A Framework for Predicting Data Breach Risk: Leveraging Dependence to Cope With Sparsity

Zijian Fang, Maochao Xu, Shouhuai Xu, *Senior Member, IEEE*, and Taizhong Hu

***Abstract*— Data breach is a major cybersecurity problem that has caused huge financial losses and compromised many individuals' privacy (e.g., social security numbers). This calls for deeper understanding about the data breach risk. Despite the substantial amount of attention that has been directed toward the issue, many fundamental problems are yet to be investigated. In this article, we initiate the study of modeling and predicting risk in *enterprise-level* data breaches. This problem is challenging because of the *sparsity* of breaches experienced by individual enterprises over time, which immediately disqualifies standard statistical models because there are not enough data to train such models. As a first step towards tackling the problem, we propose an innovative statistical framework to leverage the *dependence* between multiple time series. In order to validate the framework, we apply it to a dataset of enterprise-level breach incidents. Experimental results show its effectiveness in modeling and predicting enterprise-level breach incidents.**

***Index Terms*— Data breach, cyber threats, cyber risk analysis, breach prediction, sparse time series, cybersecurity data analytics.**

I. Introduction

DATA breaches are devastating threats to computer systems. The Privacy Rights Clearinghouse (PRC) [1] reports 9,015 data breaches between 2005 and 2019, accounting for 11,690,762,146 breached records. The Identity Theft Resource Center and Cyber Scout [2] reports 1,244 data breach incidents in 2018, exposing 446,515,334 records, which are much higher (or a 126% jump) from the 197,612,748 records exposed in 2017. The cost of data breach is also substantial. According to NetDiligence [3], for small-to-medium enterprises (i.e., less than $2 billion in annual revenue), the average breach cost from 2014 to 2018 is $178K, not including a $112K average crisis service cost and a $181K legal cost; for large companies (i.e., $2 billion or more in annual revenue), the average breach cost from 2014 to 2018 is $5.6M.

This problem is naturally studied at the enterprise level because risk management is often conducted at the enterprise level to shed light on better mitigation strategies (e.g., designing insurance policies to mitigate the damage of data breaches). This problem is challenging because enterprise-level time series are very *sparse*. This sparsity immediately disqualifies most, if not all, exiting statistical time-series models and certainly deep learning-based time-series models, simply because there are not enough data to train such models. In this article, we make a significant first step towards the tackling the data sparsity problem in this context.

A. Our Contributions

We make four contributions. First, we initiate the study of modeling and predicting multivariate time series with *sparse* events and propose a framework to tackle this problem. Existing techniques cannot tackle this problem because there are not enough data to train a statistical model for each time series, let alone deep learning models. The research problem is based on real data breach events abstracted in a time series. Moreover, our framework can be easily adapted to model and predict other types of multivariate time series with sparse events (e.g., penetrations into networks or 0-day attacks).

Second, the novelty of our framework is the idea of leveraging the statistical *dependence* between multivariate time series to cope with the *sparsity* of events. The framework can be characterized as follows: (i) It uses a two-part mixture structure to accommodate the excessively many zeros (i.e., event sparsity); (ii) it uses the heavy-tail distribution to accommodate the often-observed skewness and extreme values of breach sizes; (iii) it uses covariates to accommodate the possible breach-size heterogeneity between the time series; and (iv) it uses the mixed D-vine copula structure to accommodate the temporal dependence of multivariate time series. Intuitively, the framework leverages the inter-enterprise relationship to accommodate more information than what is accommodated when considering the time series separately.

Third, we conduct a case study on modeling and predicting enterprise-level multivariate cyber breach time series of *sparse* events. By applying our framework to this dataset, we draw a number of insights, such as: (i) Data breach sizes exhibit large variability and large skewness (i.e., heavy tails), and should be modeled by different distributions. Business-related

Manuscript received August 12, 2020; revised December 3, 2020; accepted January 4, 2021. Date of publication January 14, 2021; date of current version February 9, 2021. The work of Shouhuai Xu was supported in part by the Army Research Office (ARO) Grant W911NF-17-1-0566 and in part by NSF under Grant 1814825 and Grant 1736209. The work of Taizhong Hu was supported in part by the National Natural Science Foundation (NNSF) of China under Grant 71871208 and in part by the Anhui Center for Applied Mathematics. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Issa Traore. *(Corresponding author: Shouhuai Xu.)*

Zijian Fang and Taizhong Hu are with the Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, China.

Maochao Xu is with the Department of Mathematics, Illinois State University, Normal, IL 61790 USA.

Shouhuai Xu was with the Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249 USA. He is now with the Department of Computer Science, University of Colorado Colorado Springs, Colorado Springs, CO 80918 USA (e-mail: sxu@uccs.edu).

Digital Object Identifier 10.1109/TIFS.2021.3051804

enterprises have the largest breach sizes, perhaps because they have many customers. Moreover, different enterprises exhibit different breach characteristics, highlighting the importance of modeling the temporal dependence (i.e., time effect) and the heterogeneity between time series. In particular, enterprise-level breach sizes exhibit a negative temporal dependence, meaning consecutive breaches are unlikely to occur to an individual enterprise within a short period of time, perhaps because a breached enterprise is no longer attractive to attackers or because a recently breached enterprise is likely to strengthen its security. (ii) Business-related enterprises show a nonlinear pattern in terms of the number (or percentage) of enterprises that have data breaches, hinting at an attack-defense arms race in business sectors. Medical enterprises show an increasing (but nonlinear) pattern over time, perhaps because their systems are more vulnerable, more desirable targets, or have fewer consequences for unsuccessful attackers. Not-for-profit enterprises (including government) show a decreasing trend, hinting potential enhancements in their cyber defense or decrease in desirability for attackers. (iii) The mixed D-vine dependence structure can accommodate the complex dependence exhibited by enterprise-level multivariate breach incident time series. The distribution of enterprise-level breach sizes can be well predicted by the proposed mixed D-vine model. This sheds light on the possibility of quantitative risk management, which we discuss as a use case later.

Fourth, in order to demonstrate the broad applicability of our framework, we generate a synthetic dataset, which exhibits different properties than the real-world dataset (e.g., positive vs. negative dependence). Experimental results show that our framework achieves satisfactory fitting/prediction accuracy.

B. Related Work

1) Prior Studies Related to the Data we Analyze: The significant threat of data breaches calls for deeper understanding about them. Facilitated by the availability of data (especially [1]), there have been a number of studies on characterizing breach incidents [4]–[13]. For example, Buckman *et al.* [4] studied the time intervals between data breaches for the enterprises that have at least two incidents between 2010 and 2016. They showed that the duration between two data breaches may increase or decrease, depending on some factors. Buckman *et al.* [5] investigated the effect of data breach notification policy on data breaches. They developed a panel regressions with fixed effects to test several hypotheses on the effect of policies, by using the PRC data between 2005 and 2016. Edwards *et al.* [6] analyzed the temporal trend of data breach size and frequency and showed that the breach size follows a log-normal distribution and the frequency follows a negative binomial distribution. They further showed that the frequency of large breaches (over 500,000 breached records) follows the Poisson distribution, rather than the negative binomial distribution, and that the size of large breaches still follows log-normal distribution. Eling and Loperfido [7] studied data breaches from the perspective of actuarial modeling and pricing. They used multidimensional scaling and goodness-of-fit tests to analyze the distribution of data breaches. They showed that different types of data breaches should be analyzed separately and that breach sizes can be modeled by the skew-normal distribution. Sun *et al.* [9] developed a frequency-severity actuarial model of aggregated enterprise-level (rather than individual enterprise) breach data to promote ratemaking and underwriting in insurance. Ikegami and Kikuchi [12] studied a breach dataset in Japan and developed a probabilistic model for estimating the data breach risk. They showed that the inter-arrival times of data breaches (for those enterprises with multiple breaches) follow a negative binomial distribution. Romanosky *et al.* [14] used a fixed effect model to estimate the impact of data breach disclosure policy on the frequency of identity thefts incurred by data breaches.

However, none of the preceding studies investigated the modeling and prediction of *enterprise-level* time series with *sparse* breach events (i.e., most enterprises have very few data breach incidents within a significant period of time). This naturally triggers the current research problem: *Is it possible to model and predict enterprise-level data breach incidents?*

2) Prior Studies Related to the Approach we Use: The most closely related prior study is Xu *et al.* [8], which studied an *aggregated* cyber *hacking* breach incidents dataset (derived from [1]) and showed that both *incidents inter-arrival time* and *breach size* should be modeled by stochastic processes rather than distributions. The time series studied in [8] is *univariate* and *dense* (i.e., many nonzero observations). In contrast, we study *multivariate* time series with *sparse* observations (i.e., excessively many 'zero' observations) because we consider individual enterprises (rather than their aggregation). As a consequence, the techniques developed in [8] and deep learning-based techniques (e.g., [15]) are not applicable to our setting because there are not enough data to train models. A loosely related prior study is Eling and Jung [16], which is different from ours for two reasons. (i) In terms of objective, they studied the aggregation of breached records among companies (i.e., one aggregated observation per month) and treating these monthly observations as sampled from a *single* distribution. By contrast, we study enterprise-level breach data (i.e., one time series per enterprise), meaning that in general the observations are *not* drawn from any single distribution. (ii) In terms of techniques, they fitted the aggregated number of breached records (i.e., dense data) by using distributions (e.g., normal, gamma, and log-normal) with D-vine copula structure, and they did not consider prediction. By contrast, we overcome two challenges: the data sparsity that is exhibited at the enterprise-level data; and the temporal dependence exhibited by the data (by employing an innovative mixed model with D-vine dependence structure). This justifies why we develop a new framework to leverage dependence to cope with the sparsity.

3) Prior Studies Related to the Problem we Tackle: The present study falls into the active field of cybersecurity data analytics (cf. [17]–[22]). Along this line, Bagchi and Udo [23] used a variant of the Gompertz model to analyze the growth of computer-related crimes. Zhan *et al.* [24] investigated *grey-box* statistical models for predicting cyber attack rates, where *grey-box* means that the models can accommodate the statistical properties exhibited by the data (e.g., long-range

dependence and extreme values). Peng *et al.* [25] showed that the point-over-threshold method can model the magnitudes of extreme attack rates. Liu *et al.* [26] investigated how to use a network's externally observable features (e.g., mismanagement symptoms) to predict its potential data breach incidents. Sen and Borle [27] studied the factors that have an impact on the contextual risk of data breach.

The matter of *dependence* has been studied in two cybersecurity contexts. In theoretical cybersecurity, dependence has been studied in [28]–[30] but it has been mostly assumed away in other models (e.g., [31]–[35]). In cybersecurity data analytics, dependence has been studied in the settings of time series with *dense* events (rather than *sparse* events). Böhme and Kataria [36] studied dependence at two levels: using the Beta-Binomial model to describe the intra-enterprise dependence and a one-factor latent model to describe the inter-enterprise dependence. They also used the Archimedean copula to model cyber risks caused by virus incidents [37]. Mukhopadhyay *et al.* [38] used a copula-based Bayesian Belief network to assess cyber vulnerability. They used the normal copula to aggregate the number of failures and losses and compute the overall loss distribution on a cyber risk portfolio. Xu *et al.* [39] used vine copulas to study the dependence between *dense* time series.

C. Paper Outline

Section II reviews statistical preliminary knowledge. Section III describes the framework. Section IV presents a case study based on a real-world dataset. Section V presents a case study based on a synthetic dataset. Section VI discusses the limitations of the present study. Section VII concludes the article with future research directions.

II. Preliminaries

In order to model the dependencies within the enterprise-level data breach time series, we propose using the *copula* technique, which is an effective and popular tool for modeling high-dimensional dependence [40]. Let $X_1, \ldots, X_d$ be continuous random variables with univariate marginal distributions $F_1, \ldots, F_d$, respectively. Denote their joint cumulative distribution function (CDF) by

$$F(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

A d-dimensional copula, denoted by C, is a CDF with uniform marginals in [0, 1], namely the joint CDF of the random vector $(F_1(X_1), \ldots, F_d(X_d))$. Sklar's theorem [40] says that when the F_i's are continuous, C is unique and satisfies

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

Let $c(u_1, \ldots, u_n)$ be the d-dimensional copula density function and f_i be the marginal density function of X_i for $i = 1, \ldots, d$. The joint density function of $(X_1, \ldots, X_d)$ is

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i).$$

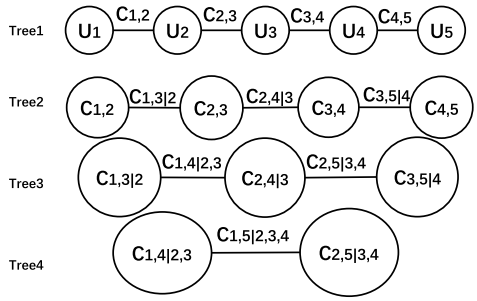


Fig. 1. Five-dimensional D-vine dependence structure.

