Cyber Risk of Healthcare Data Breaches via GARCH Model

Stefano Chiaradonna[a]

[a]*School of Mathematical and Statistical Sciences, Arizona State University, Wexler Hall, Tempe, 85287, Arizona, USA*

Abstract

Cyber risk due to data breach is defined as the risk of a financial loss due to a breach of a company's IT infrastructure by unauthorized parties. Cyber risk remains one of the most disparaging and problematic risks that companies, particularly those in the healthcare sector, continuously face. Unfortunately, there has been very little research conducted on the historical trend of cyber risk, particularly in the healthcare sector. In this paper, we shall review one of the few and more recent papers that study the historical behavior of cyber risk.

1. Introduction

Motivation. According to The Institute of Risk Management, *cyber risk* is defined as "any risk of financial loss, disruption or damage to the reputation of an organization from some sort of failure of its information technology (IT) systems" (The Institute of Risk Management, 2018). In the first six months of 2021, there were 2.5 billion malware attacks and 2.5 trillion intrusion attempts (SonicWall, 2021) in which an intruder gains or attempts to gain unauthorized access to a system or its network. These intrusions have created significant and escalating losses. From 2020 to 2021, the cost of a data breach due to a single cyberattack has risen from \$3.86 million to \$4.24 million (IBM Security, 2020; Ponemon Institute, 2021). While these losses provide a small glimpse across a wide range of companies, those in the healthcare sector in particular have been seeing unparalleled losses. In fact, from 2016 to 2020, there were 270 ransomware attacks on U.S. healthcare organizations that resulted in a total estimated cost of over \$20.8 billion impacting 2,196 hospitals, clinics, and other medical facilities (Bischoff, 2021). To address such concerns, there has been growing research.

Literature Review. There have been a wide array of cyber risk datasets leveraged to study cyber risk. Many leading studies on cyber risk consider the privacy rights clearinghouse (PRC) dataset, which started to report data-breach cases publicly in 2005 (see e.g. (Carfora and Orlando, 2019; Edwards et al., 2016; Farkas et al., 2021; Eling and Jung, 2018)). In contrast, some actuarial works consider publicly available datasets such as the Advisen dataset (Aldasoro

et al., 2020). Moreover, other studies leverage the SAS OpRisk Database (see e.g. (Biener et al., 2015; Eling and Wirfs, 2019)). In particular, Kim and Song (2023) investigated the insurability of cyber risk by analyzing 994 cases of cyber losses.

In this paper, we explore the methodology developed Kim and Song (2023) to model cyber risk losses across time. We first discuss the peaks over threshold (POT) technique to capture the heavy-tailed characteristic of cyber risk data. Next, we discuss the process of the loss distribution approach (LDA) to model monthly cyber losses, the value-at-risk (Var) measure, and the copula approach for dependence. Lastly, following their methodology, we consider the generalized autoregressive conditional heteroskedasticity (GARCH) model for reflecting the time dependence of monthly cyber losses.

While Kim and Song (2023) leveraging the SAS OpRisk Global Database, we shall use the healthcare-specific database from the U.S. Department of Health and Human Resources Office for Civil Rights (HHS) (U.S. Department of Health and Human Services, 2023). It has only been recently that studies have leveraged this dataset for healthcare-specific cyber risks (see e.g. (Ronquillo et al., 2018; Dolezel and McLeod, 2019)). In particular, Li and Mamon (2023) showed that the data-breach incidents are adequately modeled by the Markov-modulated non-homogeneous Poison process. However, there still remains a gap in the exploration of the historical trend of cyber risk. To explore this gap, we fit two models to the healthcare provider and business associated data both with an ARMA(1,1)-GARCH(1,1).

The remainder of this paper is organized as follows. Section 2 introduces the methods and models. Section 3 describes the HHS data used while Section 4 provides a time-series analysis of the data. Finally, Section 5 concludes the paper.

2. Methods and models

In the model proposed by Kim and Song (2023), they leverage various techniques to construct and ascertain the cyber risk loss distribution. In this section, we provide a brief description of each technique.

