2022

Statistical Modeling of Data Breach Risks: Time to Identification and Notification

Quynh Nhu Nguyen
Illinois State University, nqnhu2209@gmail.com

STATISTICAL MODELING OF DATA BREACH RISKS: TIME TO IDENTIFICATION AND NOTIFICATION

QUYNH NHU NGUYEN

40 pages

It is very challenging to predict the cost of a cyber incident owing to the complex nature of cyber risk. However, it is inevitable for insurance companies to offer cyber insurance policies. The time to identifying an incident and the time to noticing the affected individuals are two important components in determining the cost of a cyber incident. In this work, we initialize the study on those two metrics via statistical modeling approaches. We propose a novel approach to imputing the missing data, and further develop a dependence model to capture the complex pattern exhibited by those two metrics. The empirical study shows that the proposed approach has a satisfactory predictive performance and is superior to other commonly used models.

KEYWORDS: Copula; CRPS; Data breach; GARCH; Value-at-Risk

STATISTICAL MODELING OF DATA BREACH RISKS: TIME TO IDENTIFICATION AND NOTIFICATION

QUYNH NHU NGUYEN

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Mathematics

ILLINOIS STATE UNIVERSITY

2022

STATISTICAL MODELING OF DATA BREACH RISKS: TIME TO IDENTIFICATION AND NOTIFICATION

QUYNH NHU NGUYEN

COMMITTEE MEMBERS:

Maochao Xu, Chair

Pei Geng

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Maochao Xu for his guidance, support, and patience during the course of my research work. I would also like to thank Dr. Pei Geng for serving on the thesis committee and for supporting my studies.

I would like to thank Dr. Fusun Akman, Dr. Fuxia Cheng, Dr. Olcay Akman, and Dr. Wenhua Zhao, from whom I have learned so much about mathematics, statistics and received valuable guidance and mentorship.

My gratitude extends to Illinois State University Department of Mathematics for my master program studies.

I would like to thank my parents, my friends for unconditional love and support.

Q.N.N.

CONTENTS

TABLES

FIGURES

CHAPTER I: INTRODUCTION AND MOTIVATION

Data breach is one of the most devastating risks to computer systems, which has become and will continue to be a big problem due to the enormous network activities. This has been witnessed by many severe cybersecurity incidents. For example, the Privacy Rights Clearinghouse (PRC) 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 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, 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.

Due to the unique nature of cyber risk, the breach is often discovered after several days, months, or even years. The longer a breach goes unaddressed, the more data gets leaked and the larger the overall impact – financial and otherwise. For example, according to the 2019 research report by Ponemon Institute, the mean time to identify a data breach is 206 days in 2019, an increase of 4.9 percent from last year's study. The report also pointed out that the faster a data breach can be identified and contained, the lower the cost. Further, organizations in the healthcare and public sector took the most time in the data breach lifecycle, 329 days and 324 days, respectively, and financial organizations took far less time to identify and contain a data breach, 233 days. This discovery coincides with the simulation

study in Hua et al. (2011) that the time to identification is the key to reduce the cost. The other important factor for determining the cost of a data breach is the time to notification. The notification allows the clients to protect themselves, e.g., by changing their passwords, or by monitoring their credit scores, which can further avoid the potential lawsuits against the organization for the loss caused by the data breach. However, in the literature, to the best of our knowledge, there is no formal statistical model for modeling those two metrics: time to identification (TTI), and time to notification (TTN). The only relavent work is in Bisogni et al. (2016) which mainly focused on negative binomial regression analysis and discussed the relationship between TTI and different sectors.

Since the data breach has become the most common and dangerous cyber risk nowadays, there exist several studies on the statistical modeling of data breaches in the literature which are loosely related to our current study. For example, Romanosky et al. (2011) used a fixed effect model to estimate the impact of data breach disclosure policy on the frequency of identity thefts incurred by data breaches. Buckman et al. (2017) 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. Edwards et al. (2016) 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 a Poisson distribution, rather than the negative binomial distribution, and that the size of large breaches still follows log-normal distribution. Eling and Loperfido (2017) studied data breaches from the perspective of actuarial modeling and pricing. They used the 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. (2021) developed a frequency-severity actuarial model of aggregated enterprise-level breach data to promote ratemaking and underwriting in insurance. Ikegami and Kikuchi (2020) 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. Xu and Zhang (2021) showed that the nonstationary extreme value model can capture the statistical pattern of the monthly maximum of data breach size very well, and they also discovered a positive time trend based on the PRC dataset. Using the same dataset, Jung (2021) compared the estimates of extreme value distributions after 2014 and before 2014, and concluded that there is a significant increase with a break in the loss severity.

The current study is different from those in the literature as we aim to study the statistical properties of TTI and TTN. We summarize our contributions in the following: i) Since there exist missing data for both TTN and TTI, and particularly, the missing percentage for TTI is very high, we introduce a novel copula approach to tackling this issue. Compared to the commonly used missing data imputation approaches such as Kalman Smoothing and MICE imputations, the proposed copula approach is simple but efficient, and leads to a better predictive performance. ii) A dependence model is developed to capture the positive dependence between TTN and TTI. The empirical study shows that the proposed model is superior to other commonly used multivariate time series models. iii) The practical implications from the model results are discussed.

The rest of the paper is organized as follows. In Chapter II, we conduct exploratory data analysis on the breach data to motivate the proposed model. Chapter III introduces some preliminaries for the statistical modeling. In Chapter IV, the copula approach for imputing the missing data is introduced. In Chapter V, we develop the dependence model for TTN and TTI, and assess the model performance. In Chapter VI, we conclude the current study and present some discussion.

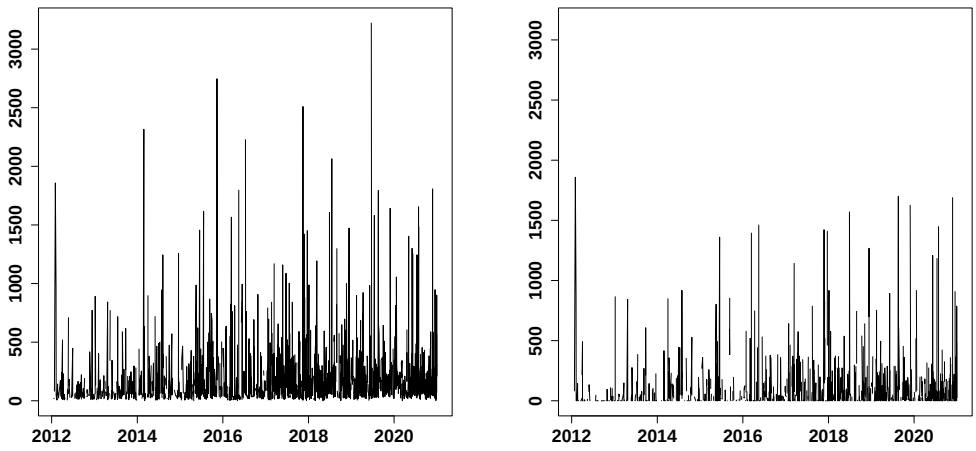
	Min	Q_1	Median	Mean	SD	Q_3	Max	NAs%	0s%
TTN	0.00	47	102	189.8	268.29	223	3222	11.35	0.09
TTI	0.00	0.00	20	101.8	231.81	96	3140	36.65	18.23
ITN	0.00	25	44	62.01	60.92	76	539	30.33	0.71

Table 1: Statistics of TTI (time to identification), TTN (time to notification), and ITN (identification to notification) where 'SD' stands for standard deviation, Q_1, Q_3 represent the first and third quantiles and NAs for missing values

The summary statistics of TTN is reported in Table 1. It is seen that the minimum of TTN is 0 which means that the breach is reported on the same day of its occurrence. However, this percentage is small, only 0.09%. The mean value of TTN is 189.8 days, and the median is 102 days, with standard deviation 268.29 days. The time series plot of TTN is shown in Figure 1(a), and we observe that there are some very large values in TTN. This indicates that TTN has a large variability. The boxplot in Figure 2(a) shows that TTN is very skewed with a large variability. The largest value of TTN is 3222, and it corresponds to the incident of Dominion National reported on 6/21/2019. The breach occurred as early as 2010*, and was the second-largest breach reported to the Department of Health and Human Services. This incident affected 2.9 million patients and caused Dominion National a 2 million settlement. Further, there exist 11.35% missing data (i.e., unknown breach dates) for the whole time period.

