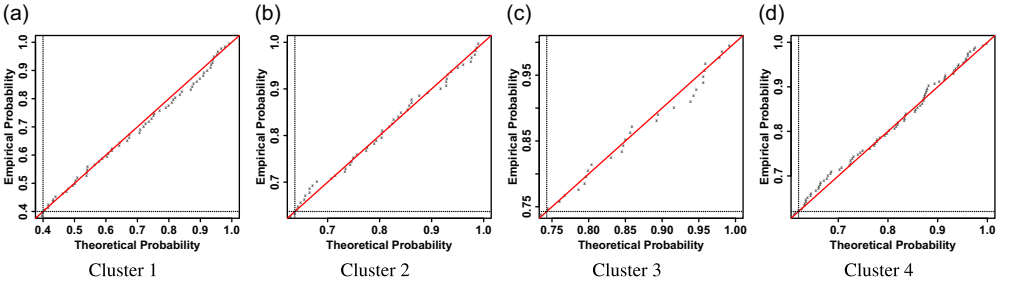

Figure 3. PP-plot of fitted mixed distribution of severity for each cluster.

accommodates the heterogeneity observed in the data. We would like to note that different clustering algorithms can yield varying clustering results. In the supplementary material, we provide two additional commonly used methods: hierarchical clustering (James et al., 2013) and Gaussian mixture models (GMM) (Maugis et al., 2009). As expected, these approaches produce different cluster assignments. Nevertheless, we find that the predictive performance of the K-means method is comparable to that of the hierarchical and GMM approaches. Given its simplicity and natural alignment with four commonly used risk groups in practice, we adopt the K-means method as the clustering approach in our analysis.

Table 5 summarizes the estimated parameters and their standard errors. We observe that all the parameters are significant at the level of 0.05 except for ξ in cluster 2. The positive values of ξ indicate that the distributions are heavy-tailed. To further examine the tail fitting performance, which is a major concern for insurance companies, Figure 3 shows the PP-plots. We observe that all the points are around the 45-degree line, which further indicates a good tail fitting for each cluster.

5.2 Estimating dependence structures

Having determined the marginal distribution of breach sizes, we use Algorithm 1 to model the temporal and cross-group dependence within each cluster, while assuming the clusters are independent of each other. The pair copula set Ω in Algorithm 1 contains all the bivariate parametric copulas in R package *rvinecopulalib* (Nagler & Vatter, 2022). Note that the Markov order of $p = 1$ is often sufficient for modeling purposes in practice (Nagler et al., 2022); in our study, we also set $p = 1$ to ensure model parsimony. In the supplementary materials, we provide additional analyses with $p = 2$ and $p = 3$ to assess the impact of higher-order dependencies. The results show only marginal improvements in distributional predictive performance in some cases, suggesting that increasing the order beyond $p = 1$ does not yield substantial benefits in our setting. Therefore, the first-order specification strikes a favorable balance between model complexity and predictive accuracy.

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

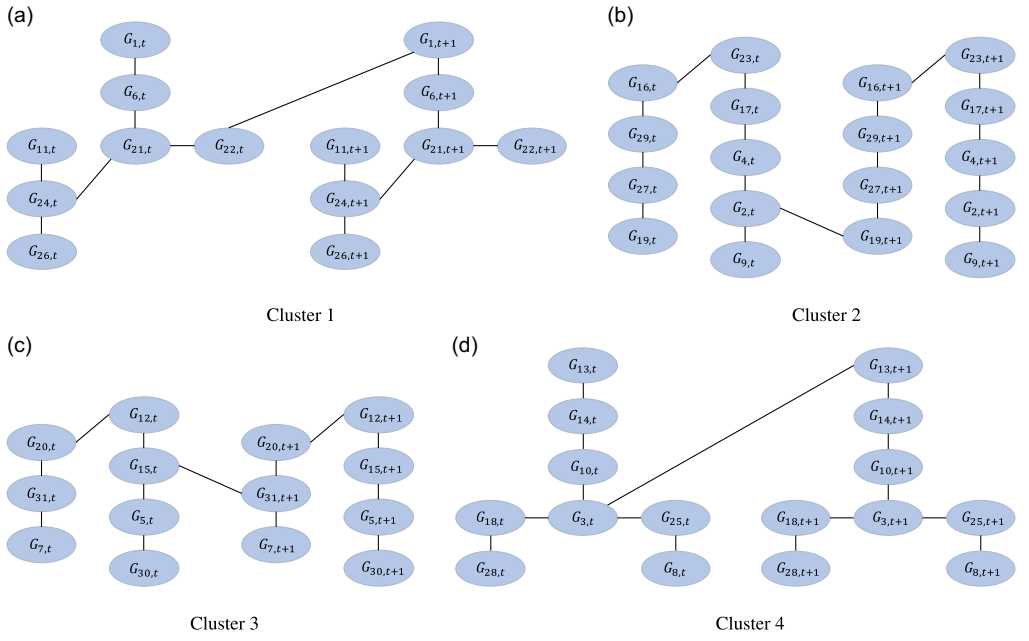

Figure 4. The first tree of the S-vine fitted on the first 15 time periods data.

