Communications for Statistical Applications and Methods
2020, Vol. 27, No. 4, 487–500

https://doi.org/10.29220/CSAM.2020.27.4.487
Print ISSN 2287-7843 / Online ISSN 2383-4757

Volatility clustering in data breach counts

Hyunoo Shim[a], Changki Kim[b], Yang Ho Choi[1,a]

[a]Department of Actuarial Science, Hanyang University, Korea;
[b]Korea University Business School, Korea University, Korea

Abstract

Insurers face increasing demands for cyber liability; entailed in part by a variety of new forms of risk of data breaches. As data breach occurrences develop, our understanding of the volatility in data breach counts has also become important as well as its expected occurrences. Volatility clustering, the tendency of large changes in a random variable to cluster together in time, are frequently observed in many financial asset prices, asset returns, and it is questioned whether the volatility of data breach occurrences are also clustered in time. We now present volatility analysis based on INGARCH models, i.e., integer-valued generalized autoregressive conditional heteroskedasticity time series model for frequency counts due to data breaches. Using the INGARCH(1, 1) model with data breach samples, we show evidence of temporal volatility clustering for data breaches. In addition, we present that the firms' volatilities are correlated between some they belong to and that such a clustering effect remains even after excluding the effect of financial covariates such as the VIX and the stock return of S&P500 that have their own volatility clustering.

Keywords: data breach, cyber risk, volatility clustering, INGARCH, covariate

1. Introduction

Personal and company information continue to grow as cyber activities expand. However, there has also been an increase in the uncontrolled and unjustified transfer of information though data breach accidents. With increasing reliance of the operations of firms on the economies occurring in cyberspace, it is widely regarded that the risk of data breach as well as other types of cyber risks would have also recently expanded in number and size (Pooser *et al.*, 2018). An increased access to private data during economic activities, or the absence of physical blockades can potentially elevate the occurrence frequency of data breach.

Data breach insurance (cyber risk insurability has been analyzed by Biener and Eling (2012), which is based on insurability criteria, Berliner (1982)), a type of cyber insurance, is an insurance that compensates for losses and expenses occurred during an information leakage accident. Data breach insurance is designed to protect insureds against direct losses from information leakage as well as indirect losses and liability for secondary damage and even losses from damage to a firm's reputation. The two fundamentals of actuarial analyses for insurance are ratemaking and estimating risk capital. While there have been several researches on ratemaking data breach risk (Eling and Loperfido, 2017), estimating solvency capital, which requires understanding the volatility of data breach risk, has yet to be fully addressed to the best of our knowledge.

[1] Corresponding author: Department of Actuarial Science, Hanyang University-ERICA Campus, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan, Gyeonggi-do 15588, Republic of Korea. E-mail: ychoi@hanyang.ac.kr

Published 31 July 2020 / journal homepage: http://csam.or.kr

We first need to find if the volatility of data breach risk is clustered in time or not to decide if the static level of solvency capital for insuring data breach risk is adequate. Volatility clustering was empirically found by Mandelbrot (1963) and he noted that "... large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes, ..." This empirical finding of volatility clustering in financial data have been mainly modeled by stochastic volatility models or ARCH-type models. Among others, ARCH models (Engle, 1982) and its generalized model - GARCH models (Bollerslev, 1986; Lee and Hwang, 2018) are commonly adopted in modeling financial time series data in econometrics. INGARCH model which was proposed by (Ferland *et al.*, 2006) describes an integer-valued time series by the GARCH with count distributions (a Poisson or negative binomial distribution) as residual distributions. Our study finds volatility clustering in data breach count time series in the INGARCH model framework.

Correlations might also exist between cyber risks (Bohme and Kataria, 2006). Accordingly, the correlations between cyber risks might generate a bias in estimating volatility clustering effects; therefore, caution needs to be taken in analyzing volatility clustering due to the possibility of correlation structure presence. If we classify cyber risks into non-epidemic risks and epidemic risks, where the latter includes damages from viruses, worms, or spywares, these epidemic risks might have a correlation structure. Eling and Wirfs (2016) argued that correlation is main characteristics of cyber risk and a research by Öğüt *et al.* (2011) investigated the nature of correlations in information security breach risk. Bohme and Kataria (2006) studied cyber risk, especially focused on information security risk with high-level correlations and argued that cyber insurance might not be best suited to managing information security risk if it has a high global correlation. In that case, since a high global correlation impedes risk-pooling in local risk groups, insurers are required to add a high safety loading to an insurance premium. Yang and Lui (2014) studied a specific epidemic model with a Bayesian network for cyber risk. In this study, our second goal is to test if there exist interclass correlations (we use the terms 'class' and 'group' interchangeably) between data breaches in different industry sectors, by including other groups' time series of data breach counts in the INGARCH model.

A more serious aspect to consider further in the INGARCH model is a model endogeneity, as exemplified by the evidence in the financial sector that volatility of large stocks can be both endogenous or exogenous (Sornette *et al.*, 2004). Achcar *et al.* (2018) used a Lévy distribution in the presence of left censored data and covariates. Third goal of our study is to test a selection of external covariates and their lagged time series for possible drivers of clustering volatility of data breach occurrences. Studies on other related cyber risks include Bojanc and Jerman-Blažič (2008) who developed an economic model for the risk.

This article is organized as follows. In Section 2, we construct an INGARCH model with covariates for data breach losses. Section 3 describes the sample of historical data breach occurrences and related considerations for our analysis. Section 4 analyzes the volatility clustering effect with and without external covariates and it tests the coincident interclass correlation between industry sectors. Section 5 summarizes our study.

