Modeling the Risk of Data Breach Incidents at the Firm Level

Kazuki Ikegami, Hiroaki Kikuchi

Abstract Many firms and organizations are at risk of cyberattack nowadays. For example, in 2018 alone, 443 data breaches in Japan compromised some 5.61 million records of personal information. To respond to this threat, firms asset a risk of cybersecurity and introduce IT security management practices. However, it is unclear whether firms are able to identifying the tradeoff between effect of development of IT security practices and the risk of data breach. To address this, we propose a probabilistic model that estimates the risk of a data breach for a given firm using the Japan Network Security Association incident dataset, being a historical collection of cyber incidents from 2005 to 2018. This model yields the conditional probabilities of a data breach given conditions, which follows a negative binomial distribution. We highlight the difference in inter-arrival time between firms with security management and one without it. Based on the experimental results, we evaluate effects of security management and discuss some reasons for these differences.

1 Introduction

Data breaches result from several causes, including malicious hacking, malware, and insiders. The Japan Network Security Association (JNSA) reports that in 2018 alone, a total of 5.61 million items of personal information, e.g., personal identifiable information and personal financial data, were compromised in some 443 cyber incidents in Japan[1]. The average damage from these incidents was approximately $6 million and this is increasing by some $860,000 dollars compared to

Kazuki Ikegami,
Graduate School of Advanced Mathematical Sciences, Meiji University, 4-21-1 Nakano Tokyo Japan, e-mail: cs192021@meiji.ac.jp

Hiroaki Kikuchi
School of Interdisciplinary Mathematical Sciences, Meiji University, 4-21-1 Nakano Tokyo Japan
e-mail: kikn@meiji.ac.jp

2017. In 2015, the Ministry of Economy, Trade and Industry (METI) published Cybersecurity Management Guidelines[2] as a way to help address this problem. These guidelines suggest three principles that every manager needs to know and the ten important security items every executive officer should have to protect the firm against cyberattack. The guideline includes some tips in recognizing risk and building mechanisms to mitigate damage damages. In addition, cyber insurance has been available from most major insurance firms since 2015. Insurance coverage can help reduce the cost of damage caused by cyber incidents. Unfortunately, the coverage of cyber insurance in Japan 17.2% is quite low compared to many other countries, as noted in a report by IDC Japan[3]. One possible reason may that managers do not always perceive the of cyberattack precisely and this can result in an overestimation of the required IT security countermeasures and underestimation of the need for cyber insurance[4].

In terms of related research, Edwards et al. revealed the tendency in data breaches in the US[5]. In [6], Kokaji et al. argue that security in financial institutions is preserved by not only general risk assessment, but also by cybersecurity assessment techniques, e.g., accurately estimating the probability of their occurrence, while in [10], Yamada et al. estimate the effect of security management based on the estimated probability of cyber incidents.

Unfortunately, these studies mostly focus on measuring the total risk of whole organizations and mainly modeled several incidents aggregated over various industries and types of security management. Accordingly, the results of these models cannot apply to a particular firm because the aggregated data is too general to model the risk of the given firm. Accordingly, in this study, we aim to reveal the risk of cyber incidents specific to a given organization and to quantify the effect of security management in reducing this risk. The following questions motivate our analysis:

- What is the probability that an incident will occur at an organization in one year?
- How long does it take before the next incident will occur at the organization?
- How much is the inter-arrival time of incidents reduced by security management?

To respond to these questions, we quantify the risk of cyber incident using the probability of inter-arrival time for the organization modeled as a negative binomial distribution (NBD). We then quantify the effect of security management by fitting the incident inter-arrival time into a generalized linear model.