2.1. Peaks over threshold (POT)

The Peaks Over Threshold (POT) approach is a statistical method used to analyze extreme events in a dataset. Essentially, the POT approach is to identify the "peaks" in a dataset that exceed a certain threshold level, which is typically set based on some predefined criteria or statistical tests. These peaks are assumed to be independent and identically distributed (i.i.d.) random variables and can be modeled using extreme value theory (EVT). One of the main advantages of the POT technique is that it allows for the modeling of extreme events that are not captured by traditional methods such as mean-variance analysis or linear regression. However, the method is quite sensitive to the choice of the threshold level, and care must be taken to choose an appropriate level based on the characteristics of the data.

2.2. Loss distribution approach (LDA)

The loss distribution approach (LDA) is the industry standard for pricing insurance risks (Jevtić and Lanchier, 2020). The approach is based on the assumption that the distribution of the losses can be decomposed into the frequency and severity of the losses. Once the frequency and severity distributions have been estimated, they can be combined to generate a joint distribution of losses for the portfolio. This joint distribution can be used to estimate a range of risk measures, such as Value-at-Risk (VaR).

2.3. Value-at-risk (VaR)

VaR is a statistical measure used to estimate the potential loss that could incur over a given time period, with a specified level of confidence. It is one of the most common methods to quantify financial risk. One limitation of VaR is that it only provides a single estimate of potential losses, and does not provide any information about the distribution of losses beyond the VaR estimate. Additionally, VaR assumes that the distribution of returns is stationary over time, which may not always be the case in practice.

2.4. GARCH model for time series

The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model is a statistical model used to describe the volatility of the time series data. The GARCH model extends the ARCH model by incorporating past values of the conditional variance as additional explanatory variables. Furthermore, the GARCH model allows for the estimation of the conditional variance of a time series, which can be used to calculate volatility measures such as the standard deviation, variance, and Value-at-Risk (VaR).

2.5. Copula

Copulas are a mathematical tool used to model the dependence between variables in a multivariate distribution. Copulas are particularly useful when modeling the dependence structure of variables with different marginal distributions. Copulas provide a way to measure and quantify the strength and direction of the dependence between variables using tools such as Kendall's tau and Spearman's rank correlation coefficient. There are several families of copulas, including Gaussian, Archimedean, and extreme-value copulas.

3. Data description

We obtained healthcare data breach data from HHS for the aggregate monthly losses due to data breaches (U.S. Department of Health and Human Services, 2023). In this dataset, we consider losses as the number of individuals affected since the dataset does not contain any financial loss amounts. Furthermore, we consider confirmed data breaches from October 2009 through February 2023, resulting in 4,448 individual data breaches aggregated across months. This results in an aggregated dataset of 159 monthly data breach observations. Finally, we

split the data between business, such as an insurance company, and healthcare-associated, such as a hospital, as the origin of data breaches (see Figures 1 - 3 for the monthly time series data).

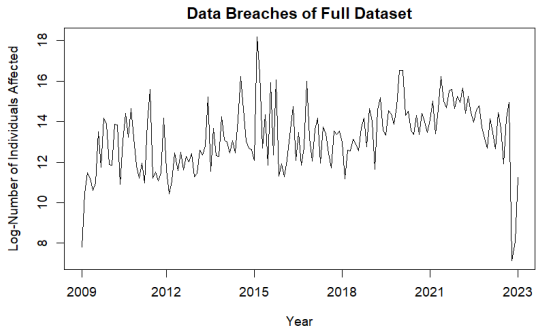

Figure 1: Time series data of log-number of individuals affected by data breaches aggregated monthly.

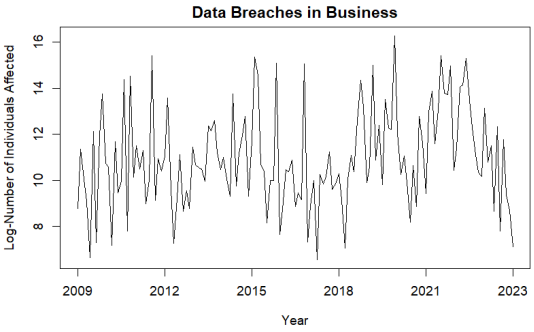

Figure 2: Time series data of log-number of individuals affected by business-associated data breaches aggregated monthly.

