Summer 8-4-2022

Debiasing Cyber Incidents – Correcting for Reporting Delays and Under-reporting

Seema Sangari

Debiasing Cyber Incidents — Correcting for Reporting Delays and Under-reporting

A Dissertation Presented for the

PhD Analytics and Data Science

Degree

Kennesaw State University, Georgia

Seema Sangari

July 2022

Dedicated to my loving parents

Acknowledgments

The completion of any academic work is inconceivable without the expert guidance of a mentor, I am lucky to have not one but many during my PhD journey. First and foremost, I take this opportunity to express my deep sense of gratitude to my advisor Dr. Eric Dallal who made me think thoroughly about the problem, correcting and bringing back on track, mentoring me to develop skills in research, and always being there to support. This research would not have been possible without his guidance.

Secondly, I would like to thank my chair Dr. Michael Whitman who courteously agreed to come on board at a later stage, guided me to ensure my degree is on track and was generous with his time to provide his valuable feedback.

Thirdly, I am grateful to Dr. Jennifer Priestley for providing valuable feedback despite her busy schedules, always being there to support since the beginning of the program, and delaying her retirement to be part of my committee.

Additionally, I would like to thank Dr. Xinyan (Abby) Zhang for being supportive and approachable.

Special thanks to Dr. Sherrill Hayes who played an instrumental role in recruiting me into the PhD program and has always been so encouraging; Dr. Xuelei (Sherry) Ni for being supportive and ensuring that I am on track to graduate on time; Dr. Herman Ray for placing me to the labs where I learned and built my knowledge; and Professor Bill Franks for helping me to improve my presentation skills.

I would like to thank all the professors and academic staff who taught me during PhD experience and laid the foundation which made me achieve this degree.

I would like to thank Ms. Cara Reeve, our program administrator, for helping me with all the logistics and being so quick to respond, and library and computing staff for their courteous disposition and for affording me the necessary facilities.

I would like to thank my cohort members and fellow PhD students who were always available to discuss anything and everything, providing feedback, asking questions, and bouncing new ideas.

I am deeply grateful to my parents and sisters who inculcated the spirit of self confidence in me, gave me enough opportunity to undertake the course of my choice, and stood by me through thick and thin. Without their encouragement and support, I would not be here.

The research is done in collaboration with Verisk Extreme Event Solutions using their proprietary cyber data. I would like to express my gratitude to Scott Stransky for giving me the opportunity and Verisk for providing support and resources to complete this research.

Abstract

This research addresses two key problems in the cyber insurance industry – reporting delays and under-reporting of cyber incidents. Both problems are important to understand the true picture of cyber incident rates. While reporting delays addresses the problem of delays in reporting due to delays in timely detection, under-reporting addresses the problem of cyber incidents frequently under-reported due to brand damage, reputation risk and eventual financial impacts.

The problem of reporting delays in cyber incidents is resolved by generating the distribution of reporting delays and fitting modeled parametric distributions on the given domain. The reporting delay distribution was found to be non-stationary and bimodal. While non-stationarity was handled by generating the monthly reporting delay distribution over the rolling two-year moving window, the bimodal aspect required an optimization algorithm to compute the parameters. The modeled parametric distribution is further extended to infinite domain to obtain the complete overview of the incidents occurred but not yet reported. The complete modeled parametric distribution provides the correction factors showing an increasing trend in recent months rather than a decline as observed from reported incidents. The correction of reporting delays is computed for the US market. This research was published as "Correcting for Reporting Delays in Cyber Incidents" in JSM Proceedings [119]. The study is further extended to highlight how reporting delays vary from industry to industry. Four different industries of US companies were compared within US market: Finance and Insurance, Educational Services, Health Care and Social Assistance, and Public Administration. The comparative study showed the corrections for reporting delays in the overall US market and by industry, with specific emphasis on the four distinct industries.

This research was published as "Modeling reporting delays in cyber incidents: an industry-level comparison" in International Journal of Information Security [120].

The problem of under-reporting in cyber incidents is addressed in context of population characteristics. The proposed solution computes the large variations in under-reporting as a function of the three variables - revenue, incident type, and industry. Three different incident types–hacking, social engineering, and ransomware– and five industries– Retail Trade, Manufacturing, Finance and Insurance, Professional Scientific Technical Services, and Wholesale Trade– were studied. The research highlighted that there is a need to address under-reporting by incident types and by industry. The research was published as "Modeling Under-Reporting in Cyber Incidents" in special edition Data Science in Insurance of Risks [121][1] .

[1]Changes made in the paper in response to feedback from the reviewers

Table of Contents

List of Tables

List of Figures

List of Algorithms

Terminology

A_{max} Maximal Age in the dataset

F_{Δ} Cumulative distribution function of delay distribution

$H_{\Delta}(t, \delta)$ Number of incidents occured at time t reported upto delay δ

Δ Delay - Time period between the incident date and the reporting date

Δ_{max} Maximal Delta in the dataset

a Age - Time between the incident date and the date last incident reported in the dataset

f_{Δ} Probability distribution function of delay distribution

$h_A(a)$ Number of incidents with age a

$h_{\Delta}(\delta)$ Number of incidents with delay δ

$h_{\Delta}(c)$ Number of incidents for the given category c

$h_{\Delta}(t, \delta)$ Number of incidents occured at time t but reported with delay δ

$h_{\Delta}(t,\delta, c)$ Number of incidents occured at time t but reported with delay δ for category c

$h_{\Delta}(t,\delta, l)$ Number of incidents occured at time t but reported with delay δ at location l

$h_{\Delta}(t)$ Number of incidents occured at time t

$h_{\Delta}(t, c)$ Number of incidents occured at time t for the given category c

$h_{\Delta}(t, c, v)$ Number of incidents occured at time t for the given category c at level v

i Cyber security breach incident

Chapter 1

Introduction

Cyber incidents have become a global risk with no geographic boundaries [136]. Although these events have grown, evolved, and challenged businesses/industries/nations over the last few decades, cyber incidents existed since the development of personal computers [102]. In 2013, the G-20 countries have described cyber incidents as "a global economy concern" [2]. In 2015, the US President Obama administration identified cyber security as one of the critical economic challenges and threat to national security [97][105]. In July, 2021, US President Biden specified that cyber incidents could result in a "real shooting war", emphasizing the gravity of such incidents [17]. US risk managers and corporate insurers recognize cyber risk as one of the top business risks [36]. Cyber attackers are becoming increasingly more sophisticated and employ state-of-the-art techniques [9]. The fact that cyber incidents are evolving is impacting every nation across the globe [2].

While some cyber incidents are detected as soon as they occur, most events are often not discovered until weeks, months, or even years after the event actually occurred, resulting in biased data - defined as a subset of data that is influenced by the incidents detection time [9][68]. For example, Marriott International, Inc. experienced a major cyber incident in 2014 that was not discovered and reported until 2018 [126]. Sometimes cyber incidents are discovered and reported by third parties [5]. For example, the Target Corporation learned about the breach three weeks later when notified by an external third party [5]. The

disclosure of cyber incidents depends on reporting requirements across locations, industries, and inspecting regulatory agencies [9]. Smaller cyber incidents may never be reported at all, or have extreme delays, as only public companies and organizations with losses of personally identifiable information may be obligated to report. Sometimes reporting can take five to ten years, due to the organization failing to realize that a cyber incident happened, failing to immediately determine the extent of accessed or stolen data, or deciding not to publicize the incident for fear of reputation risk and consequent financial impacts. As a result of knowing and not disclosing in a timely manner, reporting delays are often observed in historical cyber event databases. Due to the increase in businesses working remotely, Coleman et al. commented that there is an increase in the number of actual incidents despite a decline in reported cyber incidents in the recent few years [9]. They further described that cyber incidents would remain undetected due to advanced and sophisticated threat techniques [9]. As a result of delayed attack detection (unintentional) or hiding attack information on the part of businesses (intentional), it is difficult to get an accurate understanding of the scope and pervasiveness of the reporting problem.

Alternatively, many organizations choose not to disclose any such information if they are not required to by law or industry regulation. They may opt not to report cyber incidents for fear of damage to their reputation or negative impact to their business [24][37][46][117][125]. This highlights the other well-established problem referred to as under-reporting of cyber incident [95]. Under-reporting occurs when the number of incidents reported is less than the number of incidents that occur [20]. In the cyber context, it can be defined as the incidents which are neither reported nor recorded into the databases. Despite regulatory requirements, organizations choose not to report, providing proof that under-reporting is still a significant problem [72][128]. In 2019, ISACA[2] reported that three out of four survey respondents believed that cyber incidents are deliberately concealed irrespective of reporting requirements [72]. Since attackers can be located anywhere around the globe, they are difficult to catch. With such a pessimistic outlook, organizations find reporting incidents as a waste of time and effort [125]. According to a U.S. Federal Bureau of Investigation (FBI) report, the recovery rate of stolen assets is high, but it becomes challenging to recover

[2]Information Systems Audit and Control Association

money in the absence of prompt action [39]. Due to required time investment, challenges in reporting, the lack of confidence that attackers will be caught, and the potential negative impact on the organizations reputation and financial status, many organizations opt not to report [125].

Cyber risk modeling firms rely upon reported cyber incident data to build their models, which are in turn relied upon by cyber insurers for underwriting, portfolio management, and risk transfer. To build robust loss estimation models, the most recent and updated information is required with as little bias[3] as possible. Correcting reporting delays and under-reporting in these databases are therefore key requirements to having trustworthy cyber insurance models. With the necessary corrections of reporting delays and under-reporting, one can more accurately identify trends in the targeting of industries or in attacker tactics.

Trends in cyber incidents have increasingly become a topic of formal research consideration. One area of research has focused on using statistical distributions for estimation. Coleman et al. investigated the distributions in terms of how many days it takes to detect cyber incidents and further how many days it takes to report them [9]. This study states descriptive statistics and highlighted the reporting delays or under-reporting problems in cyber incidents but did not address it. However, they expressed concern about how attackers are applying cutting edge techniques and remain unexposed.

The current research proposes to develop a novel approach to debias the reported cyber incidents by addressing reporting delays and under-reporting problems, respectively. To resolve reporting delays, the debiased delay distributions are modeled, and parameters of modeled distributions are applied to correct the bias in the reported cyber incidents. The corrections will reflect the increase in cyber incidents in recent periods rather than the diminishing trend shown by the reported incidents.

This research is divided into five chapters. Chapter 1 initially explains the problem of reporting delays and under-reporting in cyber incidents and is further divided into five sections - Sec. 1.1 provides the overview of the related concerned literature for the two problems: reporting delays and under-reporting, Sec. 1.2 presents the gaps and open

[3]Subset of data influenced by the incidents detection time

questions in the literature, Sec. 1.3 describes the proprietary data collected from multiple sources used in the succeeding research and Sec. 1.4 presents our contribution to the literature, closing the gaps, and resolving the questions. Chapter 2 presents the completed article, "Correcting for Reporting Delays in Cyber Incident", and discusses the innovative approach, generated from present research, to debias reporting delays in US cyber incidents. Chapter 3 applies the approach developed in Chapter 2 and discusses the dynamics of four key industries with the US market as whole. Chapter 4 discusses the approach developed to correct the under-reporting in US cyber incidents. Chapter 5 summarizes the previously presented research in cyber incidents to debias for reporting delays and under-reporting, and presents future research.

1.1 Literature Review

1.1.1 Reporting Delays

Reporting delays is a well-known problem in insurance and epidemiology models [85]. Most of the existing literature on reporting delays is from the medical domain. In 1986, Brookmeyer and Gail suggested the problem be addressed by excluding recent data from the analysis [22]. Reporting delays were studied as a statistical problem for the first time by Harris in 1987 [55]. Brookmeyer and Damiano, as well as Harris, stated the reporting delays problem as an incomplete multinomial distribution [21][55]. Harris attributed under-reporting to reporting delays [56]. The failure to report an incident immediately is statistically referred to as a "right truncation problem" [23]. Mathematicians and actuaries define it as a "run-off" or "reporting triangle problem" [65][90]. Actuaries categorize reporting delays as an incurred but not reported (IBNR) or an occurred but not reported (OBNR) problem[4]. However, actuaries focused on correcting losses associated with claims rather than the claim counts [7][85]. The focus of this research is to correct for the reporting delays in the absence of covariates i.e. considering incident and reporting dates only.

Late 1980s

Acquired Immunodeficiency Syndrome (AIDS) cases remained undetected for period of time and reported even later. One possible reason could be that patients were reluctant to report their case in light of societal/cultural norms. In the context of AIDS cases, it takes months or years to report after the diagnosis [85]. Brookmeyer and Damiano, and Heisterkamp along with various researchers in three separate studies, addressed the reporting delays in the number of AIDS cases, assuming the number of cases to be infected following a Poisson distribution [21][61][62][63]. In extension of Harris' work, Brookmeyer and Damiano applied a Poisson regression model to correct the reporting delays in the number of AIDS patients. The adjusted number of cases occurred at time, t, until delay, δ, computed as shown in

[4]In the following sections, it is referred as "not reported" to be consistent with industry language

Eq.1.1 and Eq.1.2 respectively.

$$H'_{\Delta}(t,\delta) \sim \text{Pois}(e^{\alpha_t+\beta_\delta}) \tag{1.1}$$

where α_t and β_δ are Maximum Likelihood Estimates

$$H'_{\Delta}(t,\delta) = \frac{H_{\Delta}(t,\delta)}{1-\sum_{i=1}^{t} p_i} \tag{1.2}$$

where p_i is the Poisson regression estimates

$H_{\Delta}(t,\delta)$ is the reported incidents at time, t, until delay, δ

However, Heisterkamp et al. mentioned that the expected number of cases is the matter of concern rather than the reported proportions as a percentage of the total reported cases; the authors posit that the focus of the research should be on the counts rather than on the proportions. Hence, they considered proportions as a distraction and classified them as a "nuisance". They applied Maximum Likelihood Estimation (MLE) equations Eq.1.3 and Eq.1.4 to obtain the adjusted number of incidents at the time, t, $h'_{\Delta}(t)$ [61][62][63].

$$h'_{\Delta}(t) = \frac{\Sigma_{d=0}^{\delta} h_{\Delta}(t,d)}{\Sigma_{d=0}^{\delta} Adjusted\, p(d)} \tag{1.3}$$

$$p'(\delta) = \frac{\Sigma_{j=1}^{t} h_{\Delta}(j,\delta)}{\Sigma_{j=1} t Adjusted\, h_{\Delta}(t)} \tag{1.4}$$

Heisterkamp et al. looked at the problem from another perspective: how many incidents occurred but not yet reported? [61]. They modeled a number of not-reported incidents as independent Poisson distributions conditioned on the $Adjusted\ h_{\Delta}(t)$, referred to as $h'_{\Delta}(t)$. Asymptotically, they found both reported and not-reported incidents were multivariate normally distributed. To capture trends, they suggested the expected number of incidents as exponential and integrated logistic models [61][62][63]. Despite capturing trends, these models failed to correct reporting delays beyond the maximum delay, $\delta_{\max}$, in the dataset. They stated two key reasons. First, the corrections depend heavily on the choice of an appropriate model (i.e. exponential or integrated logistic model). Second, exponential models lack the ability to handle delays when number of incidents doubles. The proposed

novel approach addresses this problem by fitting the modeled distribution on the domain $[0, \delta_{\max}]$and then using the fitted modeled parameters to find the correction factors on domain $[0, \infty]$.

Morgan and Curran applied Brookmeyer and Damiano's approach to investigate AIDS cases in the U.S. but found trends in the reported incident distribution over the described period of time. Under the trend stationarity[5][6] assumption for the incidents distribution, Morgan and Curran computed the number of incidents detected within a given month out of reported incidents i.e., the number of incidents with delay,$h_\Delta(\delta = 0)$, and then applied the Box-Cox transformation to fit the quadratic polynomial with weighted regression [21][100]. Downs et al. assumed the pattern of reporting over constant time and adjusted the counts at time t with the sum of proportions and number of incidents at t but reported with δ, for all delays, as shown in Eq.1.5 and Eq.1.6 [30][31].

$$p(\delta) = \frac{h_\Delta(t, \delta)}{h_\Delta(t)} \tag{1.5}$$

$$h'_\Delta(t) = \frac{h_\Delta(t)}{\Sigma_{d=0}^{\delta_{max}} p(d)} \tag{1.6}$$

where d refers to delays from 0 to $\delta_{\max}$

Downs et al. introduced trends as exponential models to the adjusted counts, as shown in Eq.1.7and Eq.1.8 [30][31].

$$Linear\ Model : \ln(h'_\Delta(t)) = \beta_0 + \beta_1 t + \epsilon_t \tag{1.7}$$

$$Quadratic\ Model : \ln(h'_\Delta(t)) = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t \tag{1.8}$$

[5]Trend Stationarity is a time dependent stochastic process and becomes stationary with function of time is removed Source: Online Glossary of Research Economics

[6]The process is considered stationary in absence of covariance Source: Online Glossary of Research Economics

The very assumptions of constant pattern over time and stationarity in trends are concerns due to non-stationarity detected in debiased delay distributions. The proposed novel approach addresses this problem by generating the debiased delay distribution[7] over a monthly, two-year rolling period.

As in the United States, AIDS was considered a social stigma in the United Kingdom in 1980s. Healy and Tillet investigated AIDS data in the United Kingdom. They fitted an exponential curve and a log-linear model with Poisson errors to monthly reported incidents to correct reporting delays, as shown in Eq.1.9 and Eq.1.10 [59]. Again, having a fixed model as a function of time is not an appropriate way to address reporting delays due to non-stationarity in the data.

$$Exponential\ Fit : \log(h_\Delta(t) + 1) = 0.08841 + 0.06265t \tag{1.9}$$

$$Log - linear\ Fit : \log(h_\Delta(t)) = -0.04495 + 0.6399t \tag{1.10}$$

Brookmeyer and Damiano studied AIDS data until January, 1988 in the US and suggested another approach based on back calculations[8]. They used the back-calculated infected estimates to compute short-term projections. They focused on number of infected and reported incidents and defined these incidents as having a multinomial distribution with an unknown sample and computed probabilities as shown in Eq.1.11 [21].

$$p_\delta = \int_{t=0}^{\delta} I(s)(F_{\delta - s} - F_{\delta' - s})ds \tag{1.11}$$

where δ' is previous delta values

$I(s)$ is likelihood upto delay δ

[7]Debiased Delay Distribution is the novel approach to generate distribution from empirical data addressing the bias towards the shorter delays.

[8]Back calculation is the method to compute previous values based on current value.

$I(s)$ is computed from unknown parameter θ. MLE, estimates of multinomial distribution, are obtained by maximizing θ , and adjusted $H_\Delta(t,\delta)$ can be computed as shown in Eq.1.12 [21].

$$H'_\Delta(t,\delta) = \frac{H_\Delta(t,\delta)}{\sum_{i=1}^{\delta} p_\delta} \tag{1.12}$$

Working on OBNR incidents is an interesting approach but does not cater longer delays. The approach works under the assumption of complete data, not an apt assumption; it is challenging to find a dataset which could be considered representative of the data being investigated and is complete.

1990 and after

Rosenberg studied AIDS data until September, 1988 in the US. He applied Brookmeyer and Damiano's statistical model with an assumption of stationary reporting delay distribution to study AIDS' incidents where a stationary distribution implies the same distribution over a period of time[9] [106]. The Poisson model was fitted to compute counts, as shown in Eq.1.13 [21][114]. Considering the nature of the cyber incidents (the delay distribution of cyber incidents is found to be non-stationary), the underlying stationarity assumption is problematic; the proposed approach mitigates this issue, as mentioned in the earlier subsection 1.1.1.

$$\log E(h_\Delta(t,\delta)) = \mu + \alpha_t + \beta_\delta \tag{1.13}$$

where $\alpha_1 = \beta_1 = 0$

α parameter varies with time, t

β parameter varies with delay, δ

[9] $F_t(x) = F_{t+\Delta_t}(x)\ \forall t, t+\Delta_t$

Probabilities for the model defined as shown in Eq.1.14:

$$p_\delta = \frac{\exp^{\beta_\delta}}{\sum_{d=0}^{\delta} \exp^{\beta_d}} \tag{1.14}$$

Rosenberg computed probabilities with non-iterative approach, as shown inn Eq.1.15 .

$$p_\delta = \frac{\sum_{t=0}^{t_{max}} h_\Delta(t,\delta)}{\sum_{t=0}^{t_{max}} h_\Delta(t)} \tag{1.15}$$

The adjusted counts are be computed as shown in Eq.1.16:

$$h'_\Delta(t) = \frac{h_\Delta(t)}{1 - \sum_{d>\delta}^{\delta_{max}} p_d} \tag{1.16}$$

Cheng analyzed two approaches on AIDS studies, suggested by Brookmeyer and Damiano as well as Rosenberg. Cheng proposed using the Lagrange multiplier, λ, to compute probabilities in Eq.1.11 and Eq.1.15 [21][26][114]. Here again, Cheng assumes stationarity. In contrast to using the Lagrange multiplier, λ, the current study proposes fitting the bimodal empirical distribution and extracting probabilities over the extended domain, $[0, \infty)$.

Brookmeyer and Liao investigated AIDS incidents in the US and suggested the Bayesian linear model based on time-dependent, reverse-time incidents to correct delays [23][45]. They highlighted the crucial and relevant limitation that the delay, δ, cannot be considered beyond the age of the dataset and hence, only the conditional distribution on delay less than or equal to age of the dataset can be estimated. They proposed two methods. First, they considered the number of incidents that occurred at time t but reported with delay, δ, to be Poisson distributed with mean, as shown in Eq.1.17 [23].

$$E(h_\Delta(t,\delta)) = e^{(a_t, b_\delta)} \tag{1.17}$$

where α-parameter varies with time, t

β-parameter varies with delay, δ

α_t and β_δ are parameters obtained from log-linear models or Poisson regression.

Second, they defined the reporting delay distribution as the product of conditional probabilities, as shown in Eq.1.20.

$$f_\Delta = \prod_{i=0}^{age}(1 - p_i) \tag{1.18}$$

$$p_{(t,\delta)} = \frac{h_\Delta(t,\delta)}{h_\Delta(t)} \tag{1.19}$$

where $h_\Delta(t,\delta)$refers to the incidents reported at time, t, with delay, δ

$h_\Delta(t)$ refers wot the incidents reported at time, t

$$F_\Delta = \prod_{\delta=0}^{age}\left(1 - \frac{h_\Delta(t,\delta)}{h_\Delta(t)}\right) \tag{1.20}$$

Unlike AIDS where the cases were reported with delay considering the social norms, the delays in multiple sclerosis(MS) incidents is due to the nature of the problem itself. The MS incidents cannot be diagnosed until a second attack which could happen after months or years, or sustained progression in symptoms over six months. Along similar lines of Brookmeyer and Liao, Esbjerg et al. investigated delays in MS incidents. They defined models based on conditional and marginal probabilities to correct delays in reporting MS incidents. However, they computed the conditional probability and marginal distributions, as shown in Eq.1.21 and Eq.1.22 [35].

$$\text{Conditional Probability: } f_\Delta(\delta|a) = \frac{f_\Delta(\delta)}{F(a)} \tag{1.21}$$

$$F_\Delta(\delta|a) = \frac{F_\Delta(\delta)}{F(a)} \tag{1.22}$$

The marginal distribution function defined using the conditional probability results, as shown in Eq.1.23 [74].

$$\text{Marginal Probability: } F_\Delta(\delta|age) = \prod_{d=\delta+1}^{age}\left(1 - \frac{f_\Delta(d|age)}{F_\Delta(d|age)}\right) \tag{1.23}$$

These models also deal with the reporting delays being addressed until age, a. The proposed algorithm addresses this concern by fitting the modeled distribution on debiased delay distribution over domain $[0, \delta_{\max}]$ and then using parameters of a modeled distribution on extended domain $[0, \infty)$ to compute corrections. Gebhardt et al. found that linear and quadratic polynomials could capture the non-stationary trend in reporting delays but resulted in over-fitting because of few data events available on recent counts [44].

Harris discussed the basic model under the assumption that the number of incidents occurred at time t but reported with delay δ, $h_{\Delta}(t,\ \delta)$ as a random variable with an independent Poisson distribution. Consequently, the monthly counts depend on a discrete time Poisson process, whereas the count distribution with the given number of incidents at time t results in a multinomial distribution [56].

Brookmeyer and Gail, Harris, and Kalbfleisch and Lawless studied AIDS data in the context of distributional challenges. They expressed concern about finding trends based on the delay distribution over time with the non-parametric approach where parametric models are time dependent; for each time step, there exists a different parametric model [22][56][74][96]. The proposed algorithm in the current study applies a similar approach when the modeled distributional parameters are computed from the debiased delay distribution.

Lawless investigated AIDS data reported to surveillance agencies and estimated the reporting delays based on the latest available event delay probabilities under a stationary assumption, and he considered the number of events occurred at time t but reported with delay δ, $h_{\Delta}(t, \delta)$, which depends on the number of known recent time periods. He referred to this assumption as "risky" where delays δ are longer over time. He suggested multinomial models with random effects based on the probability vector over a period of time extracted from Poisson distribution (under the assumption that number of occurred events is independent of time), Dirichlet (to represent randomness in reporting delay probabilities where no systematic trend observed with respect to time) and Gamma(introduces random effects independent of time) distributions to represent the variation in reporting delays [85]. As discussed earlier, the stationarity assumption is not practical.

Bastos et al. studied disease surveillance data and discussed two common frameworks for the reporting delays [15]. A Bayesian or Hierarchical approach [65] where the $h_{\Delta}(t, \delta)$ is conditional on $h_{\Delta}(t)$ where $h_{\Delta}(t)$follows a Poisson or Negative Binomial distribution. As a result, $h_{\Delta}(t, \delta)$ is a multinomial distribution with a vector of probabilities for the given time steps. The second approach is the chain-ladder approach [90][111] where $h_{\Delta}(t, \delta)$ does not follow any distribution and has a linear relation with constant overall mean and random effects to capture when the mean varying with time and with delay δ are normally distributed. This approach can be expanded to address the requirements of parametric and non-parametric forms [14][34]. The conditional multinomial approach motivated the chain-ladder approach [116]. Bastos et al. extended the chain ladder approach to correct the delays to introduce spatio-temporal (locations) within the counts and the co-variate effects [15]. However, Hohle used the negative binomial distribution to tackle count data and handled the delay distribution as a dirichlet distribution in conjugate prior posterior form; when the process is homogeneous, i.e., irrespective of the time steps between $t = 0$ and $t = t$, the probability evolves the same way [65]. Chitwood et al. analyzed coronavirus disease (COVID-19) virus data and proposed the Bayesian Nowcasting approach where they adjusted counts using the negative-binomial distribution [27].

The extant literature summarized above demonstrates the use of multiple distributional approaches to capture the delayed reporting of incidents. Taken together in cyber incidents context, no one distributional assumption adequately captures diverse phenomena. The current study considers two alternative distributions to capture the two distinctive characteristics of the time dimension of reporting incidents. Specifically, this research uses an exponential distribution to address shorter delays and the normal distribution to address longer delays.

White et al. suggested delayed counts adjustment at time step t by dividing the actual count with the sum of delayed reported event probabilities beyond t [134]. Weinberger et al. suggested the adjustment to COVID deaths by dividing the actual counts for the given day t by the proportion of complete records during that week [132]. While these approaches addressed the reporting delays in COVID incidents, the require an assumption of parallel

complete data. Such adjustments would be meaningless if the data is complete. On the other hand, these approaches are inapplicable on incomplete data.

Noufaily et al. proposed the log-likelihood function that considers left and right truncation to model reporting delays to detect the outbreak of infectious diseases. They computed the delays as the difference between the event date (specimen collection date for testing) and three fixed reporting time steps (time at which the report was sent to the database), τ_1, τ_2, and τ_3. The log likelihood function would vary depending on where the event time lies between the fixed reporting time steps. They suggested that the likelihood contributions, L_c varies based on where the event date falls with respect to τ_1, τ_2, and τ_3, as shown in Eq.1.24 [104].

$$L_c = \begin{cases} \frac{f_\Delta(\delta)}{1-F_\Delta(\delta_{\tau_1})} & \text{when Event Date} \leq \tau_1 \\ f_\Delta(\delta) & \text{when } \tau_1 < \text{Event Date} \leq \tau_2 \\ \frac{f_\Delta(\delta)}{F_\Delta(\delta_{\tau_3})} & \text{when } \tau_2 < \text{Event Date} \leq \tau_3 \end{cases} \tag{1.24}$$

$$\text{where } \delta_t = t - \text{Event Date}$$

Accordingly, the log-likelihood function can be defined as shown in Eq.1.25

$$\begin{aligned} l = \sum_{i=1}^{n_i} \log \Big\{ & \frac{f_\Delta(\delta_i)}{1 - F_\Delta(\delta_{\tau_1})} I(\text{Event Date} \leq \tau_1) + \\ & f_\Delta(\delta_i) I(\tau_1 < \text{Event Date} \leq \tau_2) + \\ & \frac{f_\Delta(\delta_i)}{F(\delta_{\tau_3})} I(\tau_2 < \text{Event Date} \leq \tau_3) \Big\} \end{aligned} \tag{1.25}$$

$$\text{where } n_i = \text{Number of Incidents}$$

$$I(\bullet) = \text{Binary indicator function}$$

Noufaily et al. parameterize based on a hazard function[10] with survival function computed as $e^{-\int_0^\Delta \lambda(j)dj}$. They suggested the statistical approach, which monitors the counts in current

[10]In survival analysis, the distribution is modeled with hazard function, $\lambda(d) = \frac{f(d)}{S(d)}$ where $S(d)$ is survival function

and certain past periods, where the number of past periods is selected based on the median reporting delay [103]. The approach is sensitive to the choice of time steps τ_1, τ_2, and τ_3.

Wang studied transfusion-associated AIDS cases reported until December 1989. He discussed the non-parametric and semi-parametric methods on survival data, because data is complete and no further reporting delays would be expected. He investigated right truncated data where the data has a known start and the end date of the event where one is aware of the start date and is not expecting any event occurring in the given context beyond the end date. For the parametric approach, he assumed $h_\Delta(t,\delta,c)$, $h_\Delta(t,c)$ and $h_\Delta(c)$ as Poisson distributed [28]. For a non-parametric approach, he derived the likelihood function based on the complete observed data[11]. However, for the semi-parametric approach, he combined both approaches: non-parametric and parametric. He applied the MLE derived from the non-parametric approach but assumed its parameters to be Poisson distributed. The non-parametric and semi-parametric approaches are not appropriate when the data is incomplete [130]. The proposed algorithm derives the complete picture of the incidents from the incomplete data.

Midthune et al. studied cancer data from 1975 until 1997. Considering that research on cancer is still underway, cancer patients are commonly not diagnosed until later stage. Since the disease is frequently not detected in a timely manner, the reporting is delayed. The authors discussed events with delayed reporting as well as events being added to the dataset more than once as the reported data comes from multiple resources. The replication of the same record is a common problem when data collected from various resources. The data used in this research is also collected from multiple data providers. To avoid this problem, a matching algorithm based on firmographic[12] data (Table 1.4) and incident date within a week are applied to dedupe the multiple records for the same incident. They suggested a model which corrects the reporting delays and removal of events from the reported events at a delayed period, and claimed the number of events reported for the given category at delay, δ, to be marginal Poisson distributed and imposed the delays with the normally

[11]The records are complete despite reporting delays and no more events expected for the given set of data [84]

[12]Firmographic data: Geographic Data, Number of Employees, Industry and Revenue

distributed random effects linearly into a truncated log-log model. The random effects shrink over-dispersion rather than eliminating it. The models assume stability- constant mean and constant variance over time in the reporting. The models do not work on data with trends or non-stationarity [96]. Again considering the nature of the cyber incidents, the stability assumption would not work.The current approach deals with trends and non-stationarity in the cyber incidents.

Harris corrected the number of COVID-19 cases using the expectation-maximization(EM) algorithm. He split the data into two parts. The first part assumed to be complete, but the second part needed to be corrected. He assumed the counts with delays follow a Poisson distribution. The corrected count depends on a likelihood parameter computed with an iterative procedure. The parameter is finally normalized to obtain the appropriate discount factor. The corrected count is computed by dividing the marginal count by the normalized discount factor [57]. As explained in section 1.1.1, a single distribution assumption is not appropriate to address both shorter and longer delays.

