JOURNAL OF APPLIED STATISTICS, 2018
https://doi.org/10.1080/02664763.2018.1436701

Modeling multivariate cybersecurity risks

Chen Peng[a], Maochao Xu[b], Shouhuai Xu[c] and Taizhong Hu[a]

[a]Department of Statistics and Finance, University of Science and Technology of China, Anhui Sheng, China; [b]Department of Mathematics, Illinois State University, Normal, IL, USA; [c]Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA

ABSTRACT

Modeling cybersecurity risks is an important, yet challenging, problem. In this paper, we initiate the study of modeling *multivariate cybersecurity risks*. We develop the first statistical approach, which is centered at a Copula-GARCH model that uses vine copulas to model the multivariate dependence exhibited by real-world cyber attack data. We find that ignoring the due multivariate dependence causes a severe underestimation of cybersecurity risks. Both simulation and empirical studies show that the proposed approach leads to accurate predictions of multivariate cybersecurity risks.

ARTICLE HISTORY
Received 22 August 2017
Accepted 31 January 2018

KEYWORDS
Copula density forecast; GARCH; prediction; vine copula; Value-at-Risk

1. Introduction and motivation

According to the eighth Emerging Risks Survey conducted by the Society of Actuaries in 2016, cybersecurity threat has become the biggest emerging risk. This is witnessed by many severe cybersecurity incidents. For example, in February 2016, the Society for Worldwide Interbank Financial Telecommunication (SWIFT) network was hacked, which caused the theft of $101 million from Bangladesh's Central Bank; in December 2015, a coordinated cyber attack against several regional distribution power companies caused power outages in Western Ukraine; in June 2015, the Japan Pension Service system was hacked, which caused the breach of 1.25 million personal records. These many cybersecurity incidents can be attributed to the fact that cyberspace is difficult to secure, which can be further attributed to the 'asymmetry' that an attacker can succeed by exploiting a single vulnerability, but the defender has to block all vulnerabilities in order to protect a system from attacks. In other words, cyber systems have a very large 'attack surface', especially in the present era of cyber physical systems and Internet of Things (IoT) because many low-end physical devices cannot afford, or even infeasible (e.g. due to the availability of very limited computer memory resources on the device), to employ advanced but costly defense mechanisms. As a result, the United States Department of Homeland Security has made a systematic research and development plan for protecting critical infrastructures from cyber attacks [13]. In contrast to the fact that many studies have been conducted to protect critical infrastructures from *natural disasters*, such as transportation networks [20], economic systems [43] and petrochemical systems [46], our capability in quantifying cyber

CONTACT Maochao Xu ✉ mxu2@ilstu.edu

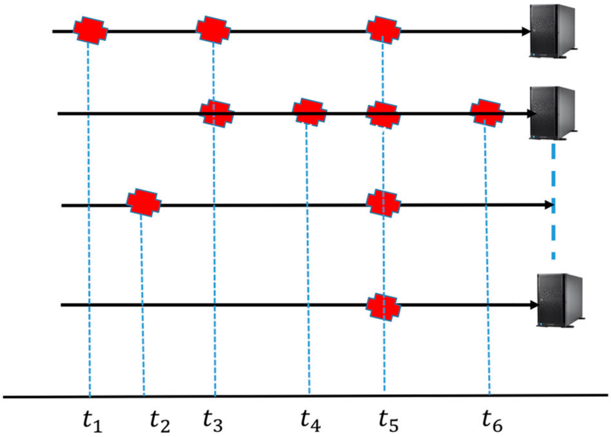


Figure 1. Cybersecurity risk against *n* servers of an enterprise.

risks is very limited. Among the many technical difficulties, one is the lack of 'good' cybersecurity risk metrics [9,40] and the other is the complexity that is encountered in dealing systemic cybersecurity risks as well as their emergent behaviors [49]. This very unsatisfactory situation calls for more research into the quantitative management of cybersecurity risks.

In this paper, we investigate how to model *multivariate* cybersecurity risks, which are manifested by multivariate dependent cyber attacks. Figure 1 illustrates the problem of multivariate dependence between the time series of cyber compromise events. Specifically, consider an enterprise of n servers (or computers, or virtual machines in a cloud) that store some secret data. Suppose there is a single server that is compromised at time t_1, t_2, t_4, and t_6; there are two servers that are compromised at time t_3; there are four servers that are compromised at time t_5. Whenever a server is compromised, a loss is incurred (e.g. the cost due to the compromise of the data in question, the recovery of the compromised data). In the real world, a compromised server can be 'cleaned up' by the defender, but the cleaned computer may get compromised again. For the purpose of assessing cybersecurity risks, we can record the compromise events and the losses they incur.

The research objective is to model the multivariate dependence among cybersecurity risks, and further predict the incoming attacks and the losses they incur. This prediction capability would help the enterprise to achieve proactive defense against the anticipated attacks [8,41,48,50–52]. This prediction capability is also useful in cybersecurity risk management via cybersecurity insurance because modeling cybersecurity risks is the first step towards determining the insurance premium, which is a challenging problem [6,18,25,30,35,48]. These real-world needs justify the importance of modeling the multivariate dependence between cybersecurity risks.

Recently, researchers have started investigating the problem of multivariate cybersecurity risks. From the perspectives of actuarial science and insurance, Böhme and Kataria [6] used the Beta-Binomial and one-factor latent risk model to describe the correlation between cybersecurity risks. Mukhopadhyay *et al.* [35] proposed a Bayesian Belief Network approach to modeling cybersecurity risks, and used the multivariate Gaussian copula to model the joint distribution and conditional distribution of each node in a network. Herath and Herath [25] proposed a copula-based actuarial model for pricing cybersecurity risks while considering three risk variables (i.e. the occurrence of an event, the time of payment, and the amount of payment). One may refer to Kosub [30] and Eling and Schnell [18] for comprehensive reviews on the modeling and management of cybersecurity risks. From the perspective of cybersecurity, Xu *et al.* [48] proposed a vine copula approach for modeling the dependence between the number of cyber attacks and the number of attacked computers. They discussed how to predict the effectiveness of a cyber defense early-warning mechanism.

Orthogonal to the focus of the present paper, there are studies that used statistical methods for detecting cyber attacks. For example, Denning [12] used statistical methods to detect attacks and introduced the concept of *intrusion detection*. Markou and Singh [33] provided a comprehensive review of intrusion detection based on statistical approaches, including Gaussian mixture models and Hidden Markov models. Neil *et al.* [36] used a scan statistic approach for intrusion detection purposes. In addition, statistical methods have been used to model the evolution of cyber threats or cybersecurity posture. Ishida *et al.* [26] proposed a Bayesian method for predicting the increase or decrease of attacks. Hidden Markov models have also been used to predict the increase or decrease of Bot agents [29]. Zhan *et al.* [50] proposed a FARIMA model to predict cyber attacks when the data exhibits the long-range dependence. This problem was further studied in [52] by using FARIMA+GARCH models for more accurate predictions. Peng *et al.* [41] studied the modeling and prediction of extreme cyber attack rates via marked point processes while using the Value-at-Risk (VaR) to measure the intensity of cyber attacks. One may refer to [21] for a survey on some computational techniques that have been used for predicting cyber attacks.

This paper aims to develop a new statistical model for describing multivariate cybersecurity risks. Specifically, we propose using the vine copula, which was first introduced in Bedford and Cooke [5], to model multivariate cybersecurity risks. Because of its flexibility and capability in estimating a large number of parameters, the vine copula has been widely used in many application settings, including econometrics, finance, insurance, weather [16,27,31]. We propose using the vine copula to model multivariate cybersecurity risks because of the following: (i) *Flexibility*. Traditional approaches to modeling a high-dimensional dependence are to use the multivariate Gaussian or t copulas because they are mathematically tractable. However, these models are restrictive in the high-dimensional settings. In contrast, vine copulas are more flexible in the high-dimensional settings because they can accommodate different dependence structures between different pairs of variables [16,27,31]. (ii) *Efficiency*. In high-dimensional setting, computation is an important factor in practice and can be challenging even when considering Gaussian or t copulas with unstructured covariances. In contrast, the truncation technique of vine copula can efficiently handle the computation in the high-dimensional settings [7].

Our research is different from the aforementioned literature: (i) The proposed vine copula approach aims to efficiently model the *high-dimensional* dependence between cybersecurity risks. In our empirical study, the number of dimensions is 69. In contrast, previous studies investigated *low-dimensional* cases. For example, most previous studies on cybersecurity risks [41,50,52] focused on one-dimensional cybersecurity risks, and the work [48], which is closely related to the present paper, focused on the *four-dimensional* dependence. (ii) We introduce a *triggering* mechanism to model cyber attack losses in the simulation study, meaning that a random loss occurs whenever the number of attacks (per time unit) exceeds a certain threshold. This mechanism reflects that an attacker may have to try multiple times before successfully penetrating into a system, and may be of independent value.

The rest of the paper is organized as follows. In Section 2, we briefly review the concept of the vine copula. In Section 3, we describe the statistical approach for analyzing multivariate cybersecurity risks, and present some simulation studies. In Section 4, we use the proposed statistical model to analyze a real-world cyber attack dataset, and evaluate both the in-sample and out-of-sample performances. In Section 5, we conclude the paper and discuss future research directions.