Figure 4 shows the first tree of the fitted S-vine structure in each cluster. The construction procedure has two steps:

- *Cross-group structure*. For each cluster, the nodes are connected according to the degree of correlation in a sequential manner. For example, for cluster 2, the two most correlated groups (among the 9 groups) are connected first; i.e., G_2 and G_4 are connected first because they are the most correlated pair with $\tau = 0.5294$, meaning that an edge is added between nodes $G_{2,t}$ and $G_{4,t}$, where $G_{i,t}$ represents the severity at time t in group i for $i = 1, 2, \ldots, 31$. Then, the two groups with the second-strongest correlation are connected. This process is repeated until all the 9 groups are connected to form the cross-group R-vine structure at time t. Note that the cross-group dependence at time $t + 1$ is the same as that of time t, and that the vine structures in clusters 1, 3, and 4 are constructed in the same manner.
- *Temporal structure*. We compare the temporal dependence at adjacent time points among two time series in the same cluster. The groups with the largest $|\tau|$ is selected to portray the temporal dependence. For example, in cluster 2, $G_{2,t}$ and $G_{19,t+1}$ are connected to specify the temporal dependence in cluster 2 because they exhibit the largest temporal dependence. From Figure 4 we observe that the nodes from the same industry are more likely to be connected. For instance, G_{21}, G_{22}, and G_{24} belonging to the MED industry, are connected in the fitted vine of cluster 1; G_2 and G_4 are connected in the fitted vine of cluster 2, and both belong to BSF. Similarly, G_{12} and G_{15} are connected in the fitted vine of cluster 3 while G_{13} and G_{14} are connected in the fitted vine of cluster 4, all belonging to BSR. Moreover, we observe that nodes belonging to the same breach type are more likely to be connected as well. For example, G_1, G_6, and G_{21} are connected in the fitted vine of cluster 1, and belong to HACK; G_{15}, G_5, and G_{30} are connected in the fitted vine of cluster 3, and belong to OTHER; G_3, G_{18} and G_{28}, belonging to INSD, are connected in the fitted vine of cluster 4.

After determining the structure of the first tree, we select the parametric copulas from Ω based on the AIC criterion and estimate their parameters. Then, we fix the first tree and determine the next tree. The second tree is constructed in the same fashion as the first, except for the constraint that

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

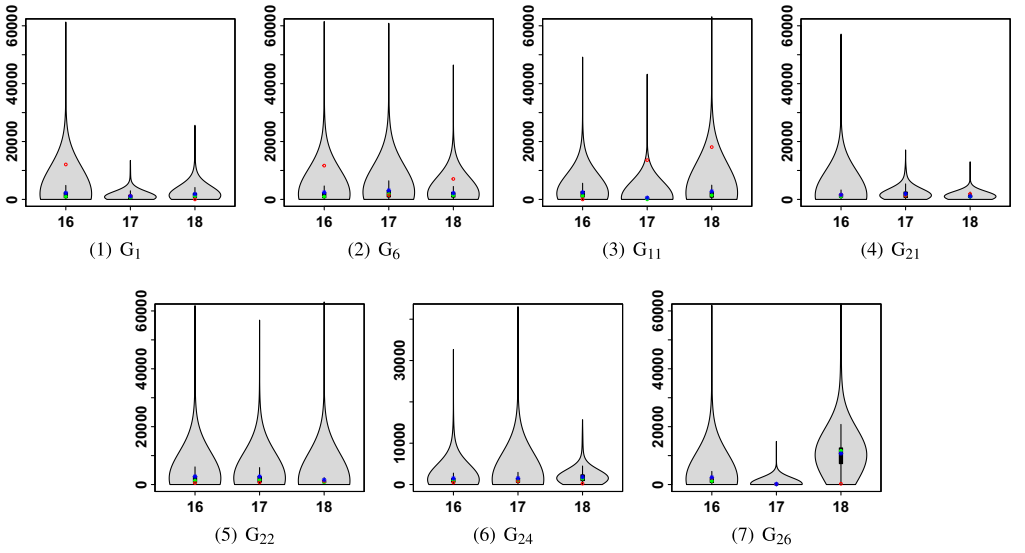

Figure 5. Violin plots of the predicted distributions of breach sizes for the *extreme* cluster, where a red circle represents an observed value; a green dot and a blue star, respectively, represent the predicted median and mean.

is set by the preceding tree. This procedure is repeated to determine the rest of the trees and estimate their copula parameters. For the first trees in the 4 clusters, the selected copulas are Gaussian, Gumbel, Joe, Clayton, Frank, BB6, and BB7, respectively. There are 13, 17, 13, and 15 edges in the first trees resulting from the 4 clusters, respectively. The most-often selected dependence structure is the Joe copula, which accounts for 53.45%. The second and third most-selected copula structures are Clayton copula with 12.07% and Gumbel copula with 8.62%, respectively. Since both Joe and Gumbel copulas indicate an upper tail dependence, there exists strong dependence among the large breach severities. For the second tree, the Joe copula is the most-selected copula structure except for the independent copula, which accounts for 9.26%. Clayton and Frank copulas account for 12.96% in total. We also observe: the higher the level of the tree, the larger the proportion of independent copulas. That is, the higher the level of the tree, the weaker the dependence. This is consistent with our principle of connecting the most correlated edges as early as possible.

5.3 Predicting the distribution of breach sizes

We use Algorithm 2 to conduct the rolling prediction, where $L = 5,000$ simulations. We use the fitted model based on $\{y_{i,t} | 1 \le i \le 31, 1 \le t \le t^*\}$ to predict $\{y_{i,t^*+1} | 1 \le i \le 31\}$ for $t^* = 15, \ldots, 17$. We depict the violin plots of the predicted distributions of breach sizes for each labeled cluster. In each violin plot, the red circle indicates the observed value; the green dot and the blue star, respectively, indicate the predicted median and mean. Recall that in a violin plot, the wide part represents a high probability, and the thinner part represents a low probability. Figure 5 shows the predicted breach size distributions for the extreme cluster. The predictive distributions appear satisfactory, as all observed severities fall within the wider sections of the violin plots, except for those in G_{11}, which corresponds to hacking incidents in BSR. This suggests that predicting hacking incidents related to BSR is particularly challenging, which is a reasonable expectation.