For TTI, the time series plot in Figure 1(b) shows there are some very large values and also small values. This indicates that there exists a large variability in TTI. It is seen

*https://www.newsbreak.com/news/2301704159983/dominion-national-reaches-2m-settlement-over-nine-year-data-breach



(a) TTN (b) TTI

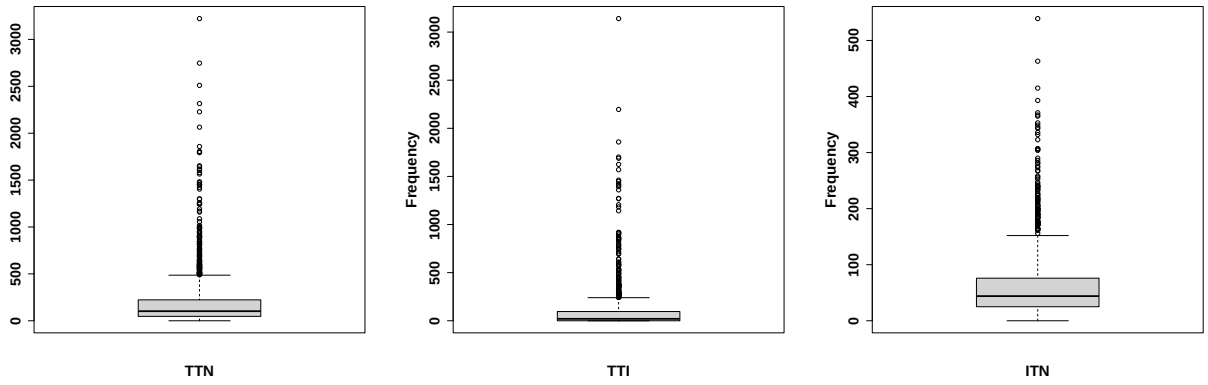
Figure 1: Time series plots of TTI and TTN, Unit: days

from Table 1 that there are 18.23% 0s, which indicates that a small percentage of incidents can be detected on the same day of their occurrences. The mean of TTI is 101.8 days, while the median, 20 days, is much smaller. This suggests that TTI is very skewed as seen from the boxplot in Figure 2(b). The percentage of missing data is very high, i.e., 36.65%. The missing data represents that the breach report does not contain either the breach date or the identification date. The largest value of 3140 days corresponds to the same incident as that of TTN.

The other quantity of interest is the time length from the identification to the notification (ITN) which is the difference of TTN and TTI. It is seen from Table 1 that the mean of ITN is 62.01 days with standard deviation 60.92 days. The median is 44 days which is smaller than the mean. California data breach laws require the most expedient time possible and without unreasonable delay for the notification*. It should be noted that the data breach laws are enacted at the state level, and some states require that the notification must be

*see Cal. Civ. Code §§ 1798.81.5, 1798.82

made within a certain time period, ranging from 24 hours to 90 days*. The boxplot in Figure 2(c) shows that there exist many large values, say, greater than 90 days. The maximum value of ITN is 539 days which corresponds to the incident reported on 6/14/2019 from UC San Diego Health. This incident was that participants' sensitive materials in an HIV research study were made accessible to everyone working at Christie's Place, a San Diego nonprofit supporting women with HIV and AIDS. The organization was criticized for being delayed to notify women affected from the breach†.



(a) TTN (b) TTI (c) ITN

Figure 2: Boxplots of TTI and TTN, Unit: days

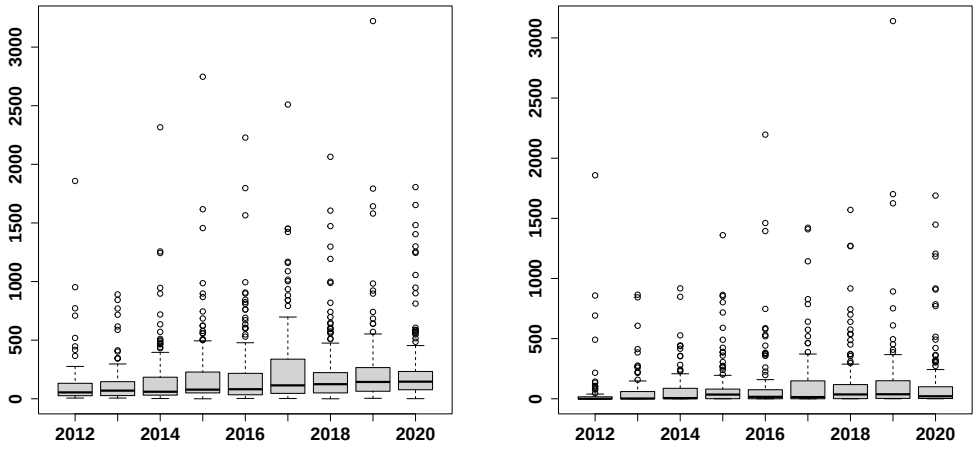
It is also of interest to see the yearly statistics of TTN and TTI. Table 2 shows the summary statistics of TTN and TTI. For TTN, we observe that the means in the first two years are smaller than those of other years, and the mean ranges from 122 days in 2012 to 223.19 days in 2017. The median of TTN shows an overall increasing trend. The standard deviations of TTN are very large for all years. This is further confirmed by the boxplot of TTN in Figure 3(a) where it shows that there exist large variabilities. The boxplot of TTN also indicates that the distribution of TTN is heavily skewed.

*https://www.itgovernanceusa.com/data-breach-notification-laws

†https://inewsource.org/2019/05/14/ucsd-data-breach-hiv-women-study/

Year	Min	Q_1	Median	Mean	SD	Q_3	Max
				TTN			
2012	6	26.00	54.50	122.00	231.21	130.50	1858
2013	6	27.25	69.50	122.42	162.45	145.25	890
2014	2	31.25	60.00	165.62	279.60	181.50	2316
2015	0	50.00	78.00	193.20	299.91	224.20	2747
2016	3	34.00	81.50	183.40	282.95	212.80	2228
2017	2	46.25	114.50	223.19	287.97	338.00	2510
2018	0	51.00	124.50	196.40	261.50	222.50	2064
2019	4	64.25	143.50	215.20	308.53	264.50	3222
2020	1	77.00	146.00	199.30	229.37	232.50	1806
				TTI			
2012	0	0	0	68.72	251.79	16.50	1858
2013	0	0	2	64.22	151.93	60.00	865
2014	0	0	5.5	84.93	166.42	85.75	918
2015	0	0.5	35.0	101.2	202.53	80.5	1360
2016	0	0	16	99.78	263.07	75.00	2196
2017	0	0	14	116.3	221.51	149.00	1422
2018	0	1	36	114.7	223.51	118.00	1570
2019	0	3	37.5	132.5	313.50	149.5	3140
2020	0	2	21	93.46	210.09	99.25	1689

Table 2: Yearly summary statistics of TTI (time to identification), TTN (time to notification), and ITN (identification to notification) where 'SD' stands for standard deviation, Q_1, Q_3 represent the first and third quantiles



(a) TTN (b) TTI

Figure 3: Boxplots of yearly TTN and TTI

For TTI, the means and medians do not show any clear patterns. There exist much more 0s in each year compared to that of TTN. The large variabities are also observed for TTI. The medians are much smaller than the means, and the boxplot of TTI in Figure 3(b) also shows that the distribution of TTI is heavily skewed.