2 Related works

Maillart and Sornette studied the statistical properties of the personal identity losses in the United States between year 2000 and 2008[7]. They found that the number of breach incidents dramatically increases from 2000 to July 2006 but remains stable thereafter. Wheatley et al. analyzed organizational breach incidents between year 2000 and 2015[8]. They found that the frequency of large breach incidents (i.e., the ones that breach more than 50,000 records) occurring to US firms is independent of

time, but the frequency of large breach incidents occurring to non-US firms exhibits an increasing trend. Martin et al. analyzed 1,579 cyber risk incidents extracted from an operational risk dataset[**?**]. They identify cyber risks of daily life and extreme cyber risks by using a new version of the peaks-over-threshold method from extreme value theory. Edwards et al. modeled the trends in cyber incidents in the US by using 2,234 separate incidents occurred from 2005 to 2015 stored in the public dataset of the Privacy Right Clearing house[5]. They employ Bayesian generalized linear models to model the number of victims in an incident and the frequency of data compromise. In [11], Ravi et al. apply the opportunity theory of crime, institutional anomie theory, and institutional theory to clarify what factors affect data breaches. They reveal that investment IT security correlates with a high risk of data breach. Elsewhere, Martin et al. use multidimensional scaling and goodness-of-fit tests to analyze the distribution of data breaches and link the model with the current discussion on goodness of fit, pricing, and risk measurement in the actuarial domain[12]. In [13], Maochao et al. investigate cyber-hacking breach incident inter-arrival time and breach size and propose some stochastic processes to predict both the inter-arrival time and the breach size.

Later, Romanosky et al. examine the extent to which identity theft decreased following the introduction of data breach disclosure laws using panel data from the US Federal Trade Commission from 2002 to 2009[14]. In [10], Yamada et al. revealed the effect of security management on incident occurrence. Using the logistic regression, they quantify the effect of security management while controlling for confounding factors like industry domain and firm scale. Their findings reveal that having a Chief Information Officer (CIO) reduces the probability of incidents by 30%.

3 Preliminary

3.1 Probability distribution

3.1.1 Negative binomial distribution

The number of times we flip a coin until it comes up heads is distributed according to the following probability mass function, known as the *NBD* $Pr(X = x) = \binom{x+r-1}{x} p^r (1-p)^x$ where X is the number of failures (incident inter-arrival time), r is the number of successes, and p is the probability of success (incident occurs).

The mean (expectation) of NBD (mean inter-arrival time) is given by $\mu = \frac{(1-p)r}{p}$

3.1.2 Poisson distribution

A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$, if, for $k = 0, 1, 2, \ldots$, the probability mass function of X is given by $Pr(X = x) = \frac{\lambda^k e^{-\lambda}}{k!}$ The Poisson distribution is widely used for modeling the number of times an event (incident) occurs.

3.1.3 Normal distribution

The general form of the normal probability distribution function is $Pr(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ where μ is a mean or expectation of the distribution (and also its median and mode); and $\sigma^2 > 0$ is its standard deviation.

3.2 Kolmogorov–Smirnov test

The Kolmogorov–Smirnov (KS) test is widely used to test if a reference probability distribution is correctly modeled for a given sample by providing a test statistic of the difference between the empirical distribution function obtained from the sample data and an estimated distribution function to be tested[15]. Let m be the size of the sample chosen from d, being a random variable following an unknown distribution function $F(D)$. Let $F_0(D)$ be an estimated distribution function based on the given sample. The null hypothesis is that the estimated distribution is identical to the unknown distribution, i.e., $H_0 : F(D) = F_0(D)$ The empirical distribution function F_m is defined by $F_m(D) = 0(D < d_1), i/m(d_{(i)} \leq D < d_{(i+1)}, \quad i = 1, \ldots, m-1), 1(D \geq d_{(m)})$ where $d_{(1)}, \ldots, d_{(m)}$ are sampled D. As m increases, the empirical distribution function $F_m(D)$ approaches to the true distribution $F_0(D)$. Therefore, the empirical distribution $F_m(D)$ is close to the true distribution $F_0(D)$. The test statistic K_m examines the distance between the estimated distribution function $F_0(D)$ and the empirical distribution function $F(D)$ is given as $K_m = \sup_D |F_m(D) - F_0(D)|$.

3.3 The generalized linear models

The generalized linear model allows to a nonlinear function to be handled as easily as a linear model and thereby extends the normal distribution to the family of exponential distributions[16]. In generalized linear models, the objective variable is not limited to quantitative data but also includes boolean values. In the linear model, $Z = \beta X + \alpha$ where Z is objective variable, Xs are explanatory variables, and α is a constant and β are the corresponding coefficients. In the generalized linear model,

a nonlinear function is converted to $g(\mu) = \beta X$ and treated as a linear model, where μ is the average of value for objective variable and g is a link function.