In this article, we will use the *vine* copula [40], because it is computationally tractable (i.e., its density can be factored in terms of bivariate linking copulas and lower-dimensional margins). A vine copula is described by a tree sequence on d elements, namely an ordered set of trees $\mathcal{V} = (\mathrm{Tr}_1, \cdots, \mathrm{Tr}_{d-1})$ where $\mathrm{Tr}_i = (N_i, E_i)$ with node set N_i and edge set E_i for $1 \leq i \leq d-1$, satisfying

a) Tr_1 is the first tree with node set $N_1 = \{1, \ldots, d\}$ and edge set E_1.
b) For $2 \leq i \leq d-1$, edge set E_{i-1} is the node set of tree Tr_i.
c) (Proximity condition) For tree Tr_i $(2 \leq i \leq d-1)$, if two nodes in E_{i-1} are connected by an edge in E_i, these two nodes share the same node in E_i as edges in Tr_{i-1}.

In general, a d-dimensional vine copula is constructed by mixing $d(d-1)/2$ bivariate linking copulas on a tree.

D-vine is a special kind of vine copula with nodes only connecting to their adjacent nodes [41]. Figure 1 presents the graphical specification of a 5-dimensional $(U_1, \cdots, U_5)$ D-vine in the form of a nested set of tree structures, where $U_1, \ldots, U_5$ are uniform random variables. A D-vine with five variables has four trees Tr_j, and tree Tr_j has $6-j$ nodes and $5-j$ edges, where $1 \leq j \leq 4$. Each edge is associated with a pair-copula density used for modeling dependence between two variables, and the edge label represents the dependence parameter in the associated pair-copula density. In Tree 1, there are four pairs of variables, namely (U_1, U_2), (U_2, U_3), (U_3, U_4) and (U_4, U_5); the pair-wise dependencies are modeled by using four copulas $c_{1,2}, c_{2,3}, c_{3,4}, c_{4,5}$, where $c_{i,i+1}$ represents the copula density between U_i and U_{i+1} with $1 \leq i \leq 4$. In Tree 2, three conditional dependencies are modeled: the one between U_1 and U_3 given U_2 using copula density $c_{1,3|2}$; the one between U_2 and U_4 given U_3 using copula density $c_{2,4|3}$; and the one between U_3 and U_5 given U_4 using copula density $c_{3,5|4}$. In Tree 3, two conditional dependence are modeled: the one between U_1 and U_4 given U_2, U_3 using copula density $c_{1,4|2:3}$; and the one between U_2 and U_5 given U_3, U_4 using copula density $c_{2,5|3:4}$. In Tree 4, only one conditional dependence is modeled, namely the one between U_1 and U_5 given U_2, U_3, U_4 using copula density $c_{1,5|2:4}$. As a result, the joint distribution density of a

TABLE I
SUMMARY OF MAIN NOTATIONS USED IN THE PAPER

Y_{it}	random variable for the breach size of enterprise i in year t
F_{it}	cdf of breach size Y_{it} of enterprise i in year t
f_{it}	probability density of breach size Y_{it} for enterprise i in year t
M_{it}	cdf of breach size Y_{it} under the condition $Y_{it} > 0$ (i.e., enterprise i is breached in year t)
m_{it}	probability density of breach size Y_{it} under the condition $Y_{it} > 0$
y_{it}	observed breach size of random variable Y_{it} in the real world
p_{it}	$p_{it} = \Pr[Y_{it} = 0]$, the probability enterprise i has no breach in year t
$\boldsymbol{\beta}$	parameter set of the logit regression
Θ	parameter set of the GPD model
C	copula structure
$C_{s,t}$	distribution function of a bivariate copula
$c_{s,t}$	density function of a bivariate copula
$\boldsymbol{T}_r$	tree set of a vine structure
τ	Kendall's tau
RPS	ranked probability score

5-dimensional D-vine is given by

$$f_{1:5}(u_1,u_2,u_3,u_4,u_5) = \prod_{i=1}^{5} f_i(u_i) \prod_{j=2}^{5}\prod_{i=1}^{j-1} c_{i,j|(i+1):(j-1)}(u_i,u_j|u_{i+1},\cdots,u_{j-1}),$$

where $f_i(u_i)$ is marginal density and $c_{i,j|(i+1):(j-1)}$ is the bivariate copula density.

In order to factor vine copulas, we need to use bivariate linking copulas. One such copula is the Frank copula with

$$C(u_1,u_2) = -\eta^{-1}\log\left(1+\frac{\left(e^{-\eta u_1}-1\right)\left(e^{-\eta u_2}-1\right)}{e^{-\eta}-1}\right),$$

where $\eta \neq 0$ is the copula parameter. The Frank copula can capture the full range of bivariate dependence and has a symmetric dependence structure [40].

Notations: Table I summarizes the notations used in the article.

III. FRAMEWORK

In this section we present a framework for modeling and predicting multivariate time series with *sparse* data breach events (i.e., mostly one or two breach incidents are observed in each time series during the lifespan of a dataset). Figure 2 illustrates the kinds of data that can be analyzed by the framework, where n entities (e.g., enterprises) are observed over a time horizon T at a certain resolution (e.g., year) and each entity has very few breach incidents (e.g., only entity 2 has 2 breach incidents). In Section VI we will discuss the other application settings of the framework.

The framework consists of 5 steps: (i) data preprocessing and exploratory data analysis; (ii) modeling the occurrence of breach incidents; (iii) modeling the breach sizes; (iv) modeling and estimating dependence structures; and (v) predicting the distributions of breach sizes.

A. Step 1: Data Preprocessing and Exploratory Data Analysis

Given a specific breach incident dataset, the first step is to preprocess the dataset according to the time series representation depicted in Figure 2, denoted by $\{(i, y_{it})|1 \leq i \leq n,\ 1 \leq t \leq T\}$, where y_{it} is the breach size (i.e., number of breached records) of entity i in year t, n is the number of entities, and T is the time horizon at a certain resolution (e.g., year). Note that $y_{it} = 0$ means entity i has no data breach during time interval t. The framework is designed to cope with *sparse* multivariate time series, meaning $y_{it} = 0$ for excessively many i's and t's, which cannot be described by the existing statistical or deep learning models because there are not enough data to train them. In order to gain initial insights into the data, an exploratory data analysis should be conducted on $\{(i, y_{it})\}$, in terms of both the breach sizes (e.g., their mean, median, standard deviations, and quantiles) and the breach occurrences (e.g., breach frequency).

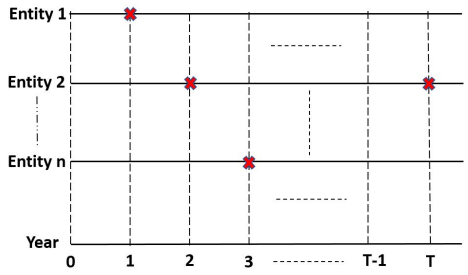


Fig. 2. Illustration of *spare* multivariate time series representation of data breach incidents over time horizon T, where an entity (e.g., enterprise) can be an enterprise and a red-cross represents an incident in a particular year.

B. Step 2: Modeling Occurrence of Breach Incidents

In order to model *sparse* multivariate time series, we propose considering the random variable breach size Y_{it} of entity i in time interval t, meaning that y_{it} is the observed value of Y_{it}. Then, we propose modeling the distribution of Y_{it}, denoted by $F_{it}(y)$, in two parts:

$$F_{it}(y) = p_{it}\mathrm{I}(y=0) + (1-p_{it})M_{it}(y), \qquad \text{(III.1)}$$

where $\mathrm{I}(\cdot)$ is the indicator function, p_{it} is the probability that entity i is *not* breached at time t (i.e., $p_{it} = \Pr[Y_{it} = 0]$) and $M_{it}(y)$ is the distribution of Y_{it} under condition $Y_{it} > 0$. The density function of Y_{it}, denoted by f_{it}, can be written as

$$f_{it}(y) = p_{it}\delta(y=0) + (1-p_{it})m_{it}(y), \qquad \text{(III.2)}$$

where $\delta(\cdot)$ is the Dirac delta function, $m_{it}(\cdot)$ is the density function of random variable M_{it}. It is worth mentioning that the strategy of studying p_{it} (as a means for coping with random variable Y_{it}) is reminiscent of what has happened in theoretical cybersecurity modeling [18], [42], [43]. In those settings, studying discrete security states would encounter the state-space explosion problem, which leads to the use of probability to represent that a network node is in a certain state at a certain point in time (cf, e.g., [30], [44]–[46]).

In order to describe p_{it}, we propose using the logit regression because it can accommodate potential *temporal trends* and *heterogeneities* [47], as follows:

$$\mathrm{logit}(p_{it}) = \boldsymbol{\beta}^T \boldsymbol{x}_i, \qquad \text{(III.3)}$$

where $\text{logit}(p_{it}) = \log(p_{it}/(1-p_{it}))$ is the logit function (widely used to model the odds ratio [47]) and $\boldsymbol{\beta}$ is a vector of coefficients of covariates $\boldsymbol{x}_i$ (for accommodating the potential temporal trend and heterogeneity mentioned above). The covariates can include time trend (e.g., t, t^2), categorical information (e.g., MED, BS, OTHER), and other information.

C. Step 3: Modeling Breach Sizes

Since previous studies showed that breach sizes may exhibit skewness and heavy tails (e.g., [8]), we propose using a mixed distribution to model them, namely using the Extreme Value Theory [48], [49] to model the extremely large breach sizes and using other distributions to model the other breach sizes. A popular approach to modeling extreme values is known as *Peaks Over Threshold* (POT) [48]. Given a sequence of i.i.d. observations $y_1, \ldots, y_n$, and a suitably-high threshold μ, the excesses $y_i - \mu$ can be modeled by, under certain mild conditions, the *generalized Pareto distribution* (GPD). The GPD distribution function can be written as

$$G(y|\mu,\sigma_\mu,\xi) = \begin{cases} 1 - \left[1 + \xi\left(\dfrac{y-\mu}{\sigma_\mu}\right)\right]_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left[-\left(\dfrac{y-\mu}{\sigma_\mu}\right)_+\right], & \xi = 0, \end{cases}$$

where $y_+ = \max(y, 0)$, μ is the threshold, $\sigma_\mu > 0$ and ξ are respectively the scale and shape parameters. If $\xi < 0$, the support is $\mu < y < \mu - \sigma_\mu/\xi$; otherwise, the support is unbounded from above. Since GPD only models the upper tail of the distribution above threshold μ, we need to model the breach sizes at or below the threshold μ. This prompts us to propose the following mixed model:

$$M_{it}(y|\boldsymbol{\Theta}) = \begin{cases} (1-\phi_{\mu_i})H_i(y|\boldsymbol{\Theta}), & y \leq \mu_i \\ (1-\phi_{\mu_i}) + \phi_{\mu_i} G_i(y|\boldsymbol{\Theta}), & y > \mu_i, \end{cases}$$

where $\boldsymbol{\Theta}$ is the parameter vector, $H_i(y|\boldsymbol{\Theta})$ is the distribution of breach sizes y's below the fitted threshold μ_i, ϕ_{μ_i} is the proportion of breach sizes above the threshold μ_i, and G_i is the GPD for entity i. This mixed model offers flexibility.

D. Step 4: Modeling and Estimating Dependence Structures

Having modeled the marginal distributions, the next step is to accommodate temporal dependence between the breach sizes, namely the dependence between $Y_{i1}, \ldots, Y_{iT}$. For this purpose, we propose using the afore-reviewed *copula* technique. For entity i, the joint distribution of breach sizes over the time horizon of T can be rewritten as

$$F_i(\mathbf{y}) = C(F_{i1}(y_1), \cdots, F_{iT}(y_T)),$$

where $\mathbf{y} = (y_1, \ldots, y_T)$, $F_{i1}, \cdots, F_{iT}$ are the marginals described in Eq. (III.1), and C is the copula structure modeling the temporal dependence across the time horizon.