To summarize, both TTN and TTI have considerable amounts of missing data. Particularly, the percentage of missing data of TTI is over 36%. Both TTN and TTI are heavily skewed and have large variabilities. This indicates that using the mean as the risk measure of TTN/TTI is unreliable. Those properties are taken into account in our modeling process.

CHAPTER III: PRELIMINARIES

In this section, we introduce some preliminaries to be used in our discussion.

1. Copula

Copula is an effective and popular tool for modeling high-dimensional dependence and has been widely employed in many areas especially in finance and insurance industry since it allows to model the margins individually (Joe and Harry, 2014). 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 (Joe and Harry, 2014) 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

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

The inverse of the Sklar's theorem also holds. We can express the copula which corresponds to multivariate distribution functions F with the marginal distribution functions F_i for $i = 1, ..., d$ as:

$$C(u_1, ..., u_d) = F(F_1^{-1}(u_i), ..., F_d^{-1}(u_d))$$

and its copula density function is defined by:

$$c(u_1, ..., u_d) = \frac{f(F_1^{-1}(u_i), ..., F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_i), ..., F_d^{-1}(u_d))}$$

To model the dependence between TTN and TTI, a bivariate extreme-value copula is used. De Haan and Resnick (1977) and Pickands (1979) laid the groundwork for probabilistic modeling of multivariate extremes, including asymptotic theory (1981). Coles et al. (2001), Beirlant et al. (2006), and McNeil et al. (2006) are three books that focus on statistical inference for extremevalue distributions (2015). Gudendorf and Segers looked at the connection between multivariate extreme-value theory and copulas (2010). A survey of employing copulas to model extremes is presented by Genest and Nelehová (2013). A bivariate copula C is called extreme value Copula if there exists a bivariate copula C_X such that for $n \to \infty$, we have:

$$[C_X(u_1^{1/n}, u_2^{1/n})]^n \to C(u_1, u_2) \forall (u_1, u_2) \in [0, 1]^2.$$

The copula C_X is said to be in copula C's domain of attraction.

From the literature review, there are many parametric families of bivariate extreme-value copulas. we recall the following two extreme-value copulas (Joe and Harry, 2014):

- Tawn type II copula.

$$C(u_1, u_2; \delta, \theta) = (1-\delta)u + \left[((1-u)\delta)^\theta + u^\theta\right]^{1/\theta},$$

where $u = \log(u_2)/\log(u_1 u_2)$, $\theta \geq 1$, and $0 \leq \delta \leq 1$ is the skewness parameter. It is known in the literature that the Tawn type II can model the right skewness of copula density.

- BB8 copula.

$$C(u_1, u_2; \delta, \theta) = 1/\delta \left[1 - \left(1 - \left(1-\bar{\delta}^\theta\right)^{-1} \left(1-(1-\delta u_1)^\theta\right)\left(1-(1-\delta u_2)^\theta\right)\right)^{1/\theta}\right],$$

where $\theta \geq 1$, and $0 \leq \bar{\delta} = 1 - \delta \leq 1$.

For a visual inspection, we use contours of the function $g(z_1, z_2)$. Contour plot is an easy way to assess departures from the Gaussian copula assumption. Normalized bivariate copula contour plots of a bivariate density is obtained from a copula density which is transformed to achieve standard normal margins. The Contour plot uses z-scale as inputs defined as:

- x-scale: original scale (X_1, X_2) with density $f(x_1, x_2)$
- u-scale: copula scale (U_1, U_2) where $U_i = F_i(X_i)$ and copula density $c(u_1, u_2)$
- z-scale: marginal normalized scale (Z_1, Z_2), where $Z_i = \Phi^{-1}(U_i) = \Phi^{-1}(F_i(X_i))$ for $i = 1, 2$ with density

$$g(z_1, z_2) = c(\Phi(z_1), \Phi(z_2))\phi(z_1)\phi(z_2)$$

where $\Phi(.)$ and $\phi(.)$ are the distribution and density function of a $N(0,1)$ variable.

Parameter estimation in bivariate Copula models can be estimated in both known (i.e up to marginal parameters) and unknown margin. If the margin is known, the copula scale can be obtained directly by using the probability integral transform:

$$(u_{i1}, u_{i2}) := (F_1(x_{i1}), F_2(x_{i2}))$$

for i = 1,...,n

When there are unknown margins, it is common to employ a two-step technique, first estimating the margins and then transforming to the copula scale by defining the pseudo-copula data using the estimated marginal distributions $\hat{F}_j$, j = 1, 2.

$$(u_{i1}, u_{i2}) := (\hat{F}_1(x_{i1}), \hat{F}_2(x_{i2}))$$

for i = 1,...,n and then formulate a copula model for the pseudo-copula data.

An inference for margins approach (IFM) is utilized when parametric marginal models are used, while a semiparametric method is employed when the empirical distribution is used. Joe (2005) explored the effectiveness of the IFM approach, while Genest et al. (1995) proposed the semiparametric approach.

2. ARMA, GARCH, DCC, and VAR models

ARMA (Auto Regressive Moving Average) and GARCH (Generalized Auto Regressive Conditional Heteroskedasticity) models are widely-used time series models (Jonathan et al., 2008).

An ARMA model is used to describe the weakly stationary stochastic time series in terms of autoregression and moving average.

Consider the pth order difference equation:

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + x_t$$

Let x_t be the MA(q) process given by:

$$x_t = \sum_{i=0}^{q} \beta_i \epsilon_{t-i}$$

Therefore, we can write:

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=0}^{q} \beta_i \epsilon_{t-i}$$

The ARMA(p,q) model has the general form:

$$Y_t = \mu + \sum_{k=1}^{p} \phi_k Y_{t-k} + \sum_{l=1}^{q} \theta_l \epsilon_{t-l} + \epsilon_t$$

where ϕ_k, θ_l are the parameters of AR and MA, and ϵ_t is the innovation of the time series.

ARMA models are applied to model a process's conditional expectation given the past data, however, the conditional variance in ARMA is constant. GARCH model allows us to model non constant variance.

Let ϵ_t be a white noise with unit variance. A GARCH process satisfies:

$$a_t = \sigma_t \epsilon_t$$

The conditional standard deviation of a_t with the past values a_{t-1}, a_{t-2},... is defined as:

$$\sigma_t = \sqrt{\omega + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2}$$

A general form of GARCH(p,q) is defined as:

$$\sigma_t^2 = w + \sum_{j=1}^{q} \alpha_j \epsilon_{t-j}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2,$$

where σ_t^2 is the conditional variance and w is the intercept.

The DCC (Dynamic Conditional Correlation) model introduced in Engle (2002) provides a very good approximation to a variety of time-varying correlation processes. Let $\mathbf{y}_t = (y_{1,t}, y_{2,t}, ..., y_{m,t})$ be a vector for m-dimensional time series at time t. A multivariate GARCH model can be defined as

$$\mathbf{y}_t = H_t^{1/2} \boldsymbol{\epsilon}_t$$

where H_t is an $m \times m$ conditional covariance matrix, and $\boldsymbol{\epsilon}_t$ is an $m \times 1$ vector of error with mean and variance given by $E(\boldsymbol{\epsilon}_t) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}_t) = I_m$ where I_m is an $m \times m$ identity matrix. The covariance matrix H_t can be decomposed into

$$H_t = D_t R_t D_t$$

where $D_t = \text{diag}(\sqrt{h_{11,t}}, ..., \sqrt{h_{mm,t}})$ is a diagonal of time varying standard deviations from a univariate GARCH(p, q) model, and R_t is a time varying positive definite conditional correlation matrix

$$R_t = \text{diag}(Q_t)^{-1/2} Q_t \text{diag}(Q_t)^{-1/2}$$

where

$$Q_t = (1 - a - b)\bar{Q} + a\mathbf{z}_{t-1}\mathbf{z}'_{t-1} + bQ_{t-1}$$

is a positive symmetric matrix, and $\bar{Q}$ is the unconditional matrix of the standardized errors $\mathbf{z}_t = D_t^{-1}\boldsymbol{\epsilon}_t$. The condition of $a + b < 1$ is imposed to ensure the stationarity and positive definiteness of Q_t. The DCC model consists of two steps: i) The first step is to estimate the univariate GARCH parameters; ii) The second step is to estimate the conditional correlation R_t. For more details on the DCC model, please refer to Engle (2002).