4 Data

4.1 JNSA dataset

The JNSA collects cyber incident information from Internet news sites and major press releases officially published each year since 2005[1]. The dataset classifies the collection of cyber incidents into several categories in terms of the number of victims, the causes of the incident, e.g., *Lost*, *Theft*, *Malware*, and ways of the data breach, e.g., *Paper media* or *Internet*. These categories help business managers to plan and revise their security measures. Over the period 2005 to 2018, the JNSA collected data on cyber incidents that occurred at 9,358 organizations representing some 16,392 data breaches in total.

Table 2 provides statistics for the number of cyber incidents occurring in the one organization using this dataset. The maximum number of incidents per organization was 195 incidents for Osaka City Hall. Table 1 details the distribution of the number of organizations sorted by the number of incidents from 2005 to 2018. Note that the total number of cyber incidents in Table 1 is less than 16,392 because multiple incidents in the one day are counted as the one incident. Over 85% of organizations reported one incident over the 13-years sample period. However, more than 40% of all incidents with two or more cyber incidents took place in just 1,386 organizations (15% of the total).

4.2 CSR dataset

Toyo Keizai Inc. conduct a survey about corporate social responsibility (CSR) for listed firms and major unlisted firms every year[17]. The CSR dataset comprises records of queries classified across three categories. The first category is "Workforce," which includes the number and average age of employee'. The second is "CSR overall," which includes the information security management system (ISMS) certification and the CIO, etc. The third is "Environment," which includes the carbon dioxide emission rate and an estimate of environmental conservation costs, etc.

The style of response varies across these questions. For example, in response to the question "Is an internal audit performed?", there are multiple responses, including "1. Perform regularly. 2. Perform occasionally ...." In this study, we combine similar responses and reclassify all questions as having a Yes or No response. We then select 17 questions related to IT security from the approximately 800 questions in the CSR dataset. Table 3 provides the statistics for CSR 2017 dataset.

Table 1 The number of data breaches in the one organization over 13 years

# data beaches	# organizations n (rate)	# data beaches B (rate)
1	7,972 (0.85)	7,972 (0.57)
2	757 (0.08)	1,514 (0.11)
3	238 (0.03)	714 (0.05)
more than 4	391 (0.04)	3,789 (0.27)
total	9,358 (1.00)	13,989 (1.00)

Table 2 Number of data beaches per organization from 2005 to 2018 ($N = 9,358$)

average	variance	max	min	total
1.5	12	195	1	16,392

Table 3 Statistics for CSR data

period	firms	questions	Common firms with JNSA
2017	1,414	840	223

5 Proposed model

Fig. 1 illustrates the possible incidents. Suppose that four cyber incidents occur in an organization at time t_1, t_2, t_3, and t_4, respectively. Let d_i be the i-th *inter-arrival time* (in days) between two consecutive cyber incidents in organization j. That is, the inter-arrival time leads to a time series $d_1 = t_1 - t_2$, $d_2 = t_3 - t_2$, and $d_3 = t_4 - t_3, \ldots 0$ For the example in Fig. 1, we have inter-arrival times of $d_1 = 130$, $d_2 = 6$, and $d_3 = 32$.

We use at least two d_i for learning and one for evaluation to evaluate inter-arrival time model. In Fig. 1, d_1, d_2 are for learning and d_3 is for evaluation. Under this condition, we investigate 391 organizations that have experienced more than three cyber incidents and identify the probability distributions using the maximum likelihood estimation.

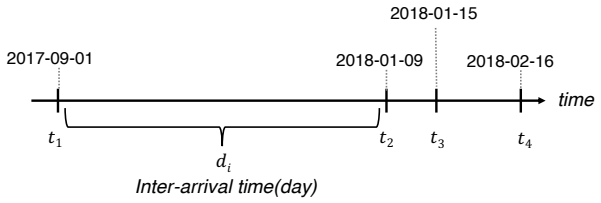

Fig. 1 inter-arrival times of data breaches

We model the series of inter-arrival time D for each of the organizations using NBD

$$D \sim F_{NB}(\mu, r)$$

where μ is the mean of the distribution. The mean μ varies across organizations. Hence our model specifies μ by

$$\mu = \mathrm{e}^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{19} x_{19}}$$

where x_1 is a dummy variable indicating each of 16 major industries and x_2 is the log firm size (number of employees). $x_3, \cdots, x_{19}$ are explanatory variables which indicates to implementation of the corresponding security management or countermeasure in Fig. 10. Implicitly, α is a constant and βs are coefficients of variable.