As observed, the majority of the literature addressed reporting delays in the medical space. In 2021, Coleman et al. discussed the time frame to discover cyber incidents and time frame to disclose them from 2016 until 2020 [9]. However, they did not address the need for a correction measure.

There are additional studies with different perspectives which are not directly related with our research but worth a brief mention. Some researchers investigated cyber claims data to correct reporting delays in claims with capital reserving perspective [11][73][137][138]. It is important to understand that reporting delays in claims is different from the problem being investigated in this research. Since organizations tend to claim against the insurance, cyber incidents are reported as and when detected. As a result, the reporting delays in claims are expected to have shorter delay periods, perhaps in few days or 1-2 weeks, unless they are not able to detect in timely manner. Unlike the current research problem, these incidents are reported to insurance companies to file for claim against insured assets and need not to be concerned about reputational risk or other factors.

Table 1.1: Reporting Delays: Literature Summary

Techniques Evaluated	Key Findings	Citations
Conditional Probabilities	• Pros: Captures trend • Cons: Model over-fitted	[23][35][43][74]
Statistical Model under assumption of Poisson distributed counts	• Pros: Easy to implement • Cons: Does not capture trends	[23][56][61][62] [63]
Exponential and Log-linear Model with Poison Errors	• Pros: Captures trend • Cons: Stationarity Assumptions	[59]
Exponential and Logistic Models	• Pros: Captures trend • Cons: Fails to correct beyond the maximum delay in the data	[30][31][61][62] [63][100]
Back-Calculation	• Pros: Easy to implement • Cons: Highly sensitive to parameters chosen	[21]

Techniques Evaluated	Key Findings	Citations
Poisson Model with stationary reporting delays	• Pros: Easy to implement • Cons: Distributional Assumption and stationary reporting delays	[26][114]
Semi-parametric with counts Poison Distributed-MLE Approach	• Pros: Easy to implement • Cons: Requires parallel complete data to train	[130]
Non-parametric-MLE approach	• Pros: No distributional assumptions • Cons: Requires parallel complete data to train	[130]
Truncated Model	• Pros: Random effects shrinks • Cons: Requires stable reporting delays	[75][96][130]
Multinomial Model with Dirichlet/Poisson/Gamma distributed random effects	• Pros: Captures trend in timely fashion • Cons: Does not work where delays are longer and distributional assumptions	[85]

Techniques Evaluated	Key Findings	Citations
Proportions	• Pros: Simplest approach • Cons: Requires parallel complete data to train	[132][134]
Bayesian/Hierarchical approach with counts Poisson/Negative-Binomial Distribution	• Pros: Easy to implement • Cons: Distributional assumptions	[15][27][65]
Log-likelihood with Truncation Model	• Pros: Data driven approach • Cons: Sensitive to choice of three reporting time-steps	[103][104]
Chain-ladder approach	• Pros: Easy to implement • Cons: Sensitive to outliers	[15]
Expectation-Maximization Approach	• Pros: Easy to implement • Cons: Requires complete data to train and Distributional assumptions	[57]

1.1.2 Under-reporting

Since 1982 the concept and statistical implications of incident, under-reporting has been a point of concern and mainly studied in the medical domain [64][122]. In the cyber domain, it is a well-known problem due to various reasons, such as reputation risk, financial impact, belief that attackers will never be caught, and incident reporting as waste of time and effort [24][37][46][95][117][125].

In 1986, Brookmeyer suggested to exclude the under-reported data from the analysis [22]. However, Wood et al. mentioned that ignoring under-reporting leads to biased statistical models [135]. Elvik and Mysen suggested that under-reporting leads to an incomplete dataset and the resulting analysis/models will be biased towards the reported data only [33]. Hence, the approach cannot be applied to other data or situation and loses generalizability.

Fletcher et al. investigated under-reporting of Acute Gastroenteritis(AcG) cases in Jamaica. AcG is a known diarrheal disease resulting in morbidity and mortality [42][77]. In Jamaica, there is ignorance about the consequences of AcG. As a result, people do not opt for appropriate medical treatment, the true incidence level recorded by National Surveillance (NSU) is under-reported. Fletcher et al. corrected the under-reporting in AcG with a proportions approach, which can be applied in the cyber domai,n but with cyber claims data. They investigated syndromic and lab-confirmed under-reporting in AcG in Jamaica. The national estimates of under-reporting for syndromic under-reported AcG were computed by adjusting the number of AcG cases reported to NSU by the data collected from surveys [42].

The syndromic under-reporting multiplier was computed by Fletcher et al. as the inverse of the proportion of patients who opted for medical care, as shown in Eq.1.26 [42].

$$\text{Syndromic Multiplier} = \frac{1}{p_{MC}} \tag{1.26}$$

where p_{MC} is the proportion of patients opted for medical care

The national estimates of under-reporting for lab-confirmed cases are computed with various proportions in an orderly manner. However, the approach remains the same; rather than one proportion, there is a product of multiple proportions, as shown in Eq.1.27 [42].

$$\text{Lab confirmed Multiplier} = \frac{1}{p_{LC}} \cdot \frac{1}{p_{Pos}} \cdot \frac{1}{p_{Tested}} \cdot \frac{1}{p_{Spec}} \cdot \frac{1}{p_{SpecReq}} \cdot \frac{1}{p_{Care}} \tag{1.27}$$

where p_{LC} is the proportion of lab-confirmed cases reported to NSU

p_{Pos} is the proportion oflpositive samples out of total samples

p_{Spec} is the proportion of specimens submitted

$p_{SpecReq}$ is the proportion of cases requested for specimen submission

p_{Care} is the proportion of l ill persons who went for care

Hazell and Shakir collected the level of under-reporting of adverse drug reactions (ADRs) from 37 studies. They gathered studies that applied the proportion approach from various countries and used descriptive statistics to estimate the overall under-reporting level. They divided the research into three categories. The first category considered the proportion of known/suspected/expected cases identified during monitoring but were not reported [4][12][25][29][40][41][47][60][66][83][87][89][91][94][99][101][107][123]. The second category considered the proportion of known/expected cases identified through data sources and also reported [8][13][18][19][32][69][71][76][82][98][108][109][118]. The third category considered the proportion of cases identified during clinical trials and reference studies but never reported [16][38][67][70][110][127]. These cases were computed within the same time period and same location. All three applied the proportion approach but from different perspectives [58].

Schuitemaker et al. also applied the same proportion method to calculate the under-reporting multiplier to correct maternal deaths during pregnancy/childbirth in the Netherlands [122]. Maternal deaths due to early pregnancy and indirect deaths are often under-reported and require corrections. Abay, Alsop and Langley, Amoros et al, Elvik and Mysen, and Wood

et al. investigated under-reporting in crash frequency, such as road accidents. They compared the reported crash data against the hospital data and applied the same proportion approach [1][3][6][33][135]. Crash data is often under-reported and gets corrected from hospital data.

Hirvonen et al. investigated the under-reporting levels and trends in energy based on dietary reference. Women and overweight adults often report less than their actual food consumption (lesser micro-nutrients intake). Since energy level is directly proportional to food consumption, under-reporting results in energy level distortion. They applied a logistic regression approach considering under-reporting a binary dependent variable and dietary factors as independent predictor variables, as shown in Eq.1.28 [64]. They showed that under-reporting should be taken into consideration when doing further advanced studies especially for women and overweight adults.

$$l = \log\left(\frac{p_{UR}}{1-p_{UR}}\right) = \beta_0 + \beta_1 Gender + \beta_2 Age + \beta_3 Area + \beta_4 BMI + \beta_5 Study\,Year \quad (1.28)$$

where p_{UR} is the probability of under-reporting

Lissener et al. investigated overweight women for overeating but under-reporting their food consumption and applied a simple regression approach with various body composition factors as independent predictors, as shown in Eq.1.29 Eq.1.30, Eq.1.31 and Eq.1.32 [88].

$$Energy = \beta_0 + \beta_1 MDWC + \beta_2 W \tag{1.29}$$

$$Energy = \beta_0 + \beta_1 MDWC + \beta_2 W + \beta_3 FM \tag{1.30}$$

$$Energy = \beta_0 + \beta_1 MDWC + \beta_2 LM + \beta_3 FM \tag{1.31}$$

$$Energy = \beta_0 + \beta_1 MDWC + \beta_2 LM \tag{1.32}$$

where MDWC = Mean Daily Weight Change

W = Weight

FM = Fat Mass

LM = Lean Mass

Lissener et al. computed the under-reporting with the standard error of mean(SEM)[13], as shown in Eq.1.33 [88].

$$UR = MDWC \pm \frac{\sigma}{\sqrt{n}} \tag{1.33}$$

Krantz et al. proposed new methods with harmonic analysis and wavelets to compute the level of under-reporting before the COVID-19 first peak. They opted for a proportion approach but mapped these proportions to time intervals. They suggested level of under-reporting at time t, $p_{UR}(t)$, to be computed as shown in Eq.1.34 [78][79].

$$p_{UR}(t) = \sum_{i=0}^{n} \frac{a_i}{a_i + b_i} \tag{1.34}$$

where a_i and b_i are the number of reported and not reported cases distributed in n time intervals, $0 \leq i \leq n$, until time t.

It is not possible to know the number of *not* reported cases at any point of time. They proposed the construction of wavelet using support of a_i to find this number, but the research is ongoing. Such methodology might not be relevant in the current research considering that

[13]SEM is computed as $\frac{\sigma}{\sqrt{n}}$ where n is sample size

the pattern is expected to start at zero, oscillate with given amplitude, and then again reduce to zero.

Table 1.2: Under-reporting: Literature Summary

Techniques Evaluated	Key Findings	Citations
Proportions	• Pros: Easy to implement • Cons: Data might not be easy to find	[1][3][4][6][8] [12][13][18][19] [25][29][32][33] [40][41][47][60] [66][69][71][76] [82][83][87][89] [91][94][98][99] [101][107][108] [109][118][123] [135]
Logistic Regression	• Pros: Easy to implement • Cons: Depends on accuracy of other independent variables	[64]
Median ± SEM with Linear Regression	• Pros: Easy to implement • Cons: Difficult to obtain such level of data	[88]

Techniques Evaluated	Key Findings	Citations
Median of Interquartile range	• Pros: Simple computation • Cons: Difficult to find the research with estimates	[58]
Harmonic Analysis & Wavelets	• Pros: Develops complete data from partial data • Cons: Complex mathematical models and computationally intensive	[78][79]

1.2 Open Research Question

1.2.1 Reporting Delays

Brookmeyer mentioned that the existing research addresses the issue of reporting delays where delays, δ, at most equates to age, a [23]. The analysis with such incomplete data would result in a biased analysis as it does not expect any events beyond the longest in the dataset. The models with such incomplete data could result in wrong decisions. The research so far either employs a back-calculation proportion or a proportion-based approach with linear/quadratic trends or a parametric approach based on distributional assumptions or a non-parametric approach based on complete datasets to derive the MLE parameters where more events are not expected to be reported in the future [130]. Under the assumption of complete data, it might be too late to take appropriate action based on the analysis.

The direct estimation of an empirical delay distribution from the raw data results in four problems:

Problem 1: The nature of reporting delays means that direct estimates from empirical data will be biased toward shorter delays, since recent events could only appear in the data set in the first place if the reporting delay is small. The data is considered to be complete in terms of shorter delays as the information is already provided, whereas data with longer delays is incomplete and requires estimation.

Problem 2: The reporting delay distribution may not be stationary, making it difficult to estimate delays beyond *age*.

Problem 3: The direct estimate from empirical data does not address any event beyond the age of the dataset, assuming zero probability of any delay longer than the age in the data set. Since the *age* of the raw data is finite, it is not possible to observe beyond the longest delay in the data set. Although defined on the domain $[0, \delta_{\max}]$, the delay distribution, f_{Δ}, is regarded as complete, which does not bank on any event to be reported beyond the longest delay in the dataset.

Problem 4: Longer reporting delays are based on a few data points. As one moves further in time, the proportion of reporting delays decreases; fewer and fewer events with longer reporting delays are expected to be reported over time. In this research, all four problems are addressed through novel proposed "Debiased Delay distribution" and "Modeled distribution" fitted with an optimization approach.

1.2.2 Under-reporting

In 1986, Brookmeyer suggested to ignore the under-reported data from the analysis. However, Abay, Alsop and Langley, Amoros et al, Kumara and Chin, and Wood et al. studied road traffic accidents and emphasized that ignoring under-reported data would result in biased estimates [1][3][6][81][135]. Elvik and Mysen also studied road accidents reporting in 13 countries and stated under-reporting as a data incomplete problem [33]. Addressing the data bias/incomplete problem, the reported data is corrected with the proportion approach. These proportions are computed from a smaller data set considered to be complete. More than 85% of the existing literature applied a proportions approach to find the level of under-reporting and considered only one dimensional domain.

One-dimensional correction cannot justify multiple directions. Corrections based on revenue might be appropriate overall but not when considering the corrections for the given incident type or given industry or vice-versa.

Problem: The direct proportion approach is applicable only with one underlying feature but might be inappropriate when multiple features are involved. As observed in reporting delays, variation in reporting delays differs from industry to industry; under-reporting level based on revenue might be applicable overall but is not appropriate at the incident type level or at the industry level for specific corrections.

1.3 Data

For reporting delays, historical incidents and firmographic data used. The historical incidents are matched with firmographic data set for the cases where name of the company cannot be matched directly based on string.

For under-reporting, two proprietary data sets, claim-exposure data and historical incident-IED data are used. Historical incident-IED data is aggregated data set constructed by combining historical incident data set with a proprietary firmographic data. The claim-exposure data is a collection of US cyber insurance claims and policies data obtained from multiple insurers. The historical incident-IED data set is an aggregated data set constructed by combining historical incident data sets with a proprietary firmographic data set of companies.

1.3.1 Historical Incidents Data

The proprietary data set used in the current research is a collection of more than 140,000 historical incidents over several decades. The aggregated data includes curated datasets from various proprietary data sources. The data set was constructed from multiple source data sets of historical incidents (collection of publicly reported incidents), with de-duping done by fuzzy matching to a firmographic data set (see section 1.3.2 for more details), as shown in Fig. 1.1.

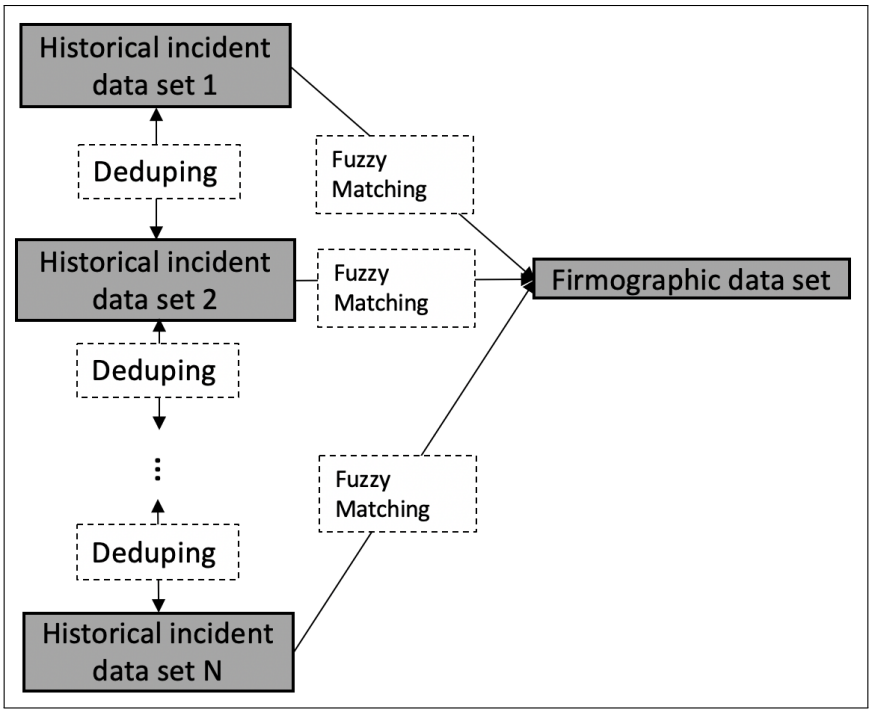


Figure 1.1: De-Duping Process

This study is done on the incidents collected from the past decade (2010-2019). The source cyber incident datasets (Table 1.3) provide information in terms of "what", "when" and "who" of an incident. Concretely, the data sets included a description of the incident (what), the name of the company to which the cyber incident occurred (who) (N.B.: "Aggregation incidents"[14] separately list each company known to be impacted), the date when the incident occurred (when), and when it was reported (when).

Table 1.3: Source Cyber Incident Dataset

Type	Variable	Information
What	Description	Description of the Incident
	Incident Type	Type of incident (extracted from Description if unavailable)
When	Incident Date	Date when incident occurred
	Reporting Date	Date when incident reported
Who	Organization	Name of the Organization which is attacked

[14]Aggregation Incidents: When single incident impacts many organizations simultaneously

Frequently, the name of the company varies from one data set to another; the data sets could not be matched with string matching directly. An alternate approach of matching the incident data sets with the firmographic data set was applied.

Limitations: Some events do not have an occurrence date listed; such events were excluded from this analysis. There are also large spikes in event counts listed as having occurred on January 1st (as observed in Fig. 1.2a), referred as "default date". The default date was assumed to be a default value when only the year of the event was known. These events were therefore re-distributed proportionally throughout the year as shown in Fig. 1.2b but excluded while developing the approach. Table 1.5 shows the number of incidents on the default date re-distributed over the year of incident.

1.3.2 Firmographic Data

Table 1.4 shows the firmographic data set which include approximately 50 million businesses in the US. The firmographic data set included information on geographic location, employee count, industry, and revenue.

Table 1.4: Firmographic Dataset

Variable	Information
Geographic Location	Location where organization is based
Employee Count	Number of employees in the organization
Revenue	Revenue from the organization
Industry	Name of the industry organization belongs to

The consolidation was done via a previously developed matching algorithm that examined company name, industry classification (e.g., via NAICS[15] codes), address information, and any other fields common to both the cyber incident data set and the firmographic data set. Incidents in distinct cyber incident data sets were identified as identical when they satisfied two criterion in terms of firmographic and time of listing:

1. They were matched to the same company in the firmographic data set; and

[15]North American Industry Classification System

2. They were listed as having occurred within 1 week of each other.

Table 1.5: Default Date Counts

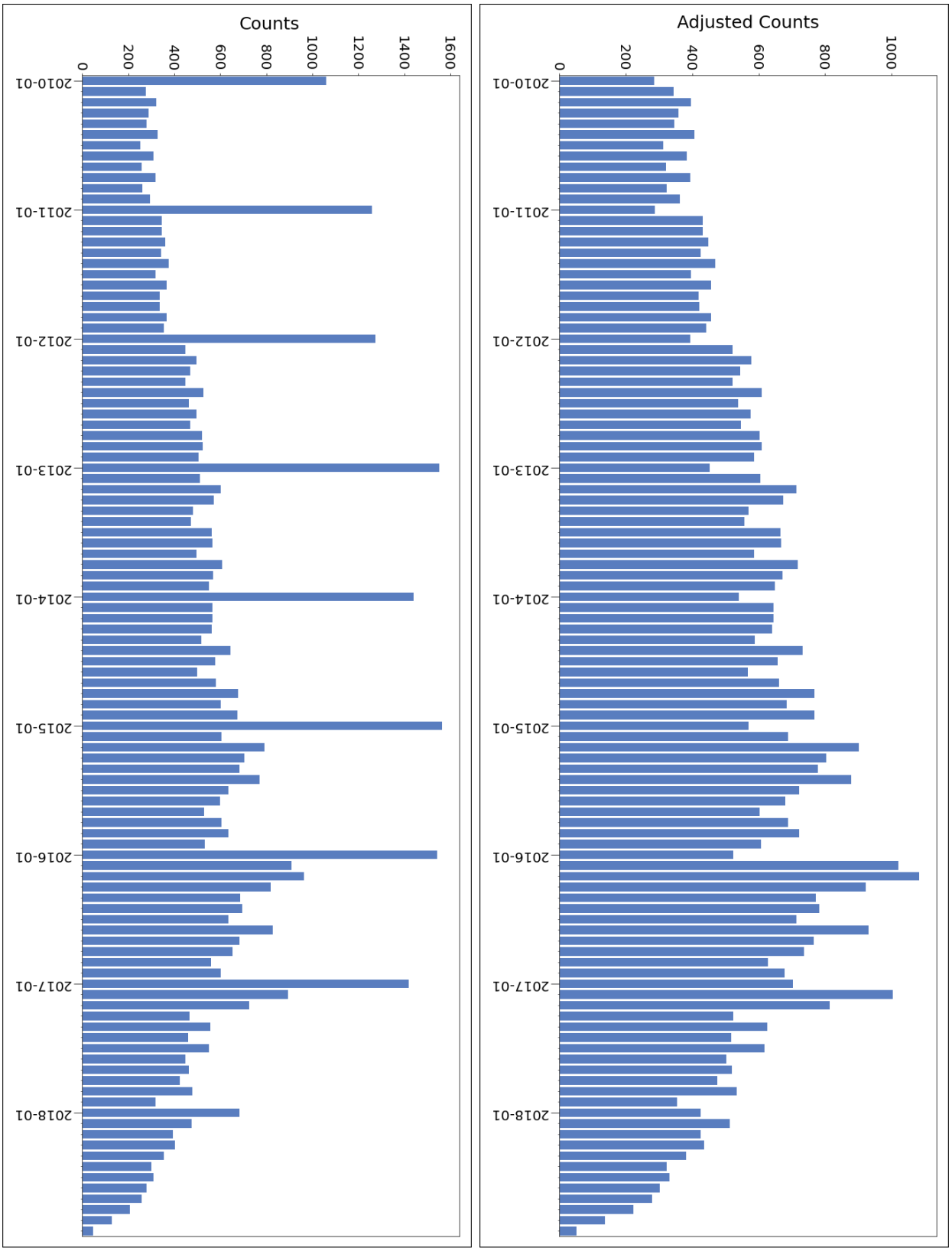
Default Date	Counts
Jan-10	830
Jan-11	1028
Jan-12	935
Jan-13	1168
Jan-14	965
Jan-15	1063
Jan-16	1076
Jan-17	793
Jan-18	287

1.3.3 Claim-exposure Data

More than 30,000 policies and their respective claims information from multiple insurers were analyzed. The proprietary data is the collection of policies underwritten by insurers and claim information if there exist claims against those policies. The dataset includes policy ID , start and end dates of policy, claim ID, claim date, claim amount, incident description, incident type (manually extracted from the incident description), employee count, geographic location, industry, and revenue.

Table 1.6: Claim Policy Database

Type	Variable	Information
Policy	ID	Policy identification number
	Start Date	Start date of the policy
	End Date	End date of the policy
Claim	ID	Claim identification number
	Date	Date of the claim filed
	Amount	Amount paid against claim
	Incident Description	Details of the incident
	Incident Type	Extracted from incident description
Organization	Employee Count	Number of Employees
	Geographic Location	Location where insured organization is based
	Revenue	Revenue of insured organization
	Industry	Industry of insured organization



(a) With Default Date (N=60380)

(b) Adjusted for Default Date

Figure 1.2: Cyber Event Counts until December 2018

Limitations: Some claims do not have policy numbers listed; such claims were excluded from this analysis. The incident type information is extracted manually from the incident description and is not an exhaustive list of incidents. There are certain claims that begin as hacking or social engineering but eventually lead to ransom demand - such incidents are classified as "ransomware". The revenue information was frequently missing and this was completed by either matching to the firmographic data set or, when employee counts were available, estimating revenue based on employee counts and industry classification.

1.4 Contributions

All four problems stated in the context of reporting delays in section 1.2.1 are addressed in the current research. Problems of under-reporting are currently under consideration.

1.4.1 Reporting Delays

The proposed solution "Debiased Delayed Distribution" algorithm addresses the reporting delays, the first problem of empirical distribution from raw data being biased towards the shorter delays. To be precise, delay and age histograms, h_Δ and h_A, are generated from the raw data.

- Delay Histogram, h_Δ, depicts the frequency distribution of a given delay over the time defined with range of bins.
- Age Histogram, h_A, depicts the frequency distribution of a given age over the time defined with range of bins.

The debiased delay distribution is then generated based on the ratio of the number of events with the given delay to the best estimate of the true number of events where age, a, is greater than or equal to delay, δ. Generating a debiased delay distribution with such an approach resolved the first open research question of delay distribution being biased to shorter delays.

The second problem of *non-stationarity* is addressed by generating the monthly debiased delay distribution over a two-year rolling window.

The third and fourth problems are addressed by fitting the modeled distribution on the debiased delay distribution with the optimization function. To be precise, the virtue of the fact that the modeled distribution is defined as a mixture distribution (combination of exponential and normal distributions) addresses the third problem of estimating delays beyond the longest delay in the data, $\delta > \delta_{max}$. Although the debiased distribution is defined on domain $[0, \delta_{\max}]$, the modeled distribution resulting from the mixture distribution has the domain defined $[0, \infty)$ allowing the delays beyond age. The fourth problem of dependency of longer delays on few data points is resolved by assigning weights to the certain terms of the optimization function: giving lesser priority to the delays further in time as marginally fewer number of events are anticipated with longer reporting delays.

1.4.2 Under-reporting

The existing research addresses the under-reporting problem from one aspect and does not consider population characteristics which involve multi-dimensional aspects. The proposed approach addresses the cyber domain attributes in terms of revenue, incident type and industry. Initially, the proposed method models under-reporting correction factor as a function of revenue. To address multi-dimension aspect, the approach proposes to find the scalar multiplier for the revenue given incident type and revenue given industry. These scalar multipliers are then multiplied to revenue correction factors to find the appropriate corrections for revenue given incident type, and revenue given industry. Similarly, the approach can be extended to three or more dimensions.

In addition, the results will make it easier for those in academia to create cyber risk models from data sets of publicly known cyber incidents, without requiring access to claims data.

Chapter 2

Correcting for Reporting Delays in Cyber Incident

2.1 Abstract

With an ever evolving cyber domain, delays in reporting incidents are a well-known problem in the cyber insurance industry. Addressing this problem is a requisite to obtaining the true picture of cyber incident rates and to model it appropriately. The proposed algorithm addresses this problem by creating a model of the distribution of reporting delays and using the model to correct reported incident counts to account for the expected proportion of incidents that have occurred but have not yet been reported. In particular, this correction shows an increase in the number of cyber events in recent months rather than the decline suggested by reported counts. The cyber models, with corrected counts for reporting delay, provides cyber modelers with better estimate of the true rate of incidents allowing them to understand the current cyber risk landscape.

2.2 Introduction

With new attack vectors emerging regularly, the cyber security domain is evolving rapidly. Hence, even the most up-to-date data cannot be considered complete. Cyber incidents take

a long time to become known and even longer to appear in online databases if known at all [131]. While some cyber events are known immediately after they occur, most events are often reported many months or years after the event actually occurred, resulting in biased data. As an example, Marriott's major cyber incident occurred in 2014 but was only reported in 2018 [126]. Major cyber events become headlines in leading newspapers when reported publicly rather than at the time of occurrence. Smaller cyber events may never be reported at all, or have extreme delays, as only public companies and those with personally identifiable information may be obligated to report. Sometimes reporting can take 5-10 years, for various intentional or unintentional reasons - failing to realize that a cyber incident happened, failing to immediately determine the extent of accessed or stolen data, or deciding not to publicize the incident for fear of reputation risk and consequent financial impacts. As a result, reporting delays are often observed in historical cyber event databases. These databases show a decrease in cyber incidents. In contrast to this, Coleman et al. raised the concern that cyber incidents would remain undetected due to advanced threat techniques [9].

Cyber risk modeling firms rely upon historical data to build their models, which are in turn relied upon by cyber insurers for underwriting, portfolio management, and risk transfer. To build robust loss estimation models for today's evolving cyber world with state-of-the-art techniques, the most recent and updated information is required, with as little bias as possible. Correcting reporting delays in these databases is therefore a key requirement to have trustworthy cyber insurance models. With the necessary corrections, one can then properly examine temporal trends in the targeting of industries or in attacker tactics.

Harris, research from medical domain, described reporting delays as a statistical problem for the first time [55]. Heisterkamp et al. and Brookmeyer and Damiano made distributional assumptions and built linear/quadratic models whereas Cheng and Ford, and Rosenberg suggested Poisson models [21][26][61] [62][63][114]. These model approaches are easy to implement but assume stationary reporting delays and do not capture trends. Downs et al., Morgan and Curan, Healy and Tillett, and Heisterkamp et al. fitted exponential, integrated logistic and log-linear models to capture trends [30][31][59]

[61][62][63][100]. Brookmeyer and Liao, Esbjerg et al., Gail and Brookmeyer, and Kalbfleisch and Lawless applied conditional probabilities to capture trends but this resulted in overfitting [23][35][43][74]. Lawless proposed a multinomial model with distributed random effects based on Dirichlet/Poisson/Gamma distributions to capture trends in a timely fashion but failed to handle longer delays [85].

Wang suggested maximum likelihood estimation (MLE) based on non-parametric and semi-parametric approaches but with complete[16] data [130]. Harris suggested correcting COVID cases with an expectation-maximization (EM) algorithm and trained the model with complete data to correct test data [57]. Weinberger et al. and White et al. proposed a simpler method based on proportions but also required complete data to train [132][134].

Keiding and Moeschberger, Midthune et al., and Wang applied truncated models to avoid random effects but require stable reporting delays [75][96][130]. Again, this approach requires stationarity not in the delays but also in overall average in delays as well. Given the non-stationary behavior in reporting delays, the stationarity assumption is not appropriate.

Bastos et al., Chitwood et al. and Hohle and An Der Heiden suggested a Bayesian and hierarchical approach with Poisson and Negative Binomial distributions. This approach is easy to implement but makes single distributional assumptions [15][27][65]. The single distribution is not suffice to capture the reporting delays distribution as if behaves differently for short-term and long-term.

Noufaily et al. suggested a log-likelihood approach with a truncation model that is data driven but sensitive to the choice of three fixed reporting time steps at which the reports are sent to the database [103][104].

Bastos et al. suggested a chain-ladder approach with spatio-temporal (locations) in counts and co-variate effects, but this approach is sensitive to outliers [15].

Avanzi et al., Jewell, Zhao et al, and Zhao and Zhou investigated cyber claims data to account for reporting delays from a capital reserving perspective [11][73][137][138]. This

[16]Complete data - No further events are expected to be reported with delays.

problem is different from the one being investigated, since reporting delays in claims are due only to detection delays.

As Brookmeyer and Liao stated, none of these approaches deal with delays longer than any data previously reported and does not allow delays beyond the maximum delay in the data. As a result, estimation of reported counts becomes challenging[23].

Most of the literature on reporting delays is found in the medical space whereas there is no literature found in cyber space (to the best of our knowledge). This might be due to unavailability of appropriate data or the cost associated to obtain the same. However, Coleman et al. does examine both the distribution of the number of days to discover cyber incidents and the number of days to disclose them [9].

The rest of this chapter is organized in five sections - Sec. 2.3 defines theoretical concepts used in the proposed approach, Sec. 2.4 describes the proposed approach, Sec. 2.5 discusses the problems faced during implementation, Sec. 2.6 discusses the interpretation of parameters, corrections for reporting delays and their validation, and Sec. 2.7 discusses the conclusion.

2.3 Theoretical Concepts

2.3.1 Mixture Distribution

Hampel credited Tukey for suggesting a mixture of distributions - a weighted linear combination of two distributions [129]. The mixture distribution is applied to address a bimodal distribution where two distributions are linearly combined based on the proportion of their explainability to describe the overall distribution [48]. In other words, overall distribution is described with two distributions where the contribution of each distribution gradually varies over time i.e. the distribution playing significant role initially plays minimal role later and vice versa. Mathematically, a mixture distribution can be defined as shown in Eq.2.1.

$$F(x) = wF_1(x) + (1 - w)F_2(x) \tag{2.1}$$

where F_1 and F_2 are two linearly combined functions with weights w and $(1-w)$.

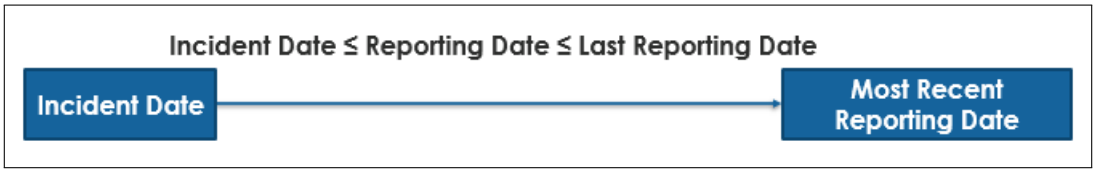
The mixture distributions are applicable where different segments of the data are modeled separately to capture different characteristics. Mathematically, it is easier to model two individual components and provide better explainability than the overall distribution.