2. Vine copula

Copula is an effective and popular tool for modeling high-dimensional dependence. A d-dimensional copula is a cumulative distribution function (cdf) with uniform marginals in $[0, 1]$. Specifically, let $X_1, \ldots, X_d$ be continuous random variables with univariate marginal distributions $F_1, \ldots, F_d$, respectively. Consider their joint cdf

$$F(x_1, \ldots, x_d) = \mathsf{P}(X_1 \leq x_1, \ldots, X_d \leq x_d)$$

and a copula C that is defined as the joint cdf of the random vector $(F_1(X_1), \ldots, F_d(X_d))$. When the F_i's are continuous, Sklar's theorem [44] says that copula C is unique and satisfies

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

The joint density function can be represented as

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

where $c(u_1, \ldots, u_n)$ is the d-dimensional copula density function, and f_i is the corresponding marginal density function for X_i for $i = 1, \ldots, d$.

In the literature, many dependence structures have been proposed [27]. A very attractive and popular dependence structure is the *vine* copula, which offers a great deal of flexibility in modeling dependence, including various tail dependences and asymmetric dependences. Moreover, the multivariate vine copula is computational tractable because its density can be factored in terms of bivariate linking copulas and lower-dimensional margins. In general, a d-dimensional vine copula is constructed by mixing $d(d-1)/2$ bivariate linking copulas on a tree. In what follows, we define regular vine (R-vine) with notations following [16].

Definition 2.1 ([5,16]): $\mathcal{V} = (T_1, \ldots, T_d)$ on d elements is called an R-vine if:

(a) T_1 is the first tree (level 1) with node set $N_1 = \{1, \ldots, d\}$ and edge set E_1.
(b) For $i = 2, \ldots, d-1$, the edge set E_{i-1} is the node set of tree T_i.
(c) (Proximity condition) For tree T_i, $i = 2, \ldots, d-1$, if two nodes in E_{i-1} are connected by an edge in E_i, then these two nodes as edges in T_{i-1} share the same node in E_i.

Definition 2.2: The following three sets will be used to study properties of R-vines [5,16].

(a) The *complete union set* of $e_i \in E_i$ is defined as

$$U_{e_i} = \{d \in N_1 \mid \exists e_j \in E_j, j = 1, \dots, i-1, \ \textit{with}\ d \in e_1 \in \dots \in e_i\} \subset N_1.$$

That is, the complete union of an edge is a set of all indices that this edge contains.

(b) For an edge $e_i = \{a, b\} \in E_i$, the *conditioning set* of edge e_i is defined as $D_{e_i} = U_a \cap U_b$.

(c) For an edge e_i, the *conditioned sets* of e_i are defined as $C_{e_i,a} = U_a \backslash D_{e_i}$, $C_{e_i,b} = U_b \backslash D_{e_i}$.

Let $(\boldsymbol{F}, \mathcal{V}, B)$ be a vine copula specification, where $\boldsymbol{F} = (F_1, \dots, F_d)$ is a vector of continuous invertible marginal distribution functions, and $B = \{B_e \mid i = 1, \dots, d-1; e \in E_i\}$ is a set of copulas with B_e being a pair-copula. There is a unique distribution that realizes this vine copula specification with density

$$f_{1\dots d}(\mathbf{x}) = \prod_{k=1}^{d} f_k(x_k) \prod_{i=1}^{d-1} \prod_{e \in E_i} c_{C_{e,a}, C_{e,b} \mid D_e}(F_{C_{e,a}|D_e}(x_{C_{e,a}} \mid \mathbf{x}_{D_e}), F_{C_{e,b}|D_e}(x_{C_{e,b}} \mid \mathbf{x}_{D_e})),$$

where $\mathbf{x} = (x_1, \dots, x_d)$, $e = \{a, b\}$, $\mathbf{x}_{D_e} = \{x_i | i \in D_e\}$, and f_i is the density function of F_i for $i = 1, \dots, d$. Moreover, $c_{C_{e,a}, C_{e,b}|D_e}$ represents the bivariate copula density for edge $e = \{a, b\}$ [16].

Two well-known subclasses of R-vines are C-vines (canonical vines) with star structures in their tree sequence, and D-vines (drawable vines) with path structures [1,5]. In fact, the class of R-vine distributions is much larger and more flexible. Since the information of vine is very large, Morales-Nápoles [34] suggested using a lower triangular M matrix to store the information of R-vines.

Definition 2.3 ([16,34]): For $i = 1, \dots, d-1$ and $k = i+1, \dots, d-1$, a lower triangular matrix $M = (m_{ij})_{i,j=1\dots d}$ is an R-vine matrix if there exist $j \in \{i+1, \dots, d-1\}$ such that

$$(m_{k,i}, \{m_{k+1,i}, \dots, m_{d,i}\}) \in B_M(j) \text{ or } \in \widetilde{B}_M(j),$$

where

$$\begin{aligned} B_M(i) &= \{(m_{i,i}, D) \mid k = i+1, \dots, d; \quad D = \{m_{k,i}, \dots, m_{d,i}\}\}, \\ \widetilde{B}_M(i) &= \{(m_{k,i}, D) \mid k = i+1, \dots, d; \quad D = \{m_{i,i}\} \cup \{m_{k+1,i}, \dots, m_{d,i}\}\} \end{aligned}$$

for $i = 1, \dots, d-1$.

When M is fixed, the density of R-vine copula can be rewritten as

$$f_{1,\dots,d}(\mathbf{x}) = \left[\prod_{k=1}^{d} f_k(x_k)\right] \left[\prod_{j=d-1}^{1} \prod_{i=d}^{j+1} c_{m_{j,j}, m_{i,j}|m_{i+1,j},\dots,m_{d,j}}(F_{m_{j,j}|m_{i+1,j},\dots,m_{d,j}}, F_{m_{i,j}|m_{i+1,j},\dots,m_{d,j}})\right],$$

where $F_{i|i_1,\dots,i_n} \triangleq F(x_i \mid x_{i_1}, \dots, x_{i_n})$.

Dißmann *et al.* [16] developed a framework for model selection and estimation based on R-vine copulas. In our empirical study, we will use the *truncated* R-vine copula [7]. An R-vine copula is called *truncated* if all pair-copulas in higher order trees are set to be bivariate independence copulas. Let tRV(K) denote a truncated R-vine copula at level K and $\boldsymbol{\theta}_{tRV(K)} = \{\boldsymbol{\theta}_{e_1,e_2|D_e} \,|\, e \in E_i,\ i = 1,\ldots,K\}$ be the pair-copula parameters, where $\boldsymbol{\theta}_{e_1,e_2|D_e}$ are the parameters of the pair copula density $c_{e_1,e_2|D_e}$. The copula density of a level-K truncated R-vine copula can be written as

$$c_{tRV(K)}(\boldsymbol{u} \,|\, \boldsymbol{\theta}_{tRV(K)}) = \prod_{i=1}^{K} \prod_{e \in E_i} c_{e_1,e_2|D_e}(F(u_{e_1} \,|\, \boldsymbol{u}_{D_e}), F(u_{e_2} \,|\, \boldsymbol{u}_{D_e})), \tag{1}$$

where $\boldsymbol{u} = (u_1,\ldots,u_d) \in [0,1]^d$. Please refer to [7] for more details of truncated R-vine copulas. For more and comprehensive discussions of vine copulas, please refer to [1,5,16,27,31].

3. Modeling multivariate cybersecurity risks

3.1. The vine copula approach to modeling high-dimensional cybersecurity risks

In this section, we propose a vine copula approach to modeling multivariate cybersecurity risks. As mentioned in the Introduction, suppose an enterprise has n servers, which respectively receive $X_{1,t},\ldots,X_{n,t}$ attacks during the tth time interval; that is, the ith server is attacked for $X_{i,t}$ times during the tth time interval, where $t = 1,\ldots,N$. In the real world, n can be large, meaning that the number of dimensions of cybersecurity risks is high. Further, the dependence among cybersecurity risks can be nonlinear. This motivates us to use the following statistical approach for analyzing cybersecurity risks, which has three steps.

Step 1: *Modeling marginal processes*. The first step is to model the attacks against individual servers, namely the marginal processes. The marginal model for the attack rate (i.e. the number of attack per time interval) can be specified as

$$X_{i,t} = \mu_{i,t} + \epsilon_{i,t},$$

where $\mu_{i,t}$ is the mean function,and $\epsilon_{i,t}$ is the error term for $i = 1,\ldots,n$ and $t = 1,\ldots,N$. Our preliminary analysis shows that the ARMA+GARCH process can model the marginal processes, that is,

$$\mu_{i,t} = \mu_i + \sum_{k=1}^{p} \phi_k X_{i,t-k} + \sum_{l=1}^{q} \theta_l \epsilon_{i,t-l} \quad \text{and} \quad \epsilon_{i,t} = \sigma_{i,t} Z_i,$$

where p is the AR order, q is the MA order, and Z_i is the innovations that are independent and identically distributed with density $g_i(\cdot|\boldsymbol{\vartheta})$ and $\boldsymbol{\vartheta}$ representing parameters of the density function.

Step 2: *Modeling the dependence structure*. The second step is to model the high-dimensional dependence. The joint distribution of $(Z_1,\ldots,Z_n)$ can be written as

$$F_z(\mathbf{z}; \boldsymbol{\vartheta}, \boldsymbol{\Theta}) = C(F_1(z_1),\ldots,F_n(z_n); \boldsymbol{\vartheta}, \boldsymbol{\Theta}), \tag{2}$$

where $\boldsymbol{\Theta}$ denotes the vector of parameters of an n-dimensional vine copula, and F_i is the marginal distribution of Z_i for $i = 1,\ldots,n$. As mentioned above, we propose using the

R-vine copula to model the dependence of cybersecurity risks, which uses bivariate copulas as building-blocks. Since the computational complexity increases exponentially as the number of dimensions of the dependence, it is infeasible to model all of the tree structures. Since the correlation between two nodes decreases along with the increasing of the tree level, the truncated vine copula is a good choice for our data [7,16].