6 Experiment

6.1 overview

In this study, to reveal risk of cyber incidents specialized for given organizations, we investigate 391 organizations that have caused more than three cyber incidents and modeled their inter-arrival time by negative binomial to estimate parameters for given inter-arrival time by the maximum likelihood estimation. Mostly, we use `fitdistr()` for the estimation. And, we tested some candidate distributions to evaluate accuracy of the model. Then, by using those models, we predict an inter-arrival time and evaluate it. Finally, we try to figure out the reduction of risk by using general liner model. we use `glm()` for the estimation.

6.2 Fitting results

Fig. 2 plots three cumulative probability distributions of cyber incidents fitted for three sample organizations selected from the 391 models. In this figure, the black and red lines represent the empirical cumulative distribution and the estimated NBD, respectively.

Closely looking at the estimated NBD for Tokyo Electric Power Company Holdings Inc. in Fig. 2 (in the middle), we model the occurrence of cyber incidents using the NBD with $\mu = 284$ and $r = 0.56$. The model allows us to predict the probability that the Tokyo Electric Power has a cyber incident in a given year, i.e., $d = 365$ such that $Pr[D \leq 365] = 0.74$, implying that *the an incident occurs within 365 days with a probability of 74%*. Table 4 provides statistics for $\Pr[D \leq 365]$. Note that we compensate for the number of organizations before we calculate the average probability of inter-arrival time because many organizations have never reported an incident.

Among organizations in the 2017 CSR dataset, 17% (223/1351) of organizations report one incident over the 13-years sample period. Using these sample statistics in the CSR dataset, we estimate the number of organizations n' that have never reported incidents with $n' = n \cdot 1351/223 = 1955$ where n is the number of organizations that

have had one data breach in the period. Concretely, the average probability D^* an organization has a data breach within one year is given as

$$\begin{aligned} Pr[D^* < 365] &= E[Pr[D_1 < 365 \text{ OR } D_0 < 365]] \\ &= E[Pr[D_1 < 365|Z = 1]Pr[Z = 1] + Pr[D_0 < 365|Z = 0]Pr[Z = 0]] \\ &= Pr[Z = 1]0.55 + Pr[Z = 0]0 = 0.11 \end{aligned}$$

where Z is a random variable indicating that the organization at least four incidents over the 13 years, and D_1 and D_0 are random variables representing the inter-arrival times of both organizations.

Table5 lists partially the estimated parameters μ, r sorted by μ. The medians of these parameters are 257 and 1.07, respectively. The average inter-arrival time μ varies by up to 94 times among organizations.

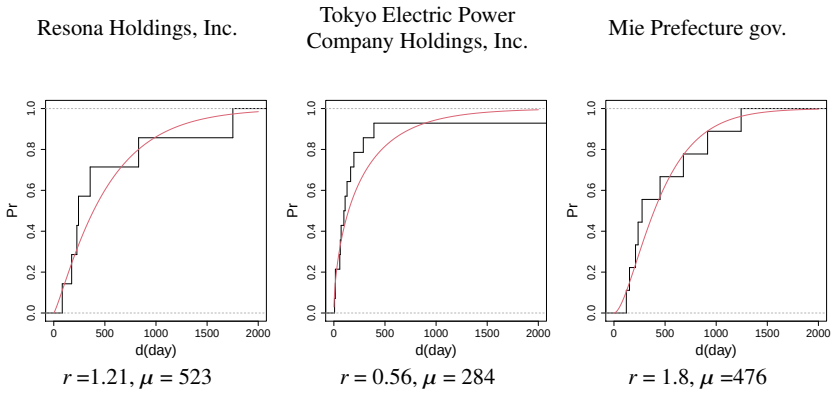

Fig. 2 The cumulative probability distribution of cyber incidents estimated using the NBD