2.4 Proposed Approach

The proposed approach consists of estimating the reporting delay distribution, f_Δ, from the empirical data. The algorithm builds on two concepts developed from incident and reporting dates: Delay and Age.

- Delay, δ, refers to the difference between incident date and reported date, as shown in Fig. 2.1a.
- Age, a, refers to the difference between incident date and most recent reporting date irrespective of when the given event is reported, as shown in Fig. 2.1b.

(a) Delay



(b) Age

Figure 2.1: Concepts: Delay and Age

Using a debiased delay distribution, a corrected count of events with age, a, computed by dividing the raw counts by $F_\Delta(a)$, where $F_\Delta(a)$ defines the proportion of events that are reported within delay, δ, of less than age, a or, equivalently, the proportion of events that are reported as of today (last reporting date/most recent reporting date). The debiased delay distribution, f_Δ, is generated from the empirical raw data.

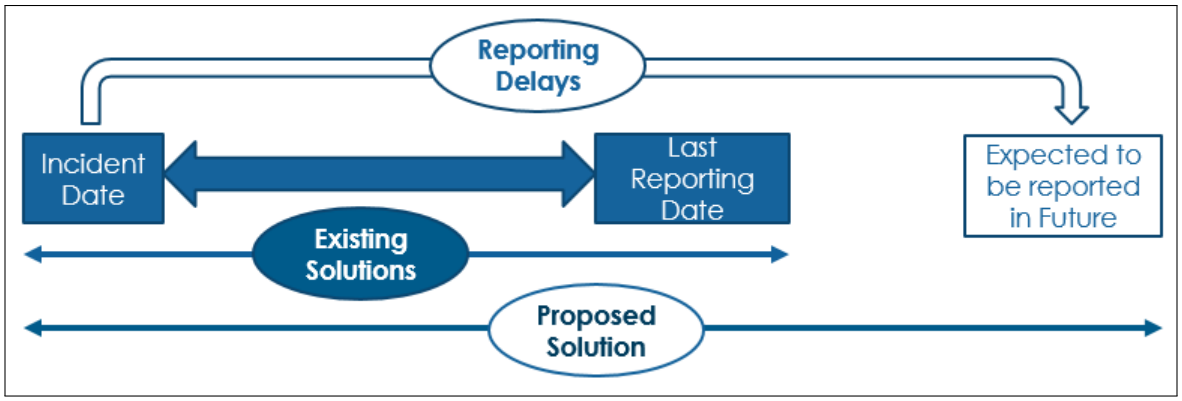


Figure 2.2: Existing Solutions Vs Proposed Solution

2.4.1 Generating the Debiased Empirical Delay Distribution

Inspired from Brookmeyer and Liao, the proposed algorithm works on the limitation that delay, δ, cannot be considered beyond age and hence only conditional distribution on delay less than or equal to age, $\delta \leq age$, can be estimated [23].

The algorithm corrects the incident counts with the cumulative distribution function, F_Δ, estimated from empirical data. The algorithm applies a top-down approach to estimate the distribution "from the outside in", accounting for the estimated proportion of events that have occurred but not yet reported as the distribution being computed. Let $A_{\max} := \max_{i \in \mathcal{I}} A_i$ be the maximal age of any event in the data set. Also, let $h_A(a)$ be the number of incidents of age a and let $h_\Delta(\delta)$ be the number of incidents with delay δ. Formally, these can be represented as shown in Eq.2.2 and Eq.2.3.

$$h_A(a) = |\{i \in \mathcal{I} : A_i = a\}| \tag{2.2}$$

$$h_\Delta(\delta) = |\{i \in \mathcal{I} : \Delta_i = \delta\}| \tag{2.3}$$

Then the delay distribution can be estimated as shown in Eq.2.4.

$$f_\Delta(\delta) = \frac{h_\Delta(\delta)}{\sum_{a=\delta}^{A_{\max}} h_A(a)/F_\Delta(a)} \tag{2.4}$$

Intuitively, the distribution is generated based on the ratio of the number of events with the given delay period, $h_\Delta(\delta)$, to the best estimate of the true number of events whose age is old enough to be seen in the incident data set i.e. either the same delay period or more, $\sum_{a=\delta}^{A_{\max}} {}^{h_A(a)}/_{F_\Delta(a)}$.

The debiased delay distribution can be implemented as shown in Algorithm 1 or pictorially Flowchart 2.3. The distribution is considered complete over $[0, \delta_{\max}]$.

Algorithm 1 Algorithm for computing the debiased empirical delay distribution

Input: The histograms, h_A and h_Δ, computed as in Eqs. (2.2) and (2.3), respectively.

Output: The distribution f_Δ.

function ComputeDelayDistribution(h_A, h_Δ)
  $A_{\max} \leftarrow \max_{i \in \mathcal{I}} A_i$
  $F_\Delta(A_{\max}) \leftarrow 1$
  $\delta_{max} \leftarrow \max_{i \in \mathcal{I}} \delta_i$
  for $\delta \leftarrow A_{\max}$ to $\delta = 0$ **do**
    $den \leftarrow 0$
    for $a \leftarrow \delta$ to $\delta_{\max}$ **do**
      $den \leftarrow den + h_A(a)/F_\Delta(a)$ ▷ Computes denominator
    end for
    $f_\Delta(\delta) \leftarrow h_\Delta(\delta)/den$ ▷ Computes PDF
    $F_\Delta(\delta - 1) \leftarrow F_\Delta(\delta) - f_\Delta(\delta)$ ▷ Updates CDF
    $\delta_{max} \leftarrow \delta$
  end for
  return f_Δ
end function

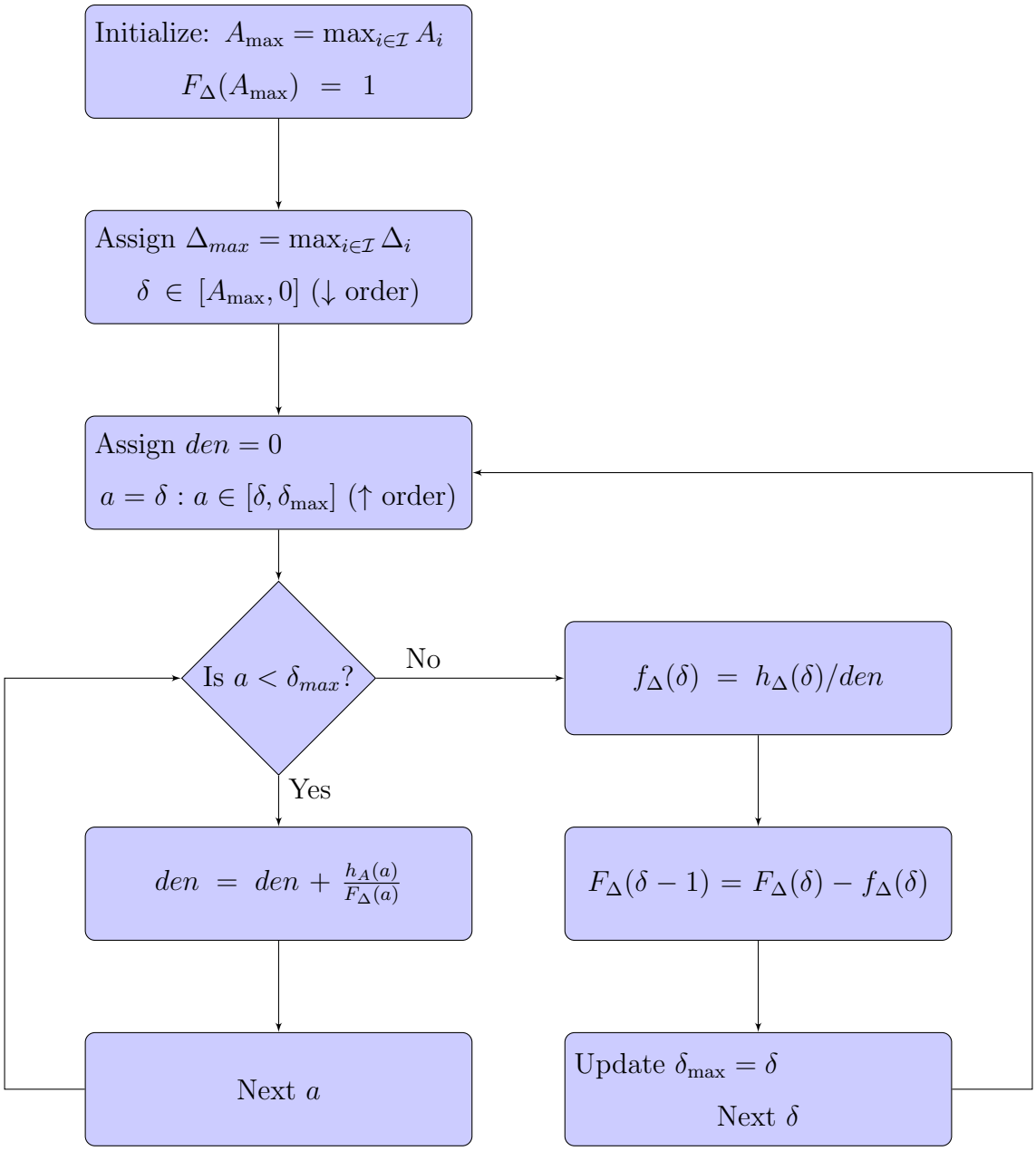


Figure 2.3: Delay Distribution

2.4.2 Why Debiased Delay Distribution

As discussed in Section 1.2.1, the direct estimation of delay distribution led to certain problems.

Biased to shorter delays

The first problem highlighted bias to shorter delays, whereas the third problem emphasized on existing methods considering reporting delays, δ, until the age of the oldest incident, a, in the data. Hence, this restricts the ability to estimate beyond age, a. Mathematically, the problem of data being biased to shorter delays can be expressed as

$$\delta \leq a \qquad \text{for given age, } a \tag{2.5}$$

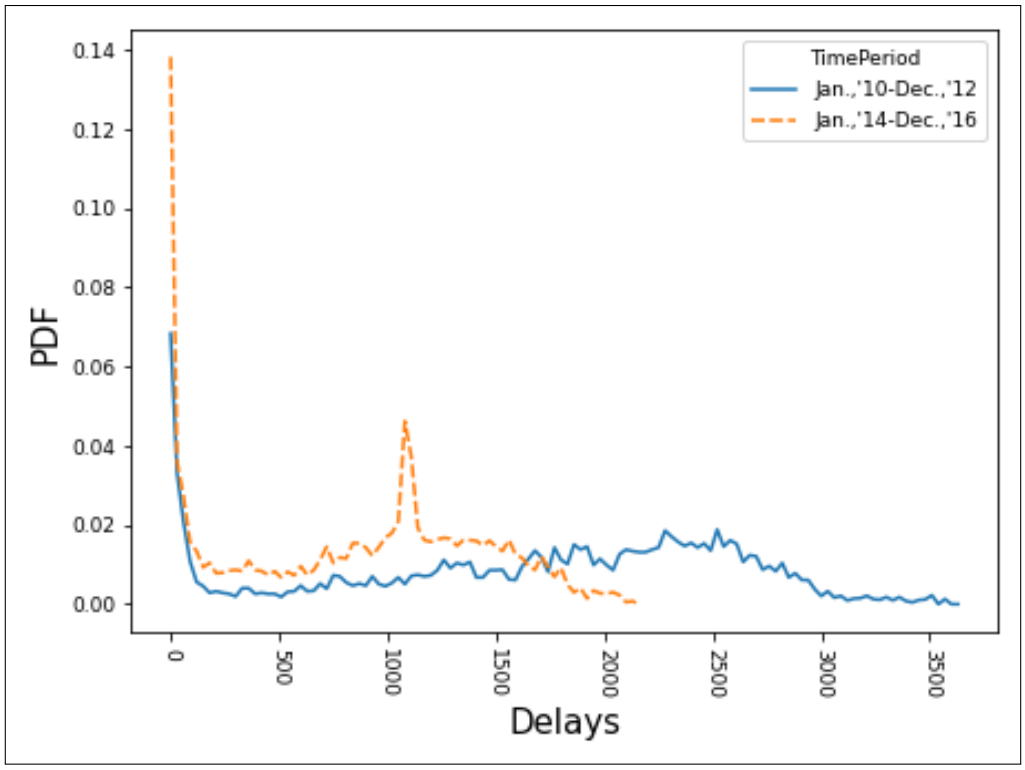
Both the problems can be resolved if the delay distribution is stationary. The distribution with stationarity allows the estimation of delays, δ, beyond the age, a. The proposed approach in Algorithm 1 corrects this problem.

While Algorithm 1 provides a solution to the bias in shorter delays, it does not resolve the assumption of stationarity - which may or may not be present.

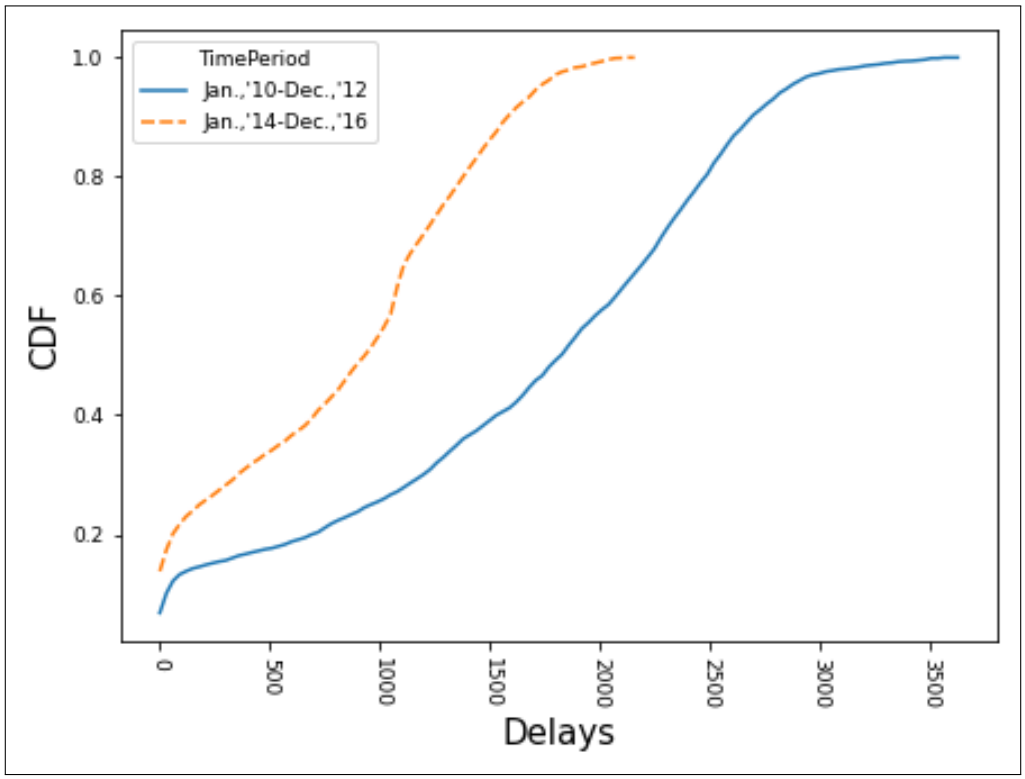
Non-stationarity

From the discussion regarding the distributional stationarity, it is clear that stationarity in reporting delays is a key requirement to estimate the delays, δ, beyond the age, a, with a debiased delay distribution.

To test stationarity, the debiased delay distribution, using Algorithm 1, was generated on two-year windows at different points of time. The given window collects events that occurred within the given two-year period irrespective of when they are reported. The first two-year window collected the events that occurred within the period Jan.,'10 to Dec.,'12 (inclusive) and the second collected events within the period from Jan.,'14 to Dec.,'16(inclusive). Both two-year windows collected events reported until Dec.'18. The two debiased delay distribution plots, shown in Fig. 2.4, highlighted two problems: The debiased delay distribution indicated the presence of a bimodal nature due to the existence of two local modes in its structure- the bimodal nature also confirmed in Mandiant recently in technical report (2022) [93]. The debiased delay distribution is found to be non-stationary, which needs to be addressed; this is in line with problem 2 from Section 1.2.1.



(a) Probability Distribution



(b) Cumulative Distribution

Figure 2.4: PDF and CDF of Delay Distribution generated for Dec.,'12 and Dec.,'16

To explain the bimodal structure of debiased delay distribution, the modeled distribution is described with two distributions. From Fig. 2.4a, the two distributions are selected informatively based on the probability distribution of the debiased delay distribution: exponential distribution and normal distribution. The exponential distribution is selected for shorter delays, considering the initial PDF structure of the debiased delay distribution in Fig. 2.4a. On the other hand, the normal distribution is selected for longer delays, considering the later structure of distribution in Fig. 2.4a.

In order to deal with non-stationarity, debiased delay distributions are estimated monthly over a two-year rolling window. The parameters of modeled parametric distribution is estimated by fitting cumulative distribution. The empirical debiased delay distribution is estimated with raw data with Algorithm 1 on each monthly two-year rolling window. For each two-year window, a parametric modeled distribution is estimated using the optimization algorithm where the parameters are computed such that it provides a good fit for the debiased delay distribution, generated with Algorithm 1. It is important to note here that the modeled delay distribution is restricted to the domain as of the debiased delay distribution, i.e. $[0, \delta_{\max}]$, where $\delta_{\max}$ is the maximum delay in the given window.

2.4.3 Generating the Modeled Delay Distribution

As mentioned in the previous section, the modeled delay distribution is determined by matching the debiased delay distribution with the modeled delay distribution on the specified domain $[0, \delta_{\max}]$. Considering the PDF and CDF plots, Fig. 2.4 indicates that a single distribution will not suffice; rather, a bimodal distribution is required. Multiple distributions were fitted on the debiased distribution.Finally, the mixture of two distributions selected based on Kullback-Leibler divergence test[17]: Exponential and Normal distributions. Mathematically, the selected bimodal distribution can be expressed in terms of mixed

[17]A statistical distance to measure how one probability distribution differs from the other [80][92].

distribution, as shown in Eq. 2.6.

$$F_\theta(\delta) = \alpha F_{Exp}(\delta, Scale) + (1-\alpha) F_N(\delta, \mu, \sigma) \tag{2.6}$$

$$\text{where } F_{Exp} = \text{Exponential CDF with parameter, } Scale = {}^1\!/\!_\lambda$$

$$F_N = \text{Normal CDF with parameters, } \mu \text{ and } \sigma$$

Notice here parameter λ is the Exponential distribution parameter and refers to the constant average rate at which the events occur. The parameter λ is different from the lagrange multiplier mentioned earlier in Section 1.1.1 in context of AIDS study by Cheng [26].

Perfect Scenario: Under the perfect scenario, α-parameter is expected to be 1 where all cyber incidents are detected immediately but take time to get reported. As a result, Eq. 2.6 reduces to the exponential component and second part associated with normal distribution reduces to zero.

Since debiased delay distribution is defined on domain $[0, \delta_{\max}]$ and Normal distribution is defined on the infinite domain (∞, ∞), the mixed modeled CDF needs to be adjusted to be on the positive domain $[0, \infty)$. The CDF in Eq. 2.6 defined on $[0, \infty)$ can be expressed as shown in Eq. 2.7.

$$F_\theta(\delta) = \frac{\alpha(F_{Exp}(\delta, Scale)) + (1-\alpha)\ \overbrace{(F_N(\delta, \mu, \sigma) - F_N(0, \mu, \sigma)}^{\text{Truncated Normal Distribution until } \delta}}{\alpha + (1-\alpha)\ \underbrace{(1 - F_N(0, \mu, \sigma)}_{\text{Truncated Normal Distribution over } [0, \infty)}} \qquad 0 \le \delta \le \infty \tag{2.7}$$

Furthermore, the corrected empirical delay distribution, denoted as $F_\theta^{'}$, defined on domain $[0, \delta_{\max}]$, in terms of Eq. 2.6 can be expressed as shown in Eq. 2.8.

$$F'_\theta(\delta) = \frac{\alpha(F_{Exp}(\delta, Scale)) + (1-\alpha)\ \overbrace{(F_N(\delta, \mu, \sigma) - F_N(0, \mu, \sigma)}^{\text{Truncated Normal Distribution until } \delta}}{\alpha(F_{Exp}(\delta_{\max}, Scale)) + (1-\alpha)\ \underbrace{(F_N(\delta_{\max}, \mu, \sigma) - F_N(0, \mu, \sigma)}_{\text{Truncated Normal Distribution over } [0, \delta_{\max}]}} \quad 0 \leq \delta \leq \delta_{\max} \tag{2.8}$$

The bimodal nature of the reporting delay distribution is a combination of two distributions where the exponential distribution estimates the events discovered almost immediately, and the normal distribution estimates events with longer delays due to both discovery time and public disclosure time. The parameter, α, can therefore be interpreted as the proportion of events that can be discovered right away by the organization but are yet to be reported. Since the modeled delay distribution defined on domain $[0, \delta_{\max}]$, the normal distribution is truncated from both sides (it is defined on the infinite domain $(-\infty, \infty)$, whereas the exponential distribution is truncated from only one side (it is defined over the positive domain $[0, \infty)$).

2.4.4 Defining the Optimization Function

There are specifically two challenges in defining an optimization function in this space.

- Multiple combinations of parameters exist that give approximately the same distribution when restricted to the domain [0, $\delta_{\max}$] but vary considerably when the differing weight of total distribution is considered in the unrestricted positive domain.
- Due to the shrinkage of data in recent two-year windows, the optimization algorithm provides increasingly unstable estimated parameters.

The optimization function used is shown in Eq. 2.9. The first term[18] $\| \log_{10} F'_\theta - \log_{10} F_\Delta \|^2$ reduces the CDF difference between debaised empirical delay distribution and modeled distribution over the domain of the window $[0, \delta_{\max}]$. The purpose of applying $\log_{10}$ weights

[18] F'_θ defined in domain $[0, \delta_{\max}]$ computed as shown in Eq. 2.8

is to give priority to the initial months while fitting the distribution. The rationale behind $\log_{10}$, is to obtain CDF values close from the point of the ratio between the two distributions (modeled and corrected empirical) and not in terms of absolute difference: $\log_{10}$ CDF difference between 0.03 and 0.06 is substantially larger than a difference between 0.93 and 0.96 even though the absolute difference is equal.

As the first factor states, there exists multiple combinations of parameters providing a good fit for the empirical debiased delay distribution but results in an overall bad fit on the domain $[0, \infty)$. This problem is not of much concern when most of the distribution is defined and captures delays beyond the second peak from a debiased delay distribution. However, it becomes a significant concern when parameters for recent two-year windows are computed. In order to avoid this, it is vital to capture modeled delay distribution beyond max in the optimization function. The second term $\| \log_{10} S_\theta - \log_{10} S_{\theta'} \|^2$ is introduced to ensure that the consecutive modeled distribution parameters are consistent with each other beyond the maximum delay, $\delta_{\max}$. In order to deal with unstable parameters as we approach recent months, another set of weights is assigned to the first two terms (Eq. 2.9), CDF matching until and beyond max, to give priority to the CDF matching to the segment of the delay distribution where the major part of the delay distribution is defined.

The third and fourth terms are penalizing terms - $F_N^2(0, \mu, \sigma)$ term penalizes for negative delays introduced by normal distribution, whereas $S_\theta^2(10Y)$ term penalizes the delays beyond 10 years. Since the domain of normal distribution is defined over $(-\infty, +\infty)$, it incorporates negative delays which would not be possible so third penalizing term was included in the optimization. The delays beyond 10 years are not expected so fourth terms was added to the optimization function.

Mathematically, the optimization is defined, as shown in Eq. 2.9.

$$\begin{aligned}\theta_{Opt} = \underset{\theta=(\alpha,Scale,\mu,\sigma)}{argmin} \frac{\delta_{\max}}{\delta_{Fix}} & \underbrace{\| \log_{10} F'_\theta - \log_{10} F_\Delta \|^2}_{\delta \in [0, \delta_{\max}]} \\ & + \left(1 - \frac{\delta_{\max}}{\delta_{Fix}}\right) \underbrace{\| \log_{10} S_{\theta'} - \log_{10} S_\theta \|^2}_{\delta \in (\delta_{\max}, \delta_{Fix}]} \\ & + \underbrace{F_N^2(0, \mu, \sigma)}_{\delta < 0} + \underbrace{S_\theta^2(10Y)}_{\delta > 10Years}\end{aligned} \tag{2.9}$$

where δ_{Fix} is the maximum value of δ in the dataset.

F'_θ can be computed using mixed distributions as shown in Eq. 2.8. The CDF beyond $\delta_{\max}$, S_θ, defined as defined as complement of F_θ, defined in Eq. 2.7 over the domain $[0, \infty)$ and can be expressed as shown in Eq. 2.10.

$$S_\theta = 1 - F_\theta \tag{2.10}$$

In Eq. 2.9, the second term reduces to zero in the absence of previous parameters at an initial two-year window, as shown in Eq. 2.11.

$$\| \log_{10} S_{\theta'} - \log_{10} S_\theta \|^2 = 0 \tag{2.11}$$

θ' refers to optimal parameters at previous step.

Optimizer

Considering the complexity of the optimization function in Eq. 2.9, it is not feasible to compute the derivative. Hence, a derivative-free algorithm is required to solve the optimization problem. The covariance matrix adaptation evolution strategy (CMA-ES) is applied to compute the modeled distribution parameters. It is a derivative-free optimization algorithm, and such algorithms are typically used when derivatives are difficult or costly to compute [49][50][51]. Compared to other optimization methods, CMA-ES makes fewer assumptions about the underlying objective function [52][53][54][112]. A number of real

world problems are successfully resolved with this algorithm[19]. The approach requires neither derivatives' nor the functions' values but ranks the potential candidate solutions to learn the sample distribution. It is particularly applicable for ill-conditioned functions[20].

Since the optimization problem involves estimation of modeled delay distribution parameters, a small change in parameters has the potential to impact the solution considerably. CMA-ES is an evolutionary algorithm with evolutionary computations which works on two concepts - Maximum Likelihood (ML) and Evolution Paths (EP).

- ML finds the candidate solutions with high probability and looks for the incremental step to further maximize the likelihood.
- EP has dual benefitwhile the covariance matrix allows for a quick variance escalation in the desired direction, the step size control may prevent convergence until the optimal solution is found [10][49][50].

The Python package cma is used to implement the CMA-ES algorithm [51].

2.4.5 Compute Correction Factors

The proposed algorithm, Algorithm 2, estimates a series of parametric reporting delay distributions generated from data. The modeled delay distribution is computed based on parameters generated on a corrected delay distribution from the optimization Eq. 2.9. The distributional parameters are extracted from monthly rolling delayed distributions generated over two-year windows. Although the entire distribution is fitted, only one point from the distribution is used to obtain the correction factor for age computed for a given month w.r.t. the most recent reporting date in the data.

[19] CMA Applications webpage lists various real world problems published until 2009
[20] The function where small change in the inputs can bring large change in solutions.

Algorithm 2 Algorithm to compute the correction factor for given month

Input: Delay Distribution for the given two-year window(f_Δ), Start date of the window (d), Start & end dates of entire dataset (D_S, D_E) and previous parameters (θ).

Output: Correction Factor (CF)

function ComputeCorrectionFactor(h_Δ, d, D_S, D_E)
  $\delta_{Fix} = \delta(D_S, D_E)$ ▷ Days between D_S and D_E
  $\delta_{\max} = \delta(d, D_E)$ ▷ Days between d and D_E
  $\theta_{Opt} \leftarrow Optimize(F_\Delta, \theta_{Fix}, \theta_{\max}, \theta')$ ▷ CMA-ES Optimization, Eq. 2.9
  $age = \delta_{\max}$
  $CF = \frac{F_{\theta_{Opt}}(age)}{F_{\theta_{Opt}}}$ ▷ Truncated CDF on $[0, \infty)$
  return CF
end function

Fig. 2.5 shows the correction factors for year 2017(Fig. 2.5a) and 2018 (Fig. 2.5b) for US market.

The debiased counts are computed with the CDF of the modeled distribution defined over a positive domain $[0, \infty)$. The modeled CDF defines the proportion of the events, which are reported until the *age* gap. The corrected counts for a month, m, are computed as shown in Eq. 2.12.

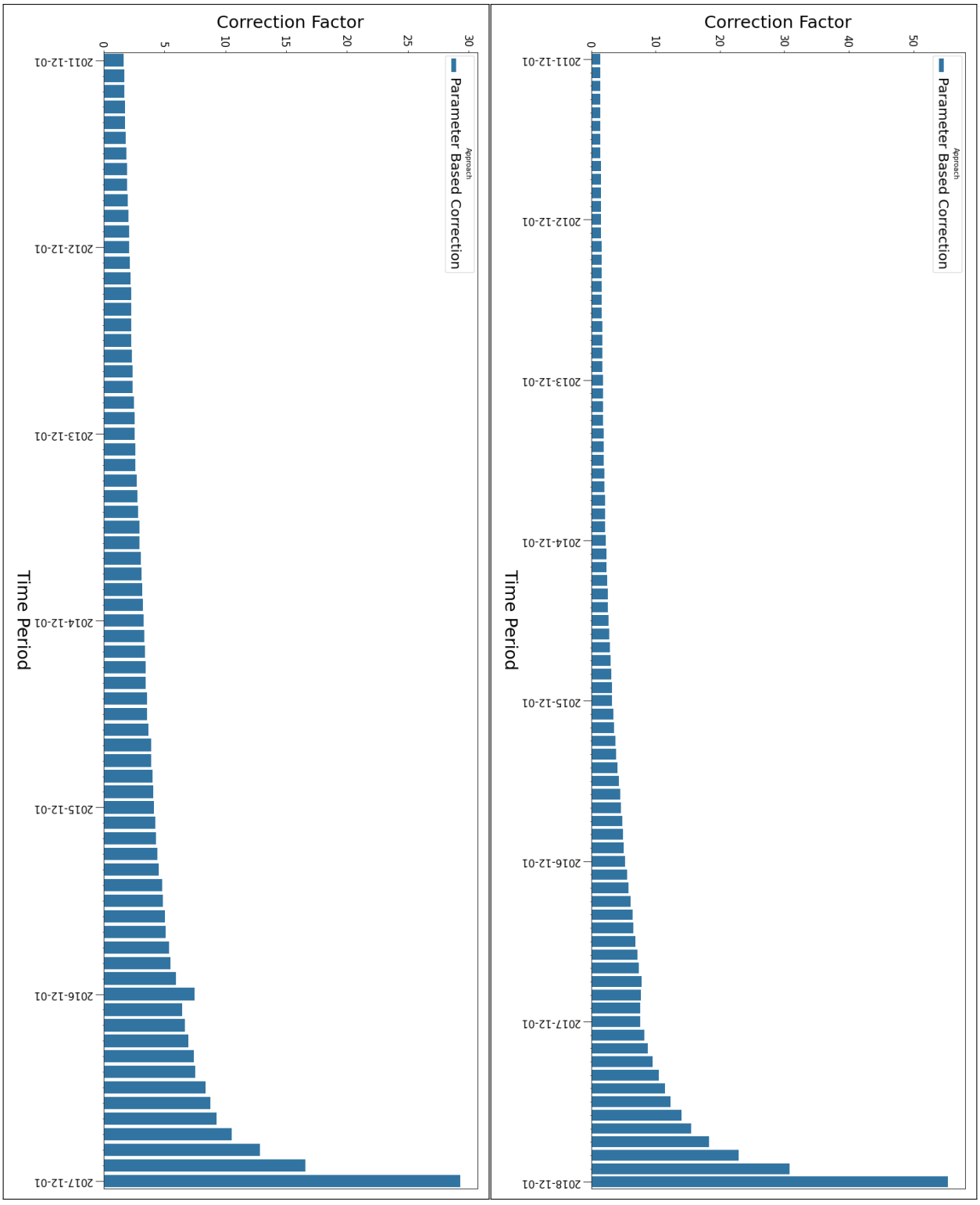
$$\text{Corrected Count for month, 'm'} = \frac{\text{Reported Counts for month, 'm'}}{F_\theta(a)} \tag{2.12}$$

where a is age of the given month, 'm'.