In order to select the truncation level according to the dataset, we use the Algorithm 1 described in [7], which is based on the Vuong test [47]. The algorithm starts from $K = 1$, which fits a truncated vine copula, and increases K one-by-one to assess whether a more complicated model should be used according to the Vuong test. In this algorithm, the pair copulas are also selected for the truncated vine.

Step 3: *Estimating model parameters*. The joint log-likelihood function of the model can be written as

$$L(\boldsymbol{\vartheta}, \boldsymbol{\Theta}) = \sum_{t=1}^{N} \Bigg[\log c\left(F_1\left(\frac{y_{1,t}-\mu_{1,t}}{\sigma_{1,t}}\right), \ldots, F_n\left(\frac{y_{n,t}-\mu_{n,t}}{\sigma_{n,t}}\right); \boldsymbol{\vartheta}, \boldsymbol{\Theta}\right) \\ - \sum_{i=1}^{n} \log(\sigma_{i,t}) + \sum_{i=1}^{n} \log\left(g_i\left(\frac{y_{i,t}-\mu_{i,t}}{\sigma_{i,t}}; \boldsymbol{\vartheta}\right)\right)\Bigg], \tag{3}$$

where $c(\cdot)$ is the copula density of $C(\cdot)$. Since the number of parameters is very large for our model, for the joint log-likelihood function $L(\boldsymbol{\vartheta}, \boldsymbol{\Theta})$, we propose using the Inference Function of Margins (IFM) method for estimating the model parameters [27]. The IFM method involves two steps: estimating the parameters of the marginal time series models, and then estimating the parameters of the copula(s) by fixing the parameters obtained in the previous step. The IFM method is a widely used approach in the literature. Moreover, it has been proved that under some regularity conditions, the IFM estimators are consistent and asymptotically normal [27]. Therefore, for modeling cybersecurity risks, we first estimate the parameters for the marginal models, thta is, ARMA+GARCH, by using maximum likelihood approach, and then model the dependence among risks by using the vine copula discussed in Step 2.

3.2. Simulation study

Now we study the proposed statistical approach based on the Monte Carlo simulation. Specifically, we discuss the out-of-sample performances of our model, and study how the dependence would affect the assessment of the cyber losses. This study would shed light on the consequence of ignoring the dependence among cybersecurity risks. For this purpose, we first introduce a triggering mechanism for simulating cyber losses, which may have an independent interest in actuarial science and risk management.

Recall that an enterprise has n servers, and the ith server is compromised when the number of attacks against it is above a threshold τ_i. This is a reasonable assumption because it reflects the number of attacks that are needed in order to compromise a server. Consider the following triggering vector

$$(\mathsf{I}(X_{1,t} > \tau_1), \ldots, \mathsf{I}(X_{n,t} > \tau_n)),$$

where $\mathsf{I}(\cdot)$ is the indicator function. Suppose the compromise of a server imposes a random loss Y_i, and the compromise of a server during the tth time interval is detected and recovered at the end of the tth time interval (equivalently, the beginning of the $(t+1)$th time interval). This can be achieved in practice by using *proactive defense* as discussed in, for example, [23,45]. The random loss at the tth time interval can be represented as

$$S(t) = \sum_{i=1}^{n} S_i(t) = \sum_{i=1}^{n} Y_i \cdot \mathrm{I}(X_{i,t} > \tau_i),$$

where $S_i(t)$ represents the random loss at time t for server i, $i = 1, \ldots, n$. We suppose that the loss Y_i follows a log-normal distribution, which is because there is empirical evidence showing that the loss of data breach follows a log-normal distribution [17]. In what follows, we discuss the performance of the proposed approach in different scenarios.

3.2.1. Simulating multivariate cyber losses

Now we discuss the simulation of multivariate cyber losses under the triggering mechanism. We are particularly interested in examining the dependence effect on the prediction performance of the proposed model. For this purpose, the experiments are set up as follows.

We generate three 10-dimensional attack processes with size 1500 and the following dependence structures: (i) Vine copula; (ii) Multivariate Gaussian copula; and (iii) Multivariate t copula. The marginal attack processes $X_{i,t}$'s are generated from the AR(1)+GARCH(1,1) model with skewed Student-t innovations. The reason for us to select this model is that, based on the empirical study in Section 4, the real attack processes can be captured by the AR(1)+GARCH(1,1) model. It is worth mentioning that the AR+GARCH process is widely used in the literature for modeling time series in practice; see, for example, [24,37,48]. Please refer to the Appendix for the details of generating the dependent multivariate attack processes with those dependence structures.

The datasets with the vine, multivariate Gaussian, and multivariate t copula structures are denoted by $\mathrm{att}_{\mathrm{VC}}$, $\mathrm{att}_{\mathrm{GC}}$ and $\mathrm{att}_{\mathrm{TC}}$, respectively.

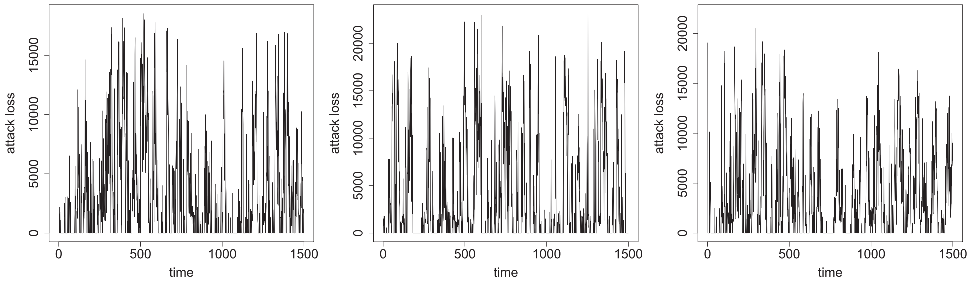
For the loss triggering mechanism, we assume that $\tau_i = 12$ for $1 \le i \le 10$. We further assume that the loss Y_i has the following log-normal distribution,

$$\log(Y_i) \sim N(\mu, \sigma^2) \tag{4}$$

with $\mu = 7.894$ and $\sigma = .3715$. To simulate the cyber losses incurred by multivariate attack process, a random log-normal loss with distribution in Equation (4) is generated when the number of attacks per unit exceeds 12 for each attack process.

Figure 2 plots the simulated total loss $s(t)$ of the enterprise during a period of time and corresponding to a specific dependence structure between multivariate cyber attack processes. It is interesting to observe that there are clusters and extreme values in the simulated data. These phenomena of clustering and extreme values are often observed in the cyber risks [52].

In the following, we examine the prediction performance of the proposed approach in different scenarios. Specifically, we study the model performances when the dependence structure is correctly specified as vine copula, and discuss the consequence of ignoring the dependence between multivariate cyber attack processes. Furthermore, we discuss the model performance when the dependence structure is misspecified as vine copula but the true dependence structures are the multivariate Gaussian and t copula structures.



(a) Multivariate Gaussian copula (b) Multivariate t copula (c) Vine copula

Figure 2. Simulated total loss $s(t)$ of the enterprise during a period of time and corresponding to a specific dependence structure between multivariate cyber attack processes.

3.2.2. Performance evaluation

For evaluation purposes, the generated attack datasets ($\mathrm{att}_{\mathrm{GC}}$, $\mathrm{att}_{\mathrm{VC}}$ and $\mathrm{att}_{\mathrm{TC}}$) are split into in-sample and out-of-sample parts with sample sizes 1000 and 500, respectively.

We perform the three steps described in Section 3.1 to estimate the parameters of the proposed model, and the log-likelihood function is calculated based on Equation (3). For the marginal processes, we use ARMA(p, q)+GARCH(1,1) with innovations from different distributions including the normal, Student-t, skewed normal, skewed Student-t, and generalized error distributions. The p and q are allowed to vary from 0 to 5, respectively. The AIC criterion is used for identifying the orders of p and q, and the simpler model is preferred when the AICs of two models are close to each. It is found that AR(1)+GARCH(1,1) model with the skewed Student-t innovation is selected for all of the marginal attack processes. The standardized residuals of the marginal models are used to estimate the dependence structure between the 10 attack processes. For selecting the vine copula structure, as suggested in [16], we use a sequential approach based on the maximum spanning tree algorithm, which has been implemented in the state-of-the-art R package VineCopula. We use this package to fit the standardized residuals and select the vine copula corresponding to Equation (3).

In order to evaluate the out-of-sample performance of the proposed approach, we use the risk measure of VaR. Specifically, the VaR at level α is defined as

$$\mathrm{VaR}_\alpha(t) = \inf\{l : P(S(t) \le l) \ge \alpha\},$$

where $S(t)$ is the cumulative loss and $0 < \alpha < 1$. The violation based on VaR_α is defined as

$$\mathrm{I}_\alpha(t) = \begin{cases} 1, & s(t) > \mathrm{VaR}_\alpha(t), \\ 0, & \text{otherwise}, \end{cases}$$

where $s(t)$ is the observed cybersecurity loss and $\mathrm{VaR}_\alpha(t)$ is the predicted VaR values of the loss. For example, $\mathrm{VaR}_{.95}(t)$ describes that there is only 5% chance that the observed attack loss S_t would exceed the predicted value of $\mathrm{VaR}_{.95}(t)$. When this happens, we say a violation occurs.

In order to assess the accuracy of the prediction, we use three widely used tests [10]. The first test is the unconditional coverage test ($\mathrm{LR}_{\mathrm{uc}}$), which evaluates whether or not the fraction of violations obtained from the model is significantly different from the theoretical one. The second test measures the independence of violations (LR_{ind}), where the null hypothesis is that the present violation has no effect on future violations. The third test, called the conditional coverage test (LR_{cc}), is a combination of the preceding two. In what follows, we discuss how to apply these tests for prediction purposes.

We use Algorithm 1 to compute the $\text{VaR}_\alpha(t)$'s of the total loss and the number of violations based on the proposed model, where α varies from .91 to .98.