In the literature, many copula structures have been proposed. An attractive structure is the afore-reviewed *vine* copula, which offers a great deal of flexibility in modeling dependence, including various kinds of tail dependencies and asymmetric dependencies [41]. One particularly attractive candidate is to use the afore-reviewed D-vine copula, which offers a great deal of flexibility in modeling pairwise dependencies. Conceptually, D-vine copula has the following attractive properties that make it suitable for the problem we study.

- *Flexibility in dealing with high-dimensional data*: Traditional multivariate copulas, such as multivariate Gaussian and exchangeable Archimedean, lack the flexibility in modeling dependence in high-dimensional data. By contrast, D-vine copula is flexible in modeling multivariate copulas via bivariate or pair-copula constructions [40], [50]. These constructions decompose a multivariate probability density into bivariate copulas, where each pair-copula can have different from and be independent of the others. That is, D-vine can flexibly model any dependence structure that can be captured by bivariate copulas (e.g., asymmetric dependence or strong joint tail behavior).
- *Efficiency*: It is known that D-vine copula has a very good fitting/prediction efficiency [39], [51], [52], because the model parameters can be efficiently estimated via the maximum likelihood estimation method [40], [50].
- *Temporal structure*: D-vine copula has a natural temporal structure, which makes it particularly suitable for time series data. As shown in Figure 1, D-vine is constructed via a specific order of variables (i.e., path structure) [40]. This temporal structure offers the prediction capabilities, which would not be offered by the other copula models that cannot accommodate such temporal structures.

Since breach sizes can be positive (when there are breaches) and zeros (when there are no breaches), the dependence structure is in fact a mixed form, leading to the use of mixed D-vine copula. When using the mixed D-vine copula structure to describe the joint density of the entities' breach sizes, the density of entity i's breach sizes can be rewritten as

$$f_i(\mathbf{y}) = \prod_{t=1}^{T} f_{i,t}(y_t) \prod_{t=2}^{T} \prod_{s=1}^{t-1} \widetilde{f}_{i,s,t|(s+1):(t-1)}(y_s, y_t | y_{(s+1):(t-1)}),$$

where $\mathbf{y} = (y_1, \ldots, y_T)$ as mentioned above, $f_{i,t}(y_t)$ is as describe in Eq. (III.2), and $\widetilde{f}_{i,s,t|(s+1):(t-1)}(y_s, y_t | y_{(s+1):(t-1)})$ is the ratio of the bivariate distribution to the product of the marginals with respect to the conditioning set described in Eq (III.4) shown at the bottom of the next page; see [52]–[54] for technical details. where $y_{(s+1):(t-1)} = (y_{s+1}, \cdots, y_{t-1})$, $C_{s,t;(s+1):(t-1)}(u_1, u_2)$ and $c_{s,t;(s+1):(t-1)}$ are respectively the distribution and density functions of a bivariate copula with conditional distributions $F_{is|(s+1):(t-1)}$ and $F_{it|(s+1):(t-1)}$, $1 \leq s < t \leq T$ with s indicating year s, $c_{j,s,t;(s+1):(t-1)}(u_1, u_2) = \partial C_{s,t;(s+1):(t-1)}(u_1, u_2)/\partial u_j$ for $j = 1, 2$,

In order to model multivariate data $\{(i, y_{it}) | 1 \leq i \leq n, 1 \leq t \leq T\}$, the log-likelihood function can be rewritten as

$$ll(\mathbf{y}_1, \ldots, \mathbf{y}_n) = \sum_{i=1}^{n} \sum_{t=1}^{T} \log f_{i,t}(y_{it}) + \sum_{i=1}^{n} \sum_{t=2}^{T} \sum_{s=1}^{t-1} \log \widetilde{f}_{i,s,t|(s+1):(t-1)}(y_{is}, y_{it} | y_{i,(s+1):(t-1)}), \quad \text{(III.5)}$$

where $\mathbf{y}_k = (y_{k1}, \ldots, y_{kT})$. In order to estimate the model parameters, we propose using the two-stage Inference Functions for Margins (IFM) approach [40]. This approach has two steps: (i) estimate the parameters of the marginal distributions; and (ii) estimate the dependence structures from the multivariate likelihood with the marginal parameters estimated in step (i). This approach is computationally efficient [39], [40].

Based on the D-vine structure, the tree set is $\mathbf{Tr} = (\mathrm{Tr}_1, \cdots, \mathrm{Tr}_{T-1})$. In principle, each tree can consist of various pair-dependence structures. However, for the purpose of prediction, a practical approach is to fix the dependence structure for each tree, meaning that we use the same copula structure within each tree but the copula structures corresponding to different trees can vary. We propose using Algorithm 1 to estimate the mixed D-vine dependence structure.

E. Step 5: Predicting Data Breach Risk

Since breach incidents are sparse, we propose predicting: (i) What is the probability that enterprise i will have a breach incident in the next time interval $t+1$, namely $1 - p_{i,t+1}$? (ii) What is the breach size, $Y_{i,t+1}$, under the condition that there will be a breach? Answering these two questions will provide more information than point prediction (i.e., what is the expected breach size at the next step?). In order to answer the preceding two questions, it is sufficient to predict the distribution of breach sizes one step ahead of time, as follows. Given historical breach sizes $\mathbf{y}_i = (y_{i1}, \cdots, y_{it})$ for entity i, the conditional density of $Y_{i,t+1}|\mathbf{y}_i$ can be rewritten as

$$f_{i,t+1|1:t}(y) = f_{i,t+1}(y) g(y, \mathbf{y}_i), \tag{III.6}$$

where $1 \le i \le n$ and

$$g(y, \mathbf{y}_i) = \prod_{s=2}^{t} \widetilde{f}_{i,s,t+1|(s+1):t}(y_s, y|y_{(s+1):t}), \tag{III.7}$$

Algorithm 1 Estimating the Mixed D-Vine Dependence Structure Between Breach Sizes in multivariate Time series

Input: Historical breach sizes $\{(i, y_{it})|1 \le i \le n, 1 \le t \le T\}$; pair copula set $\mathbf{\Omega}$.

Output: The full mixed D-vine copula structure

```
Estimate the marginal distributions described in Eq. (III.1) and the density functions described in Eq. (III.2)
for j = 1, ..., t do
  if j = 1 then
    For copula structure Tr_j, select the copula in Ω that leads to the maximum likelihood in Eq. (III.5)
  else
    Fix the copula structures in Tr_1, ..., Tr_{j-1}, and select the copula that leads to the maximum likelihood value in Eq. (III.5)
  end if
end for
Return Tr = (Tr_1, ..., Tr_{t-1})
```

where $\widetilde{f}_{i,s,t+1|(s+1):t}$ is defined in Eq (III.4). We propose using Algorithm 2 to predict the distributions of entities' breach sizes one-step ahead of time. Given a predicted distribution of $Y_{i,t+1}|\boldsymbol{y}_i$ for enterprise i in time interval $t+1$, one can easily answer the preceding two motivating questions.

Metrics for evaluating the accuracy of predicted distributions. For evaluating the accuracy of predicted distribution, most metrics (e.g., mean square error) are not competent. Instead, we can use the following two statistical approaches: *ranked probability score* and *uniform test*. The ranked scoring rule provides a summary measure for evaluating probability forecasting by assigning a numerical score based on the predicted distribution and observations. One popular scoring

$$\begin{aligned}
&\widetilde{f}_{i,s,t|(s+1):(t-1)}(y_s, y_t|y_{(s+1):(t-1)}) \\
&\quad = \begin{cases}
\dfrac{C_{s,t;(s+1):(t-1)}(F_{is|(s+1):(t-1)}(0|y_{(s+1):(t-1)}), F_{it|(s+1):(t-1)}(0|y_{(s+1):(t-1)}))}{F_{is|(s+1):(t-1)}(0|y_{(s+1):(t-1)}) F_{it|(s+1):(t-1)}(0|y_{(s+1):(t-1)})}, & y_s = 0, y_t = 0, \\
\dfrac{c_{1,s,t;(s+1):(t-1)}(F_{is|(s+1):(t-1)}(y_s|y_{(s+1):(t-1)}), F_{it|(s+1):(t-1)}(0|y_{s+1}, \cdots, y_{t-1}))}{F_{it|(s+1):(t-1)}(0|y_{(s+1):(t-1)})}, & y_s > 0, y_t = 0, \\
\dfrac{c_{2,s,t;(s+1):(t-1)}(F_{is|(s+1):(t-1)}(0|y_{(s+1):(t-1)}), F_{it|(s+1):(t-1)}(y_t|y_{(s+1):(t-1)}))}{F_{is|(s+1):(t-1)}(0|y_{(s+1):(t-1)})}, & y_t > 0, y_s = 0, \\
c_{s,t;(s+1):(t-1)}(F_{is|(s+1):(t-1)}(y_s|y_{(s+1):(t-1)}), F_{it|(s+1):(t-1)}(y_t|y_{(s+1):(t-1)})), & y_t > 0, y_s > 0,
\end{cases}
\end{aligned} \tag{III.4}$$

$$\begin{aligned}
&F_{is|(s+1):(t-1)}(y_s|y_{(s+1):(t-1)}) \\
&\quad = \begin{cases}
\dfrac{C_{s,t-1;(s+1):(t-2)}(F_{is|(s+1):(t-2)}(y_s|y_{(s+1):(t-2)}), F_{i(t-1)|(s+1):(t-2)}(0|y_{(s+1):(t-2)}))}{F_{i(t-1)|(s+1):(t-2)}(0|y_{(s+1):(t-2)})}, & y_{t-1} = 0, \\
c_{2,s,t-1;(s+1):(t-2)}(F_{is|(s+1):(t-2)}(y_s|y_{(s+1):(t-2)}), F_{i(t-1)|(s+1):(t-2)}(y_{t-1}|y_{(s+1):(t-2)})), & y_{t-1} > 0,
\end{cases}
\end{aligned}$$

and

$$\begin{aligned}
&F_{it|(s+1):(t-1)}(y_t|y_{(s+1):(t-1)}) \\
&\quad = \begin{cases}
\dfrac{C_{t,s+1;(s+2):(t-1)}(F_{it|(s+2):(t-1)}(y_t|y_{(s+2):(t-1)}), F_{i(s+1)|(s+2):(t-1)}(0|y_{(s+2):(t-1)}))}{F_{i(s+1)|(s+2):(t-1)}(0|y_{(s+2):(t-1)})}, & y_{s+1} = 0, \\
c_{2,t,s+1;(s+2):(t-1)}(F_{it|(s+2):(t-1)}(y_t|y_{(s+2):(t-1)}), F_{i(s+1)|(s+2):(t-1)}(y_{s+1}|y_{(s+2):(t-1)})), & y_{s+1} > 0.
\end{cases}
\end{aligned}$$

Algorithm 2 Predicting Distributions of Breach sizes

Input: Historical data $\{\mathbf{y}_i\}_{i=1,\ldots,n}$ where $\mathbf{y}_i = (y_{i1}, \ldots, y_{it})$; mixed D-vine structure; sample size $B = 50,000$.
Output: Predicted distribution of $Y_{i,t+1}|\mathbf{y}_i$.

1: **for** entity $i = 1$ to n **do**
2: Estimate p_{it} according to Eq. (IV.1)
3: Randomly generate B samples from $\{0, 1\}$ with probability p_{it} for 0 and with the number of 1's in the B samples denoted by M_i
4: Draw B_i samples from $f_{i,t+1}$ according to Eq. (III.6) using the rejection sampling approach [55], and denote them by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iM_i})$
5: Record the simulated vector $\hat{\boldsymbol{y}}_{i,t+1} = (\mathbf{0}, \boldsymbol{x}_i)$
6: **end for**
Return $\hat{\boldsymbol{y}}_{i,t+1}$, $i = 1, \ldots, n$.

rule is the ranked probability score (RPS) [56], [57]:

$$\mathrm{RPS}(x) = \sum_{k=0}^{\infty} \left(P_k - \mathrm{I}(x \le k)\right)^2, \tag{III.8}$$

where P_k is the predicted distribution and x is the observation. In terms of expectation, the RPS can be represented as

$$\mathrm{RPS}(x) = E|X - x| - \frac{1}{2}E|X - X'|,$$

where X and X' are independent copies of a random variable following the predicted distribution. Therefore, the RPS can compare both the point forecasts and predicted distributions; the smaller the RPS, the more accurate the prediction.