2.5 Implementation Problems

2.5.1 Optimization Function

The first attempt to optimize the function, showed in Eq.2.9, was based on $\|F'_\theta - F_\Delta\|^2$, however a significant portion of the parametric distribution is shifted to the left due to



(a) For data collected until 2017

(b) For data collected until 2018

Figure 2.5: Correction Factors for US

the impact of the normal distribution defined over (∞, ∞). As the mean parameter of the normal distribution moves to zero and the standard deviation increases, the distribution gets shifted below zero. However, the distribution below zero is ignored, and the truncation simply redistributes any part of the normal distribution that has a negative reporting delay to the rest of the distribution. To minimize the impact, the solution is to add a penalty term for the portion of the distribution that is below zero. One could easily set the negative values as zero but it is not appropriate as it would impact the CDF value in the optimization function, resulting in unexpectedly higher values at zero. The second attempt included the inclusion of term $F_N^2(0, \mu, \sigma)$. So, the squared CDF term for the untruncated distribution evaluated until zero was included. The purpose behind inclusion of $F_N^2(0, \mu, \sigma))$ term is that the normal distribution has an infinite support, and to make an approximation small for delays below zero $\delta < 0$. On the contrary, if the approximation is not small, then that would indicate the model itself is wrong. In the current scenario, the optimizer provides radically different parameters that give approximately the same values, so the model is correct, but the appropriate set of constraints are not placed in the optimization problem.

The inclusion of $(S_{\theta'} - S_\theta)^2$ term, beyond $\delta_{\max}$, helps in two ways. First, it minimizes the distribution defined beyond $\delta_{\max}$ where $S_\theta = 1 - F_\theta$ and θ' is the previous month's optimized parameters. Second, it ensures that the estimated $S_{\theta'}$ beyond $\delta_{\max}$ based on the previous month is in line with the current month estimated S_θ beyond $\delta_{\max}$. Despite all corrections, the distribution fit, based on the shape of optimization function (Eq. 2.9), was lower than expected for the most recent months. Another attempt was made with $\log_{10}$; this would allow the algorithm to place more emphasis on a good fit for the first months weights for first term, $\|F'_\theta - F_\Delta\|^2$. This gave an acceptable fit but resulted in very high correction factors.

The purpose of having $\log_{10}$ weights was to compare CDF values from a ratio perspective rather than the absolute differences between the two distributions (parametric and empirical) - $\log_{10}$ differences. The absolute difference between 0.03 and 0.06 is the same as between 0.93 and 0.96, whereas the absolute $\log_{10}$ difference between 0.03 and 0.06 is higher as compared

to the absolute $\log_{10}$ difference between 0.93 and 0.96. The aim here is to give priority to fitting CDF for shorter reporting delays as compared to longer ones.

All the above measures provided a good fit for the initial months where the CDF is defined for most of the period, but not otherwise resulting provided unacceptably higher corrections for recent period where CDF is defined for smaller period of time. Finally, another set of weights were assigned to the CDF matching until and beyond $\delta_{\max}$ to give priority to the CDF matching to the segment of the delay distribution where the major part of the delay distribution is defined.

2.5.2 CMA-ES Optimizer: Initial set of values

The choice of initial parameters becomes challenging because optimization algorithms frequently converge to different solutions when started from different initial conditions of local minima. Sometimes the direction the optimization algorithm moves the solution to is poor for some initial conditions but good for others. As a result, trying multiple initial conditions is a common approach with optimization algorithms.

In the current study, multiple sets of initial parameters were selected and tested with all converging to a similar solution. Finally, the simpler set of initial values was chosen.

2.6 Results

The objective of the experiment is to correct the cyber incidents for the reporting delays. Monthly cyber incidents are collected from 2010 onward until certain year (2017 and 2018) and corrected beyond the age of given data period. The proposed algorithm, Algorithm 1, was used to compute corrected delay distribution. The corrected count of events with given age, a, are computed by dividing the raw counts by the proportion of events that are reported within delay, δ, of less than age, a, $F_{\Delta}(a)$. The approach is validated by correcting 2017 event counts for a year ahead against 2018 counts and 2018 year ahead counts against 2019 counts.

The year ahead corrections for month, m, can be computed by modifying Eq. 2.12, as shown in Eq. 2.13.

$$\text{Year Ahead Correction(m)} = \text{Counts(m)} \times \frac{F_\theta(age + 360)}{F_\theta(age)} \tag{2.13}$$

2.6.1 US level Corrections

Fig. 2.6 shows two examples of plots comparing PDFs of the fitted parametric modeled distribution, its truncation to the domain $[0, \delta_{\max}]$, and the debiased empirical delay distribution. Fig. 2.6a shows this comparison for the two year window starting from July 2012 until June 2016 and Fig. 2.6b shows this comparison for the most recent window starting from January, 2017 until December 2018.

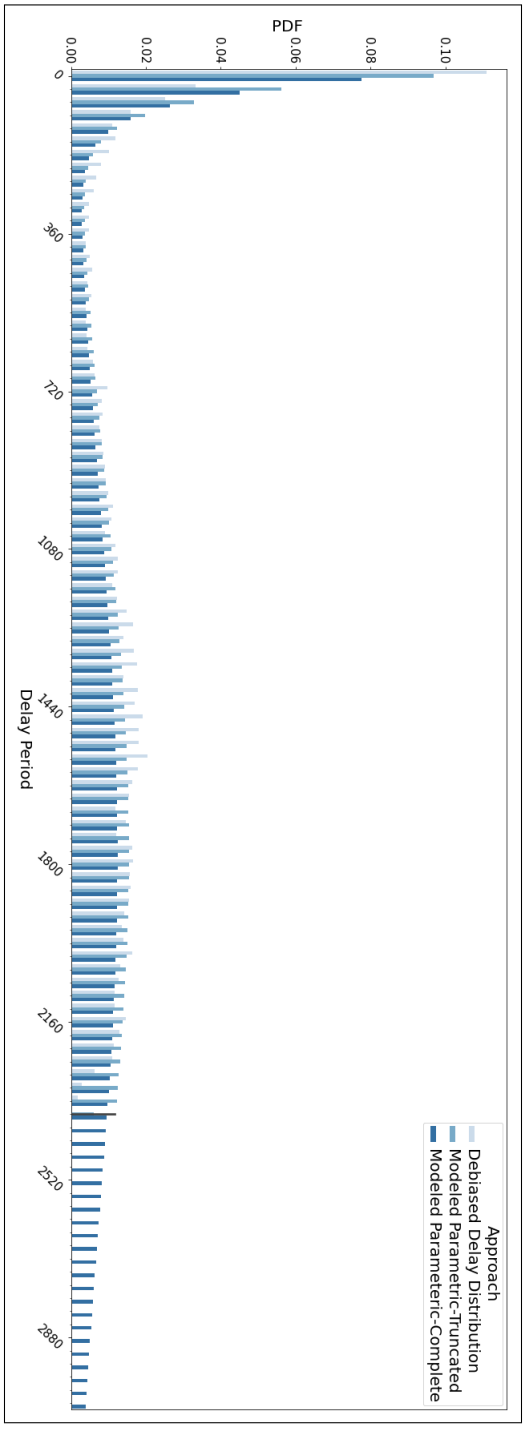
2.6.2 Parameters and Interpretation

Fig. 2.7 shows the parameter plots of the delay distribution generated for each monthly two year rolling window.

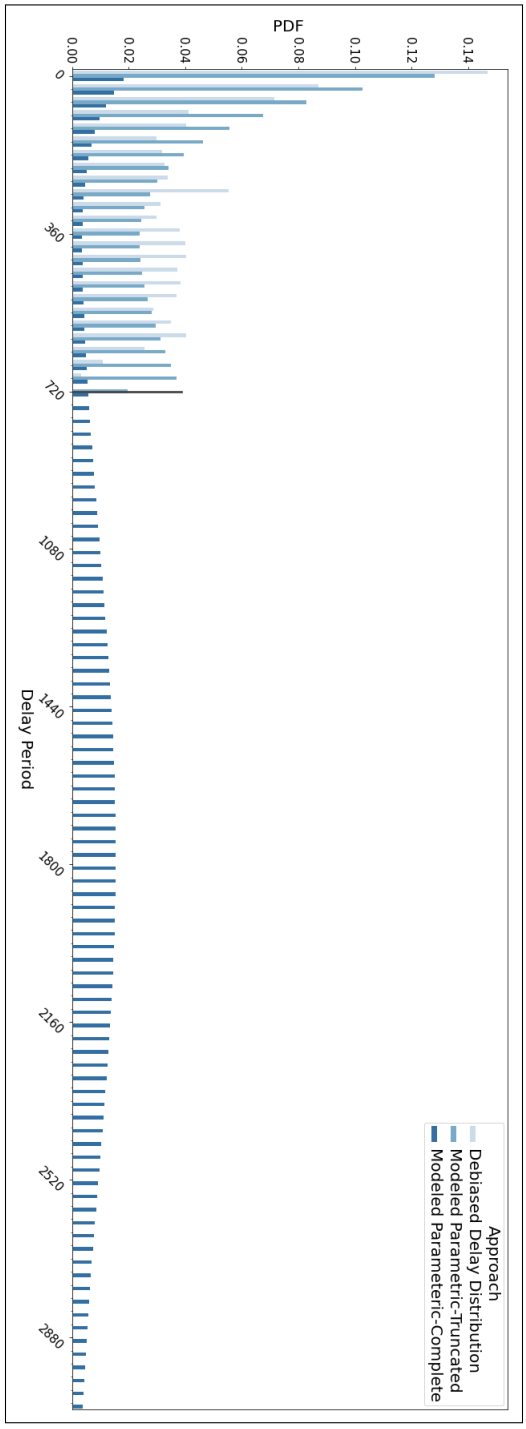
The alpha plot (Fig. 2.7a) suggests that the organizations discover 8-18% of cyber events right away ($8\% \leq \alpha \leq 18\%$).

The scale plot (Fig. 2.7b) suggests that the short delays modeled by the exponential distribution had a mean of less than 60 days delay until early 2016 but increased rapidly to around 140 days in early 2018.

The normal distribution mean, μ, (Fig. 2.7c) and standard deviation, σ, (Fig. 2.7d) parameter plots suggest that the longer delays modeled by the normal distribution remained consistent over time. The period of longer delays remain consistent varying within a10% range. At 90% confidence level, the longer delays are expected to range between 1.4 to 9.4 years approximately. Since the parameters μ and σ are computed monthly for 2-year rolling period, the confidence bounds at the given confidence level, c, are computed with truncated normal distribution, as shown in Eq. 2.14 and Eq. 2.15.



(a) From July 2012 to June, 2014



(b) From January, 2017 to December, 2018

Figure 2.6: Comparing PDFs of Debiased Delay Distribution with Parametric Modeled Distribution

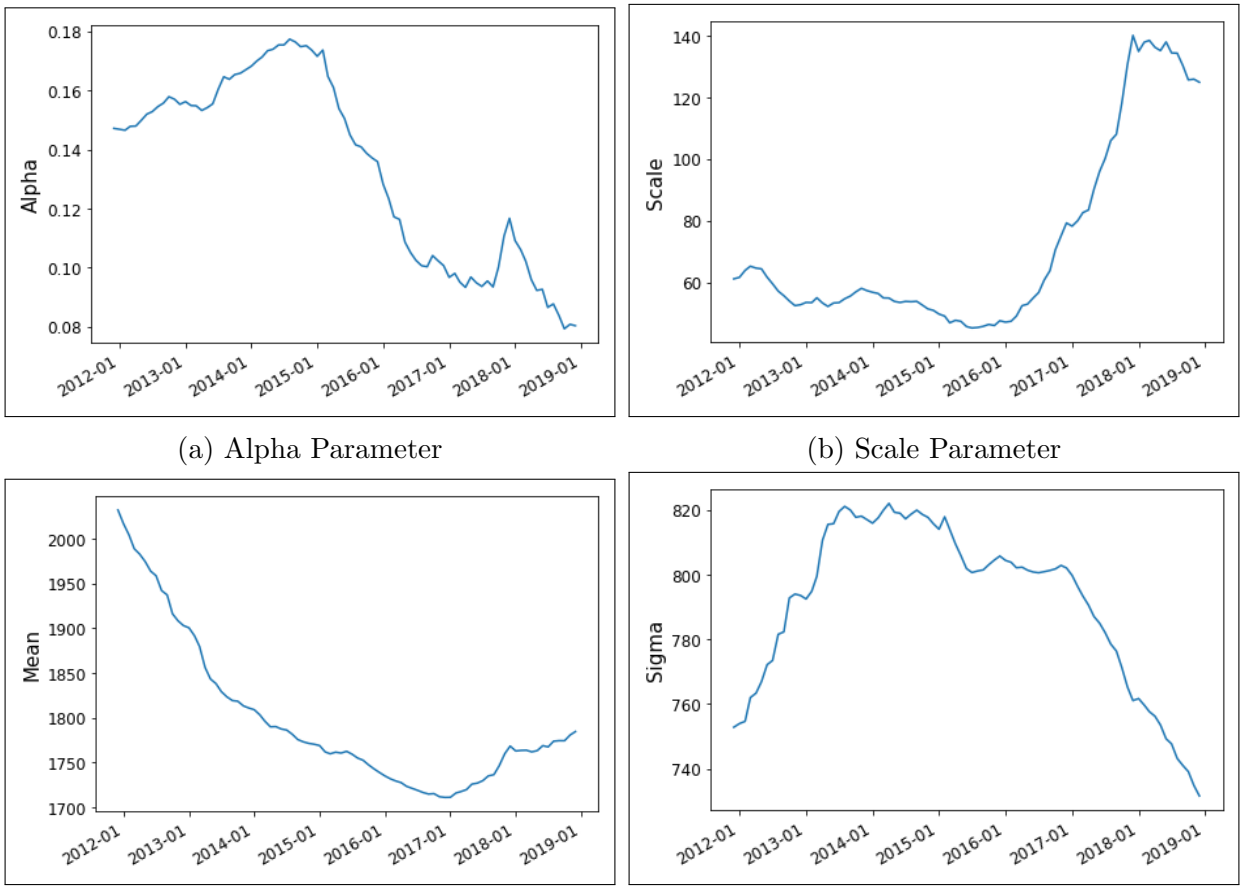
$$\text{Lower Limit(t)} = F_N^{-1}\left(F_0 + (1-F_0)\frac{1-c}{2}\right) = F_N^{-1}\left(1-(1-F_0)\frac{1+c}{2}\right) \tag{2.14}$$

$$\text{Upper Limit(t)} = F_N^{-1}\left(F_0 + (1-F_0)\frac{1+c}{2}\right) = F_N^{-1}\left(1-(1-F_0)\frac{1-c}{2}\right) \tag{2.15}$$

where $F_0 = F_N(\delta = 0) = N(0, \mu_t, \sigma_t)$

c is the confidence interval

$F_N$is the normal distribution function



(a) Alpha Parameter

(b) Scale Parameter

(c) Mean Parameter

(d) Sigma Parameter

Figure 2.7: Plots of Modeled Distribution Parameters based on Empirical Debiased Delay Distribution

Fig. 2.8 shows lower and upper bounds for the longer delays over the period of time.

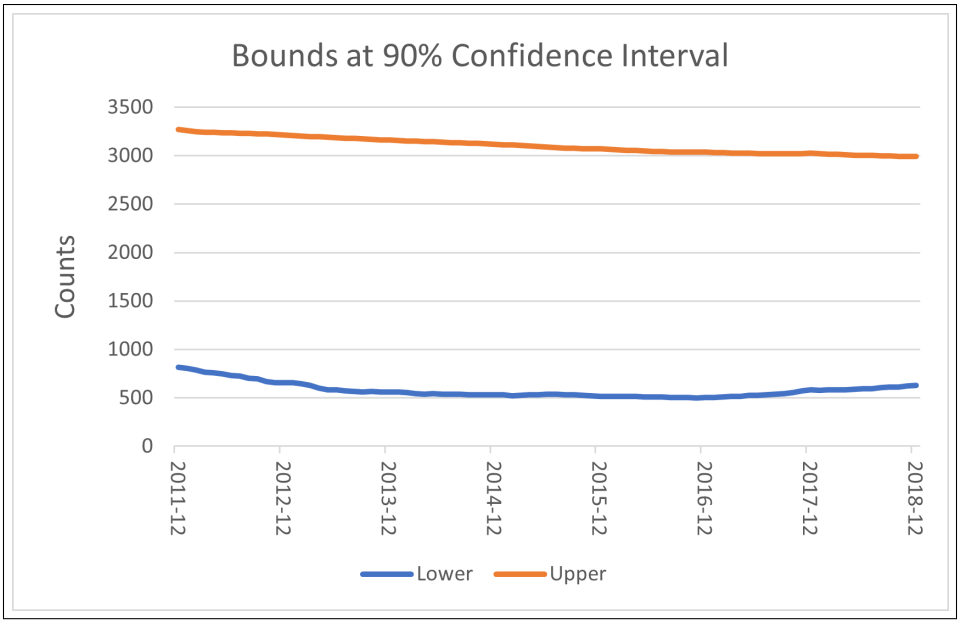


Figure 2.8: Lower and upper bounds for longer delays in US

2.6.3 Corrections and Validation

Fig. 2.9 shows the corrected incident counts based on the proposed methodology. Figs. 2.9a and 2.9b show the corrected counts for the events reported by Dec. 2017 and by Dec. 2018, respectively. Although the corrections follow similar trends in both, the correction factors vary substantially.

To validate the proposed algorithm, the counts reported until December 2018 were corrected for a year ahead and compared against the counts reported as of December 2019. Notice 1 Year = 360Days, computed based on 30 days per month in a year, the 30/360 convention is chosen to allow uniform discretization among bins of delays and age.

The year ahead correction factor is computed as shown in Eq. 2.16.

$$\text{Year ahead } F_\theta(a, a+1\,Year) = \frac{F_\theta(a)}{F_\theta(a+1\,Year)} \tag{2.16}$$

where a is the age of the event counts being corrected.

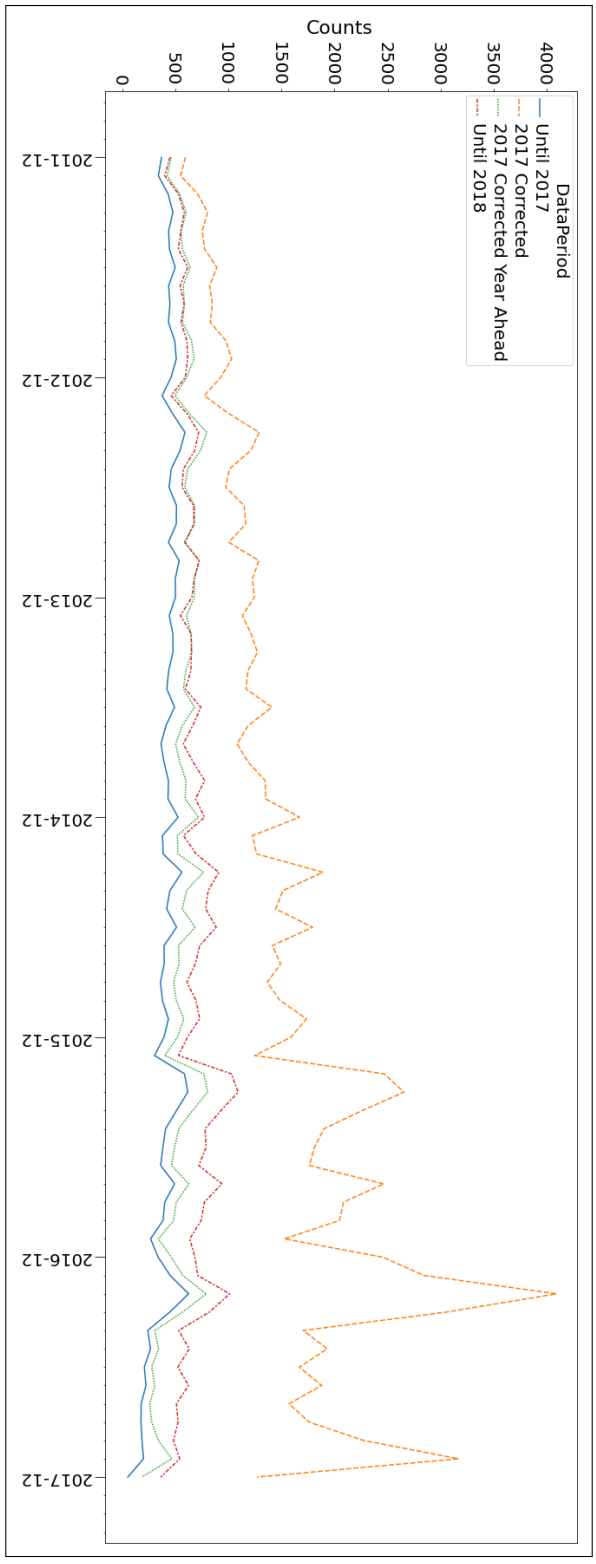
Whereas the 2017 year ahead corrections (Fig. 2.9a) initially show close agreement with the 2018 counts, more recent year ahead corrections underestimate the 2018 counts. On

the other hand, the 2018 year ahead corrections (Fig. 2.9b) generally overestimate the 2019 counts, except for the most recent months, which show close agreement. As stated earlier in subsection 2.4.4, the debiased empirical delay distribution has fewer data points for more recent two year windows so weights of $(\delta_{\max}/\delta_{Fix})$ and $(1-\delta_{\max}/\delta_{Fix})$ are used in the optimization function to dynamically adjust the weight given to the CDF before and after $\delta_{\max}$ respectively. By removing these weights, better estimates for recent months might be obtained but would come at the cost of more parameter instability and worse validation plots (overfitting). In either Fig. 2.9a or Fig. 2.9b, the corrected counts (dashed line) show a trend of increasing incident counts since 2016, which is contrary to the diminishing trend seen in the raw counts. The trend in corrected count is therefore much more in line with reports from insurers and other organizations that release reports on cyber risk.

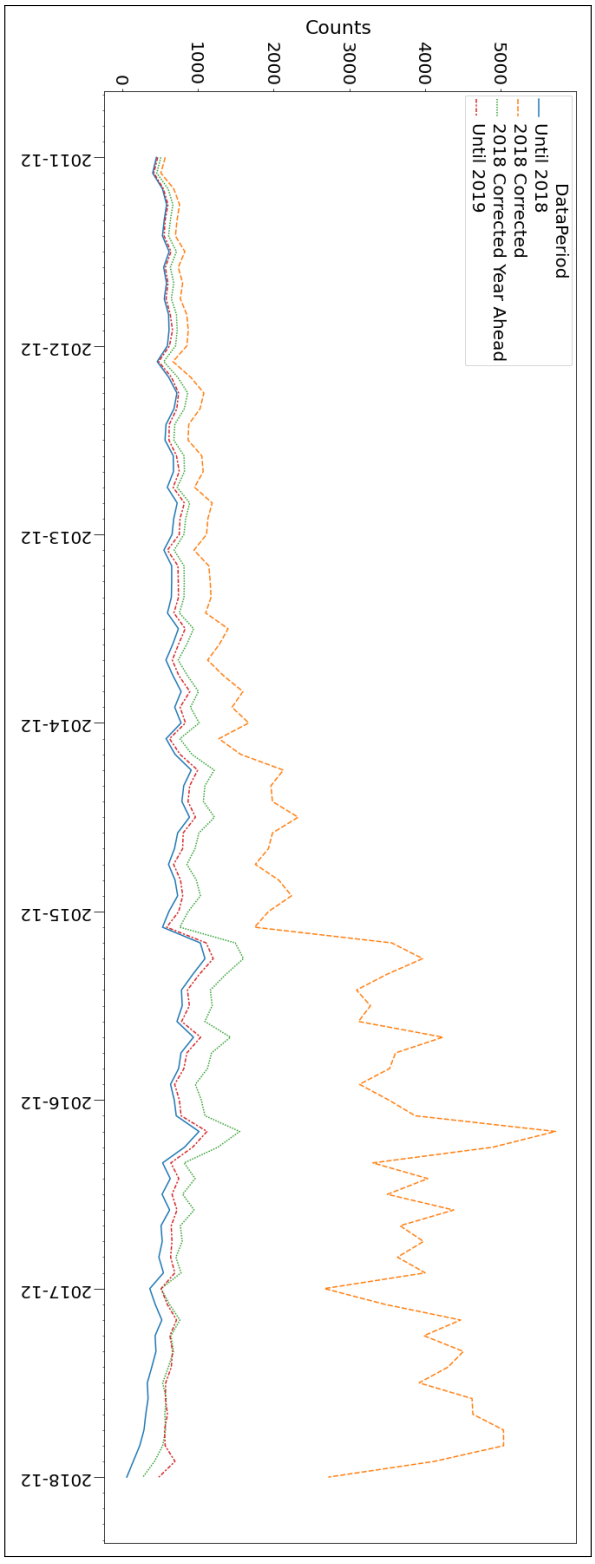
2.7 Conclusion

This work examined the long known problem of reporting delays in historical cyber events databases and proposed an algorithm to correct these delays. Interestingly, the true distribution of reporting delays appears to be bi-modal, which we have interpreted as a mixture of two distributions: one for incidents that are discovered immediately, modeled by an exponential distribution, and one for incidents that are not immediately discovered, modeled by a normal distribution. With this form of reporting delay distribution, we obtained non-stationary modeled delay distributions via optimization. These modeled delay distributions were used to estimate the total number of cyber incidents that will eventually be reported from the current counts. The approach was validated by estimating year ahead corrections.

To understand the current cyber threat landscape and to create robust cyber risk models, one needs accurate historical data. While it is not possible to get the exact count of cyber events, the proposed algorithm aims to correct for reporting delays approximately. The reported cyber incident counts in recent times show a decreasing trend simply because incidents have not been reported yet, even though they have actually already occurred. However, in reality, the rate of cyber incidents is increasing, and that is what the algorithm reveals.



(a) 2017 corrections Vs 2018 cumulative counts



(b) 2018 corrections Vs 2019 cumulative counts

Until 201X - Counts reported as of 201X adjusted for the default date of January 1, proportionally
201X Corrected - "Counts until 201X" corrected based on Eq. 2.12
201X Corrected Year Ahead - "Counts until 201X" corrected based on Eq. 2.16

Figure 2.9: Validation Plots

Chapter 3

Modeling reporting delays in cyber incidents: an industry level comparison

3.1 Abstract

Reporting delays in cyber incidents are a common problem. Incidents often take time to be detected and even more time to be reported. Due to reporting delays, the proportion of recent incidents to have been reported is smaller than for older incidents, resulting in the false impression of a diminishing frequency of cyber incident counts in recent years when examining databases of (publicly) reported cyber incidents. Obtaining an accurate view of the true trend therefore requires correcting for reporting delays. Complicating matters is the fact that the distribution of reporting delays differs from industry to industry. This paper investigates four distinct US industries, as defined by the NAICS classification system: Finance and Insurance, Educational Services, Health Care and Social Assistance, and Public Administration. This paper presents a method of modeling and correcting for cyber incidents reporting delays overall and by industry, with specific emphasis on the four distinct industries

listed above. Finally, this work compares the derived reporting delay models, which show differences in reporting delays from one industry to another.

3.2 Introduction

Cyber incidents are a global concern irrespective of region, industry, incident type or public/private sector categorization [2]. It frequently takes months or years to detect cyber incidents and further time before they are reported [9]. These reporting delays make any analysis of trends in incident rates from publicly reported data sets a challenge. A recent study contended that it is even more challenging to detect incidents now with attackers using advanced state-of-the-art techniques [9]. Because of reporting delays, the proportion of incidents that are publicly known is smaller for recent incidents than older ones, causing the appearance of a diminishing trend in incident rates when incidents are in reality increasing. Cyber insurance models are built in part from this incomplete data and, hence, their credibility requires correcting cyber incident counts for the problem of reporting delays. The disclosure of cyber incidents depends on disclosure requirements that can vary by location, industry and inspecting regulatory agency [9]. In addition, various organizations prefer not to disclose cyber incidents, fearing reputational damage which could eventually result in lost business. For these and other reasons, cyber incidents frequently have reporting delays. Furthermore, the distribution of these delays differs from one industry to another [115]. Hence, it would be erroneous to apply the same correction factors to all industries when estimating true incident rates.

Most of the research in reporting delays has been in the medical domain. Brookmeyer and Liao(1990) mentioned that the existing methods only consider delays less than or equal to the maximum age of incidents in the data [23]. That is, they take the oldest incidents to be fully reported. Harris(1987) categorized reporting delays as a statistical problem [55]. Some researchers assumed incidents to be Poisson distributed and some fitted statistical models but failed to capture trends [21][26][61][62][63][114]. Others suggested exponential, integrated logistic and log-linear models to capture trends [30][31][59][61][62][63][100]. Brookmeyer and Liao, Gail and Brookmeyer, Kalbfleisch and Lawless and Esbjerg et al. applied conditional

probabilities but this resulted in over-fitting [23][35][43][74]. White et al. and Weinberger et al. suggested proportions and Wang proposed a semi-parametric approach with maximum likelihood estimation (MLE), but both approaches require complete[21] data [130][132][134].

Sangari and Dallal introduced an approach based on debiasing the empirical delay distribution (Fig.2.3) and fitting a modeled distribution to it [119]. There are two main advantages to the approach: it considers delays beyond the maximum delay observed in the data set, $\delta_{\max}$; and it captures trends which are yet to be seen in the reported counts.

This paper applies the debiased delay distribution approach independently to four industries and investigates differences in the resulting models of reporting delays [119]. Section 3.3 describes the data. Section 3.5 presents the parameters and implied correction factors of the reporting delay models for each industry and for the US as a whole. Section 3.6 concludes the comparative study.

The following terminology will be used throughout this paper:
i: An incident
I: The set of all incidents
$delay, \delta_i$: Time period between incident and reporting dates
age, A_i: Time period between incident date and last incident reporting date in the data
f_Δ: Probability density function (PDF) of the delay distribution
F_Δ: Cumulative distribution function (CDF) of the delay distribution

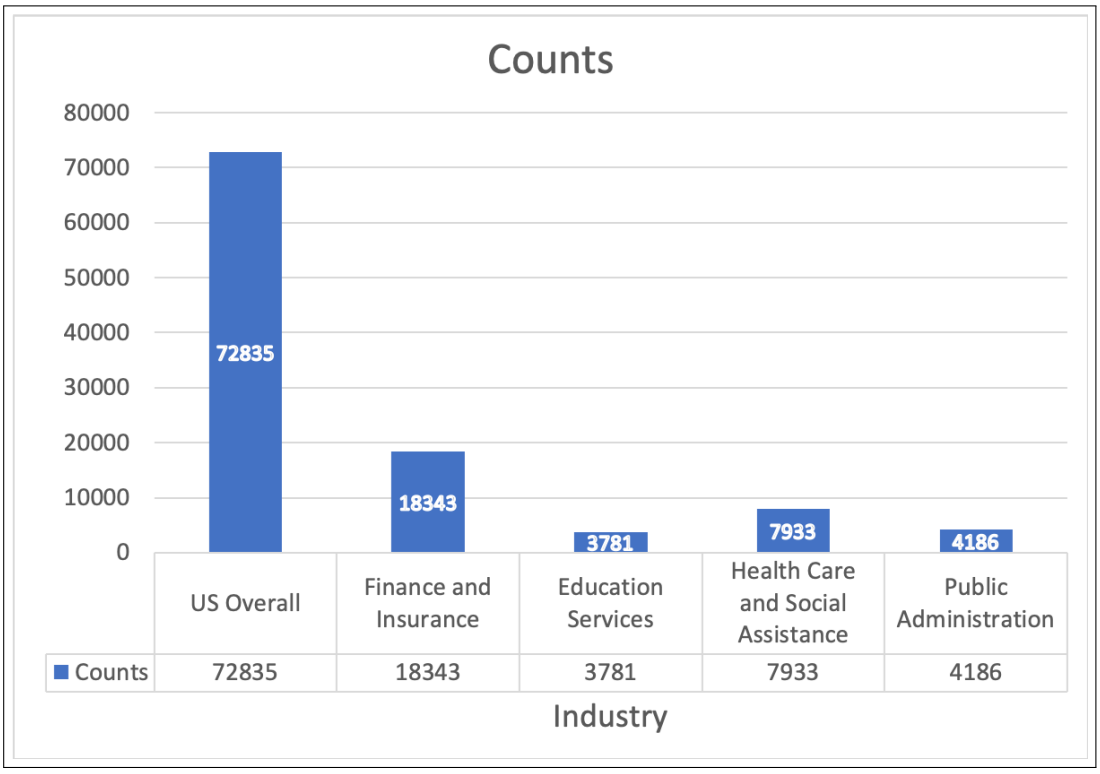
3.3 Data

The proprietary data set used consists of a collection of more than 140,000 publicly reported historical incidents over a period of decades. The incidents in this data set were gathered via numerous collection methods, including scraping of technology and news websites, securities exchange commission (SEC) filings, and other sources. An aggregated data set was constructed by combining historical incident data sets with a proprietary data set of

[21] No additional incidents are expected to be reported

companies that includes, among other things, their name, location, industry, and revenue. This study is done on the incidents collected in the decade from 2010-2019. The cyber incident data is comprised of incident information providing: an incident description, the name of the affected organization, the organization's industry code as defined by the NAICS classification system, and the occurrence and reporting dates. In the absence of an exact incident date, incident dates are listed as January 1 of the incident year. Records with such a default date were excluded before analyzing the data.