Algorithm 1 Algorithm for rolling prediction of the VaR_α of cyber attack losses

Input: Multivariate historical time series data $\{(i, x_{i,t}, s(t))|i = 1, \ldots, d; t = 1, \ldots, m + n\}$; in-sample dataset $\{(i, x_{i,t})|i = 1, \ldots, d; t = 1, \ldots, m\}$; out-of-sample dataset $\{(i, x_{i,t})|i = 1, \ldots, d; t = m + 1, \ldots, n\}$; α; loss triggering threshold $\tau = 12$.

```
for $t = m + 1, \ldots, n$ do
    Estimate the Copula-GARCH model based on $\{(i, x_{i,s})|s = 1, \ldots, t - 1\}$ by using steps 1-3 in Section 3.1, and predict the next mean $\hat{\mu}_i$ and standard error $\hat{\sigma}_{i,t}$;
    Based on the estimated vine copula, simulate 5000 $d$-dimensional copula samples $u_{i,t}^{(k)}$, $i = 1, \ldots, n$ by using the algorithm in 1, $k = 1, \ldots, 5000$;
    Convert the simulated dependent samples $u_{i,t}^{(k)}$'s into the $z_{i,t}^{(k)}$'s by using the inverse of the estimated Student-$t$ distribution, $k = 1, \ldots, 5000$;
    Compute the predicted 5000 $d$-dimensional time series attack data $x_{i,t}^{(k)}$ by using Equation (A5), $k = 1, \ldots, 5000$;
    if $x_{i,t}^{(k)} > \tau$ then
        Generate a log-normal random loss $y_{i,t}^{(k)}$ based on Equation (4);
    else
        The random loss is $y_{i,t}^{(k)} = 0$;
    end if
    Calculate the total loss $s^{(k)}(t) = \sum_{i=1}^{d} y_{i,t}^{(k)}$, $k = 1, \ldots, 5000$;
    return $\text{VaR}_\alpha(t)$ based on predicted 5000 total losses.
    if The observed total loss $s(t) > \text{VaR}_\alpha(t)$; then
        An violation occurs;
        return 1.
    else
        return 0.
    end if
end for
```

Output: The $\text{VaR}_\alpha(t)$'s of the total loss, and the number of violations.

Table 1 reports the prediction results for the att_{VC} dataset. Note that in this case, the dependence structure is correctly specified as the vine copula. For comparison purposes, we also list the testing results based on the benchmark model, that is, the independence model. We observe that the prediction performances are very satisfactory based on the proposed vine model, while the benchmark model performs poorly. In fact, none of the

Table 1. Assessing prediction performance based on the simulated out-of-sample data for att$_{\mathrm{VC}}$.

α	Obs.	Exp.	LR$_{\mathrm{uc}}$	LR$_{\mathrm{ind}}$	LR$_{\mathrm{cc}}$	$\overline{\mathrm{Var}}$
			Vine model			
.91	47	45	.756	.180	.388	6931.550
.92	43	40	.625	.458	.674	7172.647
.93	40	35	.391	.895	.686	7448.875
.94	34	30	.460	.814	.741	7772.882
.95	26	25	.838	.571	.834	8164.451
.96	19	20	.818	.215	.452	8644.734
.97	12	15	.415	.439	.532	9269.095
.98	10	10	1	.520	.813	10,147.308
			Benchmark model			
.91	73	45	0	.104	0	6359.749
.92	69	40	0	.174	0	6510.675
.93	62	35	0	.162	0	6678.511
.94	55	30	0	.072	0	6869.484
.95	45	25	0	.290	0	7094.033
.96	35	20	.002	.706	.007	7359.968
.97	29	15	.001	.796	.005	7700.567
.98	22	10	.001	.974	.004	8161.252

Notes: Obs. represents the observed violations, and Exp. represents the expected violations. The α of Var$_{\alpha}(t)$ level varies from 0.91 to 0.98, and $\overline{\mathrm{Var}}$ represents the average VaR values.

benchmark models can pass the test at the .05 level. Furthermore, we observe that the benchmark model always underestimates the VaRs of total losses as the observed number of violations are always larger than the expected ones. This can also be observed from the average VaRs because the VaRs based on the proposed vine model are always greater than the VaRs based on the independence model. This phenomenon can be explained by the fact that the positive dependence between the losses leads to a larger VaR. We conclude that the out-of-sample performance of proposed vine model is much better than that of the benchmark model.

Table 2 reports the prediction results for the att$_{\mathrm{GC}}$ and att$_{\mathrm{TC}}$ datasets, respectively, where the dependence structures of Gaussian copula and t copula are misspecified as the vine copula.

- *Gaussian copula*. From Table 2, we observe that for the multivariate Gaussian dependence, the prediction performance of misspecified vine copula model is very satisfactory. The observed number of violations are very close to the expected ones at all α's levels. In particular, the p-values of all three tests are larger than .05. Therefore, we conclude that the proposed approach works well for the attack processes with the multivariate Gaussian copula dependence.
- *t copula*. From Table 2, we observe that for multivariate t dependence, the overall prediction performance of misspecified vine copula model is also satisfactory. The p-values of all three tests are larger than .05 for all α's levels except that the p-value of LR$_{\mathrm{ind}}$ test is .034 at $\alpha = .97$. The observed number of violations are fairly close to the expected ones at all α levels. When compared to the Gaussian case, the observed number of violations is less accurate overall. The misspecification seems to slightly overestimate the total loss.

We conclude, based on the simulation study, that the proposed vine model can effectively predict the total cyber loss, and that the model still has a satisfactory prediction

Table 2. Assessing prediction performance of the proposed approach based on the simulated out-of-sample data when the real dependence structures are multivariate Gaussian and *t* copulas.

α	Obs.	Exp.	LR_{uc}	LR_{ind}	LR_{cc}	$\overline{VaR}$
			Vine model for multivariate Gaussian copula			
.91	43	45	.753	.667	.868	5965.124
.92	38	40	.740	.536	.781	6214.522
.93	34	35	.860	.289	.561	6504.463
.94	30	30	1	.480	.779	6839.169
.95	25	25	1	.100	.258	7237.705
.96	18	20	.643	.241	.452	7751.667
.97	13	15	.592	.401	.609	8409.049
.98	10	10	1	.520	.813	9349.560
			Vine model for multivariate *t* copula			
.91	39	45	.339	.552	.530	6757.069
.92	33	40	.235	.211	.226	7044.525
.93	27	35	.145	.214	.160	7388.984
.94	26	30	.441	.175	.297	7798.743
.95	22	25	.530	.067	.153	8305.730
.96	18	20	.643	.144	.308	8956.474
.97	13	15	.592	.034	.092	9866.702
.98	8	10	.508	.102	.211	11,254.689

Notes: Obs. represents the observed violations, and Exp. represents the expected violations. The α of $Var_\alpha(t)$ level varies from 0.91 to 0.98, and $\overline{Var}$ represents the average VaR values.

performance when the dependence structure is misspecified as the vine structure when the true dependence structure is multivariate Gaussian or t copula.

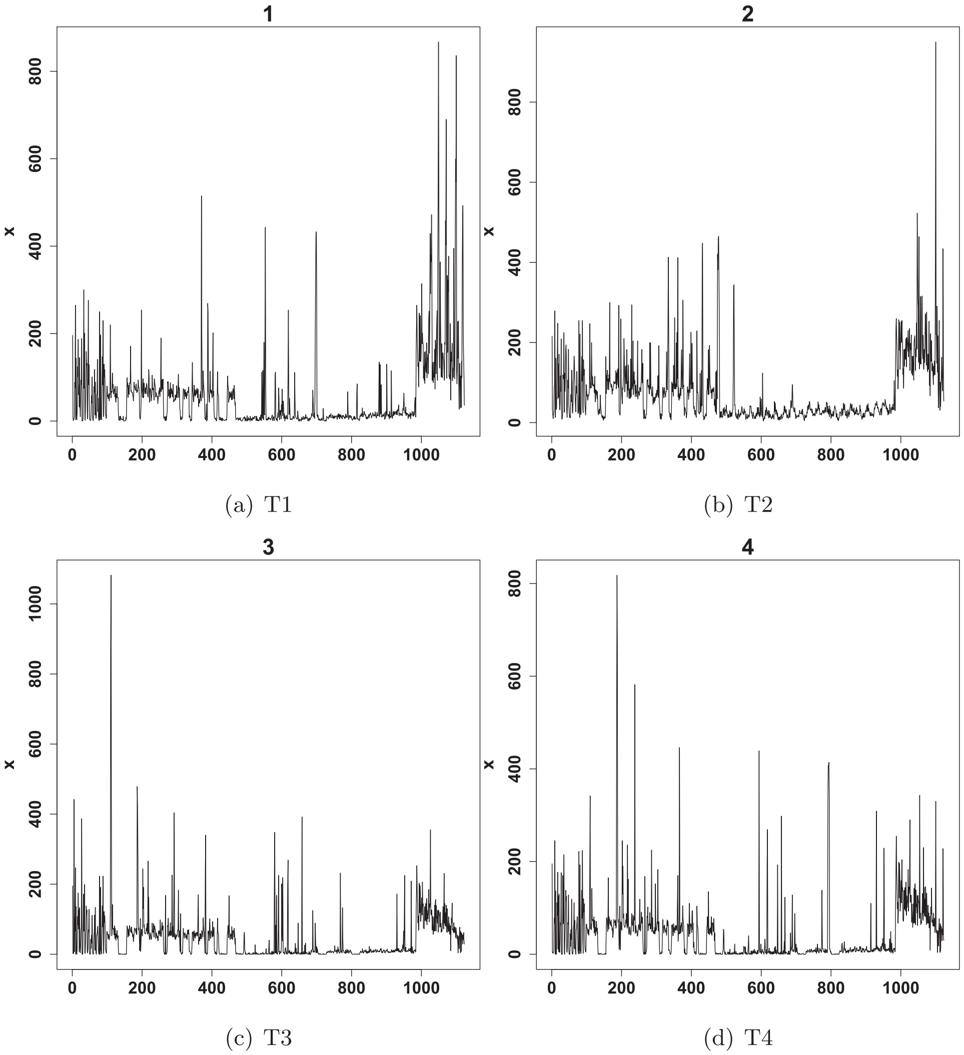
4. Empirical study

In this section, we present an empirical study on a cyber attack dataset.