Figure 6 illustrates the predicted breach size distribution for the large cluster. We again observe that the proposed model demonstrates satisfactory performance across all group members, with observed values falling within the predicted distribution with high probability. However, an exception is noted for period 18 in G_{16}, which pertains to hacking incidents in the EDU sector. This

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

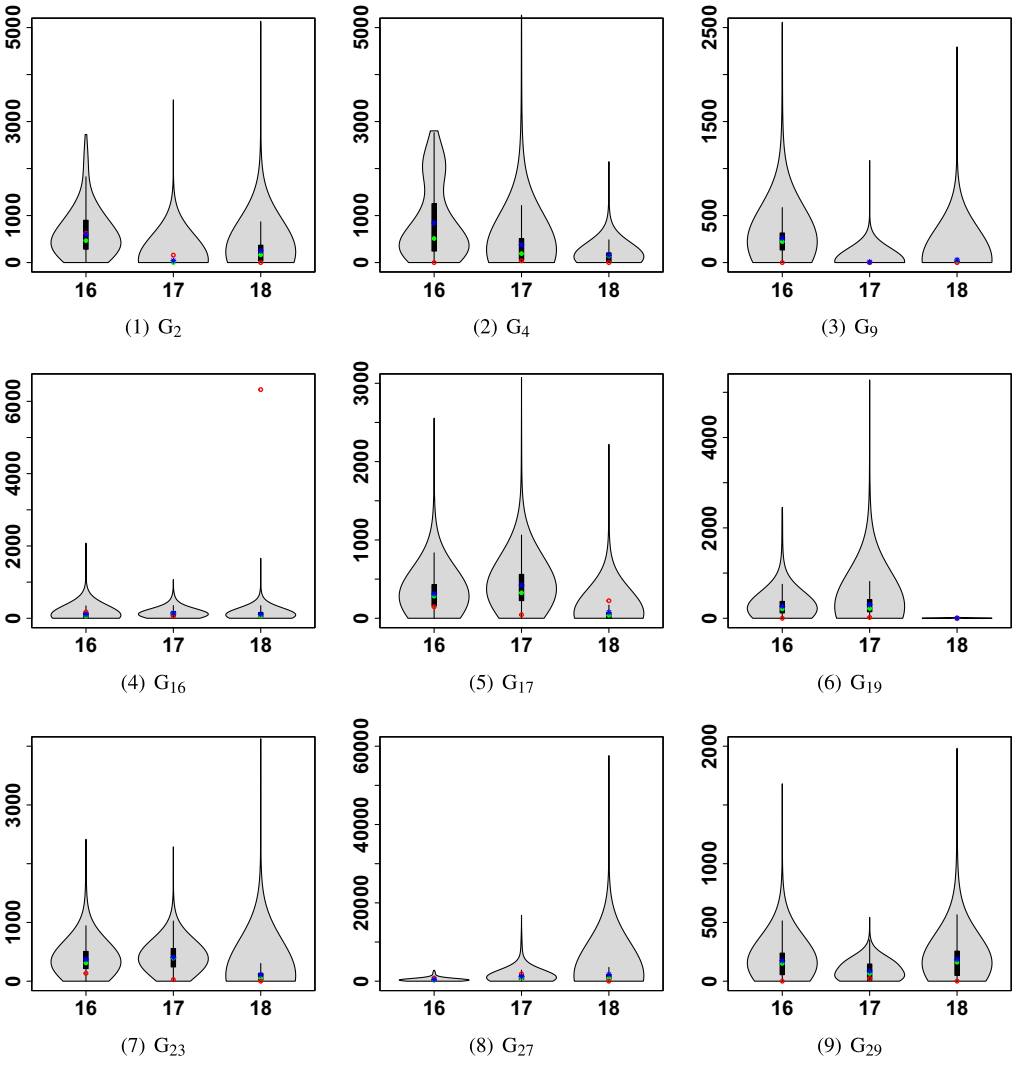

Figure 6. Violin plots of the predicted distributions of breach sizes for the *large* cluster, where a red circle represents an observed value; a green dot and a blue star, respectively, represent the predicted median and mean.

suggests that while the model generally performs well, there may be particular challenges in accurately predicting hacking incidents within the educational sector during specific periods. These challenges could be attributed to unique factors within the EDU sector, such as varying levels of cybersecurity measures, differences in incident reporting, or the unpredictable nature of hacking activities targeting educational institutions.

For the medium cluster, the model exhibits strong predictive performance across most of the groups over the three periods, as shown in Figure 7. The predicted breach sizes align well with the observed values, indicating the model's robustness in capturing the underlying patterns within this cluster. However, exceptions are observed in G7 and G12, where the actual values significantly exceed the predicted ones. Both of these groups pertain to businesses involved in unintended disclosure (i.e., DISC) incidents, which do not involve hacking. This discrepancy suggests that the model may have limitations in accurately predicting the severity of incidents related to unintended disclosures within the business sector in the medium cluster. The underestimation could be attributed to the unpredictable nature of DISC incidents, which might involve

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

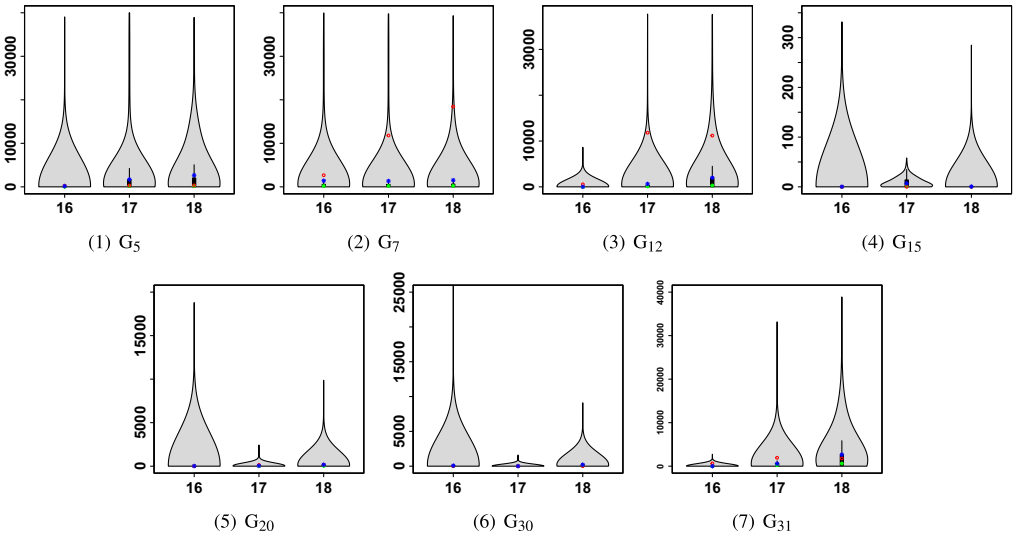

Figure 7. Violin plots of the predicted distributions of breach sizes for the *medium* cluster, where a red circle represents an observed value; a green dot and a blue star, respectively, represent the predicted median and mean.