Analysis of these data was completed first on all incidents overall, and then by each of the four specific industry classifications: Finance and Insurance (FnI), Educational Services (ES), Health Care and Social Assistance (HnS), and Public Administration (PA). Fig.3.1 shows incident counts for each industry.



Total number of US cyber incidents = 72,835

Figure 3.1: Number of Cyber incidents across industries between 2010-2019

3.4 Why Industry Level Comparison

Rosinska et al. suggested that the trends vary between categories [115]. They recommended that the overall distribution fit might not be applicable to individual sectors/industries/-places, so each and every category needs to be modeled individually. The regulatory requirements across the location and industry often impacts the reporting of cyber incidents [9]. In addition, reputational damage and losing business are also contributing factors that various organizations opt not to report incidents. Such reasons result in frequent reporting delays. Moreover, the reporting delays vary across various industries [115]. Hence, applying same correction factors to all industries to estimate true incident rates would not be fair.

The frequency distribution of cyber events within four major industries of the US market collected until 2019 is shown in Fig. 3.2. These frequency distributions shows the counts adjusted for the default date[22] as mentioned in section 1.3.

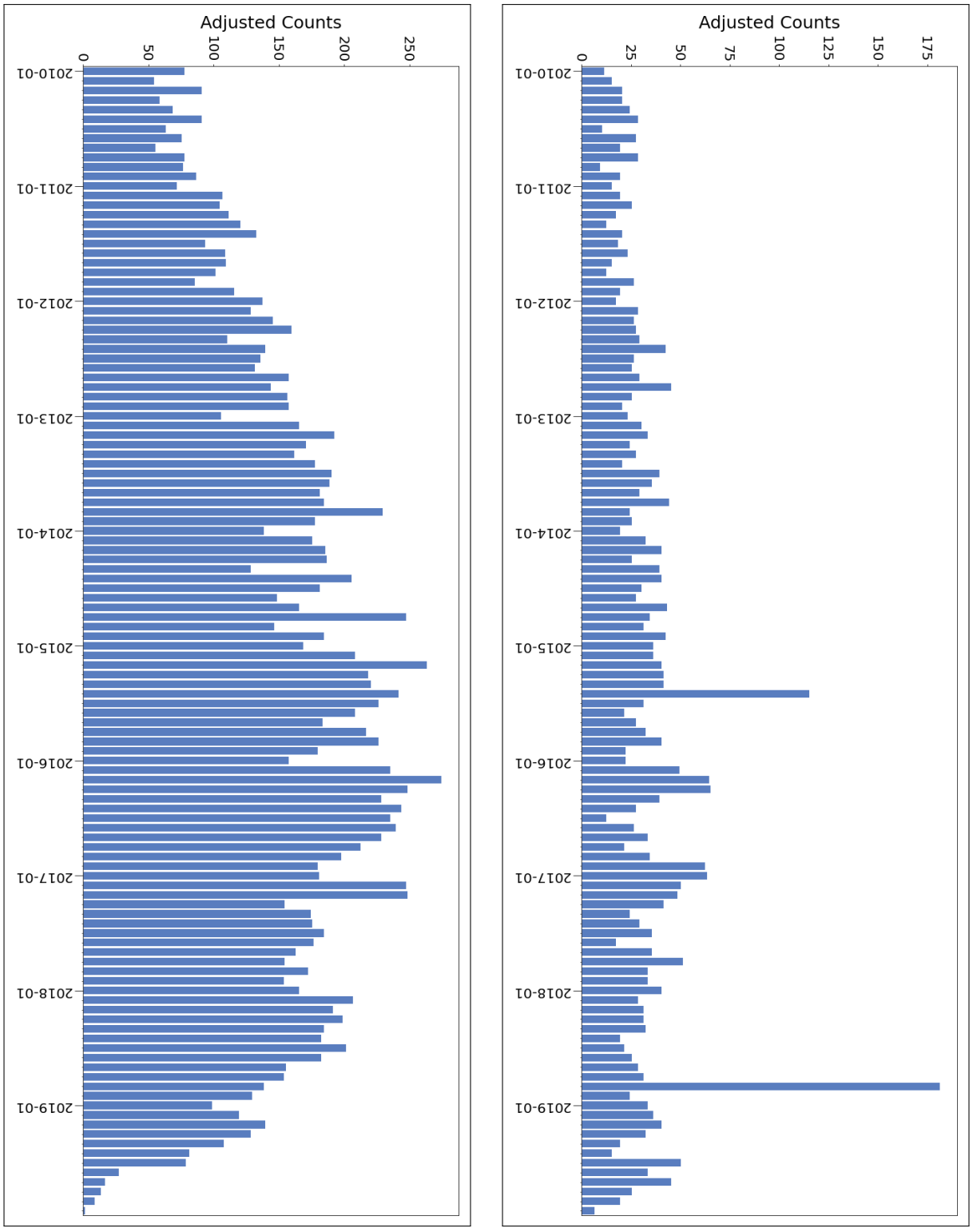
3.5 Results

3.5.1 Industry Parameters

In this section, we present plots showing the evolution of the best fit modeled distribution parameters α, $Scale$, μ, and σ over time, for the four industries of interest and the US as a whole.

The alpha parameter plot (Fig. 3.3), representing the immediate detection rate, shows that the overall US market detects 15-50% of cyber incidents right away. This immediate detection rate increased overall and all four modeled industries have higher values for this parameter than the US market as a whole in recent years. (N.B.: The US parameters are based on all industries, not merely the four separately modeled ones shown here.) For the US market, the increasing trend starts by the end of 2016 or the beginning of 2017. This jump coincides with a substantial increase in the frequency of ransomware incidents. However, three of the

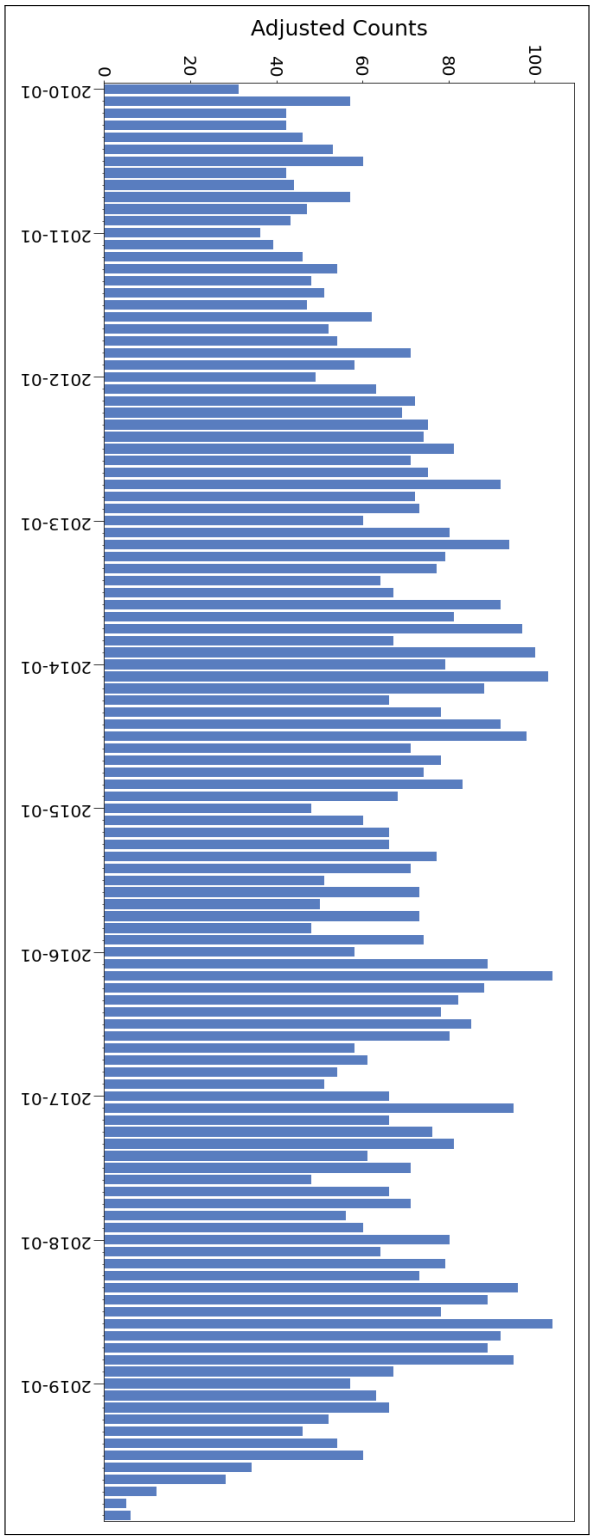
[22]When the occurrence date is not known, the incident date is set to the default date on January 1 of the given incident year; such incidents are proportionally distributed among all 12 months



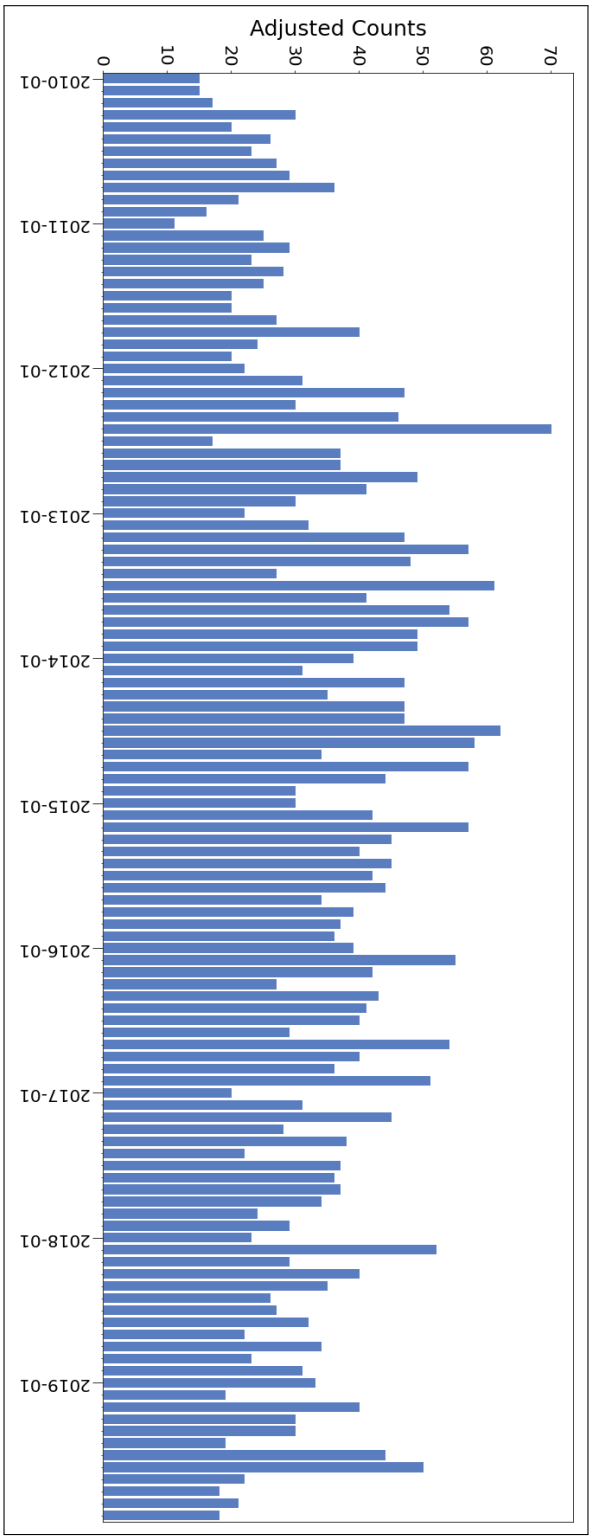
(a) Finance and Insurance (N = 18337)

(b) Educational Services (N = 3784)

Figure 3.2: Adjusted Counts for four major industries collected until 2019



(c) Health Care and Social Assistance (N = 7935)



(d) Public Administration (N = 4186)

Figure 3.2: Adjusted Counts for four major industries collected until 2019

four industries show an increase starting years earlier. While this parameter increased by a factor of 12 for the FnI industry, this should not be considered a real effect–such a jump is observed in the other three optimal parameters too and is explained after the discussion of the optimal scale parameter plot. However, other industries showed a more moderate increase, by a factor of 1.6 to 2.5. Multiple interpretations are possible both for the overall increasing trend and for the differences among industries. As for the large difference between FnI and the other three industries prior to 2016, this may indicate that greater care was taken by hackers of financial institutions to avoid timely detection of their intrusion, so as to allow more time in which to make fraudulent purchases.

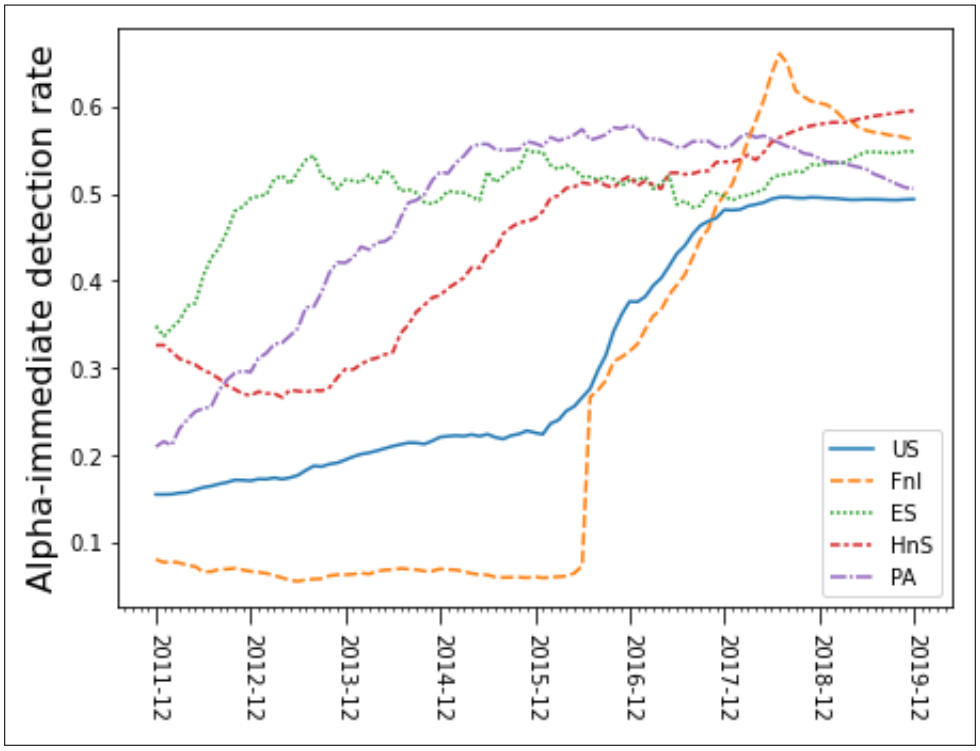


Figure 3.3: Plot of optimal Alpha parameter over time

The exponential scale parameter plot (Fig. 3.4), representing shorter delays, suggests that the time for US companies to report cyber incidents that were immediately detected was quite short until 2016–less than two months–but increased rapidly to 10 months by early 2018. Once again, the rapid increase seems to coincide with the substantial increase in the frequency of ransomware incidents that occurred starting around 2016 [86]. This explanation is plausible, since reporting requirements have traditionally been focused on incidents of data

compromise. However, the increase starts much earlier in HnS industry, and only a weak increase is seen for PA and ES industries.

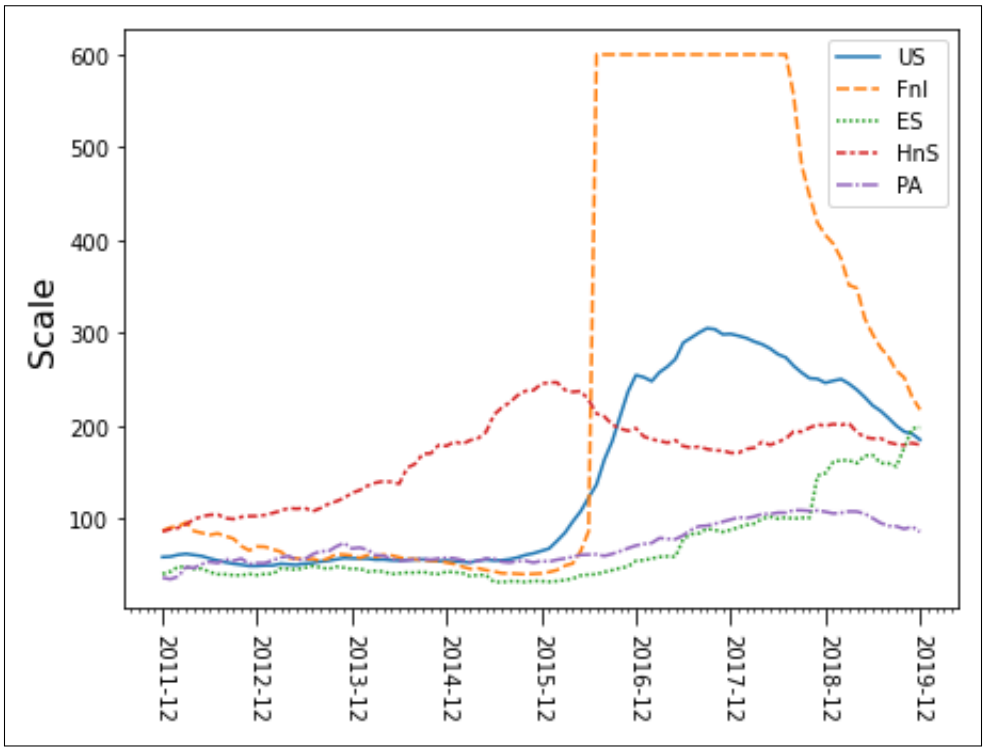


Figure 3.4: Plot of optimal Scale parameter over time

As for the FnI industry, the optimization algorithm struggled to fit parameters beyond Q1 2016 appropriately. Indeed, a sudden change in parameters is observed for the FnI industry in all four parameter plots for the July 2016 two-year window[23]. Figures 3.5 and 3.6 present the debiased empirical delay distributions for the two-year windows before and after the sudden parameter change. Contrary to the mild changes between the two plots (Fig. 3.5 and 3.6), there is an unanticipated shift noticed in the parameters. This suggests that the optimization function is switching from one local minimum to another, resulting in radically different parameters. This impacted the scale parameter immensely, which is observed in Table 3.1 and Fig. 3.4. The other three parameters also shift to compensate, though less dramatically than the scale parameter, as can be seen in Table 3.1.

Despite a substantial change in parameters, there is only a very subtle change in the objective function (Eq. 2.9) values observed, as shown in Table 3.2. Table 3.2 confirms that the set

[23]This corresponds to incidents that occurred from August 2014 through July 2016

Table 3.1: Optimal Parameters

Parameters	Jul.'14-Jun.'16	Aug.'14-Jul.'16
Alpha(α)	0.07	0.27
Scale	86.64	600.00
Mean(μ)	1452.05	1617.75
Sigma(σ)	798.56	742.18

Table 3.2: Optimization function values

	Debiased Delay Distribution	
Parameter used $\theta = [\alpha, Scale, \mu, \sigma] \downarrow$	Jul.'14-Jun.'16	Aug.'14-Jul.'16
Jul.'14-Jun.'16 θ = [0.07, 86.64, 1452.05, 798.56]	1.0771	1.0768
Aug.'14-Jul.'16 θ = [0.27, 600, 1617.75, 742.18]	1.0642	1.0640

of optimal parameters is actually a good fit for the Aug.,'14-Jul.,'16 window. However, the optimization algorithm failed to find the best possible parameters for the Jul.,'14-Jun.,'16 window.

Essentially, the mixture of an exponential and a normal distribution ceases to fit the data for the FnI industry for the two-year windows ending in the range of months from July 2016 until approximately June 2018. Afterwards, the parameters gradually drift back to the range of values seen for the US as a whole and the other three industries. It should be noted, however, that there is no substantial change in the computed correction factor itself between June and July 2016 (Fig. 3.10).

The normal distribution mean (Fig. 3.7) and standard deviation (Fig. 3.8) plots, representing longer delays, indicate that the distribution of reporting delays for incidents not immediately detected is fairly consistent over the period of time considered, both for the four industries examined and the US market as a whole. 90% of the longer reporting delays ranged between one to eight years for the ES, PA and HnS industries and between one to nine years for the FnI industry and the US market.

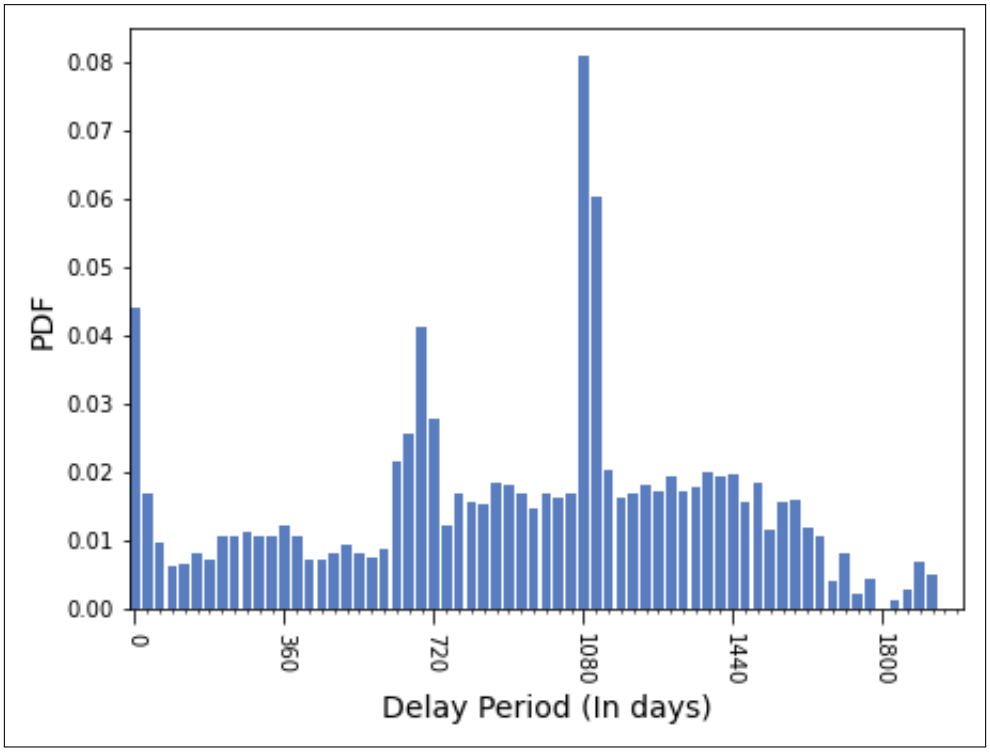


Figure 3.5: PDF comparison for period July, 2014 to June, 2016

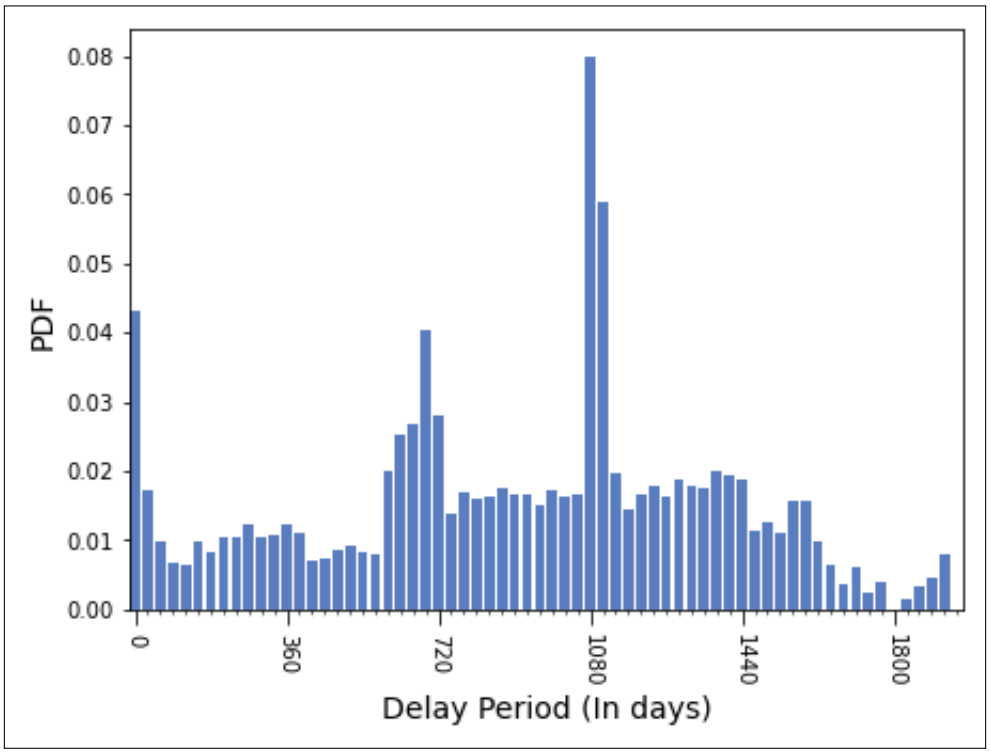


Figure 3.6: PDF comparison for period August, 2014 to July, 2016

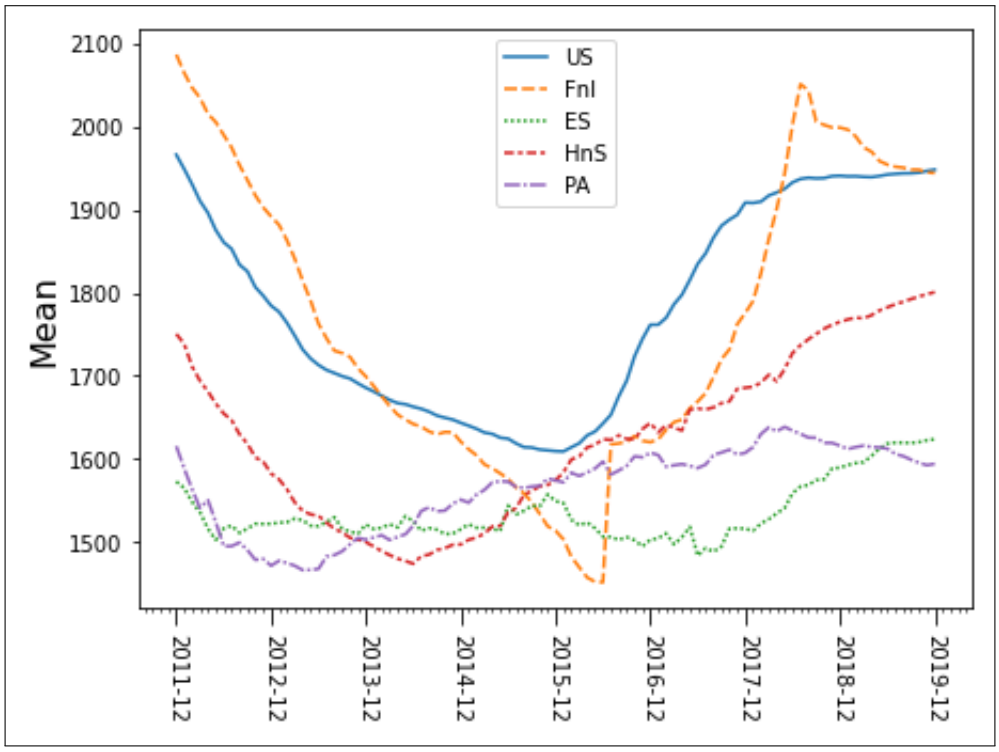


Figure 3.7: Plot of optimal Mean parameter over time

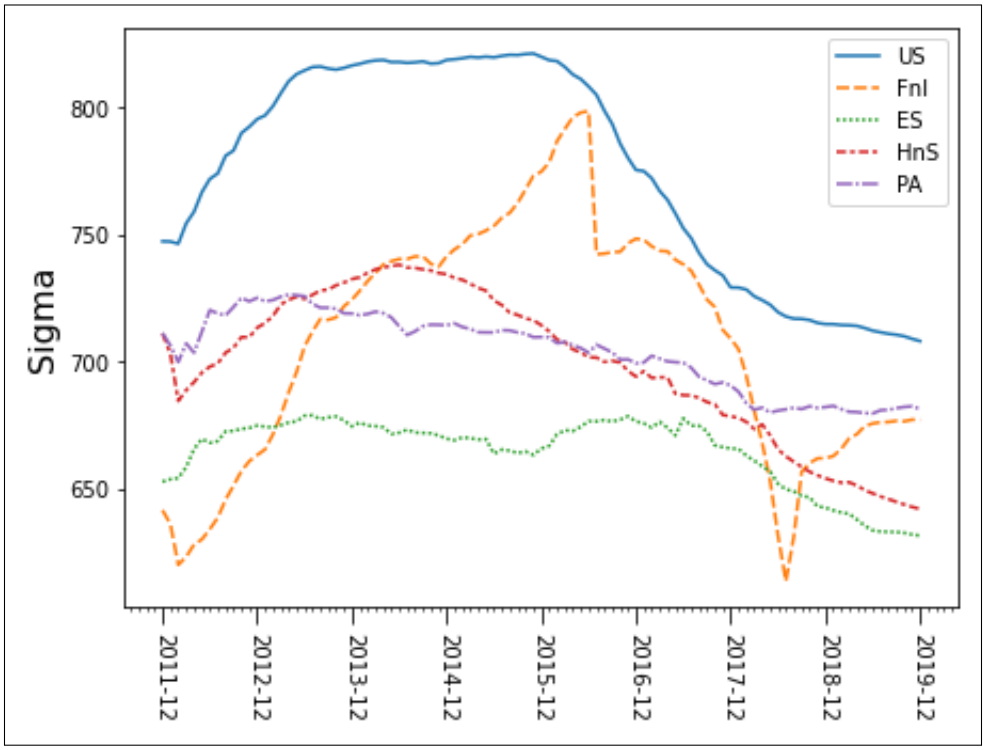


Figure 3.8: Plot of optimal Sigma parameter over time

3.5.2 Reported Counts Proportion

Fig. 3.9 shows the modeled proportion of incidents, F_θ, that have been reported as of the end of 2019, as a function of incident occurrence date. As expected, the proportion of incidents is close to one in the beginning, indicating that not many more older incidents are expected to be reported in the future. For ease of interpretation, horizontal lines are included to mark the estimated point in time where fewer than 1/2, 1/3, and 1/5 of incidents have been reported. This shows a marked difference between the FnI industry and the US as a whole on the one hand, and the ES, PA and HnS industries on the other: even as far back as early 2016, only half of incidents in the FnI industry and for the US as a whole have been reported by the end of 2019, whereas the corresponding point in time is late 2018 for the ES, PA and HnS industries. The modeled reporting proportions hit 33.3% in early 2019 for the FnI industry and the US as a whole, but only in the third quarter of 2019 for the ES, PA and HnS industries. From Eq. (2.12), we see that the correction factors are simply the reciprocal of $F_\theta(a)$. Hence, the correction factor is expected to be highest as we approach more recent months. The correction factors are shown in Fig. 3.10 and demonstrate the expected pattern: higher correction factors for industries with longer reporting delays.