4.1. Data description

The dataset was collected by a low-interaction honeypot [42], which has 69 consecutive IP addresses. The dataset is the same as the one analyzed in [50], but the present analysis is at a much higher resolution (i.e. at the *computer* level, meaning there are 69 servers, while the main analysis in [50] is at the *network* level, meaning that the 69 servers are treated as a whole). The dataset was collected during 4 November 2010 to 21 December 2010 with a total number of 1123 hours. As described in [50], each TCP flow initiated by a remote computer is treated as an attack because the honeypot offers no legitimate service. An unsuccessful TCP handshake is also deemed as an attack because a handshake can be dropped by a honeypot program [3,50].

Since we have a 69-dimensional time series of cyber attack data, we randomly select four of them and plot them in Figure 3. We observe that there exist clusters and extreme attacks in the time series. Furthermore, we observe that there are some time intervals (say the 200th and 400th) during which the numbers of attacks are large for most of the time series. This indicates that there may exist positive dependence among cyber attacks. The experience in studying multivariate time series datasets that exhibit high clusters of volatilities suggests us to use copula-GARCH models to fit the data. Indeed, a preliminary analysis of the residuals obtained after removing the means shows that a GARCH model is preferred for describing



(a) T1

(b) T2

(c) T3

(d) T4

Figure 3. Randomly selected four servers under cyber attacks, where the *y*-axis represents the number of attacks per hour, and the *x*-axis represents the time (unit: hour).

the volatilities. Please refer to [4,28,37,39,48] and the references therein for an overview and recent developments in the field of multivariate time series analysis.

Since the multivariate cyber attack dataset is 69-dimension with size 1123, we divide it into an in-sample part of 900 samples for modeling and an out-of-sample part of the rest 223 samples for out-of-sample evaluation.

4.2. Model fitting

We follow the approach described in Section 3.1 to perform the modeling process. Our preliminary analysis on the residuals shows that GARCH(1,1) is sufficient to capture the volatilities in the residuals. Therefore, we fix the GARCH part as GARCH(1,1). For the

Table 3. The first level tree with their pair copulas and AICs, where the numbers represent the copula families: 2-Student-*t*, 7-BB1, 9-BB7 and 10-BB8.

Block	AIC	Family	Block	AIC	Family	Block	AIC	Family
(12, 11)	−1674.44	2	(10, 11)	−1747.85	2	(23, 21)	−1684.66	7
(33, 37)	−2040.26	2	(37, 38)	−2152.34	2	(40, 38)	−2092.25	2
(9, 10)	−1937.52	2	(15, 19)	−1884.72	2	(21, 22)	−2020.69	2
(5, 4)	−1698.69	2	(6, 11)	−1532.73	2	(16, 28)	−1987	2
(4, 11)	−1584.98	2	(11, 16)	−1383.74	2	(44, 43)	−2249.2	2
(39, 38)	−2376.07	2	(17, 28)	−2046.29	2	(36, 35)	−2015.44	2
(3, 6)	−1568.24	2	(18, 19)	−1847.08	2	(49, 43)	−2123.29	2
(8, 11)	−1628.1	2	(51, 49)	−1978.98	2	(45, 43)	−1957.02	2
(65, 68)	−136.726	10	(50, 46)	−2207.87	2	(47, 43)	−2353.93	2
(35, 27)	−1977.83	2	(38, 22)	−2031.49	2	(22, 25)	−2117.04	2
(67, 66)	−399.532	2	(46, 43)	−2436.25	2	(42, 41)	−2250.56	2
(69, 68)	−526.255	2	(19, 16)	−2036.85	2	(48, 38)	−1915.24	2
(66, 68)	−335.894	2	(20, 17)	−1924.63	2	(41, 43)	−2257.07	2
(24, 26)	−2046.92	2	(26, 25)	−2202.01	2	(25, 27)	−2237.87	2
(68, 17)	−441.061	2	(14, 25)	−1536.61	2	(2, 27)	−293.85	2
(7, 6)	−1627.53	2	(13, 24)	−1611.19	2	(43, 40)	−2186.43	2
(54, 60)	−146.787	9	(60, 55)	−159.447	9	(57, 64)	−114.864	9
(55, 53)	−114.295	9	(64, 53)	−120.098	9	(34, 32)	−2111.71	2
(53, 61)	−120.763	2	(1, 63)	−36.601	2	(61, 63)	−122.612	9
(27, 28)	−2304.89	2	(31, 34)	−2113.72	2	(30, 32)	−1563.83	7
(32, 28)	−1991.96	2	(29, 28)	−2063.75	2	(28, 1)	−496.522	2
(62, 54)	−89.081	2	(56, 52)	−129.643	2	(52, 54)	−131.083	9
(58, 61)	−129.832	2	(59, 53)	−115.419	2			

mean part, we allow the mean of the time series to vary in the form of autoregressive and moving average processes. We find that AR(1)+GARCH(1,1) model with skewed Student-*t* innovation is sufficient for modeling all the marginal time series by using the same criterions described in the simulation study. Furthermore, the Ljung–Box tests on the residuals of marginal time series pass the significant level at 0.05 for all the time series. Therefore, for the marginal time series, the AR(1)+GARCH(1,1) model with skewed Student-*t* innovation is sufficient to remove the dependence among the marginal time series.

Next, we discuss how to model the dependence between the time series via vine copulas. We use the Algorithm 1 described in [7] to select the truncated vine copulas based on the Vuong test with a significant level 0.05 [47]. The candidate bivariate copulas include Gaussian, Student-*t* , Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7, BB8 and their rotated copulas of 90, 180 and 270 degrees.

For the cyber attack dataset, the truncated level is determined as $K = 26$, and we use the M matrix to store all the information of vine structure. Table 3 shows the first level tree blocks, pair copulas, and pair AICs. For the first level, we observe that the Student-*t* copula is preferred by most blocks (around 85%), and the second preferred one is BB7 (around 10%). This might be because both copulas are flexible to capture the tail dependence, which is exhibited by the cyber attack data in the form of simultaneous extreme cyber attacks.

In order to further illustrate the selected dependence structure, Figure 4 plots a part of the copula blocks selected at level 1, and the pair AIC values are on the edges. For the pair copulas selected on the other levels, we observe that the Student-*t* copula is still the most preferred structure (454 times, 19.35%), the second preferred is the Frank copula (285 times, 12.15%), and the third one is the Gaussian copula (91 times, 3.88%).

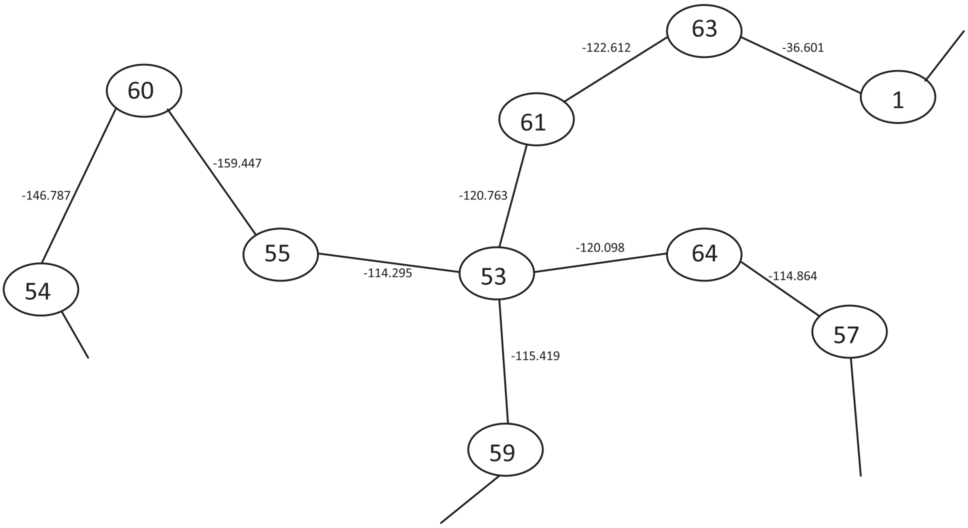


Figure 4. Part of the R vine structure at level 1: pairs (57, 64), (53, 64), (53, 61), (61, 63), (1, 63), (53, 59), (53, 55), (55, 60) and (54, 60) with pair AIC values on their edges.

4.3. Out-of-sample performance

The out-of-sample performance of the proposed model is evaluated based on two metrics: (i) the number of VaR violations (see Algorithm 2); and (ii) the predicted multivariate density [14,22].

Number of VaR violations. As discussed in Section 3, we use Algorithm 2 to evaluate the out-of-sample performance of the proposed dependence model.

Table 4 reports the testing results of the proposed model based on $\mathrm{LR}_{\mathrm{uc}}$, $\mathrm{LR}_{\mathrm{ind}}$ and $\mathrm{LR}_{\mathrm{cc}}$. For comparison purposes, we also present the testing results based on the benchmark model (i.e. the independence model). We observe that the proposed vine copula model passes all the tests for α levels between .95 and .97, and that the tests based on the benchmark model fail at all levels. In particular, we observe that ignoring the dependence among multivariate attacks underestimates the VaR's in terms of the number of both violations and average VaRs. This observation resonates the conclusion drawn in Section 3.