In the literature of time series, the VAR (Vector Auto Regressive) models are commonly used to investigate the dynamic interactions among multivariate time series (Tsay, 2005). They are also essential forecasting tools employed by the most of macroeconomic or policy-making institutions. A VAR(p) model can be represented as

$$\mathbf{y}_t = A_1\mathbf{y}_{t-1} + ... + A_p\mathbf{y}_{t-p} + \boldsymbol{\gamma}_t$$

where A_i is $m \times m$ coefficient matrix for $i = 1, 2, .., p$ and $\boldsymbol{\gamma}_t$ is a m-dimensional error process with zero mean and time-invariant positive definite covariance matrix.

3. Accuracy metrics

To evaluate the accuracy of the predictive distribution, we use the following two metrics: i) The first one is the most commonly used MAE (mean absolute error), which can be represented as

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|,$$

where y_i represents the observed values and $\hat{y}_i$ represents the predicted values, $i = 1, \ldots, m$. ii) The second is the CRPS (continuous ranked probability score). The dissimilarity of probability forecast and an outcome is measure by loss function (scoring rule). Scoring rules are used to evaluate the validity of probabilistic forecasts by providing a numerical score to the forecast and the event or value that occurs. When the forecaster gives the probabilistic forecast F, rather than $G \neq F$, it maximizes the anticipated score for an observation selected from the distribution F. If the maximum is unique, it is strictly proper. Proper scoring standards encourage the forecaster to make meticulous assessments and be accurate in prediction issues. Strictly appropriate scoring systems give appealing loss and utility functions in estimating issues that can be adjusted to the scientific subject at hand. For continuous outcome, a popular scoring rule is the continuous ranked probability score which is defined as

$$\text{CRPS}(F, s) = \int_{\mathbb{R}}(F(y) - \mathbb{1}\{s \leq y\})^2 dy,$$

where $F(\cdot)$ denotes the predictive cumulative distribution function (CDF) and $\mathbb{1}\{\cdot\}$ denotes the indicator function. The CRPS is a quadratic measures of the difference between the predicted CDF and the empirical CDF of observed values, and is a widely-used accuracy

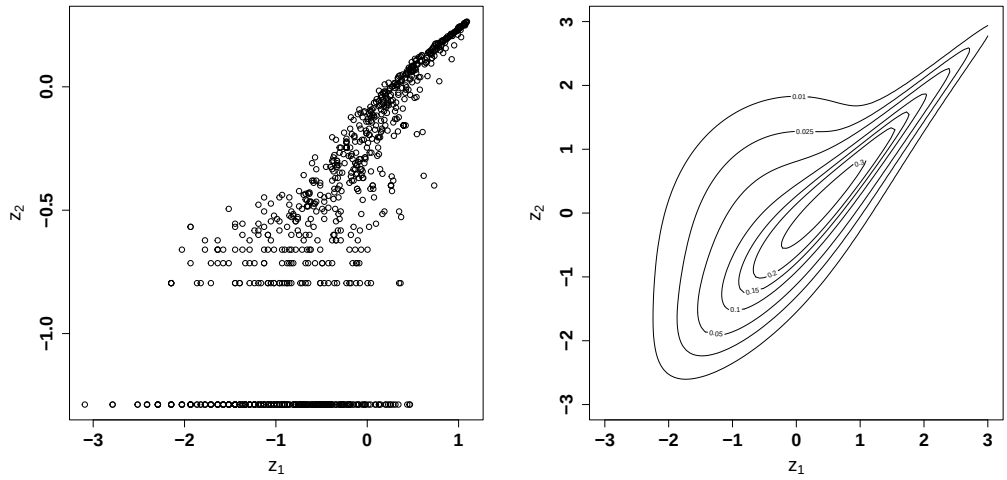
measure dealing with probability forecasts (Epstein, 1969). A smaller score indicates a better prediction.

CHAPTER IV: COPULA APPROACH TO MISSING DATA IMPUTATION

Let $x_{1,t}$s and $x_{2,t}$s be the observed time series values of TTN and TTI, respectively, $t = 1, \ldots, T$. Since both TTN and TTI have missing data, we use $x^*_{1,t}$s and $x^*_{2,t}$s to represent the missing observations for TTN and TTI, respectively. For modeling purpose, we use the time period from January 20, 2012 to December 31, 2018 as the in-sample data (i.e., 1505 pairs of observations with 596 NAs), while use the data from 2019 to 2020 as the out-of-sample data (i.e., 618 pairs of observations with 282 NAs). We propose a novel copula approach to imputing the missing data in Algorithm 1 for the in-sample data. In the following, we briefly describe the procedure to impute the missing data for TTN and TTI.

a) *Model the dependence based on the complete pairs of TTNs and TTIs.* In our study, there are 910 completely observed pairs of TTNs and TTIs among the in-sample data. To model the dependence between TTN and TTI, the empirical marginals are used. We select the best copula structure from various bivariate copula families in the VineCopula package by the AIC criterion (Schepsmeier et al, 2015). It is found that the Tawn type 2 copula is selected for modeling the joint dependence, and the estimated parameters are $\theta = 3.93$, $\delta = 0.67$, and $\tau = 0.54$. In Figure 8, we display the normal score plot and the fitted contour plot. We observe that there exist the strong right tail dependence among TTN and TTI which fits the fact that the longer TTI is, the longer TTN.

b) *Impute missing data for both TTNs and TTIs.* There are 203 completely missing pairs of TTN and TTI because of unknown breach dates. Based on the estimated copula structure, we simulate $N = 5000$ observations in Algorithm 1, where the symbol * represents the missing value. The missing values are imputed by using the means of



(a) Normal score (b) Fitted contour

Figure 4: Normal score plot and fitted contour plot of TTI and TTN. Unit: days

simulated observations

c) *Impute missing data for TTIs with observed TTNs.* Since there are 392 records with TTNs only because the identification dates are unreported, we impute the missing TTIs by using the conditional copula approach, i.e., sampling from the conditional Tawn type 2 copula with given TTNs (line 9 in Algorithm 1).

For comparison purpose, we also use the following two commonly used approaches for missing data imputation.

- Kalman Smoothing (KS) imputation and State Space model. The KS imputation and State Space model is the commonly used algorithm and often produces the best performance in the literature (Grewal et al., 2014). The state space model is specified as structural time series model fitted by maximum likelihood. The imputation process are performed by the following steps: recognizing trend and seasonality in structural state space model, specifying the state space form, estimating the unobserved state

Algorithm 1 Copula approach for imputing the missing data.

INPUT: Complete pairs $\{(x_{1,t}, x_{2,t})|t = i_1, \ldots, i_{T_1}\}$; Incomplete pairs $\{(x_{1,t}, x^*_{2,t})|t = i_{T_1+1}, \ldots, i_{T_2}\}$; Missing pairs $\{(x^*_{1,t}, x^*_{2,t})|t = i_{T_2+1}, \ldots, i_{T_3}\}$; $N = 5000$.