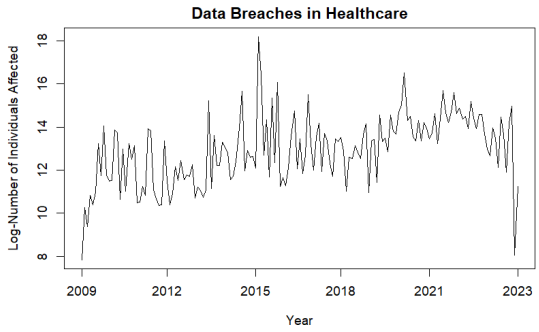

Figure 3: Time series data of log-number of individuals affected by healthcare-associated data breaches aggregated monthly.

4. Data analysis

Following Kim and Song (2023), we fit ARMA(1,1) to explain the conditional mean and GARCH(1,1) to describe the conditional variance. Let X_t be the monthly loss at time t. We have

$$X_t = \mu_t + \sigma_t \epsilon_t, \tag{1}$$
$$\mu_t = \phi_0 + \phi_1 X_{t-1} - \phi_1 a_{t-1}, \tag{2}$$
$$a_t = \sigma_t \epsilon_t, \tag{3}$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{4}$$

where ϵ_t is a white noise with zero mean and unit variance μ_t is the mean term of the time-series and σ_t^2 is the conditional variance. We fit the monthly HHS data to an ARMA(1,1)-GARCH(1,1) (see Figures 4 - 6). In particular, we see that this model is a good fit and models the stochastic behavior well. Furthermore, we construct 1% VaR intervals for the time series (see Figures 4 - 6). Here, we see that for each given time period a 99% probability that the actual loss will not exceed the VaR limit over the specified time horizon. Furthermore, the number of individuals affected in the business-associated data (see Figure 4) is comparatively lower than that of the healthcare (see Figure 6). This could indicate that the distribution of records in the healthcare sector is more widespread, with a healthcare provider such as a hospital having access to information on a larger pool of patients than a business associated with the provider, which may only have access to a subset of the data. Therefore, the monthly number of individuals affected showed characteristics of asymmetry depicted between business and healthcare-associated.

5. Conclusion

In this paper, we investigated one of the more recent papers on modeling the historical trend of cyber risk. In the paper by Kim and Song (2023), they leveraged the SAS OpRisk Global Database to model the financial losses for financial and non-financial companies across time. To accomplish this, they used the loss distribution approach coupled with the peak over threshold technique to model the long tail of the losses. Furthermore, they fitted an ARMA(1,1)-GARCH(1,1) model for the historical trend. In light of these techniques, we utilized these approaches to model the cyber risk losses presented in healthcare. Using the U.S. Department of Health and Human Resources Office for Civil Rights database, we fit the data to an ARMA(1,1)-GARCH(1,1) model, which provides a reasonable fit.

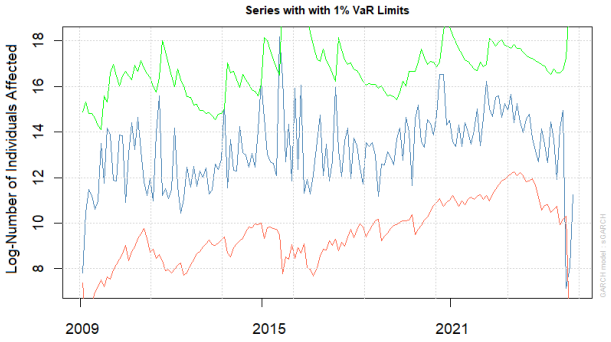

Figure 4: ARMA(1,1)-GARCH(1,1) fit to time series data of log-number of individuals affected by data breaches aggregated monthly 1% VaR limit intervals.