We conclude that the dependence between cyber attacks cannot be ignored. The consequence of ignoring the dependence would significantly underestimate cyber risks. We further compare the prediction performance of the proposed vine model to that of two other widely used models: the Gaussian and t copula models. For the Gaussian and t copula models, unstructured covariances are implemented because they are flexible enough to accommodate various kinds of dependence. Table 4 shows that the estimated VaR's are larger than the that of benchmark model. This further indicates that the dependence should not be ignored for multivariate attack processes. The prediction performances of multivariate Gaussian and t models are comparable to that of the proposed vine model in terms of the VaR violations at levels $\alpha = .96, .97$. At level $\alpha = .95$, the prediction performance of proposed vine model is, in terms of the number of violations, slightly better than that of multivariate Gaussian and t models.

Algorithm 2 Algorithm for rolling prediction of VaR$_\alpha$'s of total number of attacks

```
Input: Multivariate historical time series dataset
$\{(i, x_{i,t}, s(t)) | i = 1, \dots, 69; t = 1, \dots, 1123\}$; in-sample dataset
$\{(i, x_{i,t}) | i = 1, \dots, 69; t = 1, \dots, 900\}$; out-of-sample dataset
$\{(i, x_{i,t}) | i = 1, \dots, 69; t = 901, \dots, 1123\}$; $\alpha$ levels;
for $t = 901, \dots, 1123$ do
    Based on the estimated AR(1)+GARCH(1,1) model, predict the next mean $\hat{\mu}_i$ and
    standard error $\hat{\sigma}_{i,t}$, $i = 1, \dots, 69$
    Based on the estimated Vine copula, simulate 5000 69-dimensional copula samples
    $u_{i,t}^{(k)}$, $i = 1, \dots, n$ by using the algorithm in 1, $k = 1, \dots, 5000$
    Convert the simulated dependent samples, namely the $u_{i,t}^{(k)}$'s, into the $z_{i,t}^{(k)}$'s by
    using the inverse of the estimated Student-$t$ distribution, $k = 1, \dots, 5000$
    Compute the predicted 5000 69-dimensional time series attack data $x_{i,t}^{(k)}$ by using
    Equation (A5), $k = 1, \dots, 5000$
    Calculate the simulated total number of attacks $T_t^{(k)} = \sum_{i=1}^{69} x_{i,t}^{(k)}$, $k = 1, \dots, 5000$
    return VaR$_\alpha(t)$ based on predicted 5000 $T_t^{(k)}$'s
    if The observed total number of attacks $\sum_{i=1}^{69} x_{i,t} > \mathrm{VaR}_\alpha(t)$ then
        An violation occurs;
        return 1
    else
        return 0
    end if
end for
Output: VaR$_\alpha(t)$'s of the total number of attacks, and the number of violations.
```

Prediction accuracy of multivariate density. In order to further evaluate the prediction performance of the proposed model, we study the prediction accuracy of multivariate density of the proposed model. We adapt the framework presented in [22], which assumes that any unknown model parameters are estimated on the basis of a moving window of fixed size. We use the Kullback–Leibler information criterion (KLIC) to evaluate the prediction performance, as described in [14]. The KLIC measures the divergence between the true probability density and a candidate density. Specifically, the KLIC of the one-step-ahead predicted density $\hat{f}_t$ for $\mathbf{Z}_{t+1} = (Z_{1,t+1}, \dots, Z_{69,t+1})$ is given by

$$\mathrm{KLIC}(\hat{f}_t) = \mathrm{E}_t[\log p_t(\mathbf{Z}_{t+1}) - \log \hat{f}_t(\mathbf{Z}_{t+1})],$$

where p_t is the true, but unknown, conditional density of $\mathbf{Y}_{t+1}$. A KLIC value is nonnegative, and the KLIC value equals zero only when $\hat{f}_t$ equals the true conditional density p_t. Moreover, the smaller the KLIC value, the closer the predicted density to the true conditional density. In practice, p_t is unknown, and hence the KLIC score $K_{t+1} = \log \hat{f}_t(\mathbf{Z}_{t+1})$ may be used. A larger $\mathrm{E}_t[S_{t+1}]$ means the predicted density is closer to the true conditional

Table 4. Assessing prediction performance based on the out-of-sample dataset, where Obs. represents the observed violations, Exp. represents the expected violations, and $\overline{\text{Var}}$ represents the average VaR values.

α	Obs.	Exp.	LR_{uc}	LR_{ind}	LR_{cc}	$\overline{\text{Var}}$
			Vine model			
.95	14	11	.399	.893	.694	17,680.158
.96	13	9	.191	.773	.408	18,157.087
.97	9	7	.388	.344	.440	18,815.904
			Benchmark model			
.95	56	11	0	.642	0	15,052.927
.96	53	9	0	.883	0	15,203.785
.97	47	7	0	.371	0	15,399.267
			Multivariate Gaussian model			
.95	15	11	.260	.988	.530	17,721.939
.96	13	9	.191	.773	.408	18,136.764
.97	9	7	.388	.344	.440	18,694.954
			Multivariate *t* model			
.95	18	11	.052	.651	.137	17,578.619
.96	13	9	.191	.773	.408	18,047.065
.97	10	7	.225	.439	.356	18,688.262

density [15]. According to Equation (3), we have

$$K_{t+1} = \log \hat{c}_t(\mathbf{u}_{t+1}) - \sum_{j=1}^{69} \log(\sigma_{j,t+1}) + \sum_{j=1}^{69} \log(g_j(u_{j,t+1})),$$

where $\mathbf{u}_{t+1} = (u_{1,t+1}, \dots, u_{69,t+1})$ with $u_{i,t} = F_i((x_{i,t} - \mu_{i,t})/\sigma_{i,t})$ for $i = 1, \dots, 69$, and $\hat{c}_t(\cdot)$ is the conditional copula density estimated from the historical data up to time t.

For two competing copulas A and B, the null hypothesis is that for each t, the two models have equal prediction performance on average. Since the conditional marginals are identically specified under both density predictions, it is equivalent to testing the following null hypothesis:

$$H_0 : \text{E}(\log \hat{c}_{A,t}(\hat{\mathbf{U}}_{t+1})) = \text{E}(\log \hat{c}_{B,t}(\hat{\mathbf{U}}_{t+1}))$$

for $t = 1, 2, \dots$. The formal KLIC score test is developed as follows. Suppose the parameters of marginals and copulas are estimated via a moving window of fixed length l, and m is the length of the testing data. Based on the observed score differences

$$d_{t+1} = \log \hat{c}_{A,t}(\hat{\mathbf{U}}_{t+1}) - \log \hat{c}_{B,t}(\hat{\mathbf{U}}_{t+1}),$$

where $t = l, \dots, l + m - 1$, the test statistic can be constructed as

$$Q_{l,m} = \sqrt{m}\frac{\bar{d}_{l,m}}{\hat{\sigma}_{l,m}},$$

where $\bar{d}_{l,m} = \sum_{t=l+1}^{l+m} d_t$ is the sample mean of the KLIC score difference, and $\hat{\sigma}_{l,m}$ is a heteroscedasticity and autocovariance consistent (HAC) estimate of the asymptotic standard deviation of $\sqrt{m}\bar{d}_{l,m}$. We use the standard HAC estimation, namely $\hat{\sigma}^2_{l,m} = \hat{\gamma}_0 + 2\sum_{k=1}^{K-1} a_k \hat{\gamma}_k$ where $\hat{\gamma}_k$ represents the lag-k sample covariance of sequence d_t for $t = l + 1, \dots, t + m$, and the a_k's are the Barlett weights, namely $a_k = 1 - k/K$ with $K = \lfloor p^{1/4} \rfloor$.

Table 5. $Q_{l,m}$ test statistics by comparing the means of the proposed vine copula model and that of the other copulas, where $l = 600$ and $m = 223$.

	Benchmark model	Multivariate Gaussian model	Multivariate t model
$Q_{l,m}$	26.812	3.734	16.923
$\Phi(Q_{l,m})$	1	.9999	1

The positive sign of $Q_{l,m}$ indicates that copula A has a better prediction performance than copula B. It is known [22] that $Q_{l,m}$ follows the asymptotically standard normal distribution, which can be used for testing the significance of the difference. Specifically, we use the following criterion: if $0.05 \leq \Phi(Q_{l,m}) \leq 0.95$, two competing models A and B are not significantly different; if $\Phi(Q_{l,m}) < 0.05$, model B is preferred; If $\Phi(Q_{l,m}) > 0.95$, model A is preferred.

Similar to previous discussion, we compare the prediction performance of proposed vine model to that of the benchmark model, while considering multivariate Gaussian and t copula models with the unstructured covariances. For the cyber attack dataset, we use the length of fixed windows $l = 600$ and $m = 223$. Table 5 describes the comparison results. It is clear that the proposed vine copula model has a better prediction performance than the other models.

Based on the empirical study, we conclude that the dependence between multivariate attack processes cannot be ignored. The consequence of ignoring the dependence is the underestimation of the cybersecurity risk. The proposed vine model has a satisfactory prediction performance, and is recommended for multivariate cyber attack data.

5. Conclusion and discussion

We have proposed a new statistical approach for modeling multivariate cybersecurity risks. The proposed approach is centered on Copula-GARCH models, where multivariate dependence is accommodated by vine copulas. We have used the proposed approach to characterize multivariate cyber losses based on the triggering mechanism. Our results show that multivariate dependence between cyber attacks has a significant effect on the total loss. The implication is that assuming away the dependence between cybersecurity attacks will cause a severe underestimation of the cybersecurity risk. The proposed model can model and predict high-dimensional cybersecurity risks at a satisfactory level, and can be adopted for practical use.