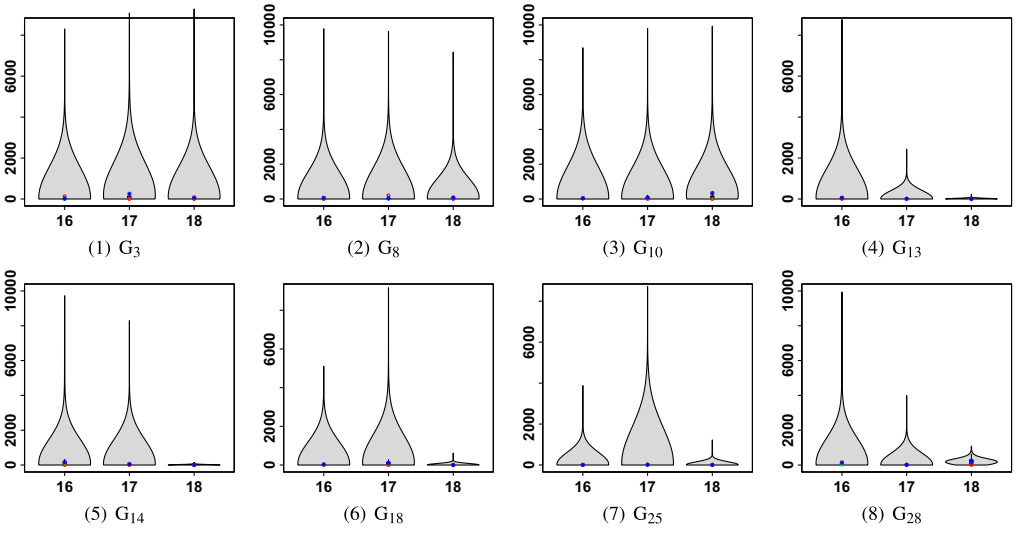

Figure 8. Violin plots of the predicted distributions of breach sizes for the *small* cluster, where a red circle represents an observed value; a green dot and a blue star, respectively, represent the predicted median and mean.

complex and variable factors such as human error, data handling practices, or lapses in internal security protocols. These factors could lead to more severe outcomes than the model anticipates.

For the small cluster in Figure 8, the proposed model demonstrates satisfactory performance across all groups and periods. The predicted distributions align well with the observed values, indicating that the model effectively captures the dynamics within this cluster. This consistency across all periods suggests that the model is robust when dealing with groups with small breach sizes, reinforcing its versatility and reliability in various scenarios.

Model comparison. To fully evaluate the prediction performance of the model, we compare the prediction performance of our approach with the other models:

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

- M1 (Proposed model). This model uses the parametric marginal distributions of Eq. (3) and S-vine dependence structure on the clustered data.
- M2 (Empirical-M-vine model). This model uses empirical marginals and the dependence structure is assumed to be M-vine (Beare & Seo, 2015).
- M3 (Kernel-gpd-M-vine model). This model uses the parametric marginal distribution of Eq. (3) and M-vine dependence structure.
- M4 (Kernel-gpd-S-vine model). This is the same as the proposed model. However, this model is applied to the non-clustered data.
- M5 (D-vine model). This is the model proposed in Fang et al. (2021), where the dependence structure is D-vine.
- M6 (Panel data model). This is the popular model for temporal and cross-group dependence in the literature (Diggle et al., 2002). Specifically, the severity is modeled as follows:

$$S_{i,t} = \beta_0 + \beta_1 t + \beta_2 \mathrm{G}_i + \beta_3 t * \mathrm{G}_i + \epsilon_{i,t},$$

where G_i represents the group i, and $\epsilon_{i,t}$ follows a multivariate normal distribution (MVN) with an AR(1) correlation structure. Specifically, it satisfies

$$(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_m) \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{V}),$$

where $\boldsymbol{S}_i = (S_{i,1}, \ldots, S_{i,n})$, $i = 1, \ldots, m$ is an n-dimensional vector representing the severity of group i in n periods, $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{V}$ is the covariance matrix with

$$\boldsymbol{V} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{nm-1} \\ \rho & 1 & \cdots & \rho^{nm-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{nm-1} & \rho^{nm-2} & \cdots & 1 \end{bmatrix}.$$

- M7 (Mixed effect model). The mixed effect model is the other popular approach to modeling dependent data (Faraway, 2016). The mixed effect model for severity accommodates the random effect among groups, and fixed effects for time and groups. That is,

$$S_{i,t} = \beta_0 + \beta_1 \mathrm{G}_i + \beta_2 t + \gamma_i \mathrm{G}_i + \epsilon_{i,t},$$

where $\beta_0, \beta_1, \beta_2$ are the fixed effect parameters, γ_i is the random effect parameter, and $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,n})$ follows an AR(1) correlation structure.

To evaluate the predictive accuracy, we use both the point assessment and the distribution assessment.

- Mean Squared Error (MSE): This is used for the point prediction assessment. Let $\hat{y}_{i,t}$ be the prediction and $y_{i,t}$ be the observed value. Then, we have

$$\mathrm{MSE} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{t=1}^{n} (y_{i,t} - \hat{y}_{i,t})^2.$$

- Mean Absolute Deviation (MAD): This is used for the point prediction assessment, with

$$\mathrm{MAD} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{t=1}^{n} |y_{i,t} - \hat{y}_{i,t}|.$$

- Continuous Ranked Probability Score (CRPS): This has been widely used in the literature as an accuracy measure for probability forecasts (Epstein, 1969; Matheson & Winkler, 1976). It is a scoring rule that is used for evaluating the distribution prediction assessment (Matheson & Winkler, 1976). Let $F(\cdot)$ be the predicted distribution and $\mathbb{1}\{\cdot\}$ be the indicator function.