1: Estimate the empirical marginals $(u_{1,t}, u_{2,t})$ based on $(x_{1,t}, x_{2,t})$, $t = i_1, \ldots, i_{T_1}$;
2: Select the preferred copula structure C via AIC criterion based on the empirical marginals;
3: **for** $j = 1, \ldots, N$ **do**
4: Simulate $T_3 - T_2$ pairs from copula C;
5: Covert the simulated observations into marginal values $\{(x^*_{1,t,j}, x^*_{2,t,j})\}$ via the sample quantile approach in (Hyndman et al., 1996);
6: **end for**
7: $x_{1,t} \leftarrow \sum_{j=1}^{N} x^*_{1,t,j}/N$, $x_{2,t} \leftarrow \sum_{j=1}^{N} x^*_{2,t,j}/N$, $t = i_{T_2+1}, \ldots, i_{T_3}$;
8: **for** $t = T_1 + 1, \ldots, T_2$ **do**
9: Simulate N samples of $u_{2,t}$'s given each $u_{1,t}$ from copula C;
10: Convert to the original value $x_{2,t}$ based on the mean of $u_{2,t}$'s using the sample quantile approach in (Hyndman et al., 1996);
11: **end for**
12: **return** $\{(t, x_{1,t}, x_{2,t})|t = 1, \ldots, T_3\}$;

OUTPUT: Imputed data $\{(t, x_{1,t}, x_{2,t})|t = 1, \ldots, T_3\}$.

equations parameters by Kalman Smoothing, and calculating the missing values (cite). These steps are implemented by the imputeTS package in R which is developed by Steffen Moritz.

For TTN and TTI, we impute the missing values on the training data by using the KS imputation. Since TTI must be no larger than TTN, we replace TTI by TTN if the imputed TTI is larger than TTN (a total of 130 imputed observations).

- Multivariate Imputation by Chained Equations (MICE) imputation. The second approach that we impose on the missing values is the MICE imputation. This method is emerged as one of the principled method to address missing values in statistical literature (Van et al., 2011). The chained equation process can be broken down into four general steps:

Step 1: Every missing value in the dataset receives a simple imputation, such as imputing the mean. These imputations are sometimes referred to as "place holders."

Step 2: Imputations for one variable ("var") are reset to missing.

Step 3: Involve regressing the observed values from the variable "var" on the other variables in the imputation model, which may or may not include all of the variables in the dataset. In other words, in a regression model, "var" is the dependent variable, whereas all the other variables are independent variables. Outside of the context of imputing missing data, these regression models function under the same assumptions that one would make when doing linear, logistic, or Poison regression models.

Step 4: The regression model's predictions are used to fill in the missing values for "var." Both the observed and imputed values will be used when "var" is used as an independent variable in regression models for other variables. Step 5: Steps 2–4 are then repeated for each missing data variable. One iteration or "cycle" is the process of cycling through each of the variables. At the end of one cycle, all of the missing values were replaced with regression predictions that reflected the data's relations. Step 6: Steps 2–4 are done a number of times, with the imputations updated each time.

We used the MICE package in R to perform the imputation. Similar to the KS method, we impute the missing values by the MICE on the training data. We replace TTI by TTN when the imputed value of TTI is greater than TTN (a total of 8 imputed observations).

Table 3 shows the summary statistics of imputed TTNs and TTIs for the proposed copula, KS, and MICE approaches. For TTN, the means are very close for different ap-

	Min	Q_1	Median	Mean	SD	Q_3	Max
TTN-copula	0.00	44.00	113.0	182.7	251.32	190.0	2747
TTN-KS	0.00	44.00	106.0	180.4	252.08	208.0	2747
TTN-MICE	0.00	39.00	87.0	181.2	262.14	213.0	2747
TTI-copula	0.00	1.00	41.00	98.43	200.10	99.19	2196
TTI-KS	0.00	3.00	51.00	89.46	171.53	104.00	2196
TTI-MICE	0.00	0.00	32.00	123.5	236.39	143.0	2196

Table 3: Statistics of imputed TTI (time to identification) and TTN (time to notification) based on different approaches where 'SD' stands for standard deviation, Q_1, Q_3 represent the first and third quantiles

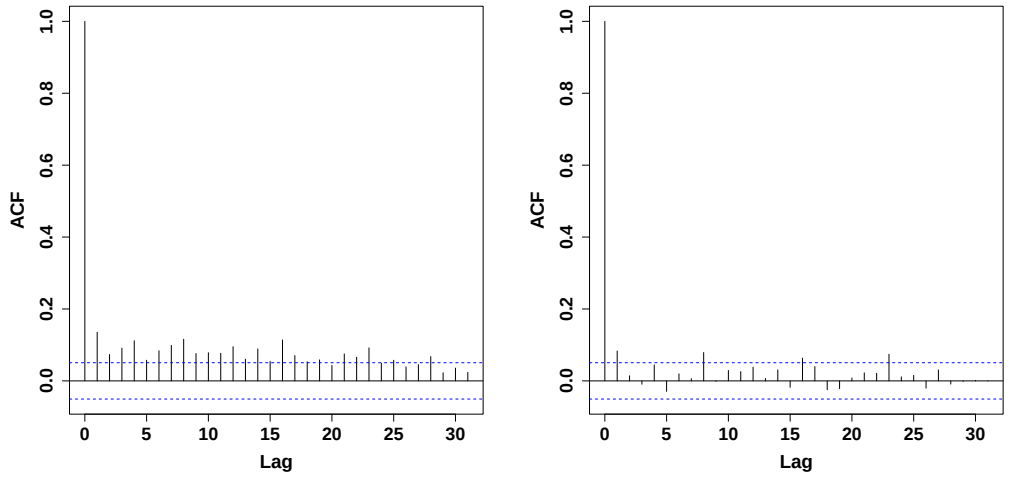
proaches while the medians are different. The MICE imputation has the smallest median 87 while the copula approach has the largest 113. In terms of standard deviation, the copula approach has the smallest value (251.32) while the MICE approach has the largest value (262.14). For TTI, the means are different, and the smallest mean is 89.46 by the KS imputation while the largest one is 123.5 by the MICE imputation. The MICE approach has the smallest median 32 while the KS approach has the largest 51. For the standard deviation, the KS approach has the smallest value 171.53 while the MICE approach has the largest value 236.39. It is interesting to note that the copula approach leads to the smallest Q_3s for both TTN and TTI. Compared with KS and MICE imputation approaches, the proposed copula approach is simple but efficient. We further note that all the imputed TTNs are no less than the corresponding TTIs by the copula approach. Therefore, there is no need to replace any observations.

Since both TTNs and TTIs are skewed and have large variabilities, the transformations are performed. For TTNs, since there are only two 0s by the copula imputation, we replace the 0s by two random values from a uniform $(0, 1)$ distribution and perform the log

transformation. For TTIs, since we have a large portion of 0s, we perform the square root transform to reduce the variability.

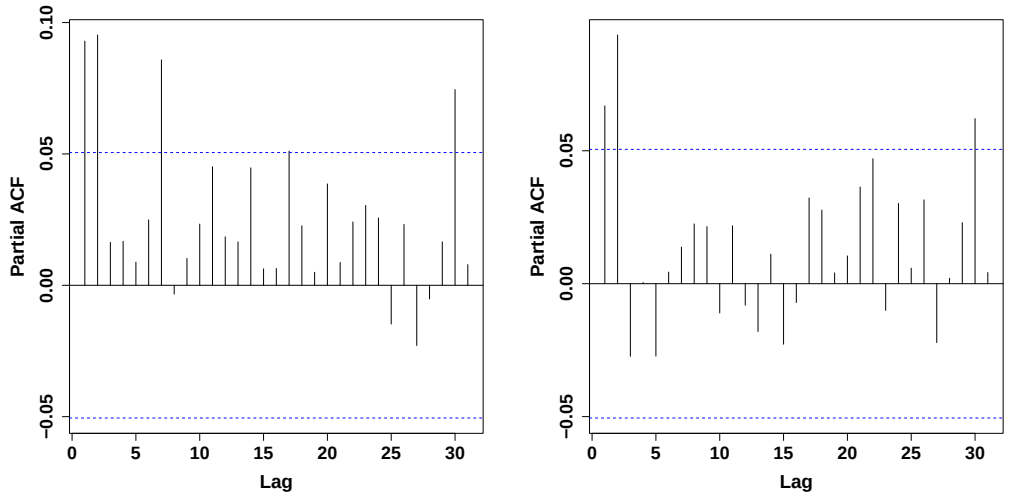
CHAPTER V: STATISTICAL MODELING

In this section, we develop a copula approach to jointly modeling the dynamics of TTN and TTI. After imputing the missing values, both transformed TTN and TTI exhibit the temporal correlations as shown by their ACFs and PACFs in Figure 5 and Figure 6.