2. INGARCH(p, q) model for data breach counts

Volatility clustering is accompanied by temporal autocorrelations. To analyze the effect in this section, we consider an INGARCH(p, q) model with covariates, i.e., an integer-valued generalized autoregressive conditional heteroskedasticity time series model for frequency counts due to data breaches occurred in a class (an industry sector). In this model, we include two types of covariates of another class's (industry sector's) coinciding (no time lag) data breach count and a time-lagged exogeneous

financial and economic index.

There are two types of correlations, i.e., interclass correlation at a group level and intraclass correlation at an individual level. We can assess the interclass correlation by including the other class's time series of data breach counts as a covariate in the model. The insurers' decision in setting a premium is also influenced if an interclass correlation exists at the class (group) level (Bohme and Kataria, 2006). Then, a simple pricing method using a ratio of losses to exposures would fail for correlated risks, since it assumes that insured risks are independent. Exogeneous covariates are also considered since they might enter the effect of volatility clustering (Samiev, 2013).

We construct an INGARCH(p, q) time series model with non-negative covariates (Heine, 2003; Ferland *et al.*, 2006), as follows. A data breach loss for the current period might depend on past observations and past means. Such a situation might occur, for example, when an individual hacker or hacker team locally seeks a specific type of identity information that can be found predominantly in a specific industry sector (Sen and Borle, 2015), e.g., the banking industry, and then they may make several continuing attempts of attack over a series of periods (Tarig *et al.*, 2006). It is then expected (with a high level of probability) that the data breach loss occurs in that industry more frequently over the extended time period in a clustered fashion. In this case, the volatility of data breach is clustered, but correlation is absent. As previously explained, the autocorrelation of a frequency time series should not be neglected in pricing insurance coverage.

Suppose we denote the time period such as one month as t, a count time series by $N_t^{(i)}$ for the i^{th} class (industry sector), and we represent the expected value of the frequency time series $E[N_t^{(i)}|F_{t-1}^{(i)}]$ conditional on $F_t^{(i)}$ by a process $\lambda_t^{(i)}$ such that $E[N_t^{(i)}|F_{t-1}^{(i)}] = \lambda_t^{(i)}$, where $F_t^{(i)}$ is a filtration of the joint process, $\{N_t^{(i)}, \lambda_t^{(i)}\}$. We then model the conditional mean, $\lambda_t^{(i)}$ with a link function, $g(\lambda_t^{(i)})$, by the following linear predictors with 5 concurrent observations in other sectors (l):

$$g\left(\lambda_t^{(i)}\right) = \beta_0^{(i)} + \sum_{j=1}^{p} \beta_j^{(i)} \tilde{g}\left(N_{t-m_j}^{(i)}\right) + \sum_{k=1}^{q} \alpha_k^{(i)} g\left(\lambda_{t-n_k}^{(i)}\right) + \sum_{l=1, l\neq i}^{5} \gamma_l^{(i)} \tilde{g}\left(N_t^{(l)}\right) + \eta^{(i)} X_{t-r}^{(i)}, \tag{2.1}$$

where g is a link function, $\tilde{g}$ is a transformation function of the frequency. Here, the $N_{t-m_j}^{(i)}, \lambda_{t-n_k}^{(i)}, N_t^{(l)}$ inside the sums, and $X_{t-r}^{(i)}$ are the j^{th} lagged observation, the k^{th} lagged conditional mean, the frequency observation in the l^{th} sector, and the covariate at lag r for the i^{th} group, respectively. Furthermore, $\beta_0^{(i)}$ is a constant, and $\beta_j^{(i)}, \alpha_k^{(i)}, \gamma_l^{(i)}$, and $\eta^{(i)}$ are linear coefficients. The first and second terms are concerned with the autocorrelation of the time series, the third terms are interclass linear correlation terms, and the fourth terms are covariate terms as explained below.

We will later show that the best-fit unconditional count distribution is the negative binomial distribution. Accordingly, we set the model by using the log link, $g(x) = \log x$, which is the canonical link function for Poisson and negative binomial distribution along with an identity transformation function, $\tilde{g}(x) = x$. Specifying a suitable set, $P = \{m_1, m_2, \ldots, m_p\}$ and $Q = \{n_1, n_2, \ldots, n_q\}$ is arbitrary, and its entire parameter space is vast. Therefore, we investigate the model by limiting analysis to the case when $P = \{m_j | m = j\}$, $Q = \{n_k | n = k\}$. The third term means interclass linear correlations: i.e., $\gamma_l^{(i)} N_t^{(l)}$ ($l = 1, \ldots, 5$, and $l \neq i$) is a term concerning with the coinciding interclass correlation between the i^{th} and l^{th} group's data breach frequencies. The fourth term represents lagged exogeneous covariates that possibly absorb volatility clustering or predict the frequency of data breach loss. In this study, we allow only one lagged exogeneous financial or economic covariate ($X_{t-r}^{(i)}$) for each class, which is chosen differently for different groups among possible candidates.

We further assume that the conditional distribution $N_t^{(i)}|F_{t-1}^{(i)}$ for residuals follows a Poisson or

negative binomial distribution. Then, given the above assumptions, the previous model reduces to the following equation:

$$\log\left(\lambda_t^{(i)}\right) = \beta_0^{(i)} + \sum_{j=1}^{p} \beta_j^{(i)} N_{t-j}^{(i)} + \sum_{k=1}^{q} \alpha_k^{(i)} \lambda_{t-k}^{(i)} + \sum_{l=1, l\neq i}^{5} \gamma_l^{(i)} N_t^{(l)} + \eta^{(i)} X_{t-r}^{(i)}. \tag{2.2}$$