3.5.3 Validation

The corrections are validated by calculating year[24] ahead corrections for the incidents reported by a given year against the incidents reported by the next year. Fig. 2.9a and Fig. 2.9b show the corrected counts of incidents for the US market reported by Dec.'17 and by Dec.,'18, respectively. Despite the same directional movements, the correction factors are considerably different.

Whereas the US 2017 year ahead corrections appear consistent with the by 2018 reported counts initially, the corrections underestimated 2018 counts in recent months (Fig. 2.9a). On the other hand, the 2018 year ahead counts are found to be consistent with the by 2019 reported counts in recent months but are an overestimate otherwise (Fig. 2.9b).

[24] $1\ Year = 360\ Days$, computed based on 30 days per month in a year

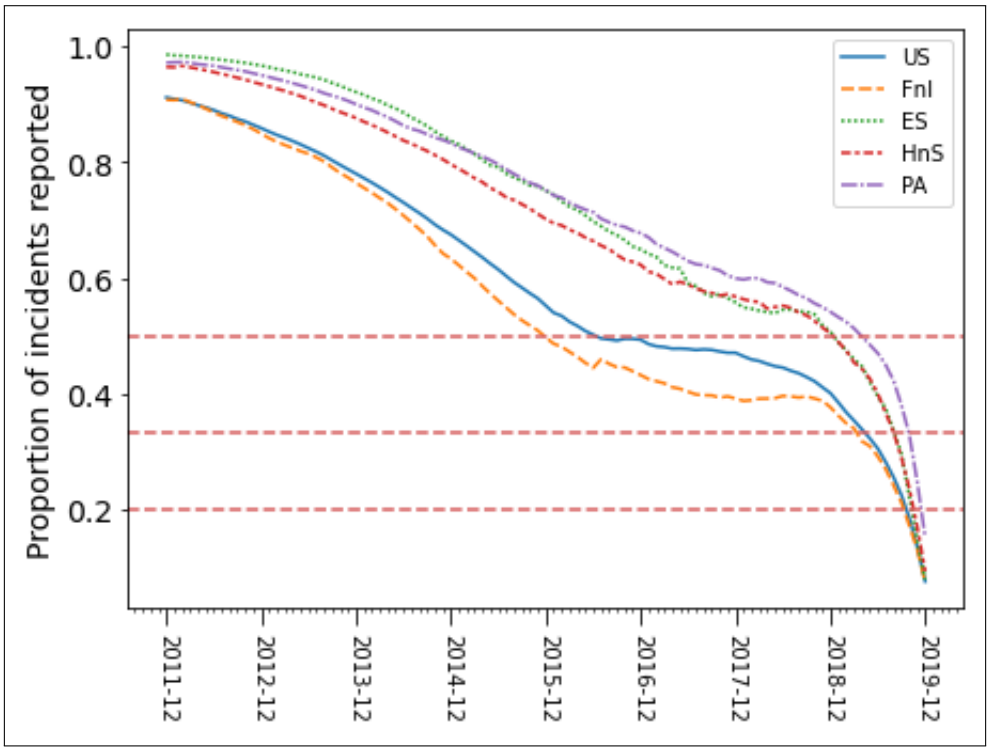


Figure 3.9: Proportion of Incidents Reported by the end of 2019

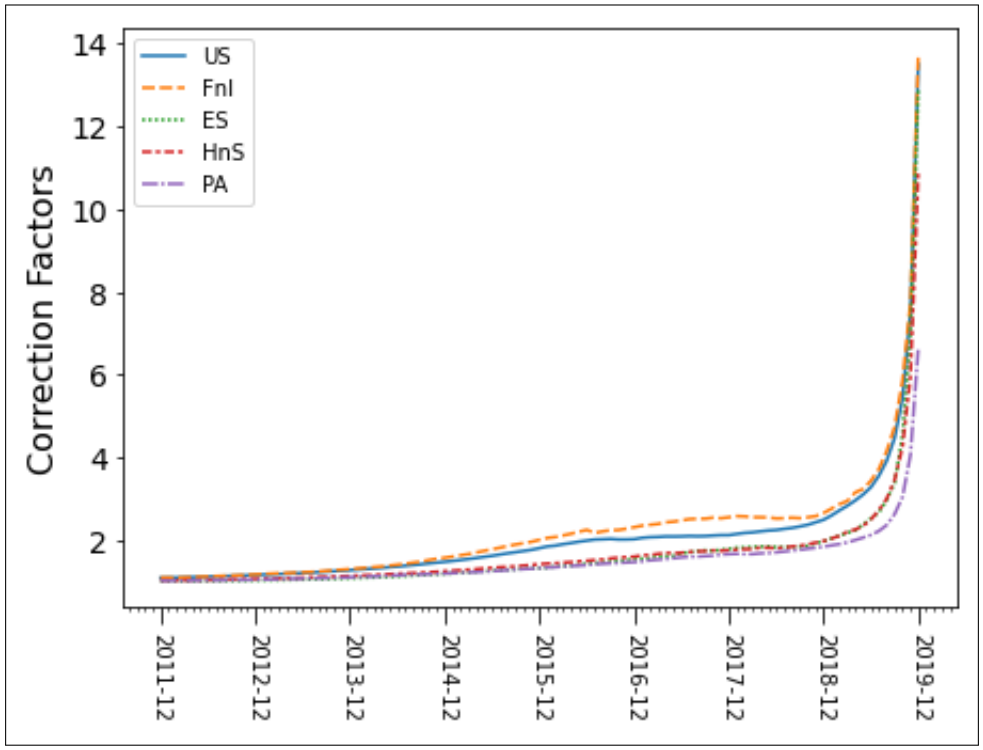


Figure 3.10: Correction Factors

For the FnI industry, the initial year ahead corrections for 2017 (Fig. 3.11a) and 2018 (Fig. 3.11b) are found to be consistent with 2018 and 2019 reported counts. Beyond mid 2014, the 2017 year ahead corrections were an underestimate and the 2018 year ahead corrections were an overestimate. The ES industry (Figs. 3.12a and 3.12b) also showed similar behavior apart for the recent few months, where the 2017 year ahead corrections were an overestimate and the 2018 year ahead corrections were an underestimate.

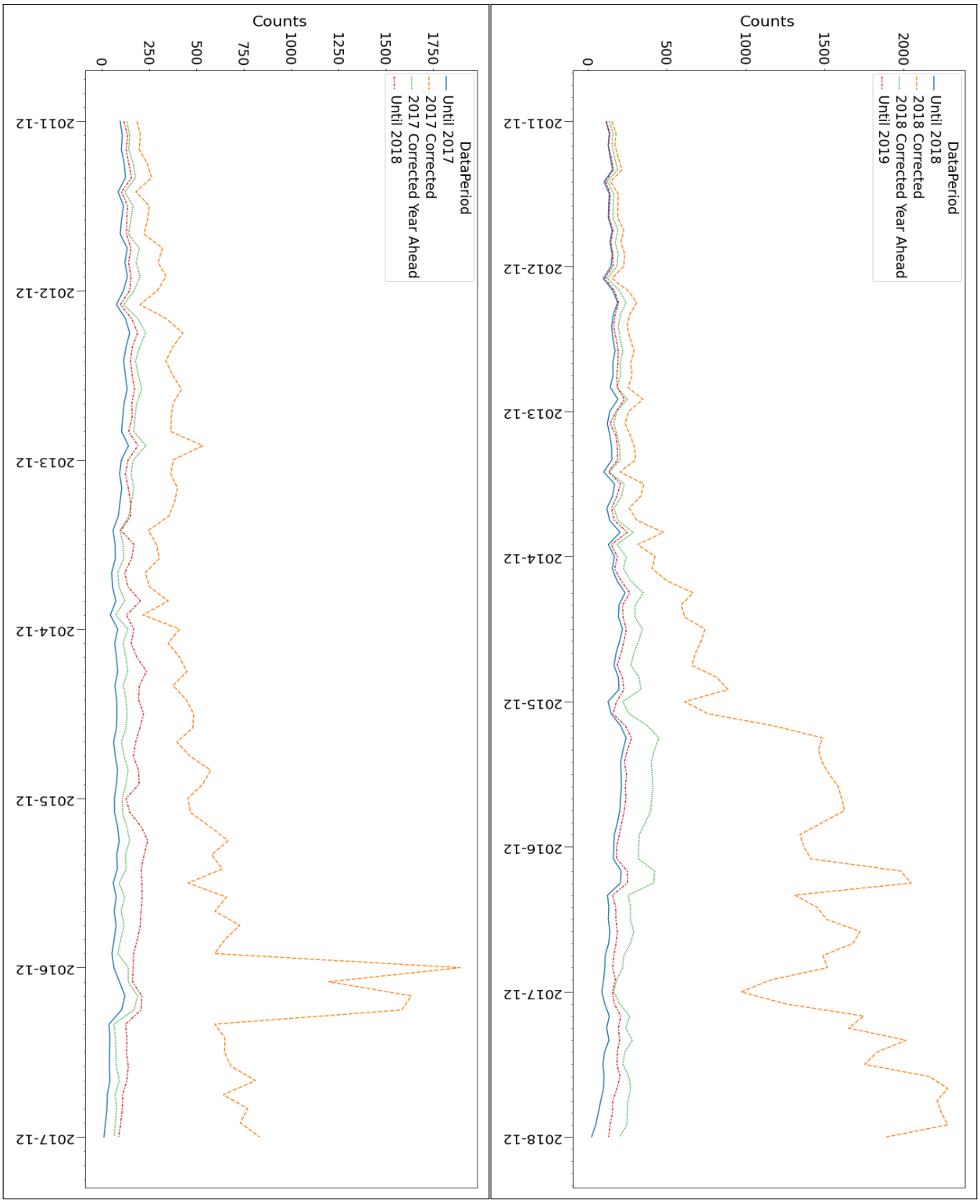
For the HnS industry, the year ahead corrections for 2017 (Fig. 3.13a) were initially an overestimate and later an underestimate. However, the 2018 year ahead corrections (Fig. 3.13b) were found to be within an acceptable range of 2019 reported counts. The year ahead corrections for the PA industry (Figs. 3.14a and 3.14b) were found to be similar to the HnS industry.

Despite variation in year ahead corrections at the industry level and for the US as a whole, both the 2017 and 2018 full corrections (dashed lines) reflect an increasing trend in counts, as expected [9].

3.6 Conclusion

In this research, we created models of reporting delays in US cyber incidents, comparing four industry sectors–finance and insurance (FnI), educational services (ES), health care and social services (HnS), and public administration (PA)–to each other and to the US as a whole. The distribution of reporting delays was found to be bi-modal, with one peak at 0-30 days and a second peak occurring after more than five years. We have interpreted this as a mixture of two distributions: one for incidents that are discovered immediately, modeled by an exponential distribution, and one for incidents that are discovered later, modeled by a normal distribution. Given the distribution of reporting delays, one can correct cyber incident counts to account for the proportion of incidents that have occurred but not yet been reported. As expected, these correction factors converge to one as incidents further and further into the past are considered. For more recent incidents, significant variation was found between the FnI industry and the US as a whole on the one hand, and the

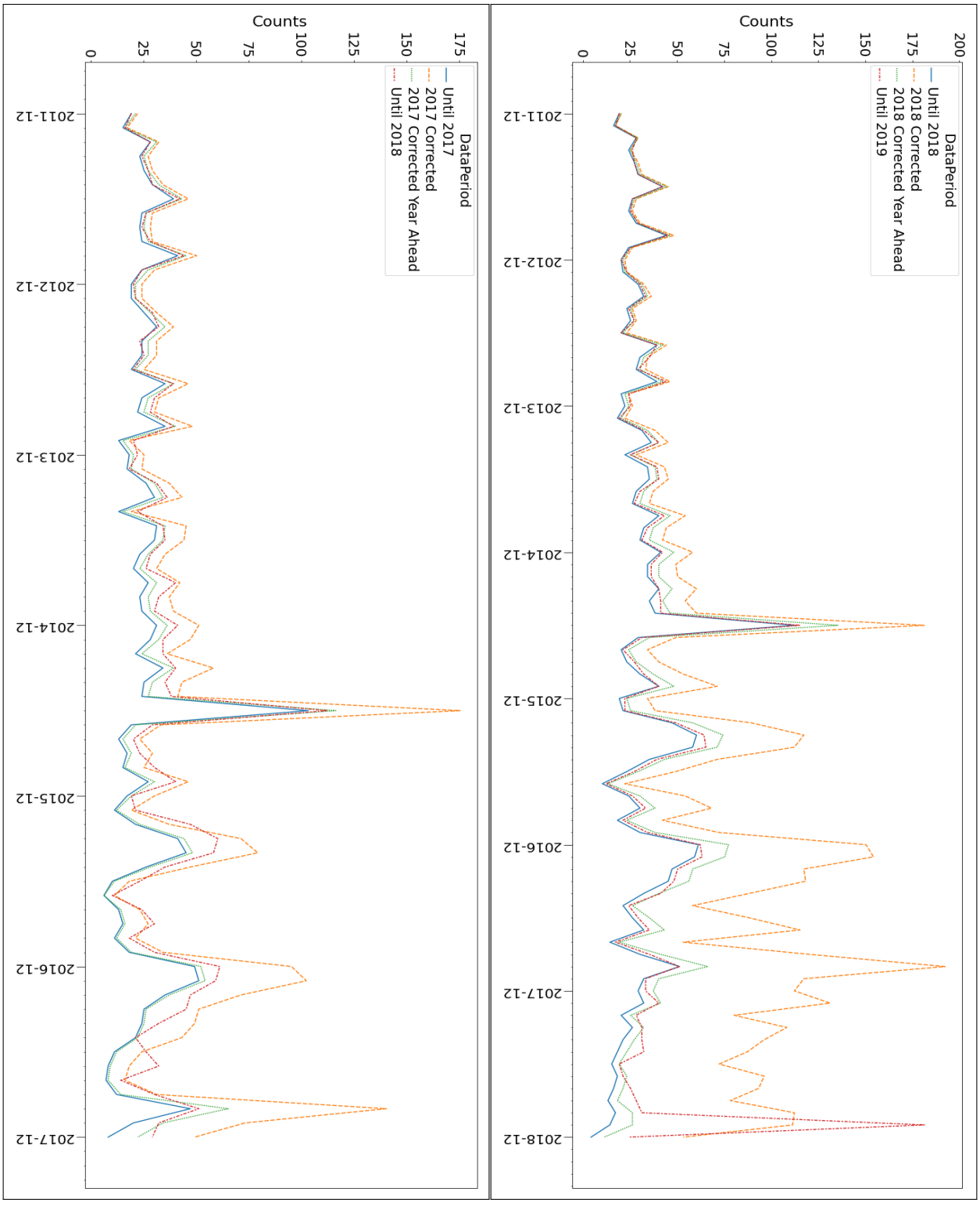
other three industries examined on the other. Specifically, the FnI industry showed a low proportion of incidents which were immediately detected and longer reporting delays in general. A possible hypothesis is that attackers took greater care to not be detected when targeting financial institutions. Overall, the proportion of incidents which were immediately detected has increased over time for the US as a whole and for the four particular industries considered. This may be a sign that companies have gotten better at detecting intrusions within a short time, or it may simply indicate a shift in attacker tactics towards "noisier" attacks like ransomware. Ultimately, the problem of reporting delays is important to model, both because of the insights to be gained from the perspective of cyber security and because accounting for as yet unreported incidents is necessary for the construction of accurate cyber risk models.



(a) 2017 corrections Vs 2018 cumulative counts (b) 2018 corrections Vs 2019 cumulative counts

Until 201X - Counts reported as of 201X adjusted for the default date of January 1, proportionally
201X Corrected - "Counts until 201X" corrected based on Eq. 2.12
201X Corrected Year Ahead - "Counts until 201X" corrected based on Eq. 2.16

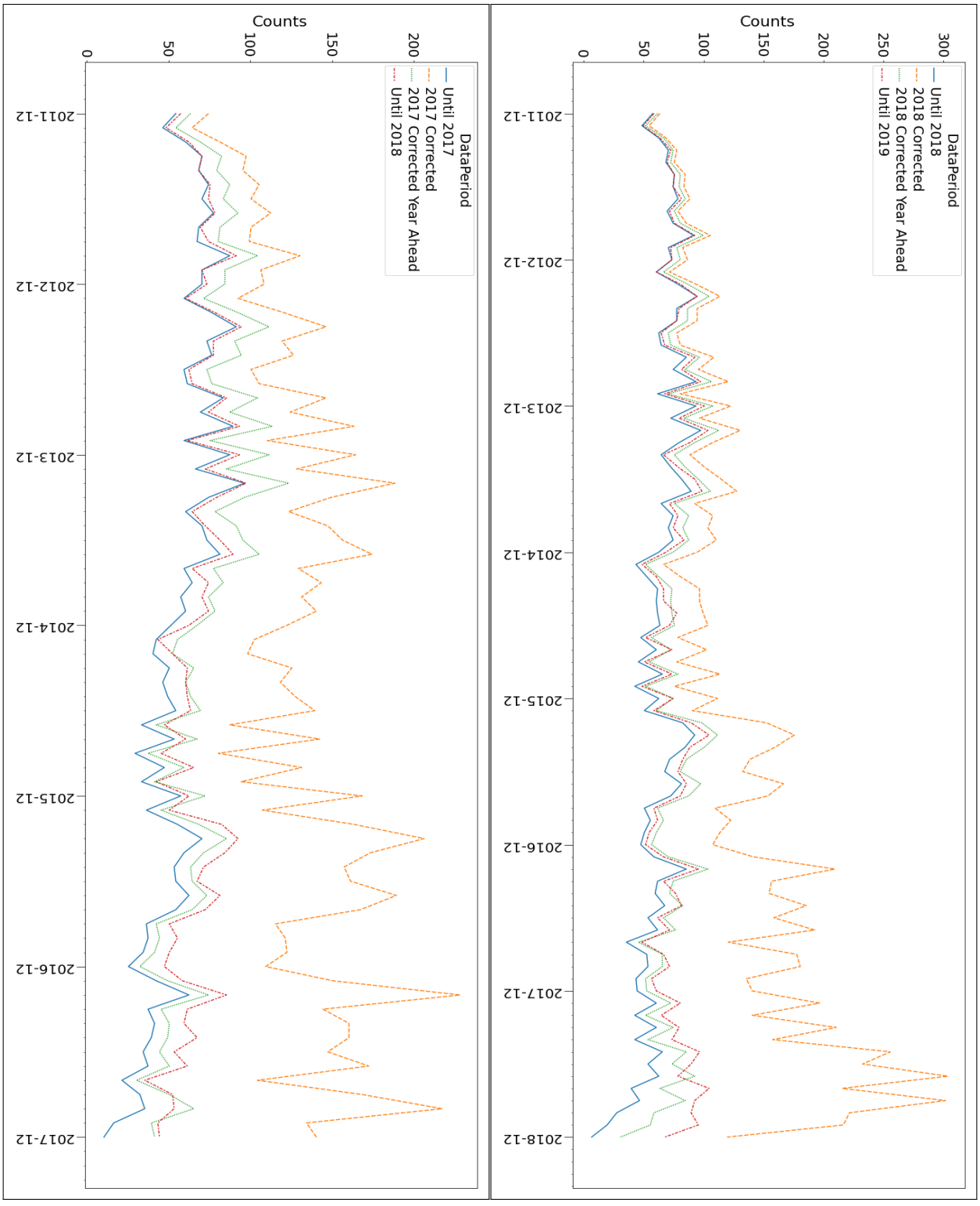
Figure 3.11: Validation Plots for Finance and Insurance Industry



(a) 2017 corrections Vs 2018 cumulative counts (b) 2018 corrections Vs 2019 cumulative counts

Until 201X - Counts reported as of 201X adjusted for the default date of January 1, proportionally
201X Corrected - "Counts until 201X" corrected based on Eq. 2.12
201X Corrected Year Ahead - "Counts until 201X" corrected based on Eq. 2.16

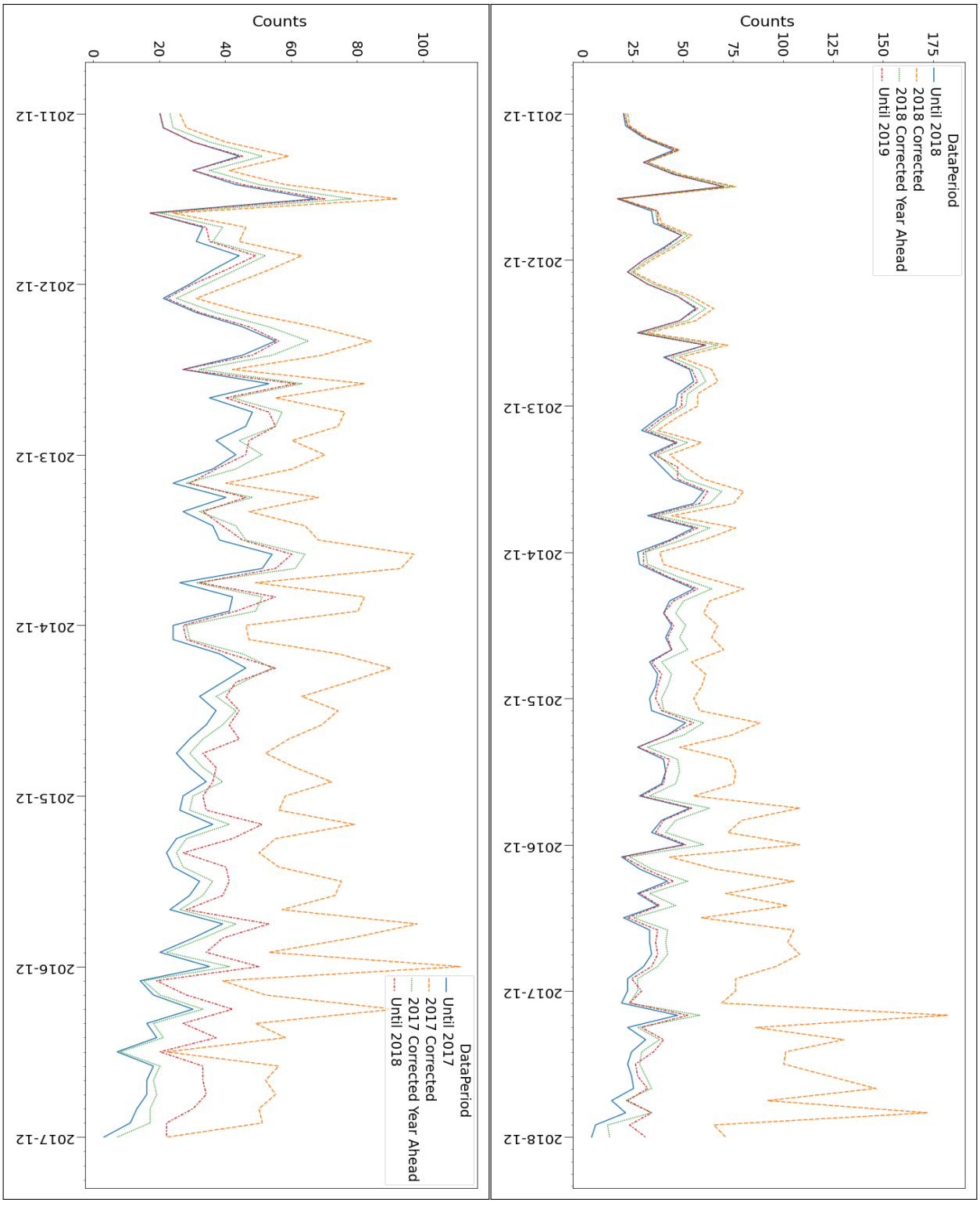
Figure 3.12: Validation Plots for Education Services Industry



(a) 2017 corrections Vs 2018 cumulative counts (b) 2018 corrections Vs 2019 cumulative counts

Until 201X - Counts reported as of 201X adjusted for the default date of January 1, proportionally
201X Corrected - "Counts until 201X" corrected based on Eq. 2.12
201X Corrected Year Ahead - "Counts until 201X" corrected based on Eq. 2.16

Figure 3.13: Validation Plots for Health Care and Social Assistance Industry



(a) 2017 corrections Vs 2018 cumulative counts (b) 2018 corrections Vs 2019 cumulative counts

Until 201X - Counts reported as of 201X adjusted for the default date of January 1, proportionally
201X Corrected - "Counts until 201X" corrected based on Eq. 2.12
201X Corrected Year Ahead - "Counts until 201X" corrected based on Eq. 2.16

Figure 3.14: Validation Plots for Public Administration Industry

Chapter 4

Modeling under-reporting in cyber incidents

4.1 Abstract

Under-reporting in cyber incidents is a well-established problem. Due to reputational risk and the consequent financial impact, a large proportion of incidents are never disclosed to the public, especially if they do not involve a breach of protected data. Generally, the problem of under-reporting is solved through a proportion-based approach, where the level of under-reporting in a data set is determined by comparison to data that is fully reported. In this work, cyber insurance claims data is used as the complete data set. Unlike most other work, however, our goal is to quantify under-reporting with respect to multiple dimensions: company revenue, industry, and incident categorization. The research shows that there is a dramatic difference in under-reporting–a factor of 100–as a function of these variables. The output of this work is an under-reporting model that can be used to correct incident frequencies derived from data sets of publicly reported incidents. This diminishes the "barrier to entry" in the development of cyber risk models, making it accessible to researchers who may not have the resources to acquire closely guarded cyber insurance claims data.

4.2 Introduction

Under-reporting is the problem of incidents being reported less often than the actual number of incidents. It is a common problem and most commonly studied in medical field [64][122]. A particularly relevant example is a sudden outbreak of COVID cases globally. Most people only get tested if they show symptoms or have been exposed to someone known to have been infected. This results in an unrepresentative sample of the full population. Medical studies are often based on a sample of patients considering the cost associated and difficulty associated in obtaining complete census data [124]. Ideally, everyone should be tested for COVID regularly but this would be prohibitively expensive. As a result, a less expensive small but accurate data set is a practical solution. The small data set would be obtained by appropriately sampling a portion of the population. This results in a sample that is unbiased by design.

The Cybersecurity field has struggled with the problem of under-reported cyber incidents. As a result, it is difficult to get an accurate estimate of the true number of cyber incidents impacting US organizations, or to accurately estimate total losses from these incidents. Organizations are reluctant to report cyber incidents since they directly impact businesses in terms of reputation and financial impact.. In addition, there is a belief that attackers will never be caught, so victims consider incident reporting a waste of time [24][37][46][95][117][125].

Correcting for under-reporting in data sets of publicly reported cyber incidents is necessary when building models from data that is typically included in these data sets, but not frequently included in claims data. This includes, for example, the number of records lost in data compromise events. As will be shown, cyber incidents impacting small companies are more under-reported than those impacting large companies. Since the number of records compromised and associated cost are typically higher for incidents impacting large companies, this skews the distributions that are directly constructed from data sets of publicly reported incidents [113].This is similar to the problem noted in the context of road accidents being under-reported where under-reported crash data is corrected from hospital data [33].

Brookmeyer recommends excluding under-reported data [22]. However, Wood et al. stated that this would lead to biased statistical models [135]. Elvik and Mysen argued that under-reporting contributed to incomplete data sets and results bias towards the reported data [33].

There are many studies found in the medical domain addressing under-reporting [64] [122]. The level of under-reporting in a data set is estimated by comparing it with fully reported but small data [1][3][4][6][8][12][13][18][19][25][29][32][33][40][41][47][60][66][69][71] [76][82][83][87][89][91][94][98][99][101][107][108][109][118][123][135]. More than 85% of the literature applied a proportions approach comparing small reported records against large population data. While this approach is easy to implement, it is difficult to obtain reliable data.

Hirvonen et al. studied under-reporting and trends in dietary data to evaluate energy level and found that women and over-weight individuals often under-report their food intake [64]. They applied logistic regression, which was easy to implement but results in concerns over the accuracy of the independent predictors. Lissener et al.performed a similar study but only on women data. They applied multiple regression with various combinations of body composition factors as an independent predictor and computed the range of under-reporting level with the mean daily weight change and standard error of mean(SEM[25]) [88]. Again, the approach was simple but obtaining such data from individuals can prove difficult.

Hazell and Shakir collected 37 different studies on adverse drug reactions. They estimated an under-reporting level as the median of the inter-quartile range [58]. This is the most simplistic and quick approach but it is difficult finding research with under-reporting estimates.

Krantz et al. studied COVID-19 data before its first peak and proposed a new method of harmonic analysis and wavelets to compute the level of under-reporting [78][79]. This approach develops complete data from incomplete partial data but involves complex mathematical models and is computationally intensive.

[25] $SEM = \frac{\sigma}{\sqrt{n}}$, where n is sample size

The proposed method models an under-reporting correction factor as a function of population characteristics. The study shows that there are extremely large differences in correction factors observed as a function of these variables.

This work presents parameters of a model of under-reporting. The frequency of cyber incidents of different types changes rapidly as attacker tactics evolve. However, the level of under-reporting of these cyber events is expected to change more slowly, as this would be primarily a consequence of legal changes. In the US, applicable laws are typically at the state level, making large changes in the level of under-reporting at the national level less likely. Therefore, the model of under-reporting presented here should have continued value for longer than a model of event frequency.

The model of under-reporting presented here is constructed by joining a number of proprietary data sets (see Section 4.3). All the constituent data sets are commercially available, with the exception of the claims and policy data. It is the intent of the authors for the results of this work to be used in conjunction with commercially available historical incident data sets and firmographic data sets in order to build unbiased cyber models *without requiring access to claims data*. Providing this model of under-reporting to the academic community should therefore help lower the barrier to entry in the development of cyber models by eliminating the need to acquire claims and policy data.

4.3 Data

Two proprietary data sets are used for this study – claim-exposure data, which is a small, unbiased and statistically representative, and historical incident-IED[26] data, which is a large but biased and statistically unrepresentative. The proprietary claim-exposure data is the collection of more than 30,000 US policies under-written by multiple insurers and claim information if there exists claims against those policies. The data set includes policy ID, start and end dates of policy, claim ID, claim date, claim amount, incident description,

[26]IED stands for Industry Exposure Database

incident type (extracted from the incident description), employee count, geographic location, industry, and revenue.

The proprietary historical incident-IED data set consists of a collection of more than 140,000 publicly reported historical incidents in the US over a period 2012-2019. The incidents in this data set were gathered via numerous collection methods, including scraped from technology and news websites, Securities Exchange Commission (SEC) filings, and other sources. An aggregated data set was constructed by combining historical incident data sets with a proprietary firmographic data set of companies that includes, name, location, industry, and revenue.

4.4 Methodology

The proposed approach aims to construct a model of under-reporting in cyber incidents as a function of revenue, incident type and industry. A model of event frequency as a function of company revenue, industry, and incident type is obtained for both the claims-exposure data set and the historical incident-IED data set. An under-reporting factor is computed as a ratio function of these variables.

Due to insufficient data, when examining combinations of revenue, industry, and incident type, separability of the models is assumed. That is, the incident frequency for a combination of variables can be expressed as a product of functions of a single variable each [133]. First, under-reporting corrections are computed as a function of revenue as shown in Eq.4.1. Assuming the revenue corrections are correct, the under-reporting corrections for revenue given incident type are computed as function of revenue and incident type as shown in Eq.4.2. Similarly, the under-reporting corrections for revenue given in any industry are computed as function of revenue and industry as shown in Eq.4.3. Extending further, the under-reporting corrections for revenue given incident type and industry can be computed as function of revenue r, incident type t, and industry i, as shown in Eq.4.4, assuming revenue

and incident type corrections are correct.

$$\text{Function of revenue: } UR(r) \tag{4.1}$$

$$\text{Function of revenue and incident type: } UR(r,t) = UR(r) \times UR(t) \tag{4.2}$$

$$\text{Function of revenue and industry: } UR(r,i) = UR(r) \times UR(i) \tag{4.3}$$

$$\text{Function of revenue, incident type and industry: } UR(r,t,i) = UR(r) \times UR(t) \times UR(i) \tag{4.4}$$

4.4.1 Revenue based corrections

The factor $UR(r)$ is computed based on the proportion of frequency of revenue from claim-exposure data, $freq_{CE}(r)$, and historical incident-IED data, $freq_{Inc-IED}(r)$, as shown in Eq.4.5.

$$UR(r) = \frac{freq_{CE}(r)}{freq_{Inc-IED}(r)} \tag{4.5}$$

For claim-exposure data, the revenue frequency, $freq_{CE}(r)$, is computed as the ratio of number of claims and the sum of policy years of the policies under-written for companies with given revenue r, as shown in Eq. 4.6. The policy year refers to the time period, in years, policy is written for.

$$freq_{CE,raw}(r) = \frac{Claims(r)}{\sum_{p \in P_r} Policy\,Years(p)} \tag{4.6}$$

where P_r refers to policies written for companies with revenue r

For historical incident-IED data, the revenue frequency, $freq_{Inc-IED}(r)$, is computed as ratio of number of incidents and the number of companies with given revenue r, as shown in Eq.4.7.

$$freq_{Inc-IED,raw}(r) = \frac{Incidents(r)}{N(r)} \tag{4.7}$$

where $N(r)$ is the number of organizations with revenue, r.