The current research can be extended in several directions. First, other kinds of cyber attack data should be studied. The present research focuses on honeypot data, but other kinds of cyber attack data may exhibit different dependence structures that worth investigation. Second, we need to collect real cyber loss data for modeling purposes in practice. The proposed approach should be evaluated against such data. Third, the present work focuses on the static copula approach for modeling high-dimensional cybersecurity risks. It would be interesting to study whether time-varying/dynamic copula models can further improve the prediction accuracy. The time-varying feature can be imposed on the copula function(s) or the dependence parameter(s). There are studies on the time-varying copula models [11,32,38,39]. Our preliminary analysis shows that the time-varying t copula with DCC(1,1) [32] can offer a good balance between the flexibility and the parsimony.

The prediction performance of the time-varying t copula is comparable to that of multivariate t copula with unstructured covariance in terms of the number of VaR violations. Dynamic factor copula, GAS (generalized autoregressive score) [11,38], or dynamic D-vine [2] models can be good candidate models for modeling high-dimensional cybersecurity risks.

Acknowledgements

The authors are very grateful to the two anonymous referees for their insightful and constructive comments which led to this improved version of the paper. We thank Marcus Pendleton for proofreading the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

M. Xu and S. Xu were supported in part by Army Research Office [W911NF-13-1-0141], C. Peng was supported by China Scholarship Council, and T. Hu was supported by National Natural Science Foundation of China [11371340].

References

[1] K. Aas, C. Czado, A. Frigessi, and H. Bakken, *Pair-copula constructions of multiple dependence*, Insur. Math. Econ. 44 (2009), pp. 182–198.

[2] C. Almeida, C. Czado, and H. Manner, *Modeling high-dimensional time-varying dependence using dynamic d-vine models*, Appl. Stoch. Models Bus. Ind. 32 (2016), pp. 621–638.

[3] S. Almotairi, A. Clark, G. Mohay, and J. Zimmermann, *Characterization of attackers' activities in honeypot traffic using principal component analysis*, Proceedings of the 2008 IFIP International Conference on Network and Parallel Computing, Shanghai, China, 2008, pp. 147–154.

[4] L. Bauwens, S. Laurent, and J.V.K. Rombouts, *Multivariate GARCH models: A survey*, J. Appl. Econometrics 21 (2006), pp. 79–109.

[5] T. Bedford and R.M. Cooke, *Vines: A new graphical model for dependent random variables*, Ann. Statist. 30 (2002), pp. 1031–1068.

[6] R. Böhme and G. Kataria, *Models and measures for correlation in cyber-insurance*, Workshop on the Economics of Information Security, 2006

[7] E.C. Brechmann, C. Czado, and K. Aas, *Truncated regular vines in high dimensions with application to financial data*, Canad. J. Statist. 40 (2012), pp. 68–85.

[8] Y.-Z. Chen, Z.-G. Huang, S. Xu, and Y.-C. Lai, *Spatiotemporal patterns and predictability of cyberattacks*, PLoS ONE 10 (2015), p. e0124472.

[9] J.-H. Cho, P.M. Hurley, and S. Xu, *Metrics and measurement of trustworthy systems*, 2016 IEEE Military Communications Conference, MILCOM 2016, Baltimore, MD, USA, 1–3 November 2016, 2016, pp. 1237–1242.

[10] P.F. Christoffersen, *Evaluating interval forecasts*, Internat. Econom. Rev. 39 (1998), pp. 841–862.

[11] D. Creal, S.J. Koopman, and A. Lucas, *Generalized autoregressive score models with applications*, J. Appl. Econometrics 28 (2013), pp. 777–795.

[12] D.E. Denning, *An intrusion-detection model*, IEEE Trans. Softw. Eng. SE-13 (1987), pp. 222–232.

[13] Department of Homeland Security, *National critical infrastructure security and resilience research and development plan*, November 2015. Available at http://publish.illinois.edu/ciri-new-theme/files/2016/09/National-CISR-RD-Plan_Nov-2015.pdf.

[14] C. Diks, V. Panchenko, and D. Van Dijk, *Out-of-sample comparison of copula specifications in multivariate density forecasts*, J. Econom. Dynam. Control 34 (2010), pp. 1596–1609.

[15] C. Diks, V. Panchenko, and D. Van Dijk, *Likelihood-based scoring rules for comparing density forecasts in tails*, J. Econom. 163 (2011), pp. 215–230.

[16] J. Dißmann, E.C. Brechmann, C. Czado, and D. Kurowicka, *Selecting and estimating regular vine copulae and application to financial returns*, Comput. Stat. Data Anal. 59 (2013), pp. 52–69.

[17] B. Edwards, S. Hofmeyr, and S. Forrest, *Hype and heavy tails: A closer look at data breaches*, Workshop on the Economics of Information Security, 2015.

[18] M. Eling and W. Schnell, *What do we know about cyber risk and cyber risk insurance?*, J. Risk Finance 17 (2016), pp. 474–491.

[19] C. Fernandez and M.F.J. Steel, *On bayesian modeling of fat tails and skewness*, J. Am. Statist. Assoc. 93 (1998), pp. 359–371.

[20] D. Freckleton, K. Heaslip, W. Louisell, and J. Collura, *Evaluation of resiliency of transportation networks after disasters*, Transp. Res. Rec. J. Transp. Res. Board 2284 (2012), pp. 109–116.

[21] E. Gandotra, D. Bansal, and S. Sofat, *Computational techniques for predicting cyber threats*, in *Intelligent Computing, Communication and Devices*, Lakhmi C. Jain, Srikanta Patnaik, and Nikhil Ichalkaranje, eds., Springer, New Delhi, IN, 2015, pp. 247–253

[22] R. Giacomini and H. White, *Tests of conditional predictive ability*, Econometrica 74 (2006), pp. 1545–1578.

[23] Y. Han, W. Lu, and S. Xu, *Characterizing the Power of Moving Target Defense via Cyber Epidemic Dynamics*, Proceeding of the 2014 Symposium and Bootcamp on the Science of Security (HotSoS'14), Raleigh, NC, USA, p. 10.

[24] P.R. Hansen and A. Lunde, *A forecast comparison of volatility models: Does anything beat a GARCH(1, 1)?* J. Appl. Econom. 20 (2005), pp. 873–889.

[25] V.S.B. Herath and T.C. Herath, *Copula-based actuarial model for pricing cyber-insurance policies*, Insur. Markets Companies: Anal. Actuar. Comput. 2 (2011), pp. 7–20.

[26] C. Ishida, Y. Arakawa, I. Sasase, and K. Takemori, *Forecast techniques for predicting increase or decrease of attacks using Bayesian inference*, PACRIM. 2005 IEEE Pacific Rim Conference on Communications, Computers and signal Processing, 2005, IEEE, Victoria, BC, Canada, pp. 450–453.

[27] H. Joe, *Dependence Modeling with Copulas*, CRC Press, Vancouver, Canada, 2014.

[28] E. Jondeau and M. Rockinger, *The copula-GARCH model of conditional dependencies: An international stock market application*, J. Int. Money Finance 25 (2006), pp. 827–853.

[29] D.H. Kim, T. Lee, S.-O.D. Jung, H.P. In, and H.J. Lee, *Cyber threat trend analysis model using HMM*, Third International Symposium on Information Assurance and Security, IEEE, Manchester, UK, 2007, pp. 177–182.

[30] T. Kosub, *Components and challenges of integrated cyber risk management*, Z. die gesamte Versicherungswissenschaft 104 (2015), pp. 615–634.

[31] D. Kurowicka and H. Joe, *Dependence Modeling: Vine Copula Handbook*, World Scientific Publishing Company, Singapore, 2011.

[32] H. Manner and O. Reznikova, *A survey on time-varying copulas: Specification, simulations, and application*, Econom. Rev. 31 (2012), pp. 654–687.

[33] M. Markou and S. Singh, *Novelty detection: A review—part 1: Statistical approaches*, Signal Process. 83 (2003), pp. 2481–2497.

[34] O. Morales-Nápoles, *Bayesian belief nets and vines in aviation safety and other applications*, Ph.D. thesis, 2008.

[35] A. Mukhopadhyay, S. Chatterjee, D. Saha, A. Mahanti, and S.K. Sadhukhan, *e-Risk management with insurance: A framework using copula aided Bayesian belief networks*, Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), Vol. 6, IEEE, Kauia, HI, USA, 2006, pp. 126a–126a

[36] J. Neil, C. Hash, A. Brugh, M. Fisk, and C.B. Storlie, *Scan statistics for the online detection of locally anomalous subgraphs*, Technometrics 55 (2013), pp. 403–414.

[37] A.K. Nikoloulopoulos, H. Joe, and H. Li, *Vine copulas with asymmetric tail dependence and applications to financial return data*, Comput. Statist. Data Anal. 56 (2012), pp. 3659–3673.

[38] D.H. Oh and A.J. Patton, *Time-varying systemic risk: Evidence from a dynamic copula model of CDS spreads*, J. Bus. Econom. Statist. (2017), pp. 1–15. doi:10.1080/07350015.2016.1177535.

[39] A.J. Patton, *Copula methods for forecasting multivariate time series*, Hand. Econ. Forecast. 2 (2012), pp. 899–960.

[40] M. Pendleton, R. Garcia-Lebron, J.-H. Cho, and S. Xu, *A survey on systems security metrics*, ACM Comput. Surv. 49 (2016), pp. 62:1–62:35.

[41] C. Peng, M. Xu, S. Xu, and T. Hu, *Modeling and predicting extreme cyber attack rates via marked point processes*, J. Appl. Statist. 44 (2016), pp. 1–30.