When $p \neq 0$ or $q \neq 0$, then there exists heteroskedasticity, i.e., volatility clustering. If we assume the Poisson distribution for residuals, i.e., $N_t^{(i)}|F_{t-1} \sim \text{Poisson}(\lambda_t^{(i)})$, then

$$P\left(N_t^{(i)} = n|F_{t-1}\right) = \frac{(\lambda_t^{(i)})^n e^{-\lambda_t^{(i)}}}{n!}, \quad n = 0, 1, \ldots,$$

and the conditional variance is $\text{Var}(N_t^{(i)}|F_{t-1}) = E(N_t^{(i)}|F_{t-1}) = \lambda_t^{(i)}$. A non-Gaussian state space approach is an alternative to this time-series model. A state-space model which was originally developed in control engineering (Kalman, 1960) represents probabilistic dependence between state process and observation process in a dynamical system. Kitagawa (1981) showed that the time series with drifting mean value can be represented as a state-space model; in addition Timmer and Weigend (1997) also revealed that dynamic volatility of time series can be described by a state-space model instead of a GARCH model. In perspective of the non-Gaussian state-space approach to nonstationary time series (Kitagawa, 1987; Durbin and Koopman, 2000); therefore, our model can be interpreted as the one, where the observed data breach counts in overall industries with dynamic latent state variables to drive the dynamics of the system that are functionally mapped from such state variables with non-Gaussian observation errors.

In the conditional Poisson response model, the conditional variance is equal to the conditional mean value, by the nature of Poisson processes. For the negative binomial distribution for residuals, i.e., $N_t^{(i)}|F_{t-1} \sim \text{NegBin}(\lambda_t^{(i)}, \phi^{(i)})$, where the distribution is parameterized by the mean $\lambda_t^{(i)}$, and the dispersion parameter, $\phi^{(i)}$, and then

$$P\left(N_t^{(i)} = n|F_{t-1}\right) = \frac{\Gamma(\phi^{(i)} + n)}{\Gamma(\phi^{(i)})\Gamma(n+1)} \left(\frac{\phi^{(i)}}{\phi^{(i)} + \lambda_t^{(i)}}\right)^{\phi^{(i)}} \left(\frac{\lambda_t^{(i)}}{\phi^{(i)} + \lambda_t^{(i)}}\right)^{n}, \quad n = 0, 1, \ldots,$$

and the conditional variance is $\text{Var}(N_t^{(i)}|F_{t-1}) = \lambda_t^{(i)} + (\lambda_t^{(i)})^2/\phi^{(i)}$. Even if the dispersion parameter $\phi^{(i)}$ is kept constant, the variance is conditional on the mean, $\lambda_t^{(i)}$. In this parameterization, $(\sigma^{(i)})^2 = 1/\phi^{(i)}$ is the overdispersion parameter.

In this paper, we choose the Bayesian information criterion (BIC) as the selection criteria defined as:

$$\text{BIC} = p\log(n) - 2\log(L),$$

where L is the likelihood, n is the number of samples, and p is the number of parameters.

3. Data

An act of data breach is an intentional or unintentional release of disclosed information to an unauthorized entity or environment. A common problem in tackling a new risk is to properly define its term and scope. The U.S. courts have tried to conceptualize data breaches (Solove and Citron, 2017).

According to the United States Department of Health and Human Services, Administration for Children and Families, a data breach is defined as "a security incident in which sensitive, protected or confidential data is copied, transmitted, viewed, stolen or used by an individual unauthorized to do so." (Information Memorandum, 2015). It is also called 'cyber liability insurance', which emphasizes a duty of an insured to protect secured information for its customers. In order to avoid confusion with the meaning of private information, the word, 'data' if used solely in our context, refers to a sample of data breach loss records.

The process of pricing cyber risk insurance for insurers has been hindered partially due to insufficient and unreliable data until recently (Biener *et al.*, 2015; Maillart and Sornette, 2010). To insurers, this environment has worsened due to the dynamic nature of cyberspace. Eling and Schnell (2016) noted that the most useful data set would be the actual data breach claims of insurers in the Cyber Claims Study (2018). Our study is based on the samples of the de-trended data breach records from Jan. 1, 2006 to Dec. 31, 2015, published by the Identity Theft Resource Center (ITRC) on its website (www.idtheftcenter.org). In a form slightly different from the above, the ITRC defines a data breach as "an incident in which an individual's name plus a social security number, driver's license number, and medical record or financial record (credit/debit cards included) is potentially put at risk of exposure." The ITRC only traces data losses with regard to four types of information: social security number, credit/debit card number, email/password/username, and protected health information (PHI). Maillart and Sornette (2010) and Wheatley *et al.* (2016) found out that the personal data breaches increased from 2001 to 2006, but afterwards, the frequency has been stable since 2006 or 2007.

The exposure under the data breach risk is the total US firms. The ITRC collects publicizes reports daily data losses from the exposure and classifies them by five industry sectors: banking/credit/financial (BCF), business (BUS), educational (EDU), governmental/military (GOV), and medical/healthcare (MED). As the daily counts of data breach accidents are low typically ranging from zero to five, we aggregate the data on the monthly basis to avoid large fluctuations of counts between periods.

With the above definition of exposures, it is implicitly assumed that during the data collection periods from 2006 to 2015, there has been no variation in the number of firms in each industry sector in the United States. The ITRC reports do not provide the explicit or estimated number of individuals or firms at risk classified by the sectors; therefore, it keeps one from attaining the exact number of exposures and applying exposure adjustment to a count time series.

4. Empirical results

4.1. Descriptive statistics and frequency distributions

In this section, we pool the longitudinal data breach count time series, and fit the unconditional pooled data to two probability distribution models for frequency count data: Poisson and negative binomial distributions. A probability density function of a negative binomial distribution for a count, k, is parametrized by a size r, which is the number of failures before the k^{th} success and a probability of success, p:

$$f(k; r, p) = {}_{k+r-1}C_k p^k (1-p)^r.$$

Frequency is defined to be the number of losses per month observed in each industry sector. We estimate parameters by the maximum likelihood estimation and perform chi-square goodness-of-fit tests for residuals. Table 1 shows the descriptive statistics, frequency distributions, p-value of the test and BIC. Table 1 indicates that Poisson distributions are adequate for most sectors except the BCF, MED sector and all sectors combined, where negative binomial distributions seem to be appropriate.