These revenue frequencies are taken on the log_{10} scale and smoothed over the rolling window of size d. The smoothed revenue frequency, $freq_{Smooth}(\log_{10} r)$, is computed as an average of frequencies in the range of revenue, $(\log_{10} r - d, \log_{10} r + d)$.

Considering the trends of revenue frequency shown in Fig.4.1, the exponential function is fitted on revenue of claim exposure (Fig.4.1a) and polynomials, with revenue frequency on log_{10} scale, fitted on historical incident-IED data (Fig.4.1b). In both functions, revenue is considered on $\log_{10}$ scale. However, the issue is more noticeable in historical incident-IED data. For historical incident-IED data, the revenue frequency on $\log_{10}$ scale is preferred considering the concentration of companies with smaller revenue. The exponential model and the polynomial, fitted with frequency being on $\log_{10}$ scale, ensure the positive values of frequency on historical incident-IED data.

The exponential function is defined as the power function of the form shown in Eq.4.8.

$$Y_{Exp}(x) = ae^{bx} \tag{4.8}$$

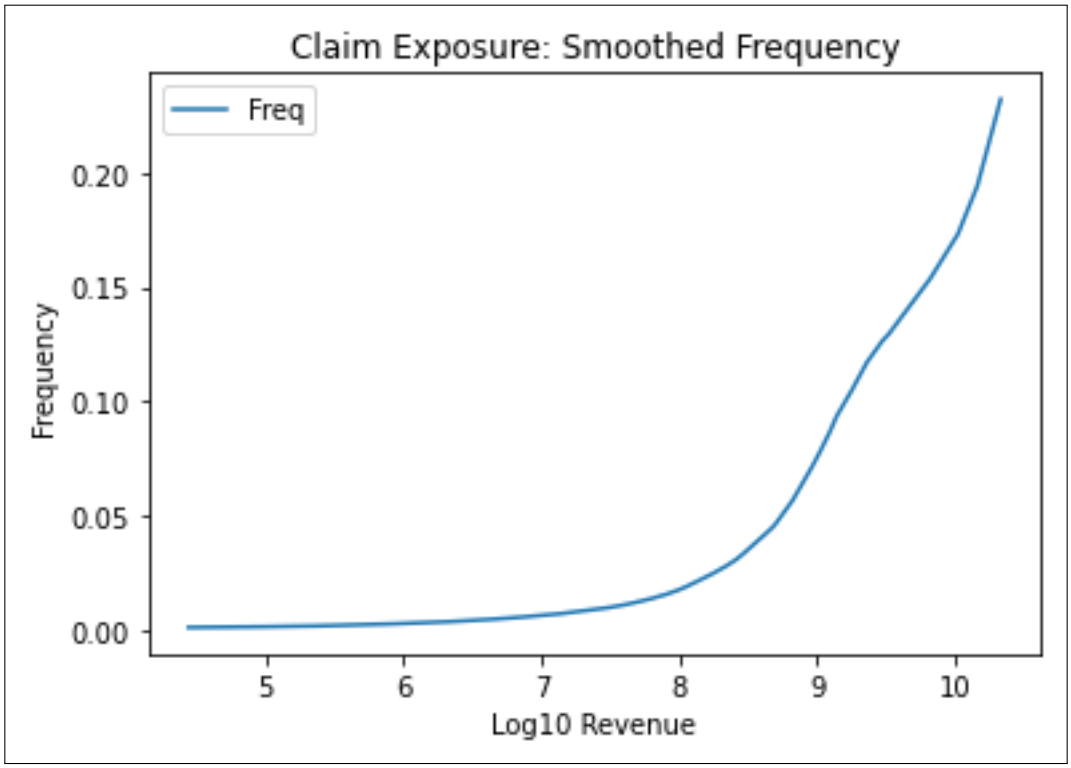
where a and b are the fitted parameters

To find the optimal degree for the polynomial, the polynomial is fitted for the multiple degrees on historical incident-IED train datasets separately and the tested on their test data sets. The degree is selected based on root mean square error(RMSE).

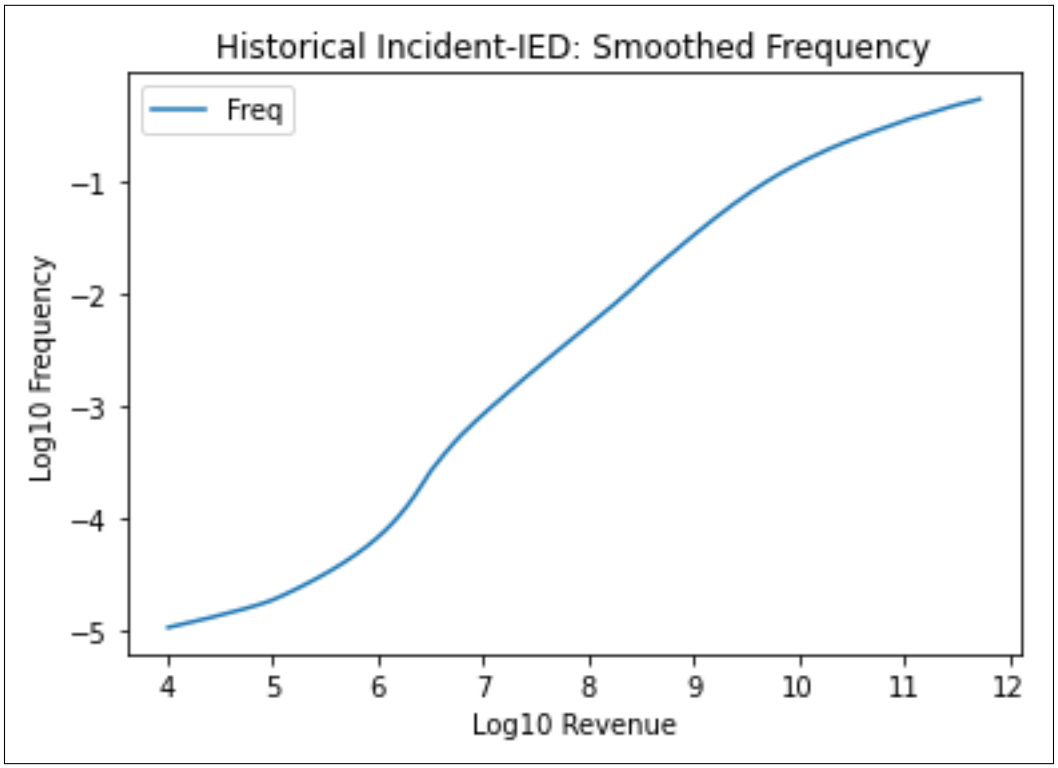
Finally, two incident frequency models are computed from claim-exposure and the historical incident-IED data sets. The under-reporting correction factors are computed as the ratio of the frequencies computed from the two models as function of revenue, as shown in Eq. 4.9.

$$UR(r) = \frac{Y_{Exp}(r)}{10^{Y_{Poly}(r)}} \tag{4.9}$$

where Y_{Exp} is the exponential model with claim-exposure data and Y_{Poly} is the polynomial model with historical incident-IED data with revenue frequency on $\log_{10}$ scale.



(a) Claim Exposure



(b) Historical incident-IED

Figure 4.1: Smoothed Frequency Plots

4.4.2 Revenue and Incident Type Corrections

The incident type factors are determined by comparing the revenue frequency for the incident type against the overall revenue frequency (irrespective of any incident type). These factors are scalar value for the given event type providing the revenue frequency of the incident type when multiplied with the overall revenue frequency.

Since policies are not specifically under-written for incident type, all policies are taken into consideration while computing revenue frequency w.r.t. incident type. For claim-exposure data, the revenue frequency for incident type, t can be computed as shown in Eq.4.10.

$$freq_{CE,raw}(r,t) = \frac{Claims(r,t)}{\sum_{p \in P_r} Policy\,Years(p)} \tag{4.10}$$

where $Claims(r,t)$ refers to number of claims with revenue r and incident type t and $PolicyYears(P_r)$ refers to the time period of the policy unwritten for organization with revenue r. Similarly, the revenue frequency with respect to incident types considers the records with given revenue from historical incident-IED data and computed as shown in Eq.4.11.

$$freq_{Inc-IED,raw}(r,t) = \frac{Incidents(r,t)}{N(r)} \tag{4.11}$$

where $Incidents(r,t)$ refers to number of incidents with revenue r and incident type t and N(r) refers to the number of policies written for organizations with revenue r.

Again, these frequencies are further smoothed over the rolling window of size d, as discussed earlier.

The factor, $f(t)$, is computed under the assumption that under-reporting correction factor of revenue, (r), is correct. The constant multiplier for the given incident type, $f(t)$, is computed such that the incident type based revenue frequencies can be determined as the proportion of overall revenue frequencies for claims exposure and historical incident-IED

dataset respectively, as shown in Eqs.4.12 and 4.13.

$$freq_{CE,Smooth}(r,t) \approx f_{CE}(t) \times freq_{CE,Fitted}(r) \tag{4.12}$$

$$freq_{Inc-IED,Smooth}(r,t) \approx f_{Inc-IED}(t) \times freq_{Inc-IED,Fitted}(r) \tag{4.13}$$

where $f_{CE}(r,t)$ and $f_{Inc-IED}(r,t)$ are constant multipliers computed by curve fitting approach such that the sum of squared difference between $freq._{,Smooth}(r,t)$ and $f.(t) \times freq._{,Fitted}(r)$ is minimized.

Accordingly, the under-reporting correction factor for given incident type t is computed as a function of both revenue and incident type, as shown in Eq.4.14-4.15.

$$UR(r,t) = UR(r) \times UR(t) \approx \frac{f_{CE,}(r,t) \times freq_{CE,Fitted}(r)}{f_{Inc-IED}r,(t) \times freq_{Inc-IED,Fitted}(r)} \tag{4.14}$$

$$\approx \frac{f_{CE}(r,t)}{f_{Inc-IED}(r,t)} \times UR(r) \tag{4.15}$$

4.4.3 Revenue and Industry Corrections

The revenue and industry corrections are computed in the same way as for revenue and incident type corrections but revenue frequencies for the given industry are computed differently.

From claim exposure data, the revenue frequency for industry i can be computed as shown in Eq.4.16.

$$freq_{CE,raw}(r,i) = \frac{Claims(r,i)}{\sum_{P_{r,i} \in P} Policy\,Years(P_{r,i})} \tag{4.16}$$

where $Claims(r,i)$ refers to number of claims with revenue r, and industry i.

From historical incident-IED data, the revenue frequency for industry i can be computed as shown in Eq.4.17.

$$freq_{Inc-IED,raw}(i) = \frac{Incidents(r,i)}{N_{r,i}} \tag{4.17}$$

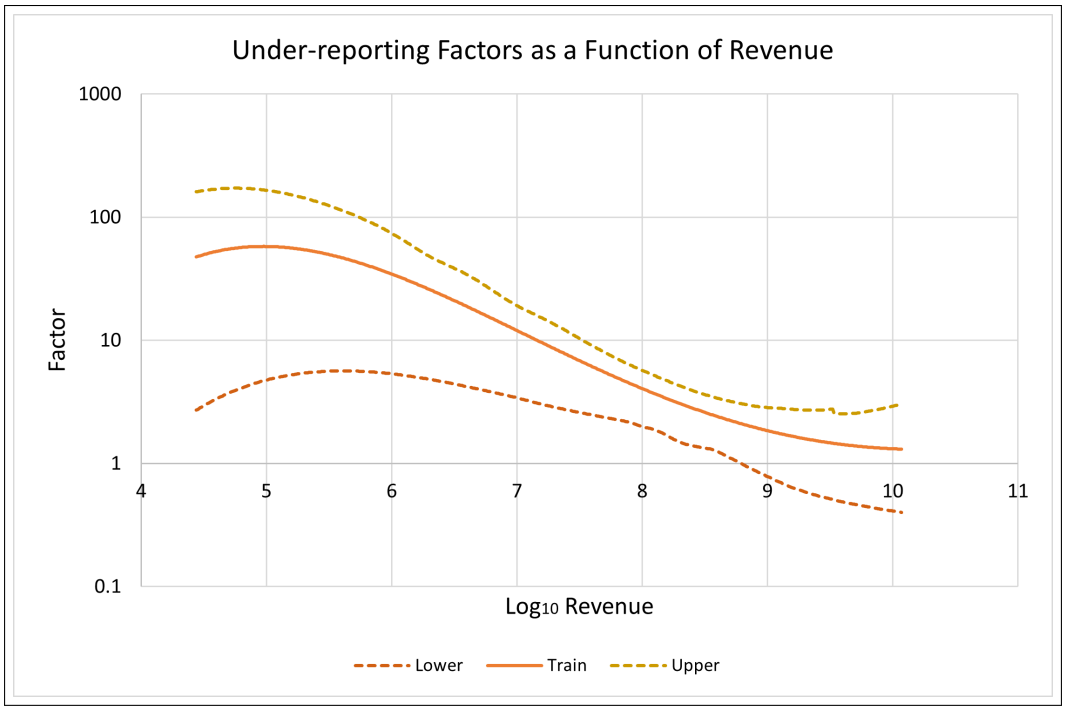


Figure 4.2: Under-reporting Factors as function of Revenue

where $Incidents(r, i)$ refers to number of incidents occurred in the industry i and $N_{r,i}$ refers to number of companies with revenue r in the industry i.

4.5 Results

In this section, under-reporting corrections are discussed for revenue, revenue and incident type, and revenue and industry.

4.5.1 Under-reporting Factors: Revenue

Fig.4.2 shows the factors and 95% confidence interval range for under-reporting as a function of revenue. The factor for low revenue companies found to be maximum as compared to the companies with high revenue.

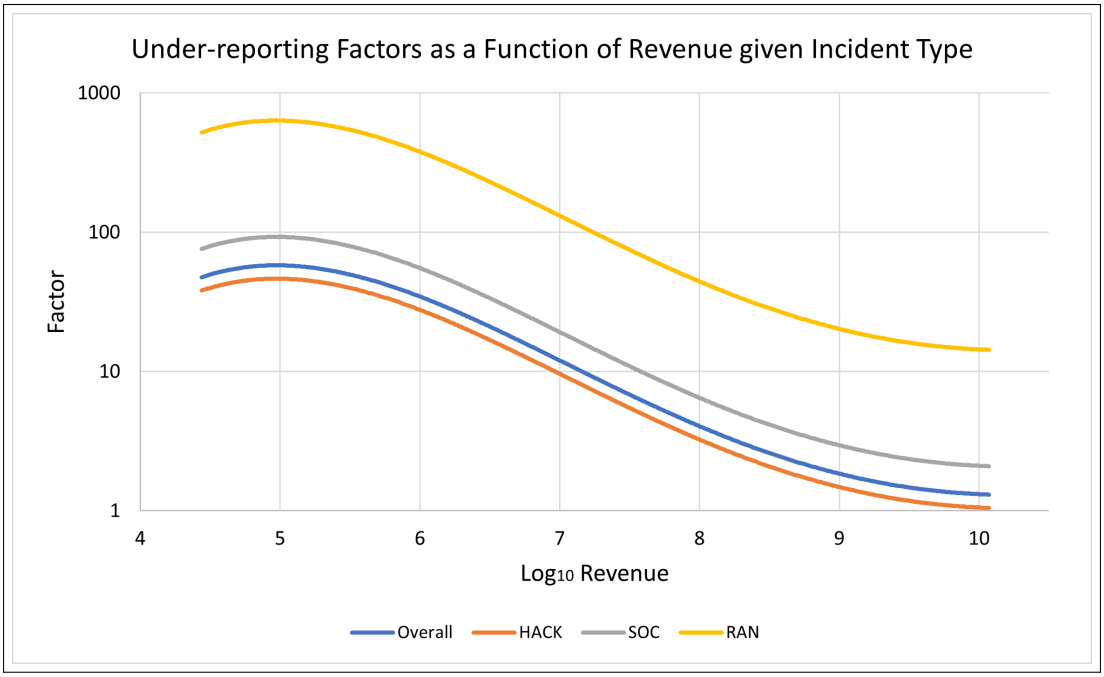


Figure 4.3: Under-reporting Factors as function of Revenue and Incident Type

4.5.2 Under-reporting Factors: Revenue and Incident Type

Table 4.1 shows the under-reporting correction factors for three incident types. Although HACK and SOC have comparatively smaller under-reporting factors when compared to RAN, the correction factor for SOC is almost double as of HACK. Fig. 4.3 shows that the

Table 4.1: Under-reporting Factors: Incident Type

Incident Type	Factor
HACK	0.8020
SOC	1.5994
RAN	10.9561

HACK incident type has least corrections for the given revenue as compared to the overall. On the contrary, SOC has higher corrections than overall and RAN requiring the highest level of corrections. The possible reason could be that there are reporting requirements for data compromise incidents whereas RAN does not have such requirements. The results emphasize on the under-reporting factors to be computed individually for incident type - one correction factor is not an ideal solution for different incident types.

4.5.3 Under-reporting Factors: Revenue and Industry

Table 4.2 shows the under-reporting correction factors for the five different industries. The RT and MFG industries have under-reporting factors less than one whereas FnI, PSTS and WT have more than one. The WT industry requires the largest correction factor of 4.6.

Table 4.2: Under-reporting Factors: Industry

Industry	Factor
RT	0.0838
MFG	0.7511
FnI	1.5472
PSTS	1.3960
WT	4.6097

Fig. 4.4 shows the under-reporting factors for five industries compared to the overall. The under-reporting factor for RT descends below one at revenues about ten million; this indicates that the separability assumption may not be adequate for this industry. Again, this emphasizes the under-reporting factors to be computed individually for industries - one correction factor is not an appropriate solution for different industries.

4.6 Validation

For validation, both claim-exposure and historical incident-IED data sets are split into training (33.33%) and test data sets(66.67%). The claim-exposure data is stratified on claims irrespective of revenue, incident type or industry. The true model determining under-reporting factors is computed using the complete training set.

For further validation purposes, 100 bootstrapped samples with replacement, each from claim-exposure and historical incident-IED, were generated from training data. In addition, 10,000 factors were computed by comparing all claim-exposure samples with each historical incident-IED sample. The 95% confidence intervals for the under-reporting factors are computed separately for $UR(r)$, $UR(t)$, and $UR(i)$ to validate against the corresponding factors obtained from the test data. Fig. 4.5 shows the validation of under-reporting

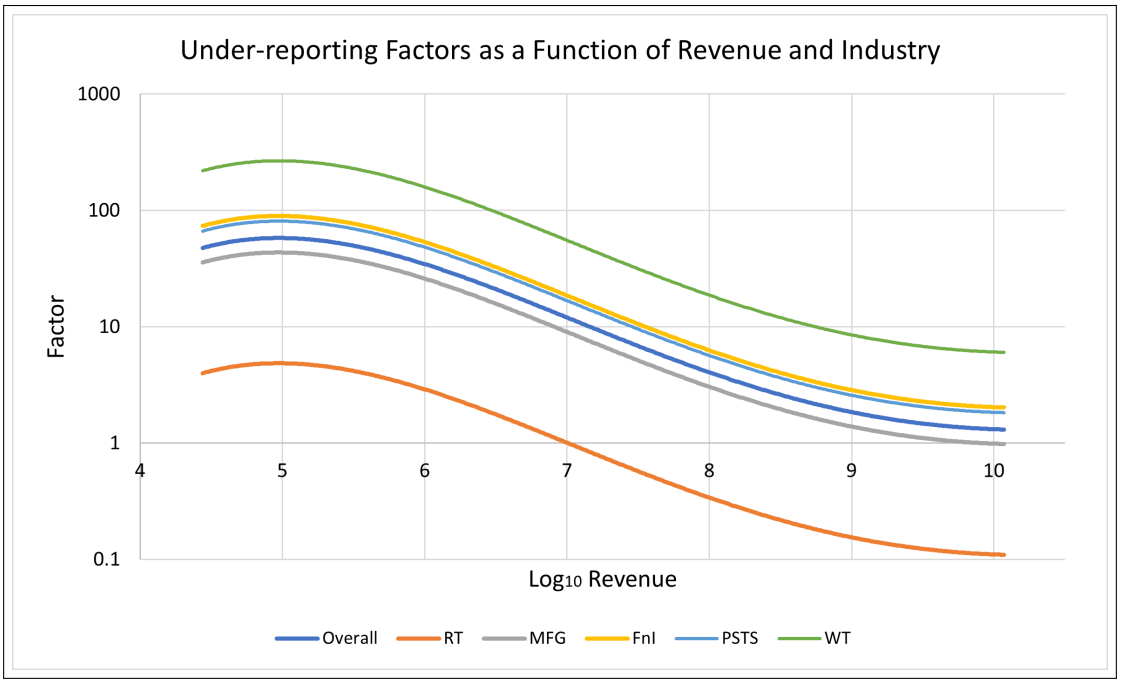


Figure 4.4: Under-reporting Factors as function of Revenue and Industry

corrections for revenue at 95% confidence interval bands. The plot indicates that there is a higher level of under-reporting for the organizations with lower revenue as compared to the ones with higher revenue. The correction factors could be more than 100 for low revenue organizations but found to be lower than one for organizations with revenues above 100 million and then increase - values lower than one is a statistical anomaly, whereas the increase is due to lack of enough data.

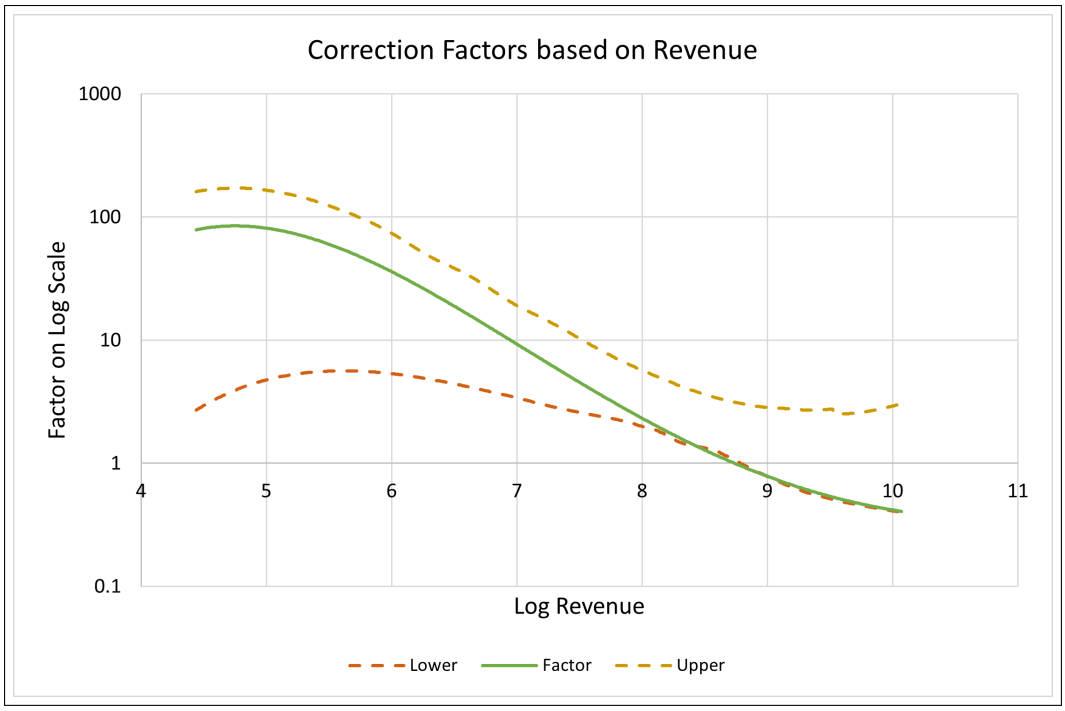
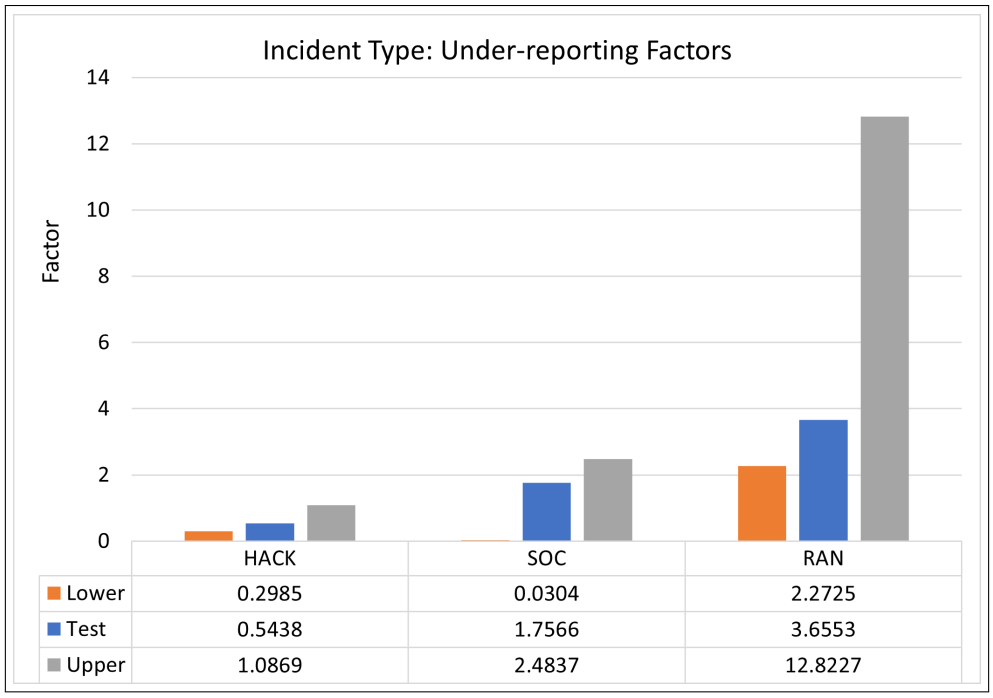


Figure 4.5: Under-reporting Factors: Revenue

Based on availability of the claim-exposure data, three incident types investigated: Hacking (HACK), Social Engineering (SOC) and Ransomware (RAN). Although the list of incident type is not exhaustive, these incident types account for the majority of incidents in both claim-exposure and historical incident data sets. HACK and SOC two different forms of data compromise incidents. An incident which begins with HACK or SOC but ultimately leads to ransom is classified as RAN.

Fig. 4.6 shows under-reporting factors for three incident types. The under-reporting factor for RAN could range from 2.3 to 12.8. The upper level of HACK is less than the lower level of RAN. Even upper level of SOC is around 10% higher than the lower level of RAN. The factors for SOC could be around 2.5 whereas there is minimum level of under-reporting for HACK.

Based on availability of the claim-exposure data, five industries investigated: Retail Trade (RT), Manufacturing (MFG), Finance and Insurance (FnI), Professional Scientific Technical Services (PSTS) and Wholesale Trade (WT). As shown in Fig. 4.7, there is contrast observed in retail and wholesale trade industries where RT has the least under-reporting factor and



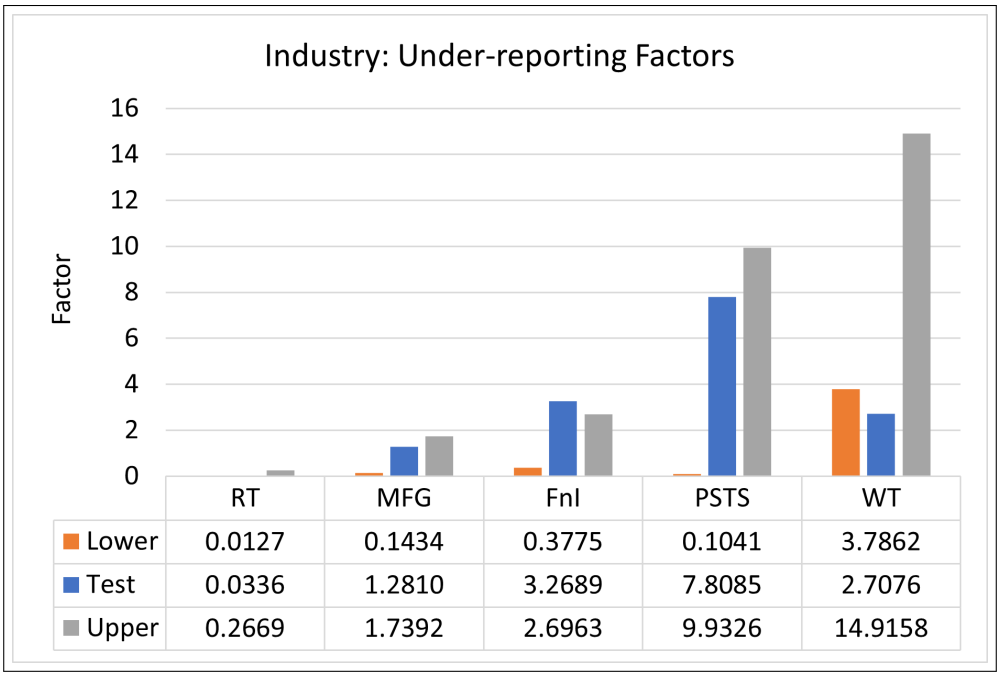
	HACK	SOC	RAN
Lower	0.2985	0.0304	2.2725
Test	0.5438	1.7566	3.6553
Upper	1.0869	2.4837	12.8227

Figure 4.6: Under-reporting Factors: Incident Type

WT has the highest. The under-reporting factor for FnI industry based on test data is found to be above 95 percentile range. On the contrary, under-reporting factor for WT industry based on test data is found to be below 95 percentile range. Since the data split for the train and test data are samples stratified on claims irrespective of revenue, incident type or industry, the virtue of the data split could result in these two anomalies. Since under-reporting factors for incident types are within 95% confidence interval range but those for industry are not always, it might be worth investigating revenue given incident types within industries; this could not be done due to lack of data.

4.7 Conclusion

The research proposed the computation of under-reporting factors in more than one dimension. The under-reporting factors were computed in cyber incidents for the organizations with varying revenue. The study shows that the organizations with lower revenues require more correction compared to those with higher revenues. Secondly, the under-reporting factors are computed for three incident types– hacking (HACK), social engineering



	RT	MFG	FnI	PSTS	WT
Lower	0.0127	0.1434	0.3775	0.1041	3.7862
Test	0.0336	1.2810	3.2689	7.8085	2.7076
Upper	0.2669	1.7392	2.6963	9.9326	14.9158

Figure 4.7: Under-reporting Factors: Industry

(SOC) and ransomware (RAN)– assuming that the revenue factors are correct,the research showed that HACK requires the least correction factors and RAN the greatest correction factor. Thirdly, the under-reporting factors are computed for five industries– Retail Trade (RT), Manufacturing (MFG), Finance and Insurance(FnI), Professional Scientific Technical Services (PSTS) and Wholesale Trade (WT)– again under the assumption that revenue factors are correct. The research showed that WT, FnI, PSTS needs to be corrected more than overall market, whereas RT and MFG requires less corrections than the overall market. The research indicates the necessity to address each industry independently for under-reporting corrections.

Chapter 5

Conclusion

The research addressed two major problems with cyber incidents– reporting delays and under-reporting. While reporting delays refers to the cyber incidents reported with delays, under-reporting refers to the difference between the true rate of claims and the rate as derived from publicly reported incidents.