(a) TTN (b) TTI

Figure 5: ACFs of transformed TTI and TTN. Unit: days



(a) TTN (b) TTI

Figure 6: PACFs of transformed TTI and TTN. Unit: days

In the following, we discuss how to capture the temporal and cross-sectional depen-

dence of TTN and TTI.

1. Model fitting

From Section Data Exploratory, it is seen that there exist large variabilities for both TTN and TTI. Therefore, we propose to use a GARCH model to model the volatilities for both TTN and TTI. The analysis on the residuals suggests that GARCH$(1,1)$ is sufficient to describe the volatilities in the residuals of both series. This in fact coincides with the conclusion drawn in the literature that higher-order GARCH models are not necessarily better than GARCH$(1,1)$ (Hansen et al., 2015). Therefore, we fix the GARCH part as GARCH$(1,1)$. To model the evolution of the means of TTN and TTI, we propose using the ARMA(p,q) process. This leads to the following ARMA+GARCH model

$$Y_t = \mu + \sum_{k=1}^{p} \phi_k Y_{t-k} + \sum_{l=1}^{q} \theta_l \epsilon_{t-l} + \epsilon_t, \tag{1}$$

where $\epsilon_t = \sigma_t Z_t$ with Z_t being the i.i.d. innovations, and the ϕ_k's and the θ_l's are respectively the coefficients of the AR and MA parts. For the standard GARCH$(1,1)$ model, we have

$$\sigma_t^2 = w + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{2}$$

where σ_t^2 is the conditional variance and w is the intercept. For model selection, we use the AIC criterion to determine the orders of the ARMA models. Note that if ARMA(p,q)+GARCH can successfully accommodate the serial correlations in the conditional mean and the conditional variance, there would be no autocorrelations left in the standardized and squared

standardized residuals. When the AIC criterion suggests to select multiple models with similar AIC values, we select the simpler model. The autoregressive p and the moving average order q are allowed to vary between 0 and 5. We find that ARMA$(1,1)$+GARCH$(1,1)$ with normal innovations is sufficient to remove the serial correlations for both TTN and TTI. Based on the Ljung-Box test (Brockwell et al., 2016), for TTN, the p-values of standardized and squared standardized residuals are 0.186 and 0.613 , respectively; for TTI, the p-values are 0.524 and 0.184, respectively.

Let $\mathbf{Z}_t = (Z_{1,t}, Z_{2,t})$ be the vector of standardized residuals of fitted models for TTN and TTI. Further, we assume that $\mathbf{Z}_t = (Z_{1,t}, Z_{2,t})$ has the following distribution

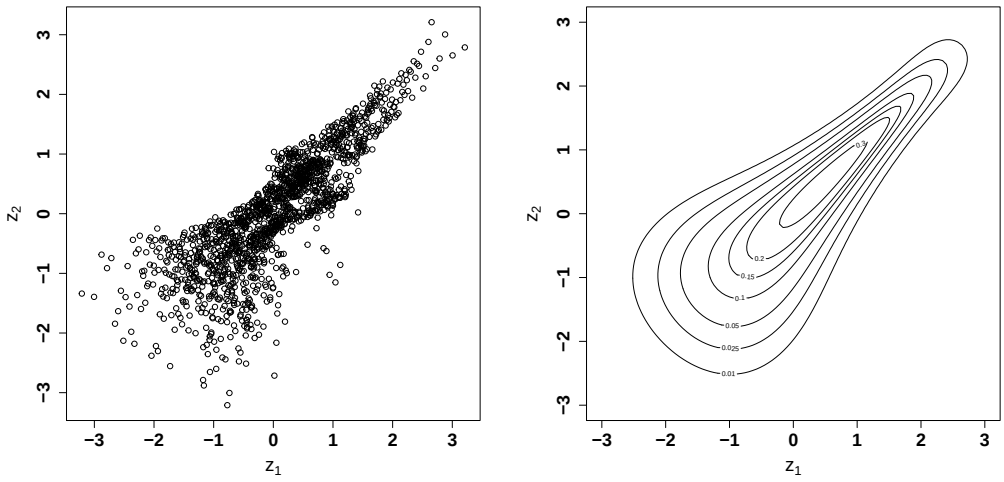
$$F\left(\mathbf{z}_t\right) = C\left(F(z_{1,t}), G(z_{2,t})\right), \tag{3}$$

where F is the marginal distribution of the residual of TTN, and G is the marginal distribution of the residual of TTI. The joint log-likelihood function of the model can be written as

$$L = \sum_{t=1}^{n} \left[\log c\left(F\left(z_{1,t}\right), G\left(z_{2,t}\right)\right) - \log(\sigma_{1,t}) - \log(\sigma_{2,t}) + \log\left(f\left(z_{1,t}\right)\right) + \log\left(g\left(z_{2,t}\right)\right)\right],$$

where $c(\cdot)$ is the copula density of $C(\cdot)$; $\sigma_{1,t}$ and $\sigma_{2,t}$ are the conditional standard deviations of TTN and TTI, respectively. $f(\cdot)$ is the density function of $Z_{1,t}$, and $g(\cdot)$ is the density function of $Z_{2,t}$. A popular method for estimating the parameters of a joint model is the Inference Function of Margins method (Joe, 1997), which is employed in our study. This method has two steps: (i) estimate the parameters of the marginal models; and (ii) estimate

the parameters of the copula by fixing the parameters obtained at step (i). Since we have identified the time series models for TTNs and TTIs, we discuss how to model the bivariate dependence in the following. Note that although we assume the normal innovations for the marginal processes to remove serial correlations, $Z_{1,t}$s and $Z_{2,t}$s are not normally distributed due to the high skewness and an excessive number of 0s. Since it is very challenging to fit parametric distributions to the marginals, we propose using the empirical marginals in Eq. (3), and then select the copula structure by using the AIC criterion. The BB8 copula is preferred to model the dependence between the standardized residuals, and the corresponding estimated parameters are $\theta = 4.54$, $\delta = 0.98$, and $\tau = 0.64$.



(a) Normal score (b) Contour

Figure 7: Normal score plot and contour plot of residuals of TTI and TTN

The normal score plot and fitted contour plot are displayed in Figure 7. It is seen that the upper tail dependence is well captured by the BB8 copula.

2. Prediction evaluation

We use Algorithm 2 to perform the rolling window prediction for the TTI and TTN. The parsimonious ARMA(1,1)+GARCH(1,1) model is applied to the sample data with window size $w = 500$. The window size is selected based on the fact that too few observations can lead to the large variability in the model and hence have a poor prediction performance, while too many observations not only cost more computational effort but also not necessarily improve the prediction performance because of the potential structure break and trends. In the rolling process, the dependence structure is allowed to vary with time. That is, the copula is re-selected during the fitting process via the criterion of AIC (see line 4 of Algorithm 2). Since the size of out-of-sample data is 618, we have $T = 1505$ and $S = 2123$. The predictive distributions of TTN and TTI are simulated based on $N = 5000$ samples. If the observed value is missing, we use the predicted mean to replace the missing value in the TTN/TTI to perform the rolling prediction. The evaluation metrics such as MAE and CRPS are computed by excluding the missing data in the out-of-sample data.

Imputation comparison. We study the predictive performance of proposed copula imputation approach. For this purpose, Algorithm 2 is applied to the KS and MICE imputed data. The predictive results are reported in Table 4. It is seen that for TTN, the predictive performances are similar for all three imputation approaches in terms of the mean of CRPSs and MAE. For TTI, the copula imputation approach leads to a slightly smaller mean of CRPSs, but the KS imputation approach has the smallest MAE. We also compute the percentages that the CRPSs of the copula approach are less than that of MICE and KS, respectively. We observe that the copula approach outperforms both of the MICE and

Algorithm 2 Algorithm for predicting the distributions of TTN and TTI.