[42] N. Provos, *A virtual Honeypot framework*, Proceedings of the 13th Conference on USENIX Security Symposium – Volume 13, SSYM'04, Berkeley, CA, USA, USENIX Association, 2004, pp. 1–1.

[43] D.A. Reed, K.C. Kapur, and R.D. Christie, *Methodology for assessing the resilience of networked infrastructure*, IEEE Syst. J. 3 (2009), pp. 174–180.

[44] A. Sklar, *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Stat. Univ. Paris 8 (1959), pp. 229–231.

[45] M. van Dijk, A. Juels, A. Oprea, and R.L. Rivest, *Flipit: The game of 'stealthy takeover'*, J. Cryptol. 26 (2013), pp. 655–713.

[46] E.D. Vugrin, D.E. Warren, and M.A. Ehlen, *A resilience assessment framework for infrastructure and economic systems: Quantitative and qualitative resilience analysis of petrochemical supply chains to a hurricane*, Process Saf. Prog. 30 (2011), pp. 280–290.

[47] Q.H. Vuong, *Likelihood ratio tests for model selection and non-nested hypotheses*, Econometrica 57 (1989), pp. 307–333.

[48] M. Xu, L. Hua, and S. Xu, *A vine copula model for predicting the effectiveness of cyber defense early-warning*, Technometrics 59 (2017), pp. 508–520.

[49] S. Xu, W. Lu, L. Xu, and Z. Zhan, *Adaptive epidemic dynamics in networks: Thresholds and control*, ACM Trans. Auton. Adapt. Syst. 8 (2014), pp. 19:1–19:19.

[50] Z. Zhan, M. Xu, and S. Xu, *Characterizing honeypot-captured cyber attacks: Statistical framework and case study*, IEEE Trans. Inf. Forensics Secur 8 (2013), pp. 1775–1789.

[51] Z. Zhan, M. Xu, and S. Xu, *A characterization of cybersecurity posture from network telescope data*, Trusted Systems – 6th International Conference, INTRUST 2014, Beijing, China, December 16–17, 2014, Revised Selected Papers, 2014, pp. 105–126.

[52] Z. Zhan, M. Xu, and S. Xu, *Predicting cyber attack rates with extreme values*, IEEE Trans. Inf. Forensics Secur 10 (2015), pp. 1666–1677.

Appendix. Simulation of multivariate cyber losses

In the following, we discuss the simulation of multivariate cyber losses under the triggering mechanism. The experiments are set up as follows.

Consider an enterprise with $n = 10$ servers, meaning the following 10-dimensional cyber attack processes

$$(X_{1,t}, \ldots, X_{10,t}),$$

where $t = 1, \ldots, 1,500$. We first generate marginal cyber attack processes from the AR(1)+GARCH (1,1) model. That is, the number of attacks $X_{i,t}$ has the following form:

$$X_{i,t} = \mu_i + \phi_i(X_{i,t-1} - \mu_i) + \epsilon_{i,t}, \quad i = 1, \ldots, 10, \tag{A1}$$

and

$$\epsilon_{i,t} = \sigma_{i,t} Z_i$$

with Z_i being the innovations that are independently and identically distributed. For the standard GARCH(1,1) model, we have

$$\sigma_{i,t}^2 = w_i + \alpha_i \epsilon_{i,t-1}^2 + \beta_i \sigma_{i,t-1}^2,$$

Table A1. Parameters of marginal models of $n = 10$ servers.

Attack process	μ	ϕ	ω	α	β	ξ	ν
1	13.982	.852	29.214	.632	.367	1.229	2.921
2	3.326	.845	2.790	.469	.530	1.122	2.689
3	7.737	.866	12.324	.703	.296	1.259	2.792
4	4.725	.840	5.560	.470	.529	1.145	2.638
5	5.438	.849	7.145	.567	.432	1.191	2.788
6	6.679	.854	9.481	.614	.385	1.237	2.832
7	6.905	.881	8.450	.562	.437	1.176	2.783
8	5.019	.892	3.511	.477	.522	1.152	3.034
9	5.775	.897	5.263	.557	.442	1.186	2.978
10	4.229	.885	2.650	.448	.551	1.131	3.031

where $\sigma_{i,t}^2$ is the conditional variance and w_i is the intercept. The density function of Z_i can be explicitly written as (see [19])

$$g_i(z;\boldsymbol{\vartheta}_i) = \frac{2}{\xi+\xi^{-1}}[t_\nu(\xi z)I(z<0) + t_\nu v(\xi^{-1}z)\mathrm{I}(z \geq 0)], \tag{A2}$$

where $\mathrm{I}(\cdot)$ is the indicator function, $\boldsymbol{\vartheta}_i = (\xi_i, \nu_i)$, $\xi_i > 0$ is the skewness parameter, and

$$t_{\nu_i}(z) = \frac{\Gamma((\nu_i+1)/2)}{\sqrt{\nu_i\pi}\,\Gamma(\nu_i/2)}[1 + z^2/\nu_i]^{-(\nu_i+1)/2}$$

with the shape parameter $\nu_i > 0$. The parameters used to generate the attack processes are listed in Table A1. These parameters are determined from the real attack data studied in Section 4.

After determining the parameters for each attack process, we generate the 10-dimensional attack process data of 1500 time intervals. Note that these 10-dimensional attack processes are independent.

Now, we discuss how to incorporate dependence into the simulated attack processes.

(a) Vine copula. To generate the vine dependence, we use the following M matrix for the vine model:

$$M = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 9 & 7 & 1 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 9 & 7 & 1 & 6 & 0 & 0 & 0 & 0 & 0 \\ 8 & 10 & 9 & 7 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 8 & 10 & 9 & 7 & 8 & 7 & 0 & 0 & 0 \\ 4 & 2 & 8 & 10 & 9 & 7 & 8 & 8 & 0 & 0 \\ 5 & 4 & 6 & 8 & 10 & 9 & 10 & 9 & 9 & 0 \\ 6 & 6 & 4 & 6 & 8 & 10 & 9 & 10 & 10 & 10 \end{pmatrix}, \tag{A3}$$

where the elements in the M matrix indicate the attack processes. The families for pair copulas in the vine tree are listed as follows:

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 9 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 2 & 23 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 19 & 4 & 14 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 9 & 5 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 5 & 2 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 20 & 5 & 20 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 2 & 5 & 2 & 2 & 0 & 0 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 0 \end{pmatrix}, \tag{A4}$$

where the numbers represent the families of pair copulas: 2-Student-*t* copula; 4-Gumbel copula; 5-Frank copula; 9-BB7 copula; 14-rotated Gumbel copula (180 degrees); 19-rotated BB7 copula (180 degrees); 20-rotated BB8 copula (180 degrees); 23-rotated Clayton copula (90 degrees). For more details on the copula families, please refer to [27]. Note that the parameters of vine structure M in Equation (A3) and H in Equation (A4) are in fact from the real attack processes studied in Section 4.

After determining the vine structure M and H, we generate 10-dimensional multivariate random samples $u_{1,t}, \ldots, u_{10,t}$, $t = 1, \ldots, 1,500$ based on the vine structure. We now convert the simulated dependent samples, namely the $u_{i,t}$'s, into the $z_{i,t}$'s by using the inverse of the Student-*t* distribution described in Equation (A2) with estimated skew and shape parameters. Now we have the multivariate dependent attack data

$$x_{i,t} = \hat{\mu}_i + \hat{\sigma}_{i,t} z_{i,t}, \tag{A5}$$

for $i = 1, \ldots, 10$, and $t = 1, \ldots, 1,500$.

(b) Multivariate Gaussian copula. The density of Gaussian copula can be written as

$$c(\mathbf{u}|\mathbf{R}) = \frac{1}{|\mathbf{R}|^{1/2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}^{\mathrm{T}} \cdot (\mathbf{R}^{-1} - \mathbf{I}) \cdot \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix} \right\},$$

where $\mathbf{u} = (u_1, \ldots, u_d)$, $d = 10$, and $\mathbf{R}$ is the correlation matrix. For the multivariate Gaussian copula, we assume a uniform correlation structure, that is,

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix} \tag{A6}$$

and $\rho = .7$. After determining the dependence structure, we use Equation (A5) to produce the multivariate attack data with the multivariate Gaussian dependence.

(c) Multivariate *t* copula. The density of multivariate t copula can be written as:

$$c(\mathbf{u}|\mathbf{R}, \nu) = \frac{\Gamma(\frac{\nu+d}{2})(\Gamma(\frac{\nu}{2})^d) \left(1 + \nu^{-1} \begin{pmatrix} t_\nu^{-1}(u_1) \\ \vdots \\ t_\nu^{-1}(u_d) \end{pmatrix}^{\mathrm{T}} \cdot \mathbf{R}^{-1} \cdot \begin{pmatrix} t_\nu^{-1}(u_1) \\ \vdots \\ t_\nu^{-1}(u_d) \end{pmatrix} \right)^{-(\nu+d)/2}}{|\mathbf{R}|^{1/2}(\Gamma(\frac{\nu+d}{2})^d)\Gamma(\frac{\nu}{2}) \prod_{i=1}^{d} (1 + \frac{(t_\nu^{-1}(u_i))^2}{\nu})^{-(\nu+1)/2}},$$

where $\mathbf{u} = (u_1, \ldots, u_d)$, $d = 10$, t_ν^{-1} is the quantile function of the student distribution with shape parameter ν, and $\mathbf{R}$ is the correlation matrix. For multivariate t copula, we assume $\nu = 2$ and the uniform correlation structure being Equation (A6) with $\rho = 0.9$. We generate the multivariate attack data according to Equation (A5).