The selection of RPS can be justified as follows. Note that we aim to conduct probability forecasts (i.e., predicting the probability for each possible breach size), rather than point forecasts (i.e., the expectation of a random variable). The other reason for using probability forecasts is that the data we study is sparse, meaning that point prediction is not informative in capturing the inherent uncertainty. In order to quantify prediction accuracy, we need to measure the distance between the observed data and the predicted distribution. Simple accuracy measures, such as MAE (Mean Absolute Error) or MSE (Mean Square Error) [58], are not applicable because they cannot capture the difference between two distributions. RPS is suitable because it generalizes MAE to accommodate distributions. Intuitively, RPS measures the difference between the forecast cumulative distribution function (CDF) and the empirical CDF of observed values. RPS is a widely-used accuracy measure dealing with probability forecasts [56], [59].

The uniform test [60] is another popular metric for assessing the prediction accuracy. It is based on the notion of Probability Integral Transform (PIT) and examines whether the predicted distribution and the observations coincide. Let $F_{i,t+1|(1:t)}(\cdot)$ be the predicted distribution of an enterprise's breach size at time $t + 1$ as discussed in Eq. (III.6). If the mixed D-vine model is accurate, then $u_{i,t+1} = F_{i,t+1|(1:t)}(y_{i,t+1})$ for $i = 1, \ldots, n$ and is uniformly distributed, where the $y_{i,t+1}$'s are the actual breach sizes observed at time $t + 1$. Since the breach sizes are mixed (i.e., many zeros and some extremely large breach sizes), the PIT widely-used in the literature cannot be directly applied. Therefore, we propose the following randomized approach to conducting the uniform test. As shown in Eq. (III.1), we have a probability $p_{i,t+1}$ of 0's for enterprise i, which results in a lower bound for $u_{i,t+1}$. Let v_i be a random sample from interval $[0, 1]$. If $u_{i,t+1} = p_{i,t+1}$, then the probability $u_{i,t+1}$ is replaced by $v_i\, p_{i,t+1}$. The uniform test is performed on the new randomized set of $u_{i,t+1}$'s. The goodness-of-fit statistics can be further used to assess the uniform test.

IV. Case Study With Real-World Data

In this section we conduct a case study by applying the framework to a specific breach dataset obtained from [1], which is, to the best of our knowledge, the most comprehensive source of such data. The case study focuses on enterprise-level breach incidents in the United States, meaning that each enterprise corresponds to an entity in the framework. This is a natural choice because cyber risk management is often conducted at the enterprise level.

A. Data Preprocessing and Exploratory Data Analysis

1) Dataset and Preprocessing: The dataset describes breach incidents corresponding to $n = 4,300$ enterprises between 2005 and 2018 (for a total time span of 14 years). We do not consider the 2019 breaches because the source (i.e., the website [1]) appears to have stopped updating its content in 2019 (as evidenced by the extremely few nonzeros). The 4,300 enterprises span across 7 industries, including: 295 businesses-financial and insurance service enterprises (BSF for short); 252 businesses-retail/merchant ones including online retail (BSR for short); 355 businesses-other ones (BSO for short); 434 educational institutions (EDU for short); 440 government and military ones (GOV for short); 2,459 healthcare, medical provider and medical insurance service ones (MED for short); and 65 nonprofit organizations (NGO for short).

In the dataset, each breach incident is described by (i) the date when the incident is reported to the website [1], rather than when the breach takes place (which may not be known to the enterprise in question), and (ii) the breach size defined in the framework, namely the number of data records that are breached in an incident. In order to structure the data into the time series representation that is required by the framework, we aggregate the breach incidents for each enterprise on a yearly basis (i.e., the time horizon is $T = 14$ years). That is, when enterprise i has multiple breaches reported in a single year, we add these breach sizes together to derive y_{it}, which is the single *virtual* incident in year t. This aggregation does not significantly "poison" the properties of the multivariate time series because among the 4,300 enterprises, only 0.6% have 2 breach incidents in a single year, 0.1% have 3 breach incidents in a single year, and 0.03% have 4 or more breach incidents in a single year. This aggregation is meant to facilitate modeling because these multivariate time series are already sparse (i.e., using a higher time resolution, such as

TABLE II
STATISTICS OF LOG-TRANSFORMED BREACH SIZES OR NONZERO y_{it}'S, WHERE 'SD' STANDS FOR STANDARD DEVIATION, Q_1 AND Q_3 REPRESENT THE FIRST AND THIRD QUANTILES

	Min	Q_1	Median	Mean	SD	Q_3	Max
BS	0.000	5.521	7.901	8.451	3.911	10.925	21.976
MED	0.000	6.740	7.783	7.859	2.319	9.146	18.182
OTHER	0.000	6.215	8.006	8.019	2.696	9.798	18.146

day or month, would make the time series even sparser and much harder to model).

The framework can model the heterogeneity between the $n = 4,300$ time series $\{(i, y_{it}) | 1 \le i \le 4,300,\ 1 \le t \le 14\}$, as long as some distinct information about these enterprises, other than $\{(i, y_{it})\}$, is available (e.g., their cyber defense postures or data server configurations). In order to overcome the lack of this kind of information, we propose grouping the 4,300 enterprises into some categories such that the enterprises in a same category may exhibit similar characteristics. This prompts us to group enterprises according to the sectors to which they belong. Specifically, we propose putting the three business-related enterprises (i.e., BSF, BSR, and BSO) into a category called BS, keeping the MED as a category because it has a large number of enterprises already, and putting the three not-for-profit enterprise (i.e., EDU, GOV, and NGO) into a category called OTHER. This leads to 902 BS enterprises, 2,459 MED enterprise, 939 OTHER enterprises.

2) Exploratory Data Analysis: The sparse time series $\{(i, y_{it}) | 1 \le i \le 4,300,\ 1 \le t \le 14\}$ are challenging to analyze because $y_{it} = 0$ for excessively many i's and t's. Specifically, there are on average (among the enterprises in a category) 91.65%, 91.91% and 90.41% 0's in the BS, MED and OTHER categories, respectively. Nevertheless, there are 92 BS, 259 MED, and 146 OTHER enterprises that have multiple breach incidents during the 14 years of time span.

Table II summarizes the log-transformed breach sizes (i.e., the nonzero y_{it}'s (i.e., discarding the 0's), where the log-scale is used because some y_{it}'s are extremely skewed (i.e., extremely large). We make the following observations. First, the BS enterprises have the largest mean breach size and standard deviation (SD). Second, the median is always smaller than the mean in each enterprise category, suggesting that the breach sizes are extremely skewed. Third, among the three categories, BS has the smallest first quantile (Q_1) and the largest third quantile (Q_3). Fourth, the MED enterprises have the smallest mean, median, and standard deviation, suggesting they have smaller customer populations of similar size.

In order to further expose their statistical properties, Figure 3 plots the histograms of the log-transformed nonzero breach sizes. We again observe the skewness and variability in each category, with extremely small and extremely large data breach sizes in each category. The distributions of the three categories are different, hinting at heterogeneities between them. In summary, there are large standard deviations and skewnesses in the breach sizes of each enterprise category and some breach sizes are especially large. This justifies the use of the log-transformation to reduce the variability and skewness exhibited by the data, which is important for modeling purposes.

Insight 1: Data breach incidents are sparse, data breach sizes exhibit large variability and large skewness, and different kinds of enterprises exhibit different breach characteristics.

Figure 4 presents the stacked bar plots of breach incidents across the 14 years. We observe that different enterprise categories exhibit different breach patterns. In the first few years, the OTHER category has the most breaches and the MED category has the least. The MED category exhibits a clear increasing trend; the BS category exhibits a fluctuating pattern; the OTHER category exhibits an increase initially and then decreases. Overall, the MED category is the highest, suggesting that they are more vulnerable; the OTHER category has been improving, perhaps because they (especially the EDU and GOV enterprises) have been investing more efforts at cyber defense. When compared to the OTHER category, the occurrence of breaches in the BS category becomes worse. These observations offer the following insight:

Insight 2: When modeling the data, the temporal effect and the inter-category *heterogeneity* must be addressed.

In what follows, we use the data between 2005 and 2017 to develop statistical models and use the data in 2018 as the out-of-sample data for assessing the prediction performance (or accuracy). This is reasonable because the time series are short (i.e., $T = 14$ years) and sparse (i.e., excessively many 0's during the 14 years). That is, we have to use the breach information as much as possible to build robust models.

B. Modeling Occurrence of Breach Incidents

Guided by the framework, we now move to model $p_{it} = \Pr[Y_{it} = 0]$, the probability that enterprise i does not have a breach incident in year t. For this purpose, we propose using the logit regression because it can accommodate temporal trends and inter-enterprise heterogeneities observed by the data. The model, denoted by $\mathcal{M}$ is:

$$\begin{aligned}\text{logit}(p_{it}) = {} & \beta_0 + \beta_1 t + \beta_2 \text{I(MED)} + \beta_3 \text{I(BS)} \\ & + \beta_4 t^2 + \beta_5 \text{I(MED)} t + \beta_6 \text{I(BS)} t \\ & + \beta_7 \text{I(MED)} t^2 + \beta_8 \text{I(BS)} t^2, \end{aligned} \tag{IV.1}$$

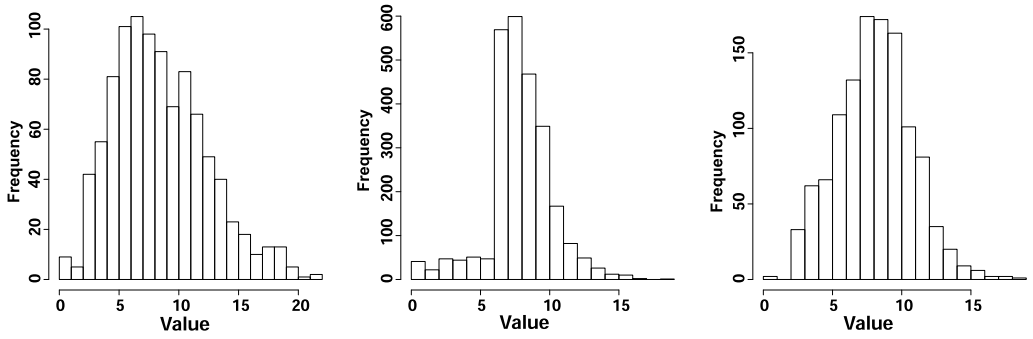
where BS and MED are two enterprise categories mentioned above while the OTHER is set to be the baseline, $\text{I}(\cdot)$ is the indicator function, $\text{logit}(p_{it}) = \log(p_{it}/(1 - p_{it}))$ is the logit function, β_1 models the linear time trend, and (β_2, β_3) models the category heterogeneity, β_4 models the quadratic time trend, (β_5, β_7) and (β_6, β_8) model the quadratic time trends within a category. For comparison purposes, we also consider the following four variants of model $\mathcal{M}$:

- $\mathcal{M}1$: Discard the second-order term in each category:

$$\begin{aligned}\text{logit}(p_{it}) = {} & \beta_0 + \beta_1\ t + \beta_2 \text{I(MED)} + \beta_3 \text{I(BS)} \\ & + \beta_4\ t^2 + \beta_5 \text{I(MED)} t + \beta_6 \text{I(BS)} t. \end{aligned}$$

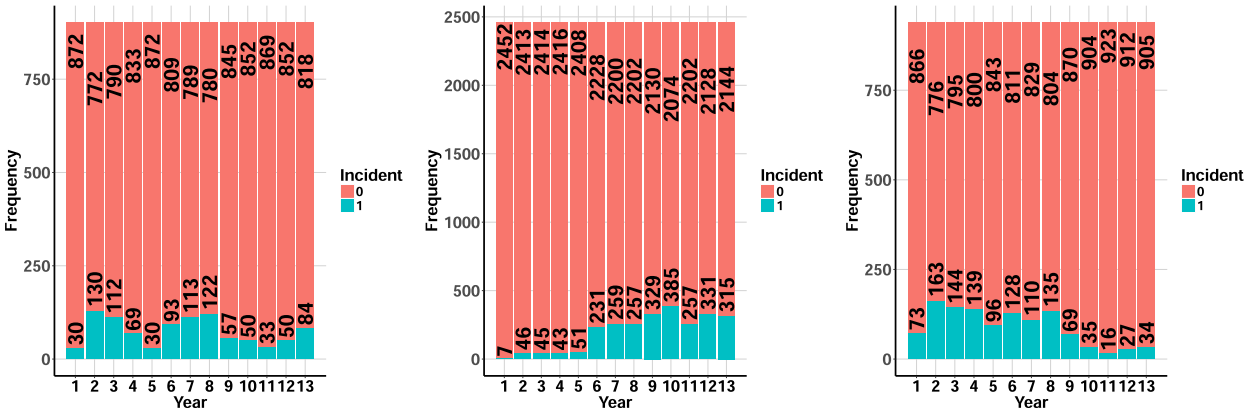
- $\mathcal{M}2$: Discard the time-related intra-category heterogeneity:

$$\text{logit}(p_{it}) = \beta_0 + \beta_1\ t + \beta_2 \text{I(MED)} + \beta_3 \text{I(BS)} + \beta_4\ t^2.$$



(a) BS (b) MED (c) OTHER

Fig. 3. Histograms of the log-transformed breach sizes in different categories, where the x-axis represents the log-transformed breach sizes.