Table 1: Descriptive statistics and frequency distributions

Sector		BCF	BUS	EDU	GOV	MED	ALL
Mean		3.358	22.158	8.625	7.450	24.717	58.642
Standard Deviation		2.307	5.109	2.596	2.813	7.014	12.114
Variance		5.324	26.101	6.741	7.913	49.196	146.753
Index of Dispersion (Variance-to-Mean Ratio)		1.585	1.178	0.782	1.062	1.990	2.503
Model 1: Poisson (P)	p-value	0.003	0.202	0.054	0.716	0.038	< 0.001
	BIC	532.622	727.137	574.283	585.801	817.903	999.074
Model 2: Negative Binomial (NB)	p-value	0.623	0.073	0.028	0.544	0.087	0.077
	BIC	524.241	730.622	579.071	590.436	794.835	939.605

BCF = banking/credit/financial; BUS = business; EDU = educational; GOV = governmental/military; MED = medical/healthcare; BIC = Bayesian information criterion.

Figure 1: *The ACF and PACF of the count time series of data breach.*

Additional evidence is that the ratio of variance to mean, which should be equal to one for a Poisson distribution, is close to one for the BUS, EDU, and GOV.

The Poisson distributions or the negative binomial distributions rather than the zero-modified distributions are adequate to fit the frequency distributions, since the probability of zero claims is not distorted during the process of data collection, and its magnitude is negligible. The fact that the negative binomial distribution is the best for some sectors indicates overdispersion in loss frequency distributions for that sector. In general, overdispersion in a generalized linear model arises when there

Table 2: The fitted INGARCH(p, q) model without covariates: goodness-of-fit, lag of autocorrelation, parameter estimates and test statistics

		Sector (i)											
		BCF (1)		BUS (2)		EDU (3)		GOV (4)		MED (5)		ALL	
Goodness of fit	Order by lowest BIC	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
	Cond. dist.	P	NB	P	P	P	P	P	P	NB	NB	NB	NB
	BIC	510.79	512.169	715.579	719.44	574.115	577.277	586.485	587.929	789.888	795.251	912.081	912.706
Lag of Autocorrelation	p	1	1	1	2	1	1	1	1	2	1	1	2
	q	1	1	1	1	1	0	0	1	0	1	1	1
Parameter (SE)	β_0	0.03	0.03	0.201	0.249	0.131	1.829	1.595	0.149	1.969	0.29	0.271	0.508
		(0.036)	(0.043)	(0.190)	(0.244)	(0.140)	(0.248)	(0.209)	(0.169)	(0.425)	(0.319)	(0.223)	(0.352)
	β_1	0.175	0.175	0.177	0.116	0.156	0.146	0.198	0.117	0.014	0.138	0.229	0.113
		(0.069)	(0.084)	(0.076)	(0.104)	(0.082)	(0.110)	(0.098)	(0.066)	(0.098)	(0.076)	(0.078)	(0.101)
	β_2				0.095					0.370			0.265
					(0.137)					(0.097)			(0.133)
	α_1	0.783	0.783	0.757	0.708	0.778			0.805		0.771	0.704	0.497
		(0.090)	(0.109)	(0.118)	(0.164)	(0.128)			(0.130)		(0.148)	(0.111)	(0.176)
	σ^2		0.135							0.027	0.035	0.017	0.015
Test statistics	Persistence	0.958	0.958	0.934	0.917	0.934	0.146	0.198	0.922	0.384	0.909	0.933	0.874
	Lagrange multiplier test (lag-1)	0.637	0.637	0.857	0.831	0.823	0.95	0.331	0.232	0.377	0.122	0.664	0.91
	Lagrange multiplier test (lag-5)	0.548	0.548	0.943	0.899	0.7	0.547	0.615	0.631	0.97	0.553	0.991	0.994
	Lagrange multiplier test (lag-9)	0.878	0.878	0.999	0.996	0.027	0.246	0.364	0.355	0.88	0.802	0.947	0.859
	Lagrange multiplier test (lag-24)	0.567	0.567	0.844	0.803	0.377	0.004	0.061	0.078	0.185	0.065	0.337	0.816
	ADF test	0.018	0.018	0.014	<0.01	0.032	0.275	<0.01	<0.01	0.048	0.017	0.018	<0.01

BCF = banking/credit/financial; BUS = business; EDU = educational; GOV = governmental/military; MED = medical/healthcare; BIC = Bayesian information criterion.

are outliers, or imprecise relationships between conditional means and regressors, or omitted explanatory variables. Maillart and Sornette (2010) and Wheatley *et al.* (2016) also argue that the frequency of cyber risk and data breaches forms a heavy-tailed distribution. However, missing information about more explanatory variables, which the ITRC did not collect, might form substructures in a predefined risk group, which is an industry sector and induce inhomogeneity in the group (Sen and Borle, 2015).

To investigate the existence of autocorrelation for the count time series, we plot the autocorrelation function and the partial autocorrelation function of all data breach counts in Figure 1. There exists autocorrelation of the time series at short lags.