Input: Imputed in-sample data $\{(t, x_{1,t}, x_{2,t})|t = 1, \ldots, T\}$; out-of-sample data $\{(t, x_{1,t}, x_{2,t})|t = T+1, \ldots, S\}$; window size $w = 500$; $N = 5000$.

for $i = T, \cdots, S-1$ **do**
  Estimate the ARMA-GARCH model based on the log-transformed TTNs with time window $[i-w+1, i]$;
  Estimate the ARMA-GARCH model based on the square root transformed TTIs with time window $[i-w+1, i]$;
  Convert the standardized residuals $z_{1,t}$s and $z_{2,t}$s to the empirical marginals;
  Select a preferred copula based on the empirical marginals via the AIC criterion;
  Based on the estimated copula, simulate N 2-dimensional copula samples $\left(u_{1,i}^{(k)}, u_{2,i}^{(k)}\right)$, $k = 1, \ldots, N$;
  For the TTN, convert the simulated dependent samples $u_{1,i}^{(k)}$s into standardized residuals $z_{1,i}^{(k)}$s via the sample quantile approach in (Hyndman et al., 1996) $k = 1, \ldots, N$;
  For the TTI, convert the simulated dependent samples $u_{2,i}^{(k)}$s into standardized residual $z_{2,i}^{(k)}$s via the sample quantile approach in (Hyndman et al., 1996), $k = 1, \ldots, N$;
  Compute the predicted N 2-dimensional $x_{1,i+1}^{(k)}$s and $x_{2,i+1}^{(k)}$s, respectively, $k = 1, \ldots, N$;
  if $x_{1,i+1}$ ($x_{2,i+1}$) is missing **then**
    $x_{1,i+1}$ ($x_{2,i+1}$) is imputed as the mean of $x_{1,i+1}^{(k)}$s ($x_{2,i+1}^{(k)}$s);
  end if
end for

Output: Predictive distributions of TTN and TTI.

KS imputation approaches in terms of CRPSs. In particular, for TTI, the copula approach improves the MICE approach by 22.71% and KS approach by 11.93%.

	CRPS-Mean	MAE	CRPS-Mean	MAE
	TTN		TTI	
Copula	100.555	141.230	84.896	130.308
MICE	100.581	141.359	85.300	134.282
KS	100.478	141.493	85.998	129.997
	Percentage			
Copula vs MICE	52.24%		72.71%	
Copula vs KS	55.86%		61.93%	

Table 4: Means of CRPSs and MAEs of the proposed copula and other imputation approaches

Therefore, the proposed copula imputation approach is preferred and used in the following discussion.

Model comparison. We compare the predictive performance of proposed model to those commonly used models of DCC and VAR. For the fair comparison, modified Algorithms 2 are applied for the DCC and VAR models. Specifically, for the DCC model, the marginals of TTN and TTI are still fitted by using ARMA(1,1)+GARCH(1,1) on the sliding window, but the DCC is used for modeling the correlation. We also simulate $N = 5000$ values for each prediction for the DCC model. For the VAR model, the order p is selected from 1 to 5 by using the AIC criterion for each sliding window, and $N = 5000$ predicted values are simulated from the selected VAR model for each prediction.

The predictive results are reported in Table 5. It is seen that for TTN, the predictive performances are comparable based on MAE and mean CRPS. However, in terms of individual CRPS, DCC is slighter better than the proposed copula approach as it improves 4.31%.

	CRPS-Mean	MAE	CRPS-Mean	MAE
	TTN		TTI	
Copula	100.555	141.230	84.896	130.308
DCC	101.022	144.146	86.422	129.550
VAR	101.233	143.234	108.532	100.465
	Percentage			
Copula vs DCC	45.69%		72.94%	
Copula vs VAR	57.24%		71.10%	

Table 5: Mean CRPSs and MAEs of the proposed copula and other model(s)

For TTI, it is seen that the VAR model has the smallest MAE. But the proposed copula approach significantly outperforms other approaches in terms of individual CRPS, namely improving 22.94% compared to DCC and 21.10% compared to VAR.

To further assess the prediction accuracy, we further use the Value-at-Risk (VaR) (McNeil et. al, 2015) metric since it is directly related to the high quantiles of interest. Recall that for a random variable X_t, the VaR at level α for some $0 < \alpha < 1$ is defined as $\text{VaR}_\alpha(t) = \inf\{l : P(X_t \leq l) \geq \alpha\}$. For example, $\text{VaR}_{.95}(t)$ means that there is only a 5% probability that the observed value is greater than the predicted value $\text{VaR}_{.95}(t)$. An observed value that is greater than the predicted $\text{VaR}_\alpha(t)$ is called a *violation*. In order to evaluate the prediction accuracy of the VaR values, we use the following three popular tests (Christoffersen, 1998): (i) the unconditional coverage test, denoted by LR_{uc}, which evaluates whether or not the fraction of violations is significantly different from the model's violations; (ii) the conditional coverage test, denoted by LR_{cc}, which is a joint likelihood ratio test for the independence of violations and unconditional coverage; and (iii) the dynamic quantile test (DQ), which is based on the sequence of 'hit' variables (Engle, 2014).

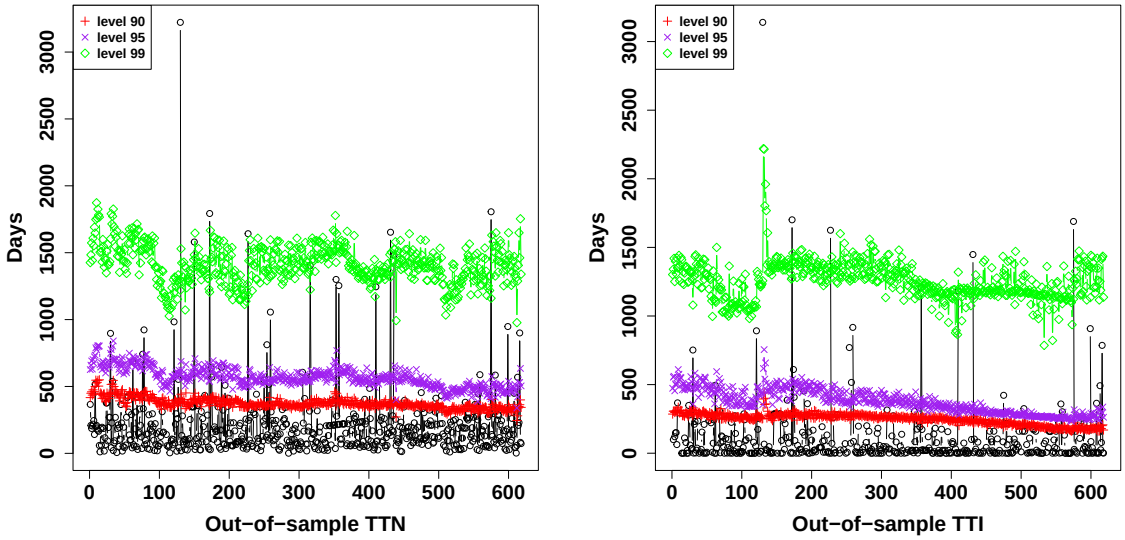
Table 6 shows the p-values of VaR tests at different levels of $\alpha = .9, .95, .99$. It

	Exp.	Ob.	LRuc	LRcc	DQ	Exp.	Ob.	LRuc	LRcc	DQ
	TTN					TTI				
	$\text{VaR}_{.9}$									
Copula	58	54	.576	.484	.999	43.6	46	.704	.795	.999
VAR	58	39	.005	.019	.410	43.6	147	0	0	0
DCC	58	39	.005	.019	.394	43.6	31	.035	.107	.758
	$\text{VaR}_{.95}$									
Copula	29	26	.561	.249	.995	21.8	22	.965	.993	.999
VAR	29	22	.164	.159	.952	21.8	139	0	0	0
DCC	29	21	.110	.126	.891	21.8	21	.860	.984	.999
	$\text{VaR}_{.99}$									
Copula	5.8	7	.001	.003	.568	4.36	5	.003	.013	.743
VAR	5.8	7	.001	.003	.566	4.36	127	0	0	0
DCC	5.8	11	.001	.003	.570	4.36	13	0	0	0