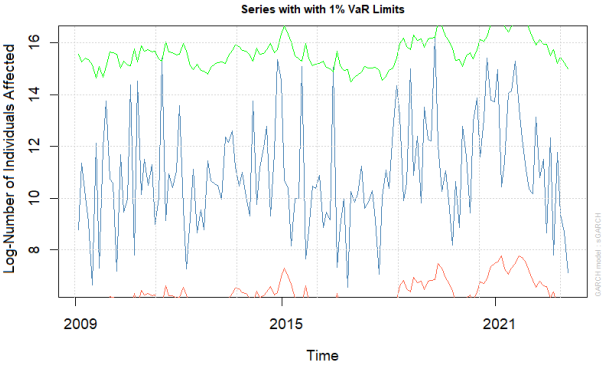

Figure 5: ARMA(1,1)-GARCH(1,1) fit to time series data of log-number of individuals affected by business-associated data breaches aggregated monthly with 1% VaR limit intervals.

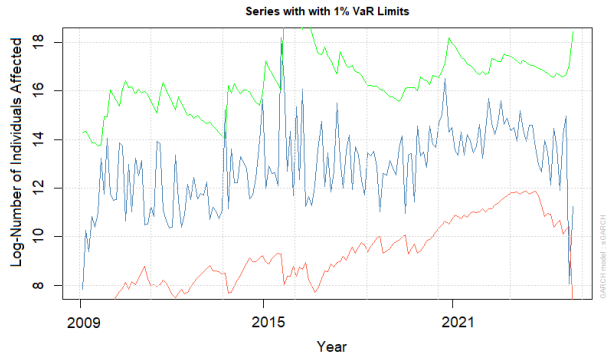

Figure 6: ARMA(1,1)-GARCH(1,1) fit to time series data of log-number of individuals affected by healthcare-associated data breaches aggregated monthly 1% VaR limit intervals.

References

Iñaki Aldasoro, Leonardo Gambacorta, Paolo Giudici, and Thomas Leach. The drivers of cyber risk. BIS Working Papers No.865, May 2020. URL `https://www.bis.org/publ/work865.htm`.

Christian Biener, Martin Eling, and Jan Hendrik Wirfs. Insurability of cyber risk: An empirical analysis. *Geneva Papers on Risk and Insurance- Issues and Practice*, 40(1):131–158, 2015.

Paul Bischoff. Ransomware attacks on u.s. healthcare organizations cost $20.8b in 2020. Comparitech, march 2021. URL `https://www.comparitech.com/blog/information-security/ransomware-attacks-hospitals-data/`.

Maria Francesca Carfora and Albina Orlando. Quantile based risk measures in cyber security. In *2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–4. IEEE, 2019.

Diane Dolezel and Alexander McLeod. Cyber-analytics: identifying discriminants of data breaches. *Perspectives in Health Information Management*, 16 (Summer), 2019.

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

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

Martin Eling and Jan Wirfs. What are the actual costs of cyber risk events? *European Journal of Operational Research*, 272(3):1109–1119, 2019.

Sébastien Farkas, Olivier Lopez, and Maud Thomas. Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics & Economics*, 98:92–105, 2021.

IBM Security. Cost of a data breach report 2020, 2020. URL `https://www.ibm.com/security/data-breach`.

Petar Jevtić and Nicolas Lanchier. Dynamic structural percolation model of loss distribution for cyber risk of small and medium-sized enterprises for tree-based LAN topology. *Insurance Mathematics & Economics*, 91:209–223, 2020.

Sanghee Kim and Seongjoo Song. Cyber risk measurement via loss distribution approach and garch model. *Communications for Statistical Applications and Methods*, 30(1):75–94, 2023.

Yuying Li and Rogemar Mamon. Modelling health-data breaches with application to cyber insurance. *Computers & Security*, 124:102963, 2023.

Ponemon Institute. Cost of a data breach report 2021, July 2021. URL https://www.ibm.com/security/data-breach.

Jay G Ronquillo, J Erik Winterholler, Kamil Cwikla, Raphael Szymanski, and Christopher Levy. Health it, hacking, and cybersecurity: national trends in data breaches of protected health information. *JAMIA open*, 1(1):15–19, 2018.

SonicWall. Mid-year update: Sonicwall cyber threat report, June 2021. URL https://www.sonicwall.com/2021-cyber-threat-report/.

The Institute of Risk Management. Cyber risk and risk management, 2018. URL https://www.theirm.org/what-we-say/thought-leadership/cyber-risk/.

U.S. Department of Health and Human Services. Breach portal: Notice to the secretary of hhs breach of unsecured protected health information, 2023. URL https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.