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

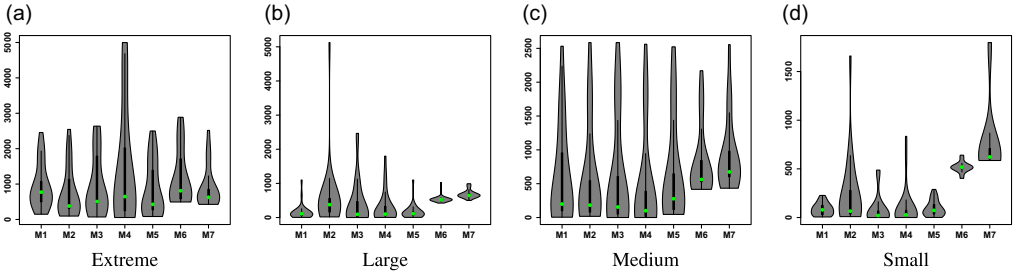

Figure 9. Violin plots of the fifth root transformed MSE for each model in various clusters, where a green dot indicates a median value.

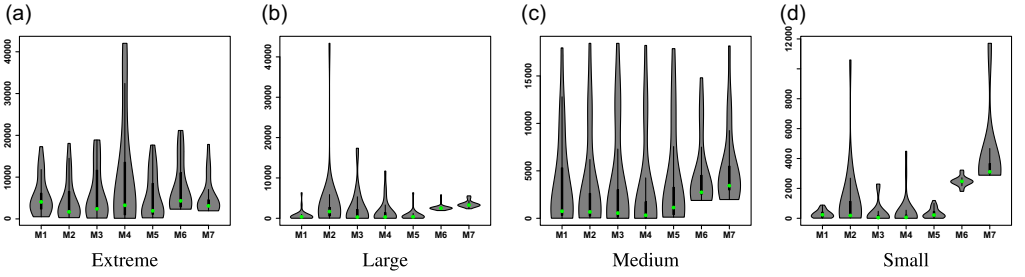

Figure 10. Violin plots of the square root transformed MAD for each model in various clusters, where a green dot indicates a median value.

Then, we have

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(w) - \mathbb{1}\{y \le w\})^2 dw.$$

Note that a smaller score indicates a better prediction.

Figure 9 presents violin plots of the transformed MSE (for visualization purposes) for each model across different clusters. In the extreme cluster, M1, M2, M3, and M5 exhibit comparable predictive performance in terms of MSE. Notably, all these models are copula-based, inherently capturing nonlinear dependence structures among variables. This suggests that incorporating nonlinear dependence significantly enhances predictive accuracy, particularly for extreme cases. For the large cluster, M1 and M5 outperform the other models. In the medium cluster, M1, M2, M3, and M4 demonstrate similar predictive performance. In the small cluster, M1 stands out as the best-performing model. A similar pattern is observed in Figure 10, which presents violin plots of MAD. In Figure 11, the CRPS results reveal that M1, M2, and M3 achieve the best performance, exhibiting low and stable CRPS values in the extreme cluster. For the large cluster, M1, M4, and M5 show comparable performance, outperforming the remaining models. In the medium cluster, M1, M2, M3, and M4 again exhibit similar predictive accuracy. Finally, for the small cluster, M1 remains the top-performing model, consistently demonstrating superior predictive capability.

In summary, the proposed M1 model demonstrates satisfactory predictive performance for each cluster and significantly outperforms other models in the small cluster.

In the following, we further compare the prediction accuracy by combining all the predictions, which are described in Table 6. We observe that the proposed model is significantly better than the other models in terms of both MSE and MAD. We also compare the prediction accuracy based on the distribution assessment metric CRPS. Table 7 shows the medians of the CRPSs for each model, and the percentages of the CRPS of M1 that are less than or equal to those of other models (i.e., M2–M7). The accuracy advantage of our model is also shown in Table 7. We observe that the

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Table 6. MSEs and MADs of predicted breach sizes for each model

Model	MSE ($\times 10^{15}$)	MAD	Model	MSE ($\times 10^{15}$)	MAD
M1	**3.557570**	**20,709,470**	M5	4.377343	22,890,626
M2	41.951300	41,669,847	M6	6.757399	36,026,918
M3	8.253925	34,727,019	M7	4.121542	33,578,668
M4	85.287860	75,849,949			

Table 7. Medians of CRPSs and percentages of CRPS of M1 less than or equal to that of the other models

Model	Median	Percentage	Improvement	Model	Median	Percentage	Improvement
M1	**37,312**	–	–	M5	38,950	51.61%	1.61%
M2	210,681	62.37%	12.37%	M6	4,668,382	88.17%	38.17%
M3	64,070	54.84%	4.84%	M7	5,594,296	90.32%	40.32%
M4	44,998	55.91%	5.91%				

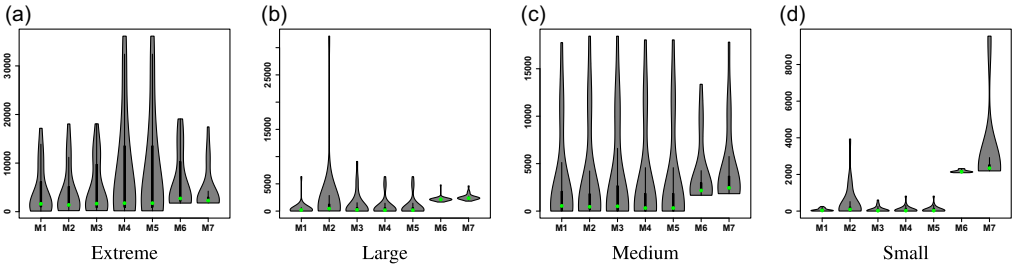

Figure 11. Violin plots of the square root transformed CPRS for each model in various clusters, where a green dot indicates a median value.

proposed model has the smallest median of CRPSs among all the models, and the improvement ranges from 1.61% to 40.32%.