5.1 Reporting Delays

To address reporting delays, delay and age were computed. Delay is computed based on time difference between the incident date and reporting date whereas age is computed as difference between the incident date and last reporting date in the data set. The debiased delay distribution is constructed from the distribution of delay and age but found to be bimodal- with one peak at 0-30 days and a second peak occurring after more than 5 years. Hence, debiased delay distribution is modeled by mixture of two distributions– exponential, for incidents discovered immediately but yet to be reported, and normal, for incidents not discovered immediately. In addition, four key problems were addressed: the distribution being biased to shorter delays, non-stationarity, delays not addressed beyond maximum delay in the data set, and longer delays are addressed based on few data points. The first problem of distribution being biased to shorter delays is addressed by generating the debiased distribution with delay and age. The second problem of non-stationarity is addressed by

generating monthly debiased distribution over two-year rolling period. The third problem of delays not being addressed beyond maximum delay is addressed by defining the bimodal distribution with mixture distribution over domain $[0, \infty)$. The last problem of longer delays based on few data points is addressed through optimization function where delays further in time are less prioritized as marginally fewer number of incidents are anticipated with longer reporting delays. The current cyber incident counts can be corrected to account for the proportion of incidents that have occurred but not yet been reported with the modeled debiased delay distribution. The approach was validated by comparing year ahead counts with year ahead corrections.

A reliable and accurate historical data is required to understand the cyber threat landscape and build robust cyber risk models. While it is not possible to obtain precise results for the number of cyber incidents, the proposed algorithm corrects for reporting delays approximately. The reported cyber incident counts in recent times indicate decreasing trend simply due to incidents not being reported yet. However, the rate of cyber incidents is actually increasing in practice which the algorithm unfolds.

Further, the reporting delays are modeled for the US market and compared for four industry sectors–finance and insurance (FnI), educational services (ES), health care and social services (HnS), and public administration (PA). These factors eventually expected and found that to converge to one for incidents in the past. The longer delays are more prominent for FnI industry as there is small percentage of the incidents being detected immediately. Generally, the research indicates that US and four industries are actively trying to detect the incidents in shorter period of time. However, it might be difficult to detect the incident in timely manner considering the hackers have access to state-of-the-art technology and are becoming increasingly sophisticated.

5.2 Under-reporting

For under-reporting, correction factors are computed for the organizations' revenue, revenue given incident type and revenue given industry for US. The research indicated that the level

of under-reporting is high for low revenue organizations. To address under-reporting, the correction factors are computed based on revenue frequency. While the correction factors are computed for the wide range of the revenue, these factors are multiplied by scalar multiplier for the given specific incident type or industry to determine the correction factors for revenue given incident type or revenue given industry.

Three incident types were investigated- hacking (HACK), social engineering (SOC) and ransomware (RAN). The research showed the under-reporting in RAN incidents require high level of corrections. This could be the result of not having reporting requirements whereas HACK and SOC do have such requirements.

Five different industries were evaluated - Retail Trade (RT), Manufacturing (MFG), Finance and Insurance (FnI), Professional Scientific Technical Services (PSTS) and Wholesale Trade (WT). The research showed the under-reporting in RT and MFG lower than the overall market whereas WT, FnI and PSTS have larger.

Data availability is major concern in conducting the research in cyber domain.

On the whole, the research shows that both reporting delays and under-reporting are important problems to be addressed. These insights can dramatically improve the cyber risk evaluation.

5.3 Future Work

5.3.1 Reporting Delays

The current research explores the correction factors for reporting delays from US market and four key industries. The future work could be extended to different markets such as Europe, Asia, Australia etc. or different industries. In addition, it would be interesting to investigate how the correction factors vary for the incident types e.g. hacking, malware, etc. Such research depends hugely on the availability of data.

5.3.2 Under-reporting

The current research focuses on determining correction factors for under-reporting for organizations with the given revenue, revenue given incident type and revenue given industry under separability assumption. The future work could be relaxing separability assumption, or extending the analysis to other incident types and industries. This would however require more data.

Bibliography

Bibliography

[1] Abay, K. A. (2015). Investigating the nature and impact of reporting bias in road crash data. *Transportation Research Part A: Policy and Practice*, 71:31–45. 22, 24, 26, 82

[2] Ackerman, G. (2013). G 20 Urged to Treat Cyber Attacks as Threat to Global Economy. 1, 61

[3] Alsop, J. and Langley, J. (2001). Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accident Analysis and Prevention*, 33(3):353–359. 22, 24, 26, 82

[4] Alvarez-Requejo, A., Carvajal, A., Bégaud, B., Moride, Y., Vega, T., and Martín Arias, L. H. (1998). Under-reporting of adverse drug reactions. Estimate based on a spontaneous reporting scheme and a sentinel system. *European Journal of Clinical Pharmacology*, 54(6):483–488. 21, 24, 82

[5] Amir, E., Levi, S., and Livne, T. (2018). Do Firms Underreport Information on Cyber-Attacks? Evidence from Capital Markets. *SSRN Electronic Journal*. 1

[6] Amoros, E., Martin, J. L., and Laumon, B. (2006). Under-reporting of road crash casualties in France. *Accident Analysis and Prevention*, 38(4):627–635. 22, 24, 26, 82

[7] Antonio, K. and Plat, R. (2012). Micro-Level Stochastic Loss Reserving for General Insurance. *SSRN Electronic Journal*. 5

[8] Arneborn, P. and Palmblad, J. (1982). DrugInduced NeutropeniaA Survey for Stockholm 19731978. *Acta Medica Scandinavica*, 212(5):289–292. 21, 24, 82

[9] Audit Analytics (2021). Trends in Cybersecurity Breaches. Technical report, Audit Analytics, Massachusetts, USA. 1, 2, 3, 16, 35, 37, 61, 64, 74

[10] Auger, A. and Hansen, N. (2011). CMA-ES: Evolution strategies and covariance matrix adaptation. In *Genetic and Evolutionary Computation Conference, GECCO'11 - Companion Publication*, pages 991–1010. 49

[11] Avanzi, B., Wong, B., and Yang, X. (2016). A micro-level claim count model with overdispersion and reporting delays. *Insurance: Mathematics and Economics*, 71:1–14. 16, 36

[12] Bäckström, M., Mjörndal, T., and Dahlqvist, R. (2004). Under-reporting of serious adverse drug reactions in Sweden. *Pharmacoepidemiology and Drug Safety*, 13(7):483–487. 21, 24, 82

[13] Bagheri, H., Michel, F., Lapeyre-Mestre, M., Lagier, E., Cambus, J. P., Valdiguie, P., and Montastruc, J. L. (2000). Detection and incidence of drug-induced liver injuries in hospital: A prospective analysis from laboratory signals. *British Journal of Clinical Pharmacology*, 50(5):479–484. 21, 24, 82

[14] Barbosa, M. T. S. and Struchiner, C. J. (2002). The estimated magnitude of AIDS in Brazil: a delay correction applied to cases with lost dates. *Cadernos de saúde pública / Ministério da Saúde, Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública*, 18(1):279–285. 13

[15] Bastos, L. S., Economou, T., Gomes, M. F., Villela, D. A., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T., and Codeço, C. T. (2019). A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine*, 38(22):4363–4377. 13, 19, 36

[16] Bégaud, B., Chaslerie, A., and Haramburu, F. (1994). [Organization and results of drug vigilance in France]. *Revue d'epidemiologie et de sante publique*, 42(5):416–41623. 21

[17] Bose, N. (2021). Biden: If U.S. has 'real shooting war' it could be result of cyber attacks. 1

[18] BÖttiger, L. E. and Westerholm, B. (1973). Drug-induced Blood Dyscrasias in Sweden. *British Medical Journal*, 3(5875):339–343. 21, 24, 82

[19] BÖttiger, M., Romanus, V., de Verdier, C., and Boman, G. (1982). Osteitis and other complications caused by generalized BCGITIS: Experiences in Sweden. *Acta Pædiatrica*, 71(3):471–478. 21, 24, 82

[20] Brill, J. E. (2013). *Encyclopedia of Survey Research Methods*. 2

[21] Brookmeyer, R. and Damiano, A. (1989). Statistical methods for shortterm projections of AIDS incidence. *Statistics in Medicine*, 8(1):23–34. 5, 7, 8, 9, 10, 17, 35, 61

[22] Brookmeyer, R. and Gail, M. H. (1986). Minimum Size of the Acquired Immunodeficiency Syndrome (Aids) Epidemic in the United States. *The Lancet*, 328(8519):1320–1322. 5, 12, 20, 82

[23] Brookmeyer, R. and Liao, J. (1990). The analysis of delays in disease reporting: Methods and results for the acquired immunodeficiency syndrome. *American Journal of Epidemiology*, 132(2):355–365. 5, 10, 17, 25, 36, 37, 39, 61, 62

[24] 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(1):70–104. 2, 20, 81

[25] Chan, T. Y. and Critchley, J. A. (1994). Reporting of adverse drug reactions in relation to general medical admissions to a teaching hospital in Hong Kong. *Pharmacoepidemiology and Drug Safety*, 3(2):85–89. 21, 24, 82

[26] Cheng, F. F. and Ford, W. L. (1991). Adjustment of aids surveillance data for reporting delay to the editor:. 10, 18, 35, 45, 61

[27] Chitwood, M. H., Russi, M., Gunasekera, K., Havumaki, J., Klaassen, F., Pitzer, V. E., Salomon, J. A., Swartwood, N. A., Warren, J. L., Weinberger, D. M., Cohen, T., and Menzies, N. A. (2020). Reconstructing the course of the COVID-19 epidemic over 2020 for US states and counties: Results of a Bayesian evidence synthesis model. *medRxiv : the preprint server for health sciences*, page 2020.06.17.20133983. 13, 19, 36

[28] Clarke, L. E. (1975). *Stochastic processes, by R. Coleman. Pp 93. £4·50 hardback, £1·95 paperback. 1974. SBN 0 04 519016 X/519017 8 (George Allen and Unwin)*, volume 59. John Wiley, New York. 15

[29] Classen, D. C., Pestotnik, S. L., Evans, R. S., and Burke, J. P. (2005). Computerized surveillance of adverse drug events in hospital patients. 1991. *Quality & safety in health care*, 14(3):2847–51. 21, 24, 82

[30] Downs, A. M., Ancelle, R., Jager, J. C., Heisterkamp, S. H., Van Druten, J. A., Ruitenberg, E. J., and J.B., B. (1988). The statistical estimation, from routine surveillance data, of past, present and future trends in AIDS incidence in Europe. In Jager, J. C. and Ruitenberg, E. J., editors, *Statistical Analysis and Mathematical Modelling of AIDS*, pages 1–16. Oxford University Press, Oxford. 7, 17, 35, 61

[31] Downs, A. M., Ancelle, R. A., Jager, H. J., and Brunet, J. B. (1987). AIDS in Europe: Current trends and short-term predictions estimated from surveillance data, January 1981-June 1986. *Aids*, 1(1):53–57. 7, 17, 35, 61

[32] Dugué, A., Bagheri, H., Lapeyre-Mestre, M., Tournamille, J. F., Sailler, L., Dedieu, G., Salvayre, R., Thouvenot, J. P., Massip, P., and Montastruc, J. L. (2004). Detection and incidence of muscular adverse drug reactions: A prospective analysis from laboratory signals. *European Journal of Clinical Pharmacology*, 60(4):285–292. 21, 24, 82

[33] Elvik, R. and Mysen, A. B. (1999). Incomplete accident reporting: Meta-analysis of studies made in 13 countries. *Transportation Research Record*, (1665):133–140. 20, 22, 24, 26, 81, 82

[34] England, P. and Verrall, R. (2002). Stochastic Claims Reserving in General Insurance. *British Actuarial Journal*, 8(3):443–518. 13

[35] Esbjerg, S., Keiding, N., and Koch-Henriksen, N. (1999). Reporting delay and corrected incidence of multiple sclerosis. *Statistics in Medicine*, 18(13):1691–1706. 11, 17, 36, 61, 62

[36] Evans, A. (2019). Managing cyber risk. *Managing Cyber Risk*, pages 1–112. 1

[37] Fafinski, S. (2009). Uk Cybercrime Report 2009. (September). 2, 20, 81

[38] Farrington, P. (1995). A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines. *Parasitology Today*, 11(5):171. 21

[39] FBI (2018). 2018 Internet Crime Report. Technical report. 3

[40] FE Skjeldestad, T Amundsen, E. H. (2000). Reporting of adverse drug reactions to the Norwegian Drug Control Agency [in Norwegian]. *Tidsskr Nor Laegeforen*, 120:336–8. 21, 24, 82

[41] Fletcher, A. P. (1991). Spontaneous adverse drug reaction reporting vs event monitoring: A comparison. *Journal of the Royal Society of Medicine*, 84(6):341–344. 21, 24, 82

[42] Fletcher, S. M., Lewis-Fuller, E., Williams, H., Miller, Z., Scarlett, H. P., Cooper, C., Gordon-Johnson, K. A., Vickers, I., Shaw, K., Wellington, I., Thame, J., Pérez, E., and Indar, L. (2013). Magnitude, distribution, and estimated level of underreporting of acute gastroenteritis in Jamaica. *Journal of Health, Population and Nutrition*, 31(4 SUPPL.1). 20, 21

[43] Gail, M. H. and Brookmeyer, R. (1988). Methods for projecting course of acquired immunodeficiency syndrome epidemic. *Journal of the National Cancer Institute*, 80(12):900–911. 17, 36, 61, 62

[44] Gebhardt, M. D., Neuenschwander, B. E., and Zwahlen, M. (1998). Adjusting AIDS incidence for non-stationary reporting delays: A necessity for country comparisons. *European Journal of Epidemiology*, 14(6):595–603. 12

[45] Ghani, A., Brookmeyer, R., and Gail, M. H. (1995). AIDS Epidemiology: A Quantitative Approach. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(2):343. 10

[46] Goucher, W. (2010). Being a cybercrime victim. *Computer Fraud and Security*, 2010(10):16–18. 2, 20, 81

[47] Hallas, J., Gram, L., Grodum, E., Damsbo, N., Brosen, K., Haghfelt, T., Harvald, B., Beck Nielsen, J., Worm, J., Jensen, K., and al, E. (1992). Drug related admissions

to medical wards: a population based survey. *British Journal of Clinical Pharmacology*, 33(1):61–68. 21, 24, 82

[48] Hampel, F. and Zurich, E. (1998). Is statistics too difficult? *Canadian Journal of Statistics*, 26(3):497–513. 37

[49] Hansen, N. (2006). The CMA evolution strategy: a comparing review, in: J.A. Lozano, P. Larranaga, I. Inza, E. Bengoetxea (Eds.), Towards A New Evolutionary Computation. Advances on Estimation of Distribution Algorithms, Springer. 192:75–102. 48, 49

[50] Hansen, N. (2016). The CMA Evolution Strategy: A Tutorial. *Computing Research Repository*. 48, 49

[51] Hansen, N. (2019). CMA - Python Package. 48, 49

[52] Hansen, N. and Kern, S. (2004). Evaluating the CMA evolution strategy on multimodal test functions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3242:282–291. 48

[53] Hansen, N., Müller, S. D., and Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18. 48

[54] Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195. 48

[55] Harris, J. E. (1987). Delay in Reporting Acquired Immune Deficiency Syndrome (AIDS). *National Bureau of Economic Research Working Paper Series*, No. 2278. 5, 35, 61

[56] Harris, J. E. (1990). Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association*, 85(412):915–924. 5, 12, 17

[57] Harris, J. E. (2020). Overcoming reporting delays is critical to timely epidemic monitoring: the case of COVID-19 in New York City. *medRxiv*, page 2020.08.02.20159418. 16, 19, 36

[58] Hazell, L. and Shakir, S. A. (2006). Under-reporting of adverse drug reactions: A systematic review. *Drug Safety*, 29(5):385–396. 21, 25, 82

[59] Healy, M. J. R. and Tillett, H. E. (1988). Short-Term Extrapolation of the AIDS Epidemic. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 151(1):50. 8, 17, 35, 61

[60] Heeley, E., Riley, J., Layton, D., Wilton, L. V., and Shakir, S. A. (2001). Prescription-event monitoring and reporting of adverse drug reactions. *Lancet*, 358(9296):1872–1873. 21, 24, 82

[61] Heisterkamp, S. H., Jager, J. C., Downs, A. M., and Van Druten, J. A. (1988a). The use of Genstat in the estimation of expected numbers of AIDS cases adjusted for reporting delays. In *Fifth Genstat Conference*, pages 4–18. 5, 6, 17, 35, 36, 61

[62] Heisterkamp, S. H., Jager, J. C., Downs, A. M., Van Druten, J. A., and Ruitenberg, E. J. (1988b). Statistical estimation of AIDS incidence from surveillance data and the link with modelling of trends. In *Statistical Analysis and Mathematical Modelling of AIDS*, pages 17–25. Oxford University Press, Oxford. 5, 6, 17, 35, 36, 61

[63] Heisterkamp, S. H., Jager, J. C., Ruitenberg, E. J., Van Druten, J. A., and Downs, A. M. (1989). Correcting reported aids incidence: A statistical approach. *Statistics in Medicine*, 8(8):963–976. 5, 6, 17, 35, 36, 61

[64] Hirvonen, T., Männistö, S., Roos, E., and Pietinen, P. (1997). Increasing prevalence of underreporting does not necessarily distort dietary surveys. *European Journal of Clinical Nutrition*, 51(5):297–301. 20, 22, 24, 81, 82

[65] Höhle, M. and An Der Heiden, M. (2014). Bayesian nowcasting during the STEC O104: H4 outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002. 5, 13, 19, 36

[66] Imbs, J. L., Pouyanne, P., Haramburu, F., Welsch, M., Decker, N., Blayac, J. P., Bégaud, B., Andréjak, M., Allain, P., Bechtel, P., Riché, C., Moulin, M., Lavarenne, J., Albengres, E., Escousse, A., Mallaret, M., Caron, J., Merle, L., Evreux, J. C., Jouglard, J., Netter, P., Larousse, C., Chichmanian, R. M., Kreft-Jais, C., Biour, M., Efthymiou,

M. L., Bavoux, F., Soubrié, C., Vandel, B., Trenque, T., Allain, H., Tuillez, C., Ollagnier, M., Montastruc, J. L., and Autret, E. (1999). Adverse drug reaction: Prevalence in French public hospitals. *Therapie*, 54(1):21–27. 21, 24, 82

[67] In 'T Veld, B. A., Van Der Linden, P. D., Feenstra, J., and Stricker, B. H. (2000). The function of a reporting system for suspected adverse drug reactions as risk indicator for Stevens-Johnson syndrome and toxic epidermal necrolysis. *Tijdschrift voor Geneeskunde*, 56(17):1258–1263. 21

[68] Infocyte (2016). The Breach Detection Gap and Strategies to Close It. Technical report. 1

[69] Inman, W. H. (1977). Study of fatal bone marrow depression with special reference to phenylbutazone and oxyphenbutazone. *British Medical Journal*, 1(6075):1500–1505. 21, 24, 82

[70] Inman, W. H. and Adelstein, A. M. (1969). Rise and fall of asthma mortality in England and Wales in relation to use of pressurised aerosols. *Lancet*, 2(7615):279–285. 21

[71] Inman, W. H. and Vessey, M. P. (1968). Investigation of Deaths from Pulmonary, Coronary, and Cerebral Thrombosis and Embolism in Women of Child-bearing Age. *British Medical Journal*, 2(5599):193–199. 21, 24, 82

[72] ISACA (2019). State of Cybersecurity 2019 - Part 2: Current Trends in Attacks, Awareness and Governance. *ISACA Press Release*, (November 2018):1–21. 2

[73] Jewell, W. S. (1990). Predicting IBNYR Events and Delays II. Discrete Time. *ASTIN Bulletin*, 20(1):93–111. 16, 36

[74] Kalbfleisch, J. D. and Lawless, J. F. (1991). Regression Models for Right Truncated Data With Applications To Aids Incubation Times and Reporting Lags. *Statistica Sinica*, 1(1):19–32. 11, 12, 17, 36, 61, 62

[75] Keiding, N. and Moeschberger, M. (1992). Independent Delayed Entry. In *Survival Analysis: State of the Art*, pages 309–326. Springer Netherlands. 18, 36

[76] Kimmel, S. E., Sekeres, M. A., Berlin, J. A., Goldberg, L. R., and Strom, B. L. (1998). Adverse events after protamine administration in patients undergoing cardiopulmonary bypass: Risks and predictors of under-reporting. *Journal of Clinical Epidemiology*, 51(1):1–10. 21, 24, 82

[77] Kosek, M., Bern, C., and Guerrant, R. L. (2003). The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bulletin of the World Health Organization*, 81(3):197–204. 20

[78] Krantz, S. G., Polyakov, P., and Rao, A. S. (2020). True epidemic growth construction through harmonic analysis. *Journal of Theoretical Biology*, 494. 23, 25, 82

[79] Krantz, S. G. and Rao, A. S. (2020). Level of underreporting including underdiagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. *Infection Control and Hospital Epidemiology*, 41(7):857–859. 23, 25, 82

[80] Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86. 44

[81] Kumara, S. S. and Chin, H. C. (2005). Application of poisson underreporting model to examine crash frequencies at signalized three-legged intersections. *Transportation Research Record*, (1908):46–50. 26

[82] La Grenade, L., Graham, D. J., and Nourjah, P. (2001). Underreporting of hemorrhagic stroke associated with Phenylpropanolamine [4]. *Journal of the American Medical Association*, 286(24):3081. 21, 24, 82

[83] Lacoste-Roussillon, C., Pouyanne, P., Haramburu, F., Miremont, G., and Bégaud, B. (2001). Incidence of serious adverse drug reactions in general practice: A prospective study. *Clinical Pharmacology and Therapeutics*, 69(6):458–462. 21, 24, 82

[84] Lawless, J. F. (1989). Estimating the incubation time distribution and expected number of cases of transfusionassociated acquired immune deficiency syndrome. *Transfusion*, 29(8):672–676. 15

[85] Lawless, J. F. (1994). Adjustments for Reporting Delays and the Prediction of Occurred but Not Reported Events. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 22(1):15. 5, 12, 18, 36

[86] Lee Mathews (2017). 2016 Saw An Insane Rise In The Number Of Ransomware Attacks. 67

[87] Lewis, M., Kühl-Habich, D., and von Rosen, J. (2009). Drug use and adverse event monitoring in German children. *Int. Journal of Clinical Pharmacology and Therapeutics*. 21, 24, 82

[88] Lissner, L., Habicht, J. P., Strupp, B. J., Levitsky, D. A., Haas, J. D., and Roe, D. A. (1989). Body composition and energy intake: Do overweight women overeat and underreport? *American Journal of Clinical Nutrition*, 49(2):320–325. 22, 23, 24, 82

[89] Lumley, C. E., Walker, S. R., and Hall, G. C. (1986). The under-reporting of adverse drug reactions seen in general practice. *Pharmaceutical Medicine*, 1(3):205–212. 21, 24, 82

[90] Mack, T. (1993). Distribution-free Calculation of the Standard Error of Chain Ladder Reserve Estimates. *ASTIN Bulletin*, 23(2):213–225. 5, 13

[91] Maistrello, I., Morgutti, M., Maltempi, M., and Dantes, M. (1995). Adverse drug reactions in hospitalized patients: An operational procedure to improve reporting and investigate underreporting. *Pharmacoepidemiology and Drug Safety*, 4(2):101–106. 21, 24, 82

[92] Mallows, C. L. and Kullback, S. (1959). Information Theory and Statistics. *Journal of the Royal Statistical Society. Series A (General)*, 122(3):380. 44

[93] Mandiant Inc. (2022). M-Trends 2022: Mandiant Special Report. Technical report. 42

[94] Martin, R. M., Kapoor, K. V., Wilton, L. V., and Mann, R. D. (1998). Underreporting of suspected adverse drug reactions to newly marketed ('black triangle') drugs in general practice: Observational study. *British Medical Journal*, 317(7151):119–120. 21, 24, 82

[95] McMurdie, C. (2016). The cybercrime landscape and our policing response. *Journal of Cyber Policy*, 1(1):85–93. 2, 20, 81

[96] Midthune, D. N., Fay, M. P., Clegg, L. X., and Feuer, E. J. (2005). Modeling reporting delays and reporting corrections in cancer registry data. *Journal of the American Statistical Association*, 100(469):61–70. 12, 16, 18, 36

[97] Milner, L. (2015). Managing Cyber Risks and Budgets. Technical report. 1

[98] Mittmann, N., Knowles, S. R., Gomez, M., Fish, J. S., Cartotto, R., and Shear, N. H. (2004). Evaluation of the extent of under-reporting of serious adverse drug reactions: The case of toxic epidermal necrolysis. *Drug Safety*, 27(7):477–487. 21, 24, 82

[99] Montastruc, P., Damase-Michel, C., Lapeyre-Mestre, M., Puget, C., Damase, L., Hurstel, J. F., Graille, V., and Montastruc, J. L. (1995). A Prospective Intensive Study of Adverse Drug Reactions in Urban General Practice. *Clinical Drug Investigation*, 10(2):117–122. 21, 24, 82

[100] Morgan, W. M. and Curran, J. W. (1986). Acquired immunodeficiency syndrome: Current and future trends. *Public Health Reports*, 101(5):459–465. 7, 17, 36, 61

[101] Moride, Y., Haramburu, F., Requejo, A. A., and Bégaud, B. (1997). Under-reporting of adverse drug reactions in general practice. *British Journal of Clinical Pharmacology*, 43(2):177–181. 21, 24, 82

[102] Muratoglu, O., Okul, ., Aydm, M. A., and Bilge, H. S. (2018). Review on Cyber Risks Relating to Security Management in Smart Cars. *UBMK 2018 - 3rd International Conference on Computer Science and Engineering*, pages 406–409. 1

[103] Noufaily, A., Farrington, P., Garthwaite, P., Enki, D. G., Andrews, N., and Charlett, A. (2016). Detection of Infectious Disease Outbreaks From Laboratory Data With Reporting Delays. *Journal of the American Statistical Association*, 111(514):488–499. 15, 19, 36

[104] Noufaily, A., Ghebremichael-Weldeselassie, Y., Enki, D. G., Garthwaite, P., Andrews, N., Charlett, A., and Farrington, P. (2015). Modelling reporting delays for outbreak

detection in infectious disease data. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 178(1):205–222. 14, 19, 36

[105] Obama, B. (2015). Remarks by the President at the Cybersecurity and Consumer Protection Summit. Technical report, The White House, The Office of the Press Secretary. 1

[106] Pishro-Nik, H. (2014). Introduction to Probability, Statistics, and Random Processes. *Kappa Research LLC*, pages 1–744. 9

[107] Pouyanne, P., Haramburu, F., Imbs, J. L., and Bégaud, B. (2000). Admissions to hospital caused by adverse drug reactions: Cross sectional incidence study. *British Medical Journal*, 320(7241):1036. 21, 24, 82

[108] Prevots, D. R., Sutter, R. W., Strebel, P. M., Weibel, R. E., and Cochi, S. L. (1994). Completeness of Reporting for Paralytic Poliomyelitis, United States, 1980 Through 1991: Implications for Estimating the Risk of Vaccine-Associated Disease. *Archives of Pediatrics & Adolescent Medicine*, 148(5):479–485. 21, 24, 82

[109] Pumphrey, R. S. and Davis, S. (1999). Under-reporting of antibiotic anaphylaxis may put patients at risk. *Lancet*, 353(9159):1157–1158. 21, 24, 82

[110] Rawlins, M. (1988). Spontaneous reporting of adverse drug reactions. I: the data. *British Journal of Clinical Pharmacology*, 26(1):1–5. 21

[111] Renshaw, A. and Verrall, R. (1998). A Stochastic Model Underlying the Chain-Ladder Technique. *British Actuarial Journal*, 4(4):903–923. 13

[112] Rios, L. M. and Sahinidis, N. V. (2013). Derivative-free optimization: A review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293. 48

[113] Romanosky, S. (2016). Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, 2(2):121–135. 81

[114] Rosenberg, P. S. (1990). A simple correction of AIDS surveillance data for reporting delays. *Journal of Acquired Immune Deficiency Syndromes*, 3(1):49–54. 9, 10, 18, 35, 61

[115] Rosinska, M., Pantazis, N., Janiec, J., Pharris, A., Amato-Gauci, A. J., Quinten, C., Schmid, D., Sasse, A., van Beckhoven, D., Varleva, T., Blazic, T. N., Hadjihannas, L., Koliou, M., Maly, M., Cowan, S., Rüütel, K., Liitsola, K., Salminen, M., Cazein, F., Pillonel, J., Lot, F., Gunsenheimer-Bartmeyer, B., Nikolopoulos, G., Paraskeva, D., Dudas, M., Briem, H., Sigmundsdottir, G., Igoe, D., O'Donnell, K., O'Flanagan, D., Suligoi, B., Konova, Š., Erne, S., Čaplinskien, I., Schmit, A. F. J. C., Melillo, J. M., Melillo, T., de Coul, E. O., van Sighem, A., Blystad, H., Rosinska, M., Aldir, I., Martins, H. C., Mardarescu, M., Truska, P., Klavs, I., Diaz, A., Axelsson, M., and Delpech, V. (2018). Potential adjustment methodology for missing data and reporting delay in the HIV surveillance system, European Union/European Economic Area, 2015. *Eurosurveillance*, 23(23). 61, 64

[116] Salmon, M., Schumacher, D., Stark, K., and Höhle, M. (2015). Bayesian outbreak detection in the presence of reporting delays. *Biometrical Journal*, 57(6):1051–1067. 13

[117] Samantha Dowling (2013). Cyber crime: a review of the evidence. Technical report, Home Office Science. 2, 20, 81

[118] Samuelsson, E., Hägg, S., Bäckström, M., Granberg, K., and Mjörndal, T. (1996). Thrombosis caused by oracl contraceptives. Underreporting to the adverse effects registry. *Läkartidningen*, 93(37):3117–8, 3121. 21, 24, 82

[119] Sangari, S. and Dallal, E. (2021). Correcting for Reporting Delays in Cyber Incidents. In *JSM Proceedings, Risk Analysis Section. Alexandria, VA: American Statistical Association*, pages 721–735. vi, 62

[120] Sangari, S., Dallal, E., and Whitman, M. (2022a). Modeling reporting delays in cyber incidents: an industry-level comparison. *International Journal of Information Security*. vii

[121] Sangari, S., Dallal, E., and Whitman, M. (2022b). Modeling Under-Reporting in Cyber Incidents. *Risks*, 10(11):200. vii

[122] Schuitemaker, N., Van Roosmalen, J., Dekker, G., Van Dongen, P., Van Geijn, H., and Gravenhorst, J. B. (1997). Underreporting of maternal mortality in the Netherlands.

Obstetrics and Gynecology, 90(1):78–82. 20, 21, 81, 82

[123] Smith, C. C., Bennett, P. M., Pearce, H. M., Harrison, P. I., Reynolds, D. J., Aronson, J. K., and Grahame-Smith, D. G. (1996). Adverse drug reactions in a hospital general medical unit meriting notification to the Committee on Safety of Medicines. *British Journal of Clinical Pharmacology*, 42(4):423–429. 21, 24, 82

[124] Stratton, S. J. (2021). Population Research: Convenience Sampling Strategies. *Prehospital and Disaster Medicine*, 36(4):373–374. 81

[125] Swinhoe, D. (2019). Why businesses don't report cybercrimes to law enforcement. 2, 3, 20, 81

[126] Tidy, J. (2020). Marriott Hotels fined £18.4m for data breach that hit millions - BBC News. 1, 35

[127] Torelló Iserte, J., Castillo Ferrando, J. R., Laínez, M. M., García Morillas, M., and Arias González, A. (1994). Adverse reactions to drugs reported by the primary care physicians of Andalusia. Analysis of underreporting. *Atencion primaria / Sociedad Española de Medicina de Familia y Comunitaria*, 13(6):307–311. 21

[128] Touhill, G. (2019). New Study Reveals Cybercrime May Be Widely Underreported Even When Laws Mandate Disclosure. 2

[129] Tukey, J. W. (1960). A Survey of Sampling from Contaminated Distributions. *Contributions to Probability and Statistics Essays in Honor of Harold Hotelling*, 2:448–485. 37

[130] Wang, M.-C. (1992). The Analysis of Retrospectively Ascertained Data in the Presence of Reporting Delays. *Journal of the American Statistical Association*, 87(418):397. 15, 18, 25, 36, 62

[131] Watson, W. T. (2020). Cyber claims analysis: Turning data into insight. Technical report, Willis Towers Watson. 35

[132] Weinberger, D. M., Chen, J., Cohen, T., Crawford, F. W., Mostashari, F., Olson, D., Pitzer, V. E., Reich, N. G., Russi, M., Simonsen, L., Watkins, A., and Viboud, C.

(2020). Estimation of Excess Deaths Associated with the COVID-19 Pandemic in the United States, March to May 2020. *JAMA Internal Medicine*, 180(