(a) BS (b) MED (c) OTHER

Fig. 4. Bar charts of the breach incident frequencies, where color '0' means there is no incident and color '1' means there is incident.

TABLE III

ESTIMATED PARAMETERS AS WELL AS THEIR STANDARD DEVIATIONS (SD), THE LOG-LIKELIHOOD VALUE $\log(L)$ AND AIC, WHERE "$**$" INDICATES A p-VALUE IS LESS OR EQUAL TO 0.001 AND "$*$" INDICATES A p-VALUE IS GREATER THAN 0.001 BUT LESS THAN OR EQUAL TO 0.01

	$\mathcal{M}$		$\mathcal{M}1$		$\mathcal{M}2$		$\mathcal{M}3$		$\mathcal{M}4$	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
β_0	2.2561$**$	.1054	2.2253$**$	.0765	3.2670$**$	.0716	1.4547$**$	.0580	2.6853$**$	.0442
β_1	-.2338$**$	.0383	-.2203$**$	.0213	-.2702$**$	.0198	.1254$**$	.0087	-.0599$**$	.0041
β_2	3.8832$**$	.1953	3.2432$**$	.1050	.1988$**$	.0371	2.5199$**$	.0839	.1985$**$	.0371
β_3	.2271	.1582	.9195$**$	.0997	.1582$**$	.0455	.7473 $**$	.0898	.1580$**$	.0455
β_4	.0285$**$	.0029	.0274$**$	.0015	.0143$**$	.0013	–	–	–	–
β_5	-.5828$**$	.0557	-.4123$**$	.0138	–	–	-.3151$**$	.0107	–	–
β_6	.1483$*$	.0548	-.1248$**$	.0144	–	–	-.0960$**$	.0125	–	–
β_7	.0099$*$	.0038	–	–	–	–	–	–	–	–
β_8	-.0201$**$	.0040	–	–	–	–	–	–	–	–
$\log(L)$	-15266.80		-15300.47		-15945.85		-15479.87		-16008.20	
AIC	30551.60		30615.00		31901.70		30971.74		32024.39	

- $\mathcal{M}3$: Discard the nonlinear time trend:

$$\text{logit}(p_{it}) = \beta_0 + \beta_1\ t + \beta_2\text{I(MED)} + \beta_3\text{I(BS)} + \beta_5\text{I(MED)}t + \beta_6\text{I(BS)}t.$$

- $\mathcal{M}4$: Discard the time-related intra-category heterogeneity and the nonlinear time trend:

$$\text{logit}(p_{it}) = \beta_0 + \beta_1\ t + \beta_2\text{I(MED)} + \beta_3\text{I(BS)}.$$

Table III summarizes the estimated parameters as well as their standard deviations, the log-likelihood values and Akaike Information Criterion (AIC [49]). Among the 5 models, M has the smallest AIC and the largest log-likelihood value, meaning that M is preferred. For model $\mathcal{M}$, we make the following observations. First, all of the parameters are significant at the .01 level, except for β_3. Second, $\beta_1 < 0$ and $\beta_4 > 0$ and the model fits well the overall nonlinear pattern observed in Figure 4. Third, $\beta_2 > \beta_3 > 0$, which matches the pattern observed in the initial periods of time in Figure 4. Fourth, $\beta_5 = -.5828$ and $\beta_7 = .0099$ suggest that the MED category shows an increasing nonlinear pattern, and $\beta_6 = .1483$ and $\beta_8 = -.0201$ suggest that the BS category shows a decreasing

TABLE IV

THE OBSERVED PROBABILITIES AND PREDICTED PROBABILITIES BASED ON MODEL $\mathcal{M}$

	BS		OTHER		MED	
Year	Predicted	Observed	Predicted	Observed	Predicted	Observed
1	0.9173	0.9667	0.8860	0.9223	0.9953	0.9972
2	0.9126	0.8559	0.8702	0.8264	0.9906	0.9813
3	0.9091	0.8758	0.8595	0.8466	0.9826	0.9817
4	0.9068	0.9235	0.8553	0.8520	0.9703	0.9825
5	0.9060	0.9667	0.8581	0.8978	0.9533	0.9793
6	0.9065	0.8969	0.8675	0.8637	0.9323	0.9061
7	0.9085	0.8747	0.8824	0.8829	0.9094	0.8947
8	0.9118	0.8647	0.9011	0.8562	0.8875	0.8955
9	0.9162	0.9368	0.9213	0.9265	0.8702	0.8662
10	0.9217	0.9446	0.9409	0.9627	0.8601	0.8434
11	0.9280	0.9634	0.9582	0.9830	0.8590	0.8955
12	0.9348	0.9446	0.9722	0.9712	0.8670	0.8654
13	0.9420	0.9069	0.9826	0.9638	0.8828	0.8719
MAE	0.0331		0.0784		0.0130	
MSE	0.0014		0.0088		0.0003	

pattern first and then an increasing nonlinear pattern. These coincide with what is shown in Figure 4.

Table IV presents the observed probabilities and the probabilities predicted by Model $\mathcal{M}$, as well as the derived MAE and MSE. We observe that the predicted values and the observed values match well, and MAE/MSE is small. Thus, Model $\mathcal{M}$ has satisfactory prediction accuracy.

Insight 3: Different enterprise categories exhibit different nonlinear patterns of breach incident occurrence. The pattern of each category can be modelled by the logit regression with temporal trend and intra-category heterogeneity.

C. Modeling Breach Sizes

Guided by the framework, we model the log-transformed nonzero breach sizes via the mixed model mentioned there. Since there are large skewness and variability (Insight 1) and inter-category heterogeneity (Insight 3), we propose using different mixed distributions for the three categories. For the BS and MED categories, we propose using nonparametric distributions to fit the breach sizes below the threshold because of the following: (i) nonparametric distributions are data-driven and offer flexibility in modeling skewed data; and (ii) breach sizes in the BS and MED categories are very skewed, meaning that parametric distributions may not be able to capture the complex pattern exhibited by them. Specifically, we propose using the following kernel density for the nonparametric distribution:

$$h(x; y, \lambda) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{y - y_i}{\lambda}\right), \tag{IV.2}$$

where $K(\cdot)$ is the kernel function (i.e., a symmetric and unimodal probability density function) and λ is the bandwidth. More specifically, for the BS category we use the Gaussian kernel function (a special case of (IV.2)); for the MED category, we use the uniform kernel function (a special case of (IV.2)) because they have better fitting performances based on our experiments. For determining the value of λ, we propose employing the cross-validation likelihood approach [61].

For the OTHER category, we observed that the breach sizes are less skewed than their counterparts in the BS and MED categories (cf. Figure 3 and the accompanying discussion). This prompts us to use the Gaussian distribution $H(y|\mu_G, \sigma_G)$ to fit the breach sizes below the threshold, where μ_G and σ_G are the mean and standard deviation of the Gaussian distribution, respectively.

TABLE V

ESTIMATED PARAMETERS OF THE MIXED MODEL AND THEIR STANDARD DEVIATIONS (SD)

Par.	λ	μ	σ_μ	ξ	ϕ_μ	μ_G	σ_G
BS							
EST.	0.096	13.490	3.351	-0.342	0.111	–	–
SD	0.032	0.003	0.417	0.082	–	–	–
MED							
EST.	0.182	10.117	1.574	-0.074	0.128	–	–
SD	0.0003	0.007	0.144	0.063	–	–	–
OTHER							
EST.	–	11.049	1.354	0.027	0.127	8.014	2.669
SD	–	0.004	0.177	0.100	–	0.079	0.061

Table V presents the estimated parameters corresponding to the three enterprise categories and their standard deviations. We observe that the estimated λ, μ, σ_μ are very significant. For the shape parameter, the model corresponding to the BS enterprises has $\xi = -0.342$ and standard deviation 0.082, meaning that there is an upper bound on breach size. For the models corresponding to the MED and OTHER categories, their shape parameters are not significant, meaning that their tails are similar to that of the exponential distribution. For the OTHER category, we further observe that the estimated μ_G and σ_G are significant.

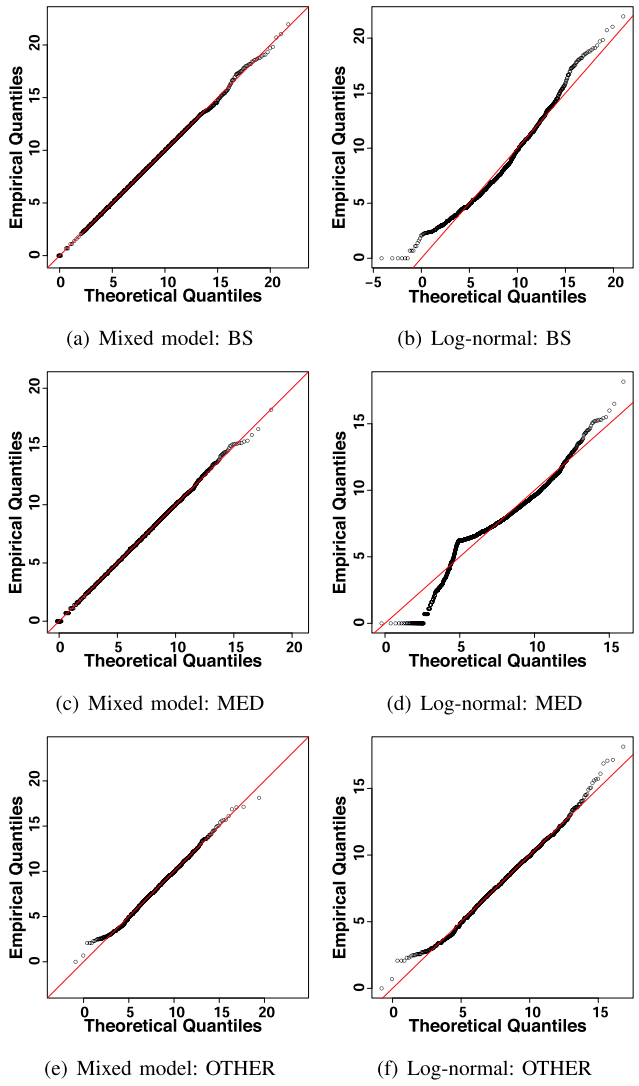
In order to further assess the fitting accuracy, Figure 5 depicts the QQ-plots of the proposed mixed models for the BS, MED, and OTHER enterprise categories, respectively. We observe that all of the points are very close to the 45-degree lines, meaning that the proposed mixed models have very satisfactory fitting accuracy. For comparison purposes, Figure 5 also depicts the QQ-plots of fittings by the log-normal distributions. We observe that the log-normal distribution has very poor fitting accuracy for both tails.

Insight 4: Breach sizes of different enterprise categories should be modeled with different distributions.

D. Modeling and Estimating Dependence Structures

1) Fitting Dependence With Vine Copula: Now we study the dependence structure between breach sizes during the first 13 years. Recall that in the D-vine structure, the tree set is $\mathbf{Tr} = (\mathrm{Tr}_1, \cdots, \mathrm{Tr}_{12})$. Guided by the framework, we use the same copula structure for one tree but different copula structures for other trees, and use Algorithm 1 to estimate the D-vine dependence structure, by setting $\Omega =$ {Gaussian, Frank, Rotated Joe, Rotated Gumbel, Rotated Clayton, Gumbel, Clayton}, which are widely-used bivariate copulas (noting that the copulas that are not selected are not reviewed in Section II but are referred to [40]).

Table VI summarizes the selected copula structures and estimated parameters. We observe that the Frank copula is selected for all of the trees. Frank copula is better than the others perhaps for two reasons. (i) Frank copula can



(a) Mixed model: BS (b) Log-normal: BS

(c) Mixed model: MED (d) Log-normal: MED

(e) Mixed model: OTHER (f) Log-normal: OTHER

Fig. 5. QQ-plots of the mixed models for the three enterprise categories.