The superior predictive performance of the proposed model can be attributed to the combined effects of clustering and flexible dependence modeling. Data breach severities exhibit significant heterogeneity across industry sectors. The clustering step in our M1 model addresses this heterogeneity by grouping industries with similar severity characteristics, enabling the model to learn more tailored marginal distributions and dependence structures within each cluster. This targeted approach enhances the model's ability to capture intra-group dynamics. Empirically, we observe that the inclusion of clustering leads to substantial improvements in prediction accuracy, as evidenced by the performance comparison between M1 and M4. Additional analysis provided in the supplementary material, where clustering is incorporated into the M3 model, further confirms the positive impact of clustering on predictive accuracy. Complementing this, the S-vine copula model offers a highly flexible framework for capturing both temporal and cross-sectional dependencies. Unlike simpler dependence structures, the S-vine is very flexible, which is essential in modeling the complex dynamics of cyber risk. By comparing the predictive performance of M1 and M3 with clustering (see supplementary material), we find that the use of a more expressive dependence structure, specifically the S-vine, leads to further improvements. These results collectively demonstrate that both clustering and the chosen dependence structure play critical roles in enhancing the predictive capability of the M1 model.

In the supplementary material, we have also compared the proposed M1 model to some deep learning models including a feedforward neural network model and a state-of-the-art deep

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

learning model-*TimeMixer* (Wang et al., 2024). We found that deep learning models, such as TimeMixer, demonstrate promising and competitive performance in terms of point prediction accuracy. Their ability to automatically learn complex temporal patterns and interactions across multiple series makes them powerful tools in time series forecasting. However, the proposed M1 model offers distinct advantages that are particularly important in risk modeling and ratemaking contexts. Specifically, M1 is designed to produce full predictive distributions rather than just point estimates, enabling a richer characterization of uncertainty. This distributional output supports important downstream tasks such as probabilistic risk assessment, Value-at-Risk, and Conditional Tail Expectation estimation, which are not straightforward to obtain from standard deep learning architectures without additional post-processing or calibration. Therefore, while deep learning approaches are indeed promising, the M1 model provides interpretable, distribution-based forecasts that are better aligned with the needs of insurance and risk management applications.

6. Insurance pricing

In this section, we evaluate the predictive performance of our model in the context of the ratemaking market. We utilize both the ordered Lorenz curve and the Gini index, as proposed by (Jed) Frees et al. (2014), to assess predictive efficiency. The ordered Lorenz curve graphically represents the relationship between loss distribution and premium distribution, with both distributions ordered by their relativities. The associated Gini index is particularly significant; insurers who adopt a rating structure with a higher Gini index are more likely to achieve a profitable portfolio (Frees et al., 2011). Since there is no standard method for converting the number of breached records into a dollar loss, we treat the number of breached records as equivalent to the dollar loss. In underwriting, the dollar loss may be determined by referencing historical data, such as the typical dollar loss associated with each breached record.

6.1 Ordered lorenz curve

In this subsection, we explore two different base premium strategies for the ordered Lorenz curve.

In the first strategy, we consider the loss predicted by each model from M2 to M7 as the base premium, while the loss predicted by our proposed model M1 serves as the competing premium. A larger area between the line of equality and the curve indicates a more favorable performance for M1. As shown in the ordered Lorenz curves in Figure 12a, the area between the line of equality and the curve is consistently substantial across all base premiums. This suggests that our proposed model, M1, is capable of generating more profitable contracts.

Our second strategy for establishing the base premium takes a simpler approach: using the sample mean of each group from the previous years as the base premium. For example, the loss at $T = 16$ is predicted based on the sample mean of losses from $T = 1$ to $T = 15$ for each group. The losses predicted by all models serve as the competing premiums. The ordered Lorenz curves in Figure 12b demonstrate that the proposed M1 model outperforms the other models by creating a significant gap between the premium and loss distributions.

6.2 Gini index

We begin by considering the Gini index with the base premiums set as the sample means of losses. Table 8 presents the computed Gini indices and their corresponding standard errors (SE) for various models of the number of breach records. The table reveals that the Gini index for the M1 model is 0.5850 with a standard error of 0.1671, making it the highest among the models and statistically significant at the 0.05 level. These results indicate that the proposed M1 model outperforms the other models.

Now, we consider the comparison of different base premiums. Table 9 provides a comparative analysis of the Gini indices for the competing premiums, each based on a different base premium.

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Table 8. Gini indices and their standard errors (SE) based on various models for the number of breach records with base premium from the sample means

Model	M1	M2	M3	M4	M5	M6	M7
Gini index	**0.5850**	0.0456	-0.2021	-0.1599	-0.1015	-0.1029	0.5641
SE	**0.1671**	0.2094	0.2729	0.3576	0.2695	0.3180	0.3435

Table 9. Gini indices of different models for the number of breach records

Base premium		M1	M2	M3	M4	M5	M6	M7
Gini indices	M1	–	0.9754	0.9385	0.9883	0.7346	0.5878	0.7891
M2	-0.0937	–	0.7103	0.5607	0.2295	0.0459	0.0873	
M3	-0.1429	0.7473	–	0.5511	0.0277	-0.1702	0.3295	
M4	0.0941	0.7982	0.6904	–	0.2702	-0.1141	0.5677	
M5	0.5197	0.9849	0.9253	0.9798	–	0.1645	0.7939	
M6	0.5129	0.9864	0.9760	0.9856	0.4084	–	0.8155	
M7	0.3930	0.9715	0.9704	0.9830	0.6273	0.3117	–	
Maximum		**0.5197**	0.9864	0.9760	0.9883	0.7346	0.5878	0.8155