Table 4 Statistics for the probabilities of a data breach occurring within a year $Pr[D \leq 365]$, $n = 391$

average	max	min	standard deviation
0.11	1	0	0.27

Table 5 Estimated parameters of incident model (partial)

name of organization	r	μ
Mitsui Fudosan Residential Co.,Ltd.	2,436	19
Asahikawa Shinkin Bank	18	20
Osaka city	0.59	25

6.3 Model evaluation

How accurate are our model estimated? Table 6 provides results for the KS tests that the given series of inter-arrival time fits the selected candidate probability distributions, including the NBD (nbinom), the Poisson distribution (pois), and the Normal distribution (norm). Fig. 3 illustrates the result of fitting the empirical distribution of cyber incidents for the Tokyo Gas Co., Ltd. Looking at the Fig. 3, we can see that the NBD model is distributed closest among the three candidates to the given distribution. The pvalue of NBD is 0.0963, which is too high to reject the null hypothesis that the data were generated using this distribution. The KS tests of the other two distributions yield pvalues of = 0.0000 < 0.05, and 0.0008 < 0.05, which tells us that the observed series were unlikely to have been generated by either distribution. Consequently, we accept only the NBD model for the observed series of cyber incident events.

Table 7 details the number of organizations for which each of the probability distributions rejects the null hypothesis. As shown, the NBD rejects the null hypothesis for the smallest number of organizations.

Table 6 Results of KS test (p-values)

	nbinom	pois	norm
Tokyo gas Co., Ltd	0.0963	0.0000	0.0008
NTT West	0.0892	4.26E-14	0.0065
UR Agency	0.088	8.82E-14	0.0004

Table 7 The share of organizations rejected by the null hypothesis

nbinom	pois	norm
0.02	0.39	0.08
(9/391)	(155/391)	(31/391)

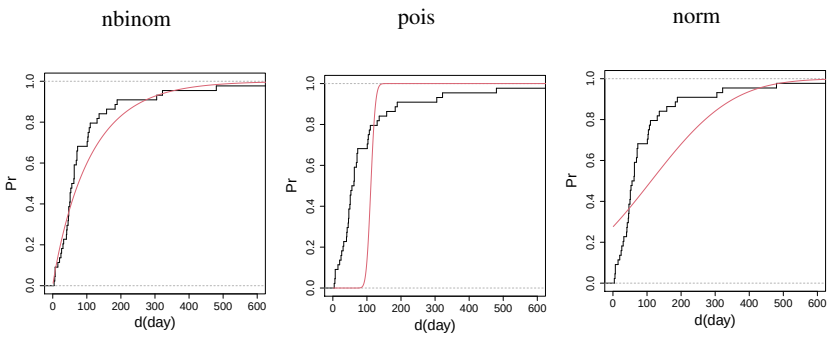

Fig. 3 Cumulative probability distribution of cyber incidents in Tokyo Gas Co., Ltd fitted with different some probability distributions

6.4 Accuracy of prediction

In this analysis, we predict when a future cyber incident occurs using the NBD model. We evaluate the accuracy of the prediction inter-arrival time with $Pr = 0.7$ as the threshold. Fig. 4 illustrates how the correctness of prediction is judged. When a predictive inter-arrival time $\hat{d}_i$ is longer than the actual inter-arrival time d_i, we regard the prediction as correct, i.e., $\hat{d}_i \geq d_i$. In the opposite case, a prediction is a failure (a miss), when $\hat{d}_i < d_i$. Further, error is defined as the difference between the correct predictive inter-arrival time and a the true (observed) inter-arrival time, i.e., $|\hat{d}_i - d_i|$. Table 8 provides the estimated inter-arrival time (in days) and the recall (the share of correctly estimated organizations in all organizations) for a threshold $Pr = 0.7$ for all industries.

The average estimated inter-arrival time for the 391 organizations is 426 days and the recall is 55%. The maximum and the minimum recalls are 64% (*Government Services*) and 17% (*Services*) in industries except some too-small categories.