TABLE VI

ESTIMATED D-VINE COPULAS FOR BREACH SIZES (SD MEANS STANDARD DEVIATION AND τ IS THE KENDALL τ)

	Copula	Est	SD	τ
Tr_1	Frank	-1.2395	0.1043	-0.1337
Tr_2	Frank	-1.7257	0.1168	-0.1814
Tr_3	Frank	-2.0365	0.12677	-0.21197
Tr_4	Frank	-1.8736	0.13212	-0.1959
Tr_5	Frank	-2.4145	0.1572	-0.2490
Tr_6	Frank	-2.5396	0.1771	-0.2613
Tr_7	Frank	-2.7292	0.2114	-0.2799
Tr_8	Frank	-3.0183	0.2785	-0.3083
Tr_9	Frank	-3.2444	0.3292	-0.3260
Tr_{10}	Frank	-3.8022	0.4264	-0.3672
Tr_{11}	Frank	-2.9971	0.4238	-0.3062
Tr_{12}	Frank	-4.2228	0.8760 3	-0.3982

accommodate both positive dependence and negative dependence, while noting that negative dependence is exhibited by the data analyzed. By contrast, copulas such as Gumbel and Clayton can only accommodate positive dependence. (ii) Frank copula has a nebulous but uniform cloud along the full correlation path. This makes it suitable for fitting the data that exhibits the uniform dependence. The dependence in the data we analyze does not concentrate at any part of the distribution (e.g., tails), perhaps because of the data sparsity issue; rather, the dependence exhibits the uniformity to some extent. This makes Frank copula able to capture the dependence well. In particular, the parameters of the Frank copula are always negatives and significant, meaning that there is a negative dependence across the years. This negative dependence hints that when there were no breaches to an enterprise in the past, a breach is anticipated to occur; when there was a breach to an enterprise in the past, it is unlikely that another breach will occur to the same enterprise within a short period of time. This negative dependence may be attributed to the fact that the attacker is not interested in breaking into the same enterprise perhaps because there is not much new data to breach, or the fact that the breached enterprise has fixed the vulnerability that was exploited by the attacker. Moreover, the Kendall's τ shows a decreasing trend when moving from the lower order trees to the higher order trees, meaning that there is an even more significant negative dependence between the higher order trees, except for Tr_4 and Tr_{11}. This suggests that higher order dependences cannot be ignored, which is contrary to the truncated dependence modeling in the literature [62].

Insight 5: Enterprise-level breach sizes exhibit a negative temporal dependence, meaning consecutive breaches are unlikely to occur to a single enterprise within a short period of time.

2) Model Comparison: Now we compare the fitting performance of the proposed mixed D-vine model to the ones that are commonly used in the literature [40], [47], including the independence (or benchmark) model, the Gaussian dependence model, and the linear mixed model. The benchmark model assumes that there is no temporal dependence between the $\mathbf{y}_t$'s. The Gaussian dependence model uses the Gaussian dependence structure to describe the temporal dependence. The linear mixed model (LMM) is widely used for longitudinal data analysis. Putting into the context of the present paper and using the AIC criterion, we select the following linear mixed model after checking various mixed models:

$$y_{it} = \beta_0 + a_i + (\beta_1 + b_i)t + \beta_2 \text{I(BS)} + \beta_3 \text{I(MED)} + (\beta_4 + c_i)\text{I(BS)}t + (\beta_5 + d_i)\text{I(MED)}t + \epsilon_{it},$$

where $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ represents the fixed effects and (a_i, b_i, c_i, d_i) represents the random effects. In particular, we allow the exchangeable dependence across the years corresponding to a single enterprise.

Table VII summarizes the fitting results. We observe that the proposed mixed D-vine model leads to the smallest AIC and

TABLE VII

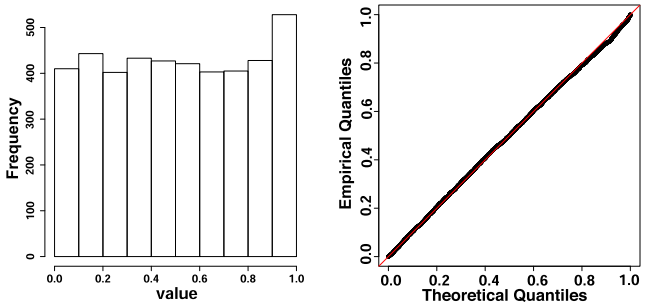
FITTING RESULTS OF VARIOUS MODELS FOR THE BREACH DATA

	log likelihood	AIC
Mixed D-vine	**-40,710.65**	**81,491.30**
Gaussian	-41,193.13	82,456.26
Benchmark	-41,395.07	82,826.14
LMM	-127,253.50	254,079.90

TABLE VIII

MEAN RPSS OF THE MIXED D-VINE WITH FRANK COPULA MODEL AND OTHER MODELS, WHERE *Percentage* IS THE % OF THE RPSS OF THE MIXED D-VINE MODEL THAT ARE LESS THAN THAT OF THE OTHER MODEL(S)

	BS	OTHER	MED	Overall
	Mean Score			
Mixed D-vine with Frank	**1.1135**	**.5521**	**2.1593**	**1.5889**
Mixed D-vine with Gaussian	1.1136	.5521	2.1600	1.5893
The benchmark model	1.1142	.5521	2.1609	1.5900
LMM model	4.7861	4.5210	5.2155	4.9738
	Percentage			
Frank vs Benchmark	66.52%	50.37%	60.80%	**59.72%**
Frank vs Gaussian	69.40%	36.95%	82.11%	**69.58%**
Frank vs LMM	95.01%	97.02%	87.72%	**91.28%**



(a) Histogram (b) QQ-plot

Fig. 6. Histogram of predicted distribution and its uniform QQ-plot.

the largest log-likelihood value, and that LMM has the worst fitting performance. Therefore, we conclude that the proposed mixed D-vine model has the best fitting performance.

Insight 6: The mixed D-vine dependence structure can accommodate the complex dependence exhibited by the enterprise-level breach data in all of the categories.

E. Predicting Data Breach Risk

Guided by the framework, we apply Algorithm 2 to predict the data breach risk for each enterprise i, namely the distribution of $Y_{i,t+1}|\mathbf{y}_i$. In what follows, we first assess the accuracy of the predicted distribution using the two metrics defined in the framework and then show how to use the predicted distribution to answer the two motivating questions.

In terms of the RPS-based evaluation of the accuracy of the predicted distribution, Table VIII summarizes the mean RPSs of the mixed D-vine model and other models. We observe that the average RPS of the mixed D-vine model is the smallest among all of the models both in each category and overall. Since the average RPSs are small, we compute the percentage of the RPSs of the mixed D-vine model that are less than that of the other models. We observe that the mixed D-vine model outperforms the benchmark model by 9.72%, the Gaussian model by 19.58%, and the LMM model by 41.28%.

In terms of the uniform test-based evaluation on the accuracy of the predicted distribution, Figure 6a plots the histogram of the distribution predicted by the mixed D-vine model, showing an almost uniform distribution. Figure 6b depicts the qq-plot of the predicted distribution, showing that all of the points are around the 45-degree line. Thus, the mixed D-vine model has a satisfactory prediction accuracy. The value of the Kolmogorov–Smirnov test statistics is small (.023) with p value .186, which is much larger than the significant level .05. All of these evidences suggest that the predicted distribution is satisfactory.

Insight 7: The proposed mixed D-vine model can predict the distribution of enterprise-level data breach sizes.

Having shown that the predicted distribution of $Y_{i,t+1}$, or $Y_{i,14}$ in the present case study, is accurate, now we show how to use the predicted distribution to answer the two motivating questions: (i) What is the probability that enterprise i will have a breach incident in the next time interval $t+1$, namely $1-p_{i,t+1}$? (ii) What is the breach size, $Y_{i,t+1}$, under the condition that there indeed will be a breach incident?

TABLE IX

STATISTICS AND PERCENTILES, WHICH ARE DERIVED FROM THE PREDICTED DISTRIBUTION OF $Y_{i,14}$, FOR THREE ENTERPRISES RANDOMLY SELECTED FROM EACH OF THE BS, OTHER, AND MED CATEGORIES

Enterprise	BS	OTHER	MED
$p_{i,14}$	0.9492	0.9897	0.9037
Min	1.0	1.0	1.0
Q_1	223.6	477.1	788.9
Mean	985,792.5	199,187.8	75,241.9
SD	111,596,686	2,592,043	1,203,079
Q_3	47,226.1	18,601.1	8,210.7
Max	2,846,163,649	53,530,320	47,288,461
10%	38.6	90.7	367.1
20%	126.3	320.6	653.9
30%	351.3	653.0	962.0
40%	969.8	1339.9	1450.0
50%	2,226.2	2,827.4	2,207.1
60%	6,237.5	5,712.8	3,470.0
70%	26,396.8	11,049.4	6,010.6
80%	91,462.7	29,583.7	10,946.7
90%	855,303	79,130.3	27,735.3
95%	6,934,631	185,169.6	77,715.9
99%	177,703,343	1,081,186.9	798,479.3

Table IX presents the statistics derived from the predicted distribution of $Y_{i,14}$, where the three enterprises are randomly selected from the three categories, respectively. Consider the BS enterprise as an example, the probability it has a breach in the 14th year (i.e., 2018) is $1-0.9492=0.0508$; under the condition that a breach incident indeed occurs in the 14th year, the min breach size is 1, the maximum is 2,846,163,649, the expected breach size is 985,792.5, the median size is 2,226.2 (i.e., the breach size is extremely skewed), the first quantile is 223.6, the third quantile is 47,226.1, the 10th percentile is 38.6, the standard deviation is 111,596,686 (i.e., the breach size has a very large variability). According to the dataset, the fraction of enterprises having breach incident in 2018 is 0.0688 for BS enterprises (vs. the predicted probability 0.0508), 0.0320 for the OTHER enterprises (vs. predicted 0.0103), and 0.1240 for the MED enterprises (vs. predicted 0.0963).

V. Applying the Framework to Synthetic Data With Positive Dependence

Recall that the real-world breach data analyzed in Section IV exhibits negative dependence. In order to demonstrate the broad applicability of the framework, we generate and use a synthetic dataset with positive dependence.

A. Data Generation and Exploratory Data Analysis

1) Generating Synthetic Data: In order to generate synthetic data with positive dependence, we first generate data with a *dependent uniform* distribution over [0, 1], denoted by $\{(i, u_{it})|1 \le i \le 1000, 1 \le t \le 5\}$, where dependence follows the multivariate Gumbel copula [40]:

$$C(u_{i1}, \ldots, u_{i5}; \alpha) = \phi^{-1}\left(\sum_{t=1}^{5} \phi(u_{it})\right)$$

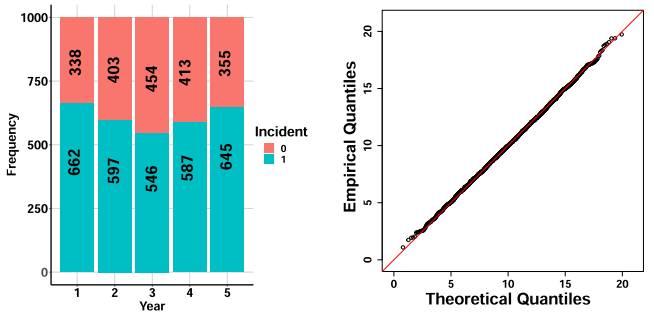
where $\phi(u_{it}) = \left[-\log(u_{it})\right]^{\alpha}$, $\alpha \ge 1$, $u_{it} \in [0, 1]$, and $\phi^{-1}(u_{it}) = e^{-u_{it}^{1/\alpha}}$. That is, we consider 1000 synthetic enterprises over 5 years. We only consider 5 years because we want to see if our framework can accommodate even shorter time series. In our simulation, we set $\alpha = 5$, which leads to positive dependence among the u_{it}'s with $1 \le t \le 5$. For the marginal distribution, we use the lognormal distribution with mean 10 and standard deviation 3. We choose the lognormal distribution because it has the heavy-tail property and has been used in the literature for fitting breach sizes [7]. The inverse of the marginal distribution is applied to each $u_{i,t}$, where $i = 1, \ldots, 1000$ and $t = 1, \ldots, 5$, to generate the marginal values. To introduce sparsity into the simulated data, we generate observation zero at time t with probability

$$p_t = -0.02t^2 + 0.125t + 0.23 + \epsilon_t, \qquad \text{(V.1)}$$

where ϵ_t is randomly generated from a Gaussian distribution with mean 0 and standard deviation 0.01 and $t = 1, \ldots, 5$. This leads to $p_{i1} = 0.339$, $p_{i2} = 0.408$, $p_{i3} = 0.441$, $p_{i4} = 0.438$, and $p_{i5} = 0.350$, for $i = 1, \ldots, 1000$. The marginal values for each t are replaced with 0's with probability p_{it}, where $t = 1, \ldots, 5$. Note that we propose using Eq. (V.1) because the real-world breach data exhibits a quadratic pattern, as shown in β_4 of Eq. (IV.1). We use $\{(i, y_{it})|1 \le i \le 1000, 1 \le t \le 5\}$ to represent the 5 year synthetic breach data, where y_{it} represents the breach size and $y_{it} = 0$ represents that there is no incident.