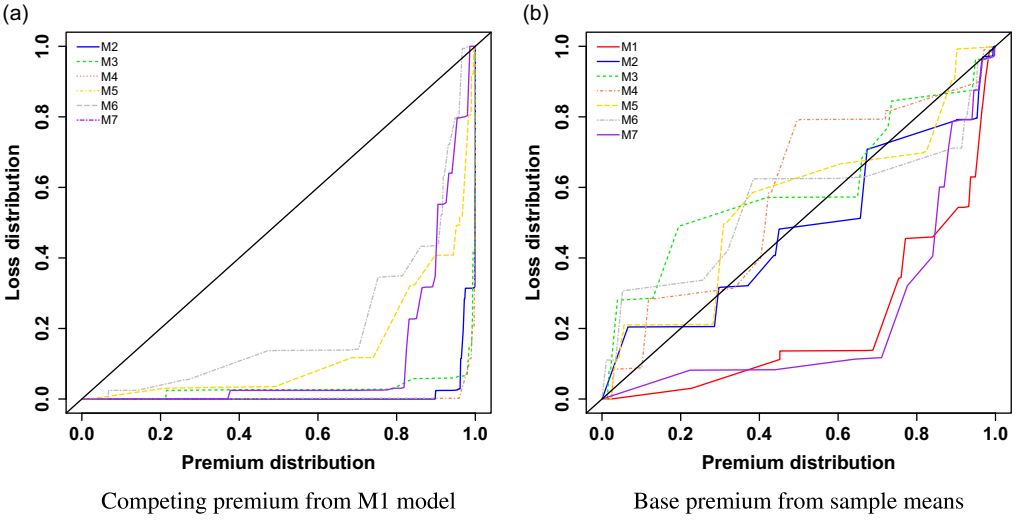

Figure 12. Ordered Lorenz curves of the number of breached records.

The last row of the table presents the maximum Gini index observed for each base premium. For model selection, we employ the "mini-max" criterion proposed by (Jed) Frees et al. (2014), which involves selecting the model with the smallest maximum Gini index across various competing models. From Table 9, we see that the M1 model has the smallest maximum Gini index value of 0.5197, indicating that it demonstrates the least vulnerability compared to the other models.

In conclusion, the above analysis has demonstrated the robustness and superior performance of the proposed M1 model in the context of insurance rating.

7. Conclusion and discussion

We have introduced a novel cluster-based multivariate dependence model that effectively captures both temporal and cross-group dependencies, providing a more accurate representation of

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

multivariate cyber breach risks. This model employs a kernel approach for modeling breach severity, while utilizing an S-vine copula to account for the multivariate dependencies. The validity of this framework was demonstrated through its application to case studies involving two real-world cyber breach data as well as synthetic data. The results consistently showed that the proposed model delivers satisfactory fitting and predictive performance. In the context of ratemaking, the model has proven to be the least vulnerable, facilitating a more equitable and profitable distribution of premiums relative to losses. This makes the proposed model a highly effective tool for insurers seeking to optimize their rating structures and achieve more reliable and profitable portfolios.

In practice, the proposed framework can be used by insurance companies who offer cyber insurance policies to perform group rating. The typical 6-month insurance policy can be created by following the empirical study in Section 5 as it predicts the loss very well. In fact, insurance companies can also customize their policies with different months (e.g., 1 month, 3 months, or 12 months) by following our approach. Since the proposed framework simulates the predictive distribution of loss, it can be used for risk management as well. For example, an insurance company may use the 95th percentile as the extreme loss scenario to prepare reserves (McNeil et al., 2015).

Our study has the following limitations: (i) The dataset used in this study may not capture all breach incidents, as some breaches might go unreported. This potential underreporting could lead to an incomplete picture of the true extent of cyber breach risks. However, it is important to note that this dataset remains the most widely analyzed and comprehensive resource available to the research community, making it a valuable tool despite its limitations. (ii) Our methodology focuses on predicting the distribution of breach sizes within a group, rather than predicting when the next breach will occur or identifying which specific enterprise within the group will be affected. A more detailed point process model could be appropriate for studies aimed at predicting the timing and location of future incidents. Developing such a model would require substantial effort in capturing and modeling the complex dynamics of incident occurrences, which we leave as an avenue for future research.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S1748499525100109.

Acknowledgments. The authors are grateful to the two anonymous referees for their insightful and constructive comments, which led to this improved version of the article.

Data availability statement. The data and code that support the findings of this study are available from the corresponding author, XW, upon reasonable request.

Funding statement. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. The authors declare none.

References

Beare, **B. K.**, & **Seo**, **J.** (2015). Vine copula specifications for stationary multivariate markov chains. *Journal of Time Series Analysis*, **36**(2), 228–246.

Bedford, **T.**, & **Cooke**, **R. M.**2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, **32**(1), 245–268.

Buckman, **J.**, **Hashim**, **M.**, **Woutersen**, **T.**, & **Bockstedt**, **J.** (2018). Fool me twice: An analysis of repeat data breaches within firms. *SSRN Electronic Journal*. doi:10.2139/ssrn.3258599.

Czado, **C.**, & **Nagler**, **T.** (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application*, **9**(1), 453–477.

Diggle, **P.**, **Heagerty**, **P.**, **Liang**, **K.-Y.**, & **Zeger**, **S.** (2002). *Analysis of longitudinal data*. Oxford University Press.

Dißmann, **J.**, **Brechmann**, **E. C.**, **Czado**, **C.**, & **Kurowicka**, **D.** (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, **59**, 52–69.

Edwards, **B.**, **Hofmeyr**, **S.**, & **Forrest**, **S.** (2016). Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, **2**(1), 3–14.

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

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

Epstein, **E. S.** (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, **8**(6), 985–987.

Fang, **Z.**, **Xu**, **M.**, **Xu**, **S.**, & **Hu**, **T.** (2021). A framework for predicting data breach risk: Leveraging dependence to cope with sparsity. *IEEE Transactions on Information Forensics and Security*, **16**, 2186–2201.

Faraway, **J. J.** (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Frees, **E. W.**, **Meyers**, **G.**, & **Cummings**, **A. D.** (2011). Summarizing insurance scores using a gini index. *Journal of the American Statistical Association*, **106**(495), 1085–1098.