Table 9 shows the top three estimated inter-arrival times $\hat{d}_i$ ($Pr = 0.7$). The maximum error in organizations is 1,938 days.

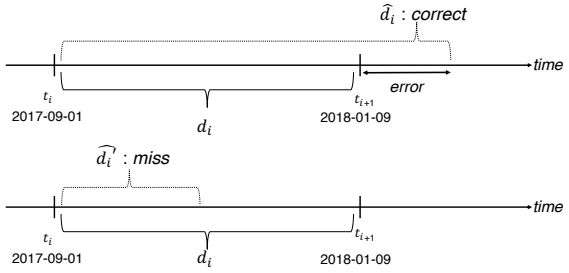

Fig. 4 Explanation of judge of prediction

Fig. 5 plots the distribution of the number of organizations in terms of the magnitude of the mean error (in days) for selected thresholds ($Pr = 0.6, 0.7, 0.8, 0.9$). As shown, half of the organizations are estimated within errors of 365 days at a threshold of 60–80%. However, there are some exceptional organizations with errors of more than 1,000 days. With a threshold of 70%, the number of organizations with errors of 365 days is at a maximum of 135, and is the highest among most other thresholds.

Table 8 Average estimated inter-arrival time and recalls

Industry	Average of predict inter-arrival-time	Recall (correct organization/ all organization)
Multi Service	364	1.00 (2/2)
Forestry	177	1.00 (1/1)
Government Services (Not Otherwise Categorized)	439	0.64 (103/162)
Telecommunications	641	0.61 (19/31)
Education/Learning Support	777	0.57 (20/35)
Finance/Insurance	614	0.53 (31/58)
Real Estate	244	0.50 (6/12)
Construction	297	0.50 (3/6)
Manufacturing	349	0.50 (2/4)
Wholesale/Retail	513	0.44 (4/9)
Health Care/Welfare	341	0.39 (12/31)
Utilities: Electricity, Gas, Heat, Water	507	0.35 (8/23)
Transportation	388	0.33 (1/3)
Services (Not Otherwise Categorized)	394	0.17 (2/12)
Hospitality (Restaurant/Hotel)	348	0.00 (0/2)
Total	426	0.55 (214/391)

Table 9 The predict inter-arrival time for each organization $\hat{d}_i$

	Organization name	$\hat{d}_i$
(Pr=0.7)	Mitsui Fudosan Residential Co.,Ltd.	23
	Asahikawa Shinkin Bank	24
	Osaka City	28

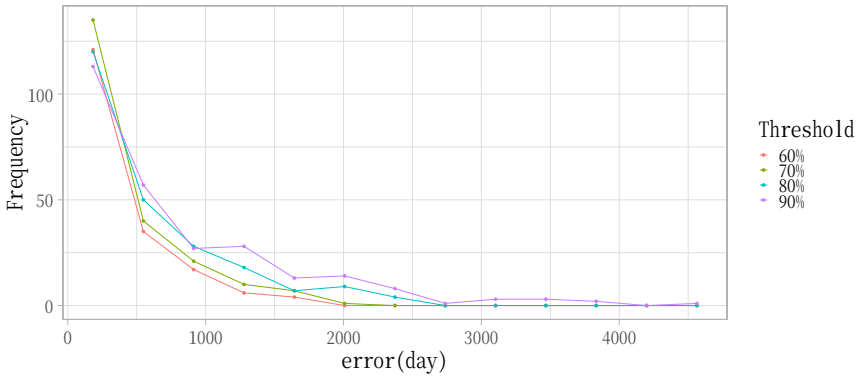

Fig. 5 Distribution of estimated error in days

6.5 Effect of security manage countermeasures in inter-arrival time

We now calculate the reduction of risk by means of the estimated inter-arrival time (the longer the better). Let mu_ℓ^+ and μ_ℓ^- be the mean inter-arrival time with/without security management (countermeasures), and *CIO*. The effect of security management *ell* is quantified by the ratio of two mean inter-arrival times, shown as,

$$\frac{\mu_\ell^+}{\mu_\ell^-} = \frac{\mathrm{e}^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{\ell-1}x_{\ell-1}+\beta_\ell x_\ell+\beta_{\ell+1}x_{\ell+1}+\cdots}}{\mathrm{e}^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{\ell-1}x_{\ell-1}+\beta_{\ell+1}x_{\ell+1}+\cdots}} = \mathrm{e}^{\beta_l}.$$