2) Exploratory Data Analysis: Figure 7a presents the synthetic breach data, and there are respectively 338, 403, 454, 413, and 355 0's among the 1000 enterprises over the 5 years. We observe that the time series are denser than the real-world dataset because we want to see whether our framework is widely applicable. For the nonzero breach sizes, Table X presents the summary statistics for each year, and shows that there exist extreme large values which reflect the heavy-tail property of the synthetic data. Note that the synthetic data is still sparse (albeit denser than the real-world data) and heavy-tailed (i.e., similar to those of the real-world breach data).

In what follows we use $\{(i, y_{it})|1 \le i \le 1000, 1 \le t \le 4\}$ as the training data to fit a model and use $\{(i, y_{it})|1 \le i \le 1000, t = 5\}$ as the test data to assess prediction accuracy.



(a) Bar chart (b) QQ-plot

Fig. 7. Bar chart of synthetic incident frequencies, where color '0' means there is no incident and color '1' means there is incident, and QQ-plot of mixed model for the marginals.

TABLE X
Statistics of Log-Transformed Nonzero y_{it}'s, Where 'SD' Stands for Standard Deviation, Q_1 and Q_3 Represent the First and Third Quantiles

t	Min	Q1	Median	Mean	SD	Q3	Max
1	1.959	8.217	10.058	10.254	3.078	12.254	19.400
2	1.086	8.116	10.004	10.188	3.097	12.343	19.733
3	2.450	8.063	10.208	10.078	2.893	12.056	19.385
4	1.937	7.795	10.041	10.092	3.061	12.110	18.380
5	0.507	7.939	9.961	10.104	3.116	12.221	18.370

B. Modeling Occurrence of Breach Incidents

Similar to Section IV, we use the logistic model proposed in the framework to fit the occurrence of zeros, i.e.,

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 t + \beta_2 t^2, \quad 1 \le t \le 4. \qquad \text{(V.2)}$$

The estimated parameters are $\hat{\beta}_0 = -1.255$, $\hat{\beta}_1 = 0.678$, and $\hat{\beta}_2 = -0.112$ with standard deviations 0.183, 0.165, and 0.032, respectively. Based on Eq. (V.2), we have the fitted probabilities as $\hat{p}_{i1} = 0.334$, $\hat{p}_{i2} = 0.414$, $\hat{p}_{i3} = 0.443$, and $\hat{p}_{i4} = 0.417$. The predicted probability based on Eq. (V.2) for $t = 5$ is $\hat{p}_{i5} = 0.339$. Thus, the fitted and predicted probabilities are close to the true probabilities.

C. Modeling Breach Sizes

For modeling the breach size, we use the proposed mixed model with the Gaussian kernel function as discussed in Section IV-C. The estimated parameters are $\lambda = 0.616$, $\mu = 14.304$, $\sigma_\mu = 1.559$, and $\xi = -0.192$ with standard deviations 0.134, 0.005, 0.135, and 0.058, respectively. We observe that all these estimates are significant at level 0.01. The proportion of breach sizes above the threshold is $\phi_\mu = 0.102$. Figure 7 depicts the QQ-plot, showing that the proposed mixed model has a very satisfactory fitting accuracy.

D. Modeling and Estimating Dependence Structures

1) Fitting Dependence With Vine Copula: Similar to Section IV-D, We use Algorithm 1 to estimate the D-vine dependence structure. Table XI summarizes the selected copula structures and estimated parameters. We observe that the Gumbel copula is selected for all of the trees. This is not

TABLE XI

ESTIMATED D-VINE COPULAS FOR BREACH SIZES (SD MEANS STANDARD DEVIATION AND τ IS THE KENDALL τ)

	Copula	Est	SD	τ
Tr_1	Gumbel	8.159	0.0436	0.877
Tr_2	Gumbel	1.417	0.0111	0.294
Tr_3	Gumbel	1.293	0.0174	0.227

TABLE XII

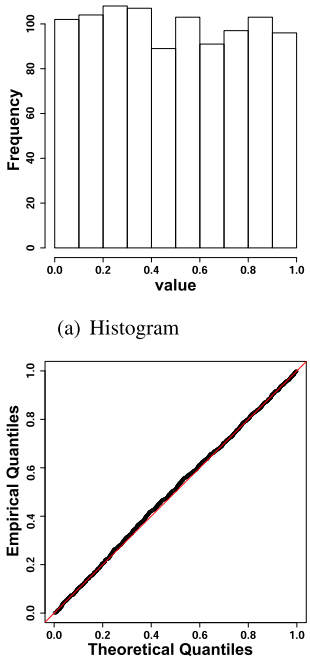
FITTING RESULTS OF VARIOUS MODELS FOR THE BREACH DATA

	log likelihood	AIC
Mixed D-vine with Gumbel	**-2061.965**	**4141.931**
Gaussian	-3892.126	7802.251
Benchmark	-6045.225	12108.451
LMM	-12466.57	24945.14

TABLE XIII

MEAN RPSS OF THE MIXED D-VINE WITH GUMBEL COPULA MODEL AND OTHER MODELS, WHERE *Percentage* IS THE % OF THE RPSS OF THE MIXED D-VINE MODEL WITH GUMBEL COPULA THAT ARE LESS THAN THAT OF THE OTHER MODEL(S)

	Mean Score	Percentage
Mixed D-vine with Gumbel	**13.633**	**—**
Mixed D-vine with Gaussian	13.637	**51.8%**
The benchmark model	15.078	**100%**
LMM model	17.518	**66.8%**



(a) Histogram

(b) QQ-plot

Fig. 8. Histogram of predicted distribution and its uniform QQ-plot.

surprising as the ground-truth dependence structure is the multivariate Gumbel copula. The parameters of the Gumbel copula are all significant with positive Kendall τ's, because the ground-truth dependence is positive.

2) Model Comparison: Similar to Section IV-D, Table XII summarizes the fitting results. We observe that the proposed mixed D-vine model with Gumbel copula leads to the smallest AIC and the largest log-likelihood value, and that LMM has the worst fitting accuracy. Thus, the proposed mixed D-vine with Gumbel copula model has the best fitting accuracy.

E. Predicting Data Breach Risk

Similar to Section IV-E, we apply Algorithm 2 to predict the data breach risk for $t = 5$. We assess the accuracy of the predicted distribution using the same two metrics.

In terms of the RPS-based metric, Table XIII summarizes the mean RPSs of the mixed D-vine model and the other models. We observe that the average RPS of the mixed D-vine model is the smallest. Since the average RPSs are small, we compute the percentage of the RPSs of the mixed D-vine model that are less than that of the other models. We observe that the mixed D-vine model outperforms the benchmark model by 100%, outperforms the Gaussian model by 1.58%, and outperforms the LMM model by 16.8%.

In terms of the uniform test-based evaluation of the accuracy of the predicted distribution, Figure 8a plots the histogram of the distribution predicted by the mixed D-vine model, showing an almost uniform distribution. Figure 8b depicts the qq-plot of the predicted distribution, showing that all of the points are around the 45-degree line. This means that the mixed D-vine model has a satisfactory prediction accuracy. The value of the Kolmogorov–Smirnov test statistic is small (.024) with a large p value .5964. All of these evidences suggest that the predicted distribution is accurate.

VI. DISCUSSION

Use case: One use case of the predicted distribution of enterprise i's breach size $Y_{i,t+1}$ is the following. The enterprise can compute the distribution of the cost that would be incurred by a breach incident in the next time interval (e.g., year), together with the unit price of each breached record. This distribution of cost can be used as an input to a quantitative risk management engine, which can decide, for example, how to spend its cybersecurity investment (e.g., buying cyber breach insurance vs. enhancing cyber defense).

Broader applications of the framework: Our framework is presented in a way geared towards modeling and predicting data breach incidents, but can be adopted in, or adapted for, other application settings. In principle, our framework is applicable to any multivariate time series with *sparse* events. The only technical restriction or assumption that must be satisfied in order to apply our framework is the Extreme Value Theory. That is, if the sparse data does not exhibit the heavy tail phenomenon, it is not necessary to use the GPD to model the tail; in this case, other parametric or non-parametric distributions should be used instead. Nevertheless, the heavy tail phenomenon is often exhibited by breach data.

Limitations of the present study: This article has limitations. First, the framework is geared towards predicting the distribution of an entity's breach size one-step ahead of time. It does not predict when the next breach will occur to an enterprise, which is a challenging open problem because most enterprises

only have one incident (i.e., most inter-arrival times between breach incidents are truncated). Second, the dataset for the case study does not provide enterprises' distinct information and may not be *complete* because some breach incidents may not be reported. Nevertheless, it is, to the best of our knowledge, the most comprehensive dataset that is publicly available and has attracted a due amount of attention [6], [8], [63]. Third, it is an interesting future work to investigate breach datasets together with enterprises' network security postures [26]. This may identify potential correlations between these two aspects.

It is known that the breach data may be under-reported [6], [63], meaning that our model may be biased downwards. For instance, the large number incidents in the MED category, when compared with the BS and OTHER categories, may be a consequence of higher reporting rates. This is so because the HIPAA breach notification rule requires to report every breach containing more than 500 records [64]. Nevertheless, the PRC data would be a reasonable resource for US breach incidents because all 50 states have enacted legislation to require private and governmental entities to notify individuals about security breaches that involve personally identifiable information. At the very least, the PRC data represents what is currently available. When higher-quality data becomes available in the future, our framework can be equally applied.

VII. Conclusion

We presented an novel framework for modeling and predicting multivariate breach incident time series with *sparse* events. The key idea behind the framework is to leverage the dependence between the multivariate time series to cope with the event sparsity. Intuitively, this is possible because the model can leverage the inter-entity or inter-enterprise relationship that accommodates more information than what is accommodated when considering the time series separately. As a case study, we applied the framework to analyze a breach dataset and showed that the statistical distribution of breach sizes can be predicted with a good accuracy. This hints at the possibility of leveraging cyber insurance to mitigate data breach risks, which is an important topic that is little understood. We hope this study will inspire many more research activities in understanding data breaches and mitigating their damages.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments that guided us in revising the article, and Yan Xu for running some experiments related to the present study and Eric Ficke for proofreading the article. The opinions expressed in the article are those of the authors' and do not reflect the funding agencies' policies in any sense.

References

[1] P. R. Clearinghouse. *Privacy Rights Clearinghouse's Chronology of Data Breaches*. Accessed: Jun. 9, 2020. [Online]. Available: https://www.privacyrights.org/data-breaches

[2] I. T. R. Center. *Data Breach Report in 2018*. Accessed: Jun. 6, 2020. [Online]. Available: https://www.idtheftcenter.org/tag/cyberscout/

[3] NetDiligence. *Cyber Claim Study*. Accessed: Jun. 16, 2020. [Online]. Available: https://netdiligence.com/cyber-claims-study-2019-report/

[4] J. Buckman, J. Bockstedt, M. J. Hashim, and T. Woutersen, "Do organizations learn from a data breach," in *Proc. Workshop Econ. Inf. Secur.*, 2017, pp. 1–22.

[5] J. Buckman, M. J. Hashim, T. Woutersen, and J. Bockstedt, "Fool me twice: An analysis of repeat data breaches within firms," *SSRN Electron. J.*, Jul. 2019, doi: 10.2139/ssrn.3258599.

[6] B. Edwards, S. Hofmeyr, and S. Forrest, "Hype and heavy tails: A closer look at data breaches," *J. Cybersecurity*, vol. 2, no. 1, pp. 3–14, Dec. 2016.