(Jed) Frees, **E. W.**, **Meyers**, **G.**, & **Cummings**, **A. D.** (2014). Insurance ratemaking and a gini index. *Journal of Risk and Insurance*, **81**(2), 335–366.

Hartigan, **J. A.**, & **Wong**, **M. A.** (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**(1), 100–108.

He, **R.**, **Jin**, **Z.**, & **Li**, **J. S.-H.** (2024). Modeling and management of cyber risk: A cross-disciplinary review. *Annals of Actuarial Science*, **18**(2), 270–309.

IBM Security. Cost of a data breach report 2024. Accessed: 2025-03-10.

Identity Theft Resource Center. 2024 data breach report. Accessed: 2025-03-10.

Ikegami, **K.**, & **Kikuchi**, **H.** (2020). Modeling the risk of data breach incidents at the firm level . *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing* (pp. 135–148). Springer.

James, **G.**, **Witten**, **D.**, **Hastie**, **T.**, & **Tibshirani**, **R.** (2013). *An introduction to statistical learning*, vol. **112**. Springer.

Joe, **H.** (2014). *Dependence modeling with copulas*. CRC Press.

Joseph Buckman, **J. B.**, **Hashim**, **M. J.**, & **Woutersen**, **T.** (2017).Do organizations learn from a data breach? *Workshop on the Economics of Information Security*.

Maillart, **T.**, & **Sornette**, **D.** (2010). Heavy-tailed distribution of cyber-risks. *The European Physical Journal B-Condensed Matter and Complex Systems*, **75**(3), 357–364.

Malavasi, **M.**, **Peters**, **G. W.**, **Shevchenko**, **P. V.**, **Trück**, **S.**, **Jang**, **J.**, & **Sofronov**, **G.** (2022). Cyber risk frequency, severity and insurance viability. *Insurance: Mathematics and Economics*, **106**, 90–114.

Matheson, **J. E.**, & **Winkler**, **R. L.** (1976). Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–1096.

Maugis, **C.**, **Celeux**, **G.**, & **Martin-Magniette**, **M.-L.** (2009). Variable selection for clustering with gaussian mixture models. *Biometrics*, **65**(3), 701–709.

McNeil, **A. J.**, **Frey**, **R.**, & **Embrechts**, **P.** (2015). *Quantitative risk management: Concepts, techniques and tools-revised edition*. Princeton University Press.

Nagler, **T.**, **Krüger**, **D.**, & **Min**, **A.** (2022). Stationary vine copula models for multivariate time series. *Journal of Econometrics*, **227**(2), 305–324.

Nagler, **T.**, & **Vatter**, **T.** (2022). rvinecopulib: High performance algorithms for vine copula modeling. R package version.

Privacy Rights Clearinghouse (2025). Privacy rights clearinghouse's chronology of data breaches. https://www.privacyrights.org/data-breaches.

Scarrott, **C. J.** (2016). Univariate extreme value mixture modelling. In *Extreme*value modeling and risk analysis: methods and applications (pp. 41–67).

Schraub, **D.**, & **Rudolph**, **M. J.** (2025). Casualty actuarial society and society of actuaries. *18th Annual Survey of Emerging Risks: Key Findings*. Accessed: 2025-03-10.

Sen, **R.**, & **Borle**, **S.** (2015). Estimating the contextual risk of data breach: An empirical approach. *Journal of Management Information Systems*, **32**(2), 314–341.

Shi, **P.**, & **Yang**, **L.** (2018). Pair copula constructions for insurance experience rating. *Journal of the American Statistical Association*, **113**(521), 122–133.

Sklar, **M.** (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP* (pp. 229–231), vol. **8**.

Smith, **M. S.** (2015). Copula modelling of dependence in multivariate time series. *International Journal of Forecasting*, **31**(3), 815–833.

Sun, **H.**, **Xu**, **M.**, & **Zhao**, **P.** (2021). Modeling malicious hacking data breach risks. *North American Actuarial Journal*, **25**(4), 484–502.

https://doi.org/10.1017/S1748499525100109 Published online by Cambridge University Press

Sun, **H.**, **Xu**, **M.**, & **Zhao**, **P.** (2023). A multivariate frequency-severity framework for healthcare data breaches. *The Annals of Applied Statistics*, **17**(1), 240–268.

Wand, **M. P.**, & **Jones**, **M. C.** (1994). *Kernel smoothing*. Chapman and Hall/CRC.

Wang, **S.**, **Wu**, **H.**, **Shi**, **X.**, **Hu**, **T.**, **Luo**, **H.**, **Ma**, **L.**, **Zhang**, **J. Y.**, & **Zhou**, **J.** (2024). Timemixer: Decomposable multiscale mixing for time series forecasting. In: *The Twelfth International Conference on Learning Representations*.

Wheatley, **S.**, **Maillart**, **T.**, & **Sornette**, **D.** (2016). The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B*, **89**(1), 7.

Woods, **D. W.**, **Moore**, **T.**, & **Simpson**, **A. C.** (2021). The county fair cyber loss distribution: Drawing inferences from insurance prices. *Digital Threats: Research and Practice*, **2**(2), 1–21.

Wu, **M. Z.**, **Luo**, **J.**, **Fang**, **X.**, **Xu**, **M.**, & **Zhao**, **P.** (2023). Modeling multivariate cyber risks: Deep learning dating extreme value theory. *Journal of Applied Statistics*, **50**(3), 610–630.

Xu, **M.**, **Schweitzer**, **K. M.**, **Bateman**, **R. M.**, & **Xu**, **S.** (2018). Modeling and predicting cyber hacking breaches. *IEEE Transactions on Information Forensics and Security*, **13**(11), 2856–2871.

Cite this article: Li Y, Wang X, Zhao P and Hu T (2025). Cyber breach risk modeling for insurance: capturing temporal and cross-group dependence, *Annals of Actuarial Science*, 1–25. https://doi.org/10.1017/S1748499525100109