4.2. INGARCH(p, q) Model without Covariates: Evidence of Volatility Clustering

In this section, we show the fitted INGARCH(p, q) models for data breach count time series without covariates. First, we investigated various INGARCH(p, q) models for all possible cases of the Poisson residual distribution and the negative binomial residual distribution with $p = 1, \ldots, 10$ and $q = 1, \ldots, 10$ lags, and we obtained the best-fit model by maximizing the MLE among the given parameter set space. In Table 2, we present the parameter estimates and BICs for the best and the second-to-best models for each class. Goodness-of-fit in Table 2 shows that for the most classes, the best model is the INGARCH(1, 1) time series with a Poisson conditional distribution. The superscript '(i)' from all the parameters, such as $\beta_j^{(i)}$, $\alpha_k^{(i)}$, and $\gamma_l^{(i)}$ are dropped for simplicity. In addition, σ^2 is an estimated overdispersion coefficient in a negative binomial model only and persistence is defined as $\sum_{j=1}^{p} \beta_j^{(i)} + \sum_{k=1}^{q} \alpha_k^{(i)}$ in the table.

In Table 2, $\beta_j (j \neq 0)$ and α_k have nonzero values. The mean, λ_t, is conditional on the lagged observations ($N_{t-j}^{(i)}$) and the lagged means ($\lambda_{t-k}^{(i)}$) and the variance is also conditional on past informa-

tion; therefore, the data breach counts are heteroscedastic. Furthermore, a volatility clustering effect persists strongly because the persistence of autocorrelation is close to one in most of models.

The model with one lagged observation ($p = 1$) and one lagged mean ($q = 1$) has a strong persistence and this model is adequate to use the information in the previous period in predicting the volatility of data breach counts. The small-lag autocorrelation implies that the expected frequency rates have short-range dependence and that the data breach attack has a period of high activity coming after a period of low activity. Comparison between β_j and α_k show that the effect of lagged means, α_k (0.70–0.81), is greater than that of lagged observations, β_j (0.01–0.37), i.e., the autoregressive component is larger than the moving average one. The weak dependence on the past observation is also valid for the two lagged models, for example, the INGARCH$(2, 1)$ model of the BUS sector. This dependency on past information may be concerned with the epidemic nature of cyber risk or the behavioral property of data breach hackers. Table 2 also contains the p-values of various tests for validity of the INGARCH(p, q) model. The Lagrange multiplier test for the following null hypothesis that the squared residuals in the current period are independent of the past squared residuals at lag-m,

$$\epsilon_t^2 = \beta_0 + \beta_1\epsilon_{t-1}^2 + \beta_2\epsilon_{t-2}^2 + \cdots + \beta_m\epsilon_{t-m}^2,$$

is not rejected at the significance level of 0.05. It shows that the residual variances are not heteroskedastic. The Ljung-Box test for the null hypothesis that the autocorrelation between the standardized residuals for a set of lags up to 24 is zero is not rejected at the significance level of 0.05. It represents that no autocorrelation remains in the residuals of the INGARCH$(1, 1)$ model. The augmented Dickey-Fuller test for the null hypothesis that a unit root is present in time series is rejected at the same significance level, and thus the time series can be regarded as stationary. In Figure 2, we also show the ACF and the PACF plot of the residuals, where no noticeable pattern of autocorrelation is observed.

4.3. INGARCH model with covariates: clustering of volatilities remained even after the inclusion of covariates

Individual exposures or group exposures under data breach risk may not be independent with each other. In that case, the data breach losses of a certain industry sector might be affected by other industry's data breach losses, or vice versa. Table 3 shows the correlation of each industry sector's data breach counts and its p-value. At the significance level of 0.01, meaningful correlations are found between the following sectors: BCF-BUS, BCF-EDU, BUS-EDU, and GOV-MED. From Table 3, the most significant is the correlation between BCF-BUS, 0.558, implying that the data breach risk in finance sector is closely related with the general business sector in terms of frequencies. In the previous section, we analyzed the data breach risk by the INGARCH model without covariates. We now include covariates in the INGARCH model and as the first kind of covariates; in addition, we consider coincident time series of other industries' counts, $\gamma_l^{(i)} N_t^{(l)}$ ($l = 1, \ldots, 5$, and $l \neq i$), for the i^{th} class in relation with the l^{th} class. We considered the candidate models with all possible combinations of correlation terms (a single-covariate model and a double-covariate model), and Table 3 shows the results of the best-fit models.

We found that the best-fit models are single-covariate models and they have $\gamma_l^{(i)} \neq 0$ ($l \neq i$) for only one external class (one industry sector). A possible explanation of correlation is that a target of threat is not limited to a preferred class's firms but instead ranges over several industries. Another cause might be the instantaneous secondary data breach using the stolen secured information from the primary data breach activity. The covariate of data breach counts for each sector are: BUS count for BCF sector, BCF count for BUS sector, BUS count for EDU sector, and MED count for GOV

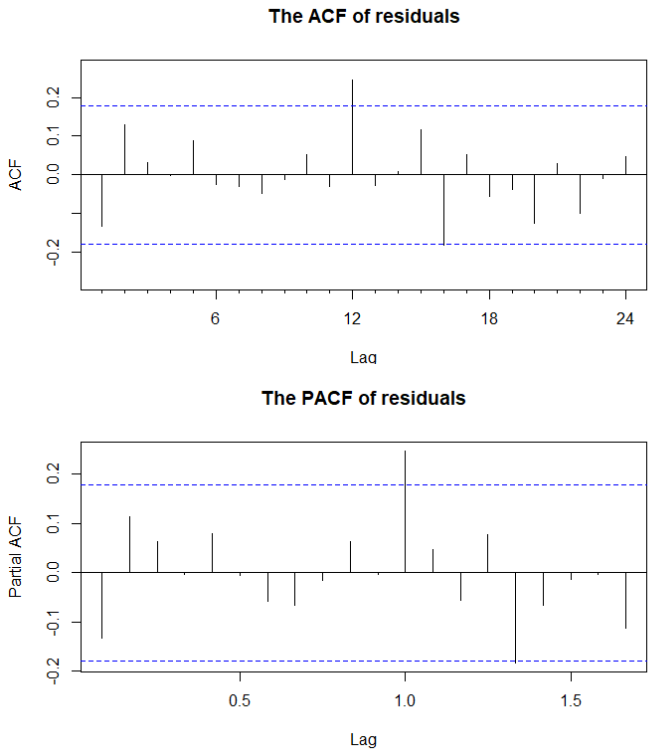

Figure 2: *The ACF and PACF of the residuals.*

Table 3: Correlation of each sector's data breach counts: the correlation and the p-value in parenthesis