[7] M. Eling and N. Loperfido, "Data breaches: Goodness of fit, pricing, and risk measurement," *Insurance: Math. Econ.*, vol. 75, pp. 126–136, Jul. 2017.

[8] M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu, "Modeling and predicting cyber hacking breaches," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2856–2871, Nov. 2018.

[9] H. Sun, M. Xu, and P. Zhao, "Modeling malicious hacking data breach risks," *North Amer. Actuarial J.*, early access, pp. 1–19, Jul. 2020, doi: 10.1080/10920277.2020.1752255.

[10] T. Maillart and D. Sornette, "Heavy-tailed distribution of cyber-risks," *Eur. Phys. J. B*, vol. 75, no. 3, pp. 357–364, Jun. 2010.

[11] R. B. Security. *Datalossdb*. Accessed: Nov. 9, 2017. [Online]. Available: https://blog.datalossdb.org/

[12] K. Ikegami and H. Kikuchi, "Modeling the risk of data breach incidents at the firm level," in *Proc. Int. Conf. Innov. Mobile Internet Services Ubiquitous Comput.*, Springer, 2020, pp. 135–148.

[13] S. Wheatley, T. Maillart, and D. Sornette, "The extreme risk of personal data breaches and the erosion of privacy," *Eur. Phys. J. B*, vol. 89, no. 1, p. 7, Jan. 2016.

[14] S. Romanosky, R. Telang, and A. Acquisti, "Do data breach disclosure laws reduce identity theft?" *J. Policy Anal. Manage.*, vol. 30, no. 2, pp. 256–286, Mar. 2011.

[15] X. Fang, M. Xu, S. Xu, and P. Zhao, "A deep learning framework for predicting cyber attacks rates," *EURASIP J. Inf. Secur.*, vol. 2019, no. 1, p. 5, Dec. 2019.

[16] M. Eling and K. Jung, "Copula approaches for modeling cross-sectional dependence of data breach losses," *Insurance, Math. Econ.*, vol. 82, pp. 167–180, Sep. 2018.

[17] S. Xu, "The cybersecurity dynamics way of thinking and landscape," in *Proc. 7th ACM Workshop Moving Target Defense*, Nov. 2020, pp. 69–80.

[18] S. Xu, "Cybersecurity dynamics: A foundation for the science of cybersecurity," in *Proactive and Dynamic Network Defense*. New York, NY, USA: Springer, 2019, pp. 1–31.

[19] S. Xu, "Cybersecurity dynamics," in *Proc. HotSoS*, 2014, pp. 14:1–14:2.

[20] P. Du, Z. Sun, H. Chen, J.-H. Cho, and S. Xu, "Statistical estimation of malware detection metrics in the absence of ground truth," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 12, pp. 2965–2980, Dec. 2018.

[21] E. Condon, A. He, and M. Cukier, "Analysis of computer security incident data using time series models," in *Proc. 19th Int. Symp. Softw. Rel. Eng. (ISSRE)*, Nov. 2008, pp. 77–86.

[22] Z. Zhan, M. Xu, and S. Xu, "A characterization of cybersecurity posture from network telescope data," in *Proc. 6th Int. Conf. Trustworthy Syst. (InTrust)*, 2014, pp. 256–266.

[23] K. Bagchi and G. Udo, "An analysis of the growth of computer and Internet security breaches," *Commun. Assoc. Inf. Syst.*, vol. 12, no. 1, p. 46, 2003.

[24] Z. Zhan, M. Xu, and S. Xu, "Characterizing honeypot-captured cyber attacks: Statistical framework and case study," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 11, pp. 1775–1789, Nov. 2013.

[25] C. Peng, M. Xu, S. Xu, and T. Hu, "Modeling and predicting extreme cyber attack rates via marked point processes," *J. Appl. Statist.*, vol. 44, no. 14, pp. 2534–2563, Oct. 2017.

[26] Y. Liu *et al.*, "Cloudy with a chance of breach: Forecasting cyber security incidents," in *Proc. USENIX Secur. Symp.*, 2015, pp. 1009–1024.

[27] R. Sen and S. Borle, "Estimating the contextual risk of data breach: An empirical approach," *J. Manage. Inf. Syst.*, vol. 32, no. 2, pp. 314–341, Apr. 2015.

[28] M. Xu and S. Xu, "An extended stochastic model for quantitative security analysis of networked systems," *Internet Math.*, vol. 8, no. 3, pp. 288–320, Aug. 2012.

[29] G. Da, M. Xu, and S. Xu, "A new approach to modeling and analyzing security of networked systems," in *Proc. Symp. Bootcamp Sci. Secur. - HotSoS*, 2014, pp. 6:1–6:12.

[30] M. Xu, G. Da, and S. Xu, "Cyber epidemic models with dependences," *Internet Math.*, vol. 11, no. 1, pp. 62–92, Jan. 2015.

[31] X. Li, P. Parker, and S. Xu, "A stochastic model for quantitative security analyses of networked systems," *IEEE Trans. Depend. Sec. Comput.*, vol. 8, no. 1, pp. 28–43, Jan. 2011.
[32] S. Xu, W. Lu, and L. Xu, "Push- and pull-based epidemic spreading in networks: Thresholds and deeper insights," *ACM Trans. Auto. Adapt. Syst.*, vol. 7, no. 3, pp. 1–26, Sep. 2012.
[33] S. Xu, W. Lu, and Z. Zhan, "A stochastic model of multivirus dynamics," *IEEE Trans. Depend. Sec. Comput.*, vol. 9, no. 1, pp. 30–45, Jan. 2012.
[34] S. Xu, W. Lu, L. Xu, and Z. Zhan, "Adaptive epidemic dynamics in networks: Thresholds and control," *ACM Trans. Auto. Adapt. Syst.*, vol. 8, no. 4, pp. 1–19, Jan. 2014.
[35] Y. Han, W. Lu, and S. Xu, "Characterizing the power of moving target defense via cyber epidemic dynamics," in *Proc. Symp. Bootcamp Sci. Secur. HotSoS*, 2014, pp. 1–12.
[36] R. Böhme and G. Kataria, "Models and measures for correlation in cyber-insurance," in *Proc. Workshop Econ. Inf. Secur. (WEIS)*, 2006, p. 3.
[37] H. Herath and T. Herath, "Copula-based actuarial model for pricing cyber-insurance policies," *Insurance Markets Companies, Analyses Actuarial Comput.*, vol. 2, no. 1, pp. 7–20, 2011.
[38] A. Mukhopadhyay, S. Chatterjee, D. Saha, A. Mahanti, and S. K. Sadhukhan, "Cyber-risk decision models: To insure IT or not?" *Decis. Support Syst.*, vol. 56, pp. 11–26, Dec. 2013.
[39] M. Xu, L. Hua, and S. Xu, "A vine copula model for predicting the effectiveness of cyber defense early-warning," *Technometrics*, vol. 59, no. 4, pp. 508–520, Oct. 2017.
[40] H. Joe, *Dependence Modeling With Copulas*. Boca Raton, FL, USA: CRC Press, 2014.
[41] E. Brechmann and U. Schepsmeier, "Cdvine: Modeling dependence with c-and d-vine copulas in r," *J. Stat. Softw.*, vol. 52, no. 3, pp. 1–27, 2013.
[42] Z. Lin, W. Lu, and S. Xu, "Unified preventive and reactive cyber defense dynamics is still globally convergent," *IEEE/ACM Trans. Netw.*, vol. 27, no. 3, pp. 1098–1111, Jun. 2019.
[43] R. Zheng, W. Lu, and S. Xu, "Preventive and reactive cyber defense dynamics is globally stable," *IEEE Trans. Netw. Sci. Eng.*, vol. 5, no. 2, pp. 156–170, Apr. 2018.
[44] W. Lu, S. Xu, and X. Yi, "Optimizing active cyber defense dynamics," in *Proc. GameSec*, 2013, pp. 206–225.
[45] S. Xu, W. Lu, and H. Li, "A stochastic model of active cyber defense dynamics," *Internet Math.*, vol. 11, no. 1, pp. 23–61, Jan. 2015.
[46] R. Zheng, W. Lu, and S. Xu, "Active cyber defense dynamics exhibiting rich phenomena," in *Proc. Symp. Bootcamp Sci. Secur.*, Apr. 2015, pp. 1–12.
[47] W. W. Stroup, *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Boca Raton, FL, USA: CRC Press, 2012.
[48] C. Scarrott, "Univariate extreme value mixture modelling," in *Extreme Value Modeling and Risk Analysis: Methods and Applications*. Boca Raton, FL, USA: CRC Press, 2016, pp. 41–67.
[49] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools-Revised Edition*. Princeton, NJ, USA: Princeton Univ. Press, 2015.
[50] K. Aas, C. Czado, A. Frigessi, and H. Bakken, "Pair-copula constructions of multiple dependence," *Insurance, Math. Econ.*, vol. 44, no. 2, pp. 182–198, Apr. 2009.
[51] M. Fischer, C. Köck, S. Schlüter, and F. Weigert, "An empirical analysis of multivariate copula models," *Quant. Finance*, vol. 9, no. 7, pp. 839–854, Oct. 2009.
[52] P. Shi and L. Yang, "Pair copula constructions for insurance experience rating," *J. Amer. Stat. Assoc.*, vol. 113, no. 521, pp. 122–133, Jan. 2018.
[53] J. Stöber, H. G. Hong, C. Czado, and P. Ghosh, "Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses," *Comput. Statist. Data Anal.*, vol. 88, pp. 28–39, Aug. 2015.
[54] M. S. Smith, "Copula modelling of dependence in multivariate time series," *Int. J. Forecasting*, vol. 31, no. 3, pp. 815–833, Jul. 2015.
[55] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York, NY, USA: Springer, 2013.
[56] E. S. Epstein, "A scoring system for probability forecasts of ranked categories," *J. Appl. Meteorol.*, vol. 8, no. 6, pp. 985–987, Dec. 1969.
[57] C. Czado, T. Gneiting, and L. Held, "Predictive model assessment for count data," *Biometrics*, vol. 65, no. 4, pp. 1254–1261, Dec. 2009.
[58] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.
[59] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007.
[60] C. Czado, *Analyzing Dependent Data with Vine Copulas: A Practical Guide With R*, vol. 222. New York, NY, USA: Springer, 2019.
[61] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London, U.K.: Chapman & Hall, 1994.
[62] E. C. Brechmann, C. Czado, and K. Aas, "Truncated regular vines in high dimensions with application to financial data," *Can. J. Statist.*, vol. 40, no. 1, pp. 68–85, Mar. 2012.
[63] A. H. Bisogni, Fabio and M. Eeten, "Estimating the size of the iceberg from its tip," in *Proc. Workshop Econ. Inf. Secur. (WEIS)*, 2017, pp. 1–29.
[64] U. D. of Health Human Services. *Breach Notification Rule*. Accessed: Jun. 9, 2020. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/breach-notification/index.html

Zijian Fang is currently pursuing the Ph.D. degree in statistics and finance with the School of Management, University of Science and Technology of China. His research interests include cybersecurity and applied statistics.

Maochao Xu received the Ph.D. degree in statistics from Portland State University in 2010. He is currently a Full Professor of statistics with Illinois State University. He is also a Cyber Insurance Advisor with CloudCover, Inc. His research interests include statistical modeling, cyber insurance, and risk modeling. He currently serves as an Associate Editor for *Communications in Statistics*.

Shouhuai Xu (Senior Member, IEEE) received the Ph.D. degree in computer science from Fudan University. He has been with The University of Texas at San Antonio. He is currently the Gallogly Chair Professor with the Department of Computer Science, University of Colorado Colorado Springs (UCCS). He pioneered the Cybersecurity Dynamics approach as foundation for the emerging science of cybersecurity, with three pillars: first-principle cybersecurity modeling and analysis (the x-axis); cybersecurity data analytics (the y-axis, to which the present paper belongs); and cybersecurity metrics (the z-axis). He co-initiated the International Conference on Science of Cyber Security and is serving as its Steering Committee Chair. He is/was an Associate Editor of the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING (IEEE TDSC), the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (IEEE T-IFS), and the IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING (IEEE TNSE).

Taizhong Hu received the Ph.D. degree in probability and statistics from the University of Science and Technology of China (USTC) in 1994. He is currently a Full Professor with the Department of Statistics and Finance, USTC. His research interests include risk measures, extreme value theory, statistical dependence, stochastic comparisons, and their applications. He also serves as an Associate Editor for *Insurance: Mathematics and Economics* and *Probability in the Engineering and Informational Science*.