We show the results of the generalized linear model (glm) in Table 10. In this study, we compensate the inter-arrival times d_i (d_i') of organizations i that have never had a data breach with $d_i = 4,745$ ($d_i' = 4,270$) days for 13years, which is the duration of the whole observation period (the maximum inter-arrival time among the observations). A larger coefficient β_ℓ means that the predicted inter-arrival time $\hat{d}$ increases when security management ℓ is deployed in an organization. For example, the mean inter-arrival time is predicted as 0.9 times longer when there is an *External Audit* ($x_{13} = 1$). Our study shows that nine of the 17 management items have a positive effect in increasing the inter-arrival time, namely, lowering the risk of cyber incidents.

6.6 Discussion

In this study, we model the probability of a cyber incident using an NBD model fitted for organizations with four or more data breaches over 13 years (2005–2018). Fig. 6 plots the distribution of the number of data breaches in terms of error size at a threshold of 70%. In general, the error is likely to be large when the number of inter-arrival times used for learning is limitedṪhe size of the error ranges from 1 to 4,279 days. Therefore, we maintain that cyber incident occurrences exhibit periodic behavior for some reason. Fig. 7 depicts the scatterplot of NBD parameters over μ and r. We observe that the inter-arrival time d is independent of the size of the model. We do not find any correlation between *mu* and r. Note that these results are limited because we used publicly collected information datasets.

Let us remark upon the divergence of incidents across industries. The frequency of cyber incidents varies widely according to the business type. For example, companies classified as "business-to-customer" (B2C) face a higher risk of privacy breaches. For instance, the inter-arrival time in Electric Power industries is the longest for the glm result, informing us that it has a significant effect in reducing data breaches.

Table 10 Effect of Security Management (countermeasure)

	Security Management ℓ	Estimate	Std. Error	P-value
	(Intercept)	8.96	0.12	< 2e−16 ***
x_1	Medicine	−0.14	0.13	0.26
	Transportation & Logistics	−0.07	0.12	0.55
	Machinery	−0.04	0.12	0.72
	Financials (excl. banks)	−0.31	0.13	0.01 *
	Bank	−1.06	0.14	0.00 ***
	Construction & Materials	−0.37	0.12	0.00 **
	Automobiles & Transportation Equipment	0.00	0.12	0.98
	Commercial & Wholesale Trade	−0.12	0.12	0.31
	Retail Trade	−0.30	0.12	0.01 *
	ITServices, Others	−0.18	0.12	0.12
	Food	−0.02	0.12	0.87
	Raw Materials & Chemicals	−0.05	0.12	0.66
	Steel & Nonferrous Metals	−0.03	0.13	0.84
	Electric Appliances $ Precision Instruments	−0.08	0.12	0.49
	Electric Power & Gas	−2.24	0.29	0.00 ***
	Real Estate	−0.46	0.13	0.00 ***
x_2	LOG(# employee)	−0.07	0.01	$< 2e-16$ ***
x_3	ISMS	0.04	0.03	0.18
x_4	CIO	−0.07	0.03	0.01 **
x_5	CFO	0.01	0.03	0.64
x_6	External Report Window	0.01	0.02	0.70
x_7	Internal Report Window	−0.07	0.05	0.14
x_8	Whistleblower Rights Protection	0.06	0.05	0.24
x_9	Establishment of Internal Control Committee	−0.01	0.02	0.65
x_{10}	Privacy Policy	0.00	0.03	0.98
x_{11}	Security Policy	−0.01	0.04	0.79
x_{12}	Internal Audit	0.01	0.03	0.75
x_{13}	External Audit	−0.07	0.02	0.00 **
x_{14}	Independent Internal Audit Department	0.02	0.04	0.61
x_{15}	Establish a Risk Management/Crisis Management System	0.03	0.04	0.42
x_{16}	Basic Risk and Crisis Management Policy	−0.08	0.04	0.03 *
x_{17}	Conduct Environmental Audits	−0.03	0.04	0.33
x_{18}	Establish Environment Management	0.10	0.03	0.01 **
x_{19}	Building an Occupational Health and Safety Management System	0.00	0.02	0.98