Correlation (p-value)	BCF	BUS	EDU	GOV	MED
BCF	1	0.558 (<0.001)	0.286 (0.002)	0.081 (0.379)	−0.025 (0.784)
BUS		1	0.291 (0.001)	0.18 (0.049)	0.161 (0.078)
EDU			1	0.111 (0.229)	−0.02 (0.827)
GOV				1	0.331 (< 0.001)
MED					1

BCF = banking/credit/financial; BUS = business; EDU = educational; GOV = governmental/military; MED = medical/healthcare.

sector. This relation bears a similarity to the previous result of correlations in many cases. Other pairs show no evidence of interclass correlation. However, it is still questionable if there exist intraclass correlations. Intraclass correlations in each sector are hard to evaluate with the given limited data set due to the firm's incomplete loss records. This result is in line with the finding of Herath and Herath (2011) that shows that cyber risks are correlated with a non-linear dependency. Hofmann and Ramaj (2011) also noted the interdependence risk structure in a cyber network, and Bashan *et al.* (2013) showed that the spatially embedded networks are interdependent and vulnerable to failure. Öğüt *et al.*

Table 4: The fitted INGARCH(p, q) models with a correlated sector's count as covariates: goodness-of-fit, lag of autocorrelation, parameter estimates and test statistics

		Sector (i)			
		BCF (1)	BUS (2)	EDU (3)	GOV (4)
Goodness-of-fit	Conditional distribution	P	P	P	P
	BIC	486.401	697.535	570.885	578.464
Lag of autocorrelation	p	1	1	1	1
	q	1	1	1	1
Parameter (SE)	β_0	−0.546	2.267	0.434	0.286
		(0.185)	(0.473)	(0.183)	(0.108)
	β_1	0.145	0.104	0.020	0.014
		(0.108)	(0.113)	(0.090)	(0.077)
	α_1	0.318	0.109	0.664	0.718
		(0.147)	(0.187)	(0.150)	(0.114)
Covariation	Correlated class (l)	BUS	BCF	BUS	MED
	$\gamma_l^{(i)}$	0.051	0.048	0.011	0.010
		(0.008)	(0.009)	(0.005)	(0.003)
Test statistics	Persistence	0.464	0.213	0.684	0.732
	Lagrange multiplier test (lag-1)	0.951	0.231	0.749	0.381
	Lagrange multiplier test (lag-5)	0.948	0.628	0.830	0.788
	Lagrange multiplier test (lag-9)	0.734	0.889	0.197	0.127
	Ljung-Box test (lag-24)	0.702	0.844	0.347	0.073
	ADF test	0.013	<0.01	0.054	<0.01

BCF = banking/credit/financial; BUS = business; EDU = educational; GOV = governmental/military; BIC = Bayesian information criterion.

(2011) examined the role of correlated risks in information security. In the presence of correlation, mean, variance, and VaR might not be sufficient for insurance ratemaking, since diversification is not efficient for reducing risk capital (Chavez-Demoulin *et al.*, 2006). However, the above analysis might be limited if the characteristics of ever-changing environment for data breaches might diminish the usefulness of the given data set (CRO Forum, 2014).

A variety of theories in other forms have also explained correlations between groups for cyber risk: multivariate distributions with the theory of copulas (Clemen and Reilly, 1999; Bohme and Kataria, 2006; Herath and Herath, 2011), copula-based Bayesian network for cyber risk (Mukhopadhyay *et al.*, 2013), and a two-step risk arrival process considering internal and external correlations by a two-tier approach (Bohme and Kataria, 2006). For example, Bohme and Kataria (2006) systematically studied the correlations using t-copula. Next, we consider the INGARCH model with a lagged exogeneous financial variables as the second kind of covariates. The effect of data breach announcements on the stock price has been shown to be negative (Cavusoglu *et al.*, 2004; Campbell *et al.*, 2003). However, we also study the reverse effect that the stock price return or other financial market status affect the volatility of data breach accidents. We selected two financial metrics: the return of S&P 500 index and the VIX, that are believed to have volatilities clustered on their own (Jacobsen and Dannenburg, 2007; Tseng and Li, 2012; Aliyu, 2012). Therefoere, it motivates us to check if the clustered volatilities of data breaches are fully or partially due to the effect of volatility clustering occurring in the covariate variable. Table 5 presents the summaries of the best-fit models with the selected covariates. Compared with the results in Table 2 for the model with no covariate, we notice that including the covariate term into the previous model marginally increases the likelihood (lowers BIC). Since the best-fit model is still the INGARCH(p, q) with $p = q = 1$, the volatility clustering persists even in the presence of volatility-clustered covariates.