Table 6: The p-values of the VaR tests of the predicted violations for $\alpha = .9, .95, .99$

is seen that at $\alpha = .9$ and $\alpha = .95$ levels, the copula approach predicts very well. For TTN, we observe that the numbers of expected violations are very close to the numbers of observed violations. The p-values are all very large for those three tests. Compared to the proposed copula approach, the discrepancies between the numbers of expected violations and the number of observed violations are large for the VAR and the DCC models. At level $\alpha = .9$, the p-values of LRuc and LRcc are also small for the VAR and the DCC models. Similarly, for TTI, the proposed copula approach significantly outperforms the other approaches. Particularly, we observe that the VAR has the worst prediction performance. At $\alpha = .99$ level, we observe that the numbers of expected violation and the numbers of observed violation are close based on the copula approach for both TTN and TTI. However, the p-values seem to be small for both LRuc and LRcc because of the small sample size. But the proposed copula approach still outperforms the other approaches. Figure 8 shows the VaR plots of TTN and TTI, and we observe that the proposed copula approach predicts the

VaRs well.



(a) TTN (b) TTI

Figure 8: VaR plots of TTN and TTI

To summarize, the proposed copula model outperforms the DCC and VAR models in terms of CRPS and VaR overall.

CHAPTER VI: CONCLUSION AND DISCUSSION

In this paper, we develop a statistical model for capturing the dependence between two important metrics related to the data breach risk, i.e., TTN and TTI. To tackle the missing data, we propose a novel copula approach. Our study shows that the developed model is superior to other commonly used multivariate imputation approaches and multivariate time series models such as DCC and VAR. From our study, it is discovered that both TTN and TTI have large variabilities, and therefore, using the mean of TTN or TTI from the public source (e.g. 2021 Cost of a Data Breach Report by IBM) as the measurement for determining the relevant cost is unreliable. We recommend to use the VaR as the measurement to reflect the risk. Taking the $\text{VaR}_{.95}$ as a representative example, according to the proposed copula model, 95% of TTNs are less than 619.5 days with a standard error 70.769 days in 2019, and it reduces to 534.9 days with a standard error 57.06 days in 2020. For TTI, 95% of TTIs are less than 456.9 days with a standard error 64.3222 days in 2019, and it reduces to 323.7 days with a standard error 55.307 days in 2020. Therefore, the mean VaR is more suitable for measuring the risk. We also find that there exists a large variability in the differences of TTNs and TTIs. We urge the authority in CA to require that the notification must be made within a certain period (e.g., 60 days). This can efficiently reduce the unnecessary delay from the identification to the notification.

Although the proposed model is based on the CA breach data, the developed approach can be equally applied to similar breach data. Further, the proposed model can help insurers to estimate the TTIs and TTNs. In the following, we briefly mention how to use the developed approach in the risk assessment from the insurer perspective:

- Pricing factor. Assume that an insurance company offers a cyber insurance policy covering the cost related to TTN/TTI (e.g., notification expense, regulatory fines and penalties, and forensic expenses, etc). For the pricing formula, the TTN/TTI should be taken into account since it is directly related to the loss. The proposed model can be used to predict the quantities of interest such as high quantiles of TTN/TTI. Those quantities can be used as a pricing factor to adjust the pricing formula.

- Individual incident. It is common in practice that when a cyber incident is identified, the breach date is unknown or needs much time or effort to investigate. The proposed model can be used to estimate/predict the missing/unknown TTN/TTI for the cost estimation.

The current study also has some limitations like other studies. First, the proposed approach is based on the CA breach reports, it is possible that the data from other states or countries may exhibit different patterns. Therefore, the current model should be cautiously used when the different pattern appears. Second, the covariates are not incorporated in our modeling process. In the future study, the text mining approach can be used to extract the key information as the covariates. Third, the severity related to the TTN/TTI is of interest. However, due to the limited loss data, this study will be pursued when more data is available.

REFERENCES

[BiAV] Bisogni, F., Asghari, H., & Van Eeten, M. J. (2016). Estimating the size of the iceberg from its tip. Regulation (EU), 2016 (679).

[BrD] Brockwell, P. J., & Davis, R. A. (Eds.). (2002). Introduction to time series and forecasting. New York, NY: Springer New York.

[BBHW] Buckman, J., Bockstedt, J., Hashim, M. J., & Woutersen, T. (2017). Do organizations learn from a data breach. In Workshop on the Economics of Information Security (WEIS) (pp. 1-22).

[Ch] Christoffersen, P. F. (1998). Evaluating interval forecasts. International economic review, 841-862.

[CrC] Cryer, J. D., & Chan, K. S. (2008). Time series analysis: with applications in R (Vol. 2). New York: Springer.

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

[El] Eling, M. (2020). Cyber risk research in business and actuarial science. European Actuarial Journal, 10(2), 303-333.

[ElL] Eling, M., Loperfido, N. (2017). Data breaches: Goodness of fit, pricing, and risk measurement. Insurance: mathematics and economics, 75, 126-136.

[En] Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics, 20(3), 339-350.

[EnM] Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. Journal of business economic statistics, 22(4), 367-381.

[Ep] Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology (1962-1982), 8(6), 985-987.

[GnR] Gneiting, T., Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association, 102(477), 359-378.

[GrA] Grewal, M. S., Andrews, A. P. (2014). Kalman filtering: Theory and Practice with MATLAB. John Wiley Sons.

[HaL] Hansen, P. R., Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?. Journal of applied econometrics, 20(7), 873-889.

[HuX] Hua, L., Xu, M. (2020). Pricing cyber insurance for a large-scale network. arXiv preprint arXiv:2007.00454.

[HTFSS] Huyghues-Beaufond, N., Tindemans, S., Falugi, P., Sun, M., Strbac, G. (2020). Robust and automatic data cleansing method for short-term load forecasting of distribution feeders. Applied Energy, 261, 114405.

[HyF] Hyndman, R. J., Fan, Y. (1996). Sample quantiles in statistical packages. The American Statistician, 50(4), 361-365.

[HyF] Hyndman, R. J., Fan, Y. (1996). Sample quantiles in statistical packages. The American Statistician, 50(4), 361-365.

[IkK] Ikegami, K., Kikuchi, H. (2020, July). Modeling the risk of data breach incidents at the firm level. In International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (pp. 135-148). Springer, Cham.

[Iy] Iyengar, S. (1998). Multivariate models and dependence concepts.

[Jo] Joe, H. (2014). Dependence modeling with copulas. CRC press.

[Ar] Arrow, K. J. (1951). Alternative approaches to the theory of choice in risk-taking situations. Econometrica: Journal of the Econometric Society, 404-437.

[MFE] McNeil, A. J., Frey, R., Embrechts, P. (2015). Quantitative risk management: concepts, techniques and tools-revised edition. Princeton university press.

[MoB] Moritz, S., Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. R J., 9(1), 207.

[RTA] Romanosky, S., Telang, R., Acquisti, A. (2011). Do data breach disclosure laws reduce identity theft?. Journal of Policy Analysis and Management, 30(2), 256-286.

[SSBGNEJ] Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, T., Erhardt, T., ... Joe, H. (2015). Package 'vinecopula'. R package version, 2(5).

[SXZ] Sun, H., Xu, M., Zhao, P. (2021). Modeling malicious hacking data breach risks. North American Actuarial Journal, 25(4), 484-502.

[Ts] Tsay, R. S. (2005). Analysis of financial time series. John wiley sons.

[VaG] Van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45, 1-67.

[XuZ] Xu, M., Zhang, Y. (2021). Data Breach CAT Bonds: Modeling and Pricing. North American Actuarial Journal, 25(4), 543-561.

[ZeS] Zeller, G., Scherer, M. (2021). A comprehensive model for cyber risk based on marked point processes and its application to insurance. European Actuarial Journal, 1-53.