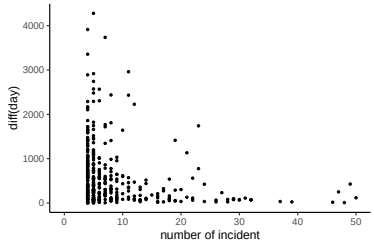

Fig. 6 Distribution of estimated errors in terms of number of incidents

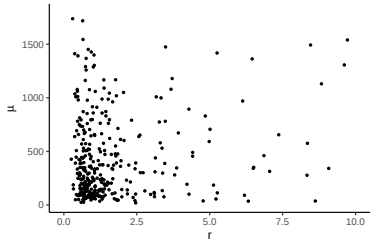

Fig. 7 Scatterplot of NBD models

7 Conclusion

We conclude our study with some new findings from our analysis. Our model reveals that there is an average probability of 0.11 of an organization suffering a data breach in a year. The minimum number of days until the next data breach is then 23 days on average across the 391 organizations in our sample, with an average duration of a cyber incident being 426 days. The effect of security management to inter-arrival-time is that the inter-arrival time is 0.9 times shorter when an external audit

is conducted. In terms of limitation, there is some inconsistency in our data sets in that we fit the CSR data from 2017 with data breaches over the period 2005–2018.

Our future research will focus on improving the prediction accuracy and revising the incident model that considers management changes over the year.

References

1. Information security incident survey report (JNSA dataset), Japan Network Security Association(2018).
2. Cybersecurity Management Guidelines Ver1.1. In: Ministry of Economy, Trade and Industry Home page.
   https://www.meti.go.jp/policy/netsecurity/downloadfiles/CSM_Guidelines_v1.1_en.pdf.
   Cited 19 Mar
3. What is cyber insurance, In: cyber-hoken.com.
   https://cyberhoken-jp.com/cyber-hoken. Cited 19 Mar
4. Sakuma J, Inomata A (2019), Proposal for Improvement of Penetration on Investigation and Analysis of Cyber Insurance. Internet and Operation Technology (IOT) (IPSJ), 1–8
5. Edwards B, Hofmeyr S, and Forrest S (2016), Hype and heavy tails: A closer look at data breaches. Journal of Cybersecurity, **2**:3–14
6. Kokaji A, Harada Y, Goto A (2019), A consideration of assessment on cyber security in financial institutions. Electronic Intellectual Property (IPSJ), 1–5
7. T. Maillart, D. Sornette (2010), Heavy-tailed distribution of cyber-risks, Eur. Phys. J. B, **75**:357–364
8. Wheatley S, Maillart T, and Sornette T (2016), The extreme risk of personal data breaches and the erosion of privacy, The European Physical Journal B, **89**
9. Eling M, Loperfido N (2019), What are the actual costs of cyber risk events? European Journal of Operational Research, **272**:1109–1119
10. Yamada M, Ikegami K, Kikuchi H, Inui K (2018), Assessment of the effect of decreasing data breach by the management situation (2). Computer Security Symposium (CSS2018), 376–384
11. Sen R, Borle S (2015), Estimating the Contextual Risk of Data Breach: An Empirical Approach. Journal of Management Information Systems, **32**:314–341
12. Eling M, Loperfido N (2017), Data breaches: Goodness of fit, pricing, and risk measurement. Insurance:Mathematics and Economics. **75**:126–136,
13. Xu M, Schweitzer K, Bateman R, Xu S (2018), Modeling and Predicting Cyber Hacking Breaches. IEEE Transactions on Information Forensics and Security, **13**:2856–2871
14. Romanosky S, Telang R and Acquisti A (2011), Do data breach disclosure laws reduce identify theft? Journal of Policy Analysis and Management, **30**:256–286
15. Massey F, Jr (1951). The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc, **46**:68–78
16. Nelder, J. A., and R. W. M. Wedderburn (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), **135**:370–384,
17. CSR DATA, Toyo Keizai Data Services (2017).