The lagged covariate with a nonzero coefficient might help in predicting the data breach count;

Table 5: The fitted INGARCH(p, q) models with a lagged exogeneous financial covariated of $X^{(i)}$: goodness-of-fit, lag of autocorrelation, parameter estimates and test statistics. The model for industry sectors that are not shown are found to have no lower BICs

		Sector (i)		
		BCF (1)	BUS (2)	ALL
Goodness-of-fit	Conditional distribution	P	P	NB
	BIC	509.831	714.782	911.107
Lag of autocorrelation	p	1	1	1
	q	1	1	1
Covariation	$X^{(i)}_{t-r}$	VIX	Return	Return
	Number of lags (r)	13	8	22
Parameter (SE)	β_0	4.42e−06	0.123	0.222
		(0.0296)	(0.122)	(0.186)
	β_1	0.179	0.158	0.218
		(0.057)	(0.071)	(0.080)
	α_1	0.821	0.800	0.727
		(0.074)	(0.097)	(0.106)
	$\eta^{(i)}$	−0.00158	0.00559	0.00519
		(0.000748)	(0.00236)	(0.00248)
	σ^2			0.015
Test statistics	Persistence	0.999	0.958	0.945
	Lagrange multiplier test (lag-1)	0.707	0.745	0.497
	Lagrange multiplier test (lag-5)	0.528	0.937	0.979
	Lagrange multiplier test (lag-9)	0.880	0.998	0.977
	Ljung-Box test (up to lag-24)	0.610	0.779	0.149
	ADF test	<0.01	0.014	0.024

BCF = banking/credit/financial; BUS = business; BIC = Bayesian information criterion.

however, caution needs to be taken since the effect is marginal. Since the coefficient of financial exploratory variable term has a positive sign, we might conjecture that the act of data breach has financial or economic motive, though its process has not been concretely studied yet. A high value in the corresponding financial or economic index might be indicative of an anticipated increase in data breach occurrence since the signs of the coefficients are positive for all models. For example, concerning the BUS sector, we guess that a high return in a stock market may motivate a malicious hacker to break into a secured data repository of any general firms, which has a higher tradable value in such a bull market. We therefore conjecture that the lagged S&P500 or VIX forecasts the count frequency of data breach due to financial or psychological volatility. It is also interesting if data breaches or denial-of-service attacks affect a company's stock prices (Campbell *et al.*, 2003; Hovav and Darcy, 2003) in the opposite way.

5. Conclusion

Data breach insurance is considered an effective and efficient means of protection from the leakage of personal information or important information. In this paper, we analyzed the volatility clustering of data breach counts, using the ITRC data classified by five industry groups from 2006 to 2015 for data breach records in US. We modeled the data breach counts by the INGARCH(p, q) model with and without covariates, assuming a Poisson or negative binomial distribution as the conditional distribution for the mean and the variance.

Empirical results in the INGARCH(1, 1) model without covariates show that the volatility of time series of data breach counts is clustered in time with short-range time dependence. The volatility has the one-lag autocorrelation and they are conditional on the one-lag observation and the one-lag mean.

Therefore, if a large number of data breach losses are observed in the previous month, then a large number of losses are expected also in the current month. In the INGARCH(1, 1) model including coincident time series of other industries' counts as covariates, we found that the volatilities are clustered across industries too and the data breach risk of some industries are correlated. However, in the INGARCH(1, 1) model with exogenous financial covariates, we observed that the part of volatility clustering in the original count time series might be marginally explained by the volatility clustering occurred in the external covariates. In that case, the best covariate is the VIX for the business industry, the stock return of S&P500 for both the finance industry and all industries combined.

A problem with managing data breach risk is the possibility of unstable losses, which requires additional solvency capital; therefore, a more delicate design in an insurance policy is needed. As Biener *et al.* (2015) noted, once more detailed data becomes available, it would be also interesting to directly examine the risk drivers of data breach volatility them include to firm sizes (Axtell, 2001; Bonfim, 2009), breach types, the quality of information, and the risk control of the firm, or entity. In addition, it would be a prospective topic to study correlation between those insured and to develop a robust multivariate copula model for data breach risk.

Acknowledgements

We thank the two reviewers for constructive comments which led to a substantial improvement in the revised version. This paper was supported by the Daesan Shin Yong Ho Memorial Society.

References

Achcar JA, Coelho-Barros EA, Cuevas JRT, and Mazucheli J (2018). Use of Lévy distribution to analyze longitudinal data with asymmetric distribution and presence of left censored data, *Communications for Statistical Applications and Methods*, **25**, 43–60.

Aliyu SUR (2010). Does inflation has an impact on stock returns and volatility? Evidence from Nigeria and Ghana, *Applied Financial Economics*, **22**, 427–435.

Axtell RL (2001). Zipf Distribution of U.S. Firm Sizes, *Science (New York, N.Y.)*, **293**, 1818–1820.

Bashan A, Berezin Y, Buldyrev SV, and Havlin S (2013). The extreme vulnerability of interdependent spatially embedded networks, *Nature Physics*, **9**, 667–672.

Berliner B (1982). *Limits of Insurability of Risks* (1st Ed), Prentice Hall, Englewood Cliffs, N.J.

Biener C and Eling M (2012). Insurability in microinsurance markets: an analysis of problems and potential solutions, *The Geneva Papers on Risk and Insurance - Issues and Practice*, **37**, 77–107.

Biener C, Eling M, and Wirfs JH (2015). Insurability of cyber risk: an empirical analysis, *Geneva Papers on Risk and Insurance-Issues and Practice*, **40**, 131–158.

Bohme R and Kataria G (2006). On the Limits of Cyber-Insurance. In *Trust, Privacy, and Security in Digital Business, Proceedings*, (S. Fischer-Hubner, S. Furnell, and C. Lambrinoudakis Eds), **4083**, Springer, Berlin, Heidelberg.

Bojanc R, and Jerman-Blažič B (2008). An economic modelling approach to information security risk management, *International Journal of Information Management*, **28**, 413–422.

Bollerslev T (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, **31**, 307–327.

Bonfim D (2009). Credit risk drivers: evaluating the contribution of firm level information and of macroeconomic dynamics, *Journal of Banking & Finance*, **33**, 281–299.

Cyber Claims Study (2018). NetDiligence. https://netdiligence.com/wp-content/uploads/2018/11/2018-NetDiligence-Claims-Study_Version-1.0.pdf.

Campbell K, Gordon LA, Loeb MP, and Zhou L (2003). The economic cost of publicly announced information security breaches: empirical evidence from the stock market, *Journal of Computer Security*, **11**, 431–448.

Cavusoglu H, Mishra B, and Raghunathan S (2004). The effect of internet security breach announcements on market value: capital market reactions for breached firms and internet security developers, *International Journal of Electronic Commerce*, **9**, 69–104.

Chavez-Demoulin V, Embrechts P, and Neŝlehová J (2006). Quantitative models for operational risk: extremes, dependence and aggregation, *Journal of Banking & Finance*, **30**, 2635–2658.

Clemen RT and Reilly T (1999). Correlations and copulas for decision and risk analysis, *Management Science*, **45**, 208–224.

CRO Forum (2014). Cyber resilience - The cyber risk challenge and the role of insurance. CRO Forum. https://www.thecroforum.org/wp-content/uploads/2015/01/Cyber-Risk-Paper-version-24-1.pdf.

Durbin J, and Koopman SJ (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **62**, Part1: 3–56.

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

Eling M and Schnell W (2016). What do we know about cyber risk and cyber risk insurance?, *The Journal of Risk Finance*, **17**, (5). Emerald Group Publishing Limited, 474–491.

Eling M and Wirfs JH (2016). Cyber risk: too big to insure? *Risk Transfer Options for a Mercurial Risk Class*. Institute of Insurance Economics I.VW-HSG.

Engle RF (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom Inflation, *Econometrica*, **50**, 987–1007.

Ferland R, Latour A, and Oraichi D (2006). Integer-valued GARCH process, *Journal of Time Series Analysis*, **27**, 923–942.

Heinen A (2003). Modelling time series count data, *An Autoregressive Conditional Poisson Model*. CORE Discussion Paper 2003062. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Herath H and Herath T (2011). Copula-based actuarial model for pricing cyber-insurance policies, *Insurance Markets and Companies: Analyses and Actuarial Computations*, **2**, 7–10.

Hofmann A and Ramaj H (2011). Interdependent risk networks: the threat of cyber attack, *International Journal of Management and Decision Making*, **11**, 312.

Hovav A and Darcy JL (2003). The impact of denial-of-service attack announcements on the market value of firms, *Risk Management and Insurance Review*, **10**, 97–121.

Information Memorandum (2015). *Information Memorandum*, United States Department of Health and Human Services, Administration for Children and Families.

Jacobsen B and Dannenburg D (2007). *Volatility Clustering in Monthly Stock Returns*, SSRN Scholarly Paper ID 1016668. Rochester, NY: Social Science Research Network.

Kalman RE (1960). A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, **82**, 35–45.

Kitagawa G (1981). A nonstationary time series model and its fitting by a recursive filter, *Journal of Time Series Analysis*, **2**, 103–116

Kitagawa G (1987). Non-Gaussian state-space modeling of nonstationary time series, *Journal of the American Statistical Association*, **82**, 1032–1041.

Lee J and Hwang E (2018). A generalized regime-switching integer-valued GARCH(1, 1) model and

its volatility forecasting, *Communications for Statistical Applications and Methods*, **25**, 29–42.

Maillart T and Sornette D (2010). Heavy-Tailed Distribution of Cyber-Risks, *The European Physical Journal B*, **75**, 357–764.

Mandelbrot B (1963). The variation of certain speculative prices, *The Journal of Business*, **36**, 394–394.

Mukhopadhyay A, Chatterjee S, Saha D, Mahanti A, and Sadhukhan SK (2013). Cyber-risk decision models: To insure IT or not?, *Decision Support Systems*, **56**, 11–16.

Öğüt H, Raghunathan S, and Menon N (2011). Cyber security risk management: public policy implications of correlated risk, imperfect ability to prove loss, and observability of self-protection: cyber security risk management, *Risk Analysis*, **31**, 497–512.

Pooser DM, Browne MJ, and Arkhangelska O (2018). Growth in the perception of cyber risk: evidence from U.S. P&C insurers, *The Geneva Papers on Risk and Insurance - Issues and Practice*, **43**, 208–223.

Samiev S (2013). GARCH(1, 1) with Exogenous Covariate for EUR/SEK Exchange Rate Volatility: On the Effects of Global Volatility Shock on Volatility. https://www.diva-portal.org/smash/get/diva2:676106/FULLTEXT01.pdf

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

Solove DJ, and Citron DK. 2017. "Risk and Anxiety: A Theory of Data-Breach Harms. *Texas Law Review* **96** (4): 737786.

Sornette D, Malevergne Y, and Muzy JF (2004). Volatility fingerprints of large shocks: endogenous versus exogenous. In *The Application of Econophysics* (Hideki Takayasu ed, 9–102), Springer, Tokyo.

Tariq U, Hong MP, and Lhee KS (2006). A Comprehensive Categorization of DDoS Attack and DDoS Defense Techniques. In *Advanced Data Mining and Applications* (Xue Li, Osmar R. Zaïane, and Zhanhuai Li eds, pp. 1025–1036). Lecture Notes in Computer Science, **4093**, Springer, Berlin, Heidelberg.

Tseng JJ and Li SP (2012). Quantifying volatility clustering in financial time series, *International Review of Financial Analysis*, **23**, 11–19.

Timmer J and Weigend AS (1997). Modeling volatility using state space models, *International Journal of Neural Systems*, **8**, 385–398.

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

Yang Z and Lui JCS (2014). Security adoption and influence of cyber-insurance markets in heterogeneous networks, *Performance Evaluation*, **74**, 1–17.

Received April 20, 2020; Revised June 20, 2020; Accepted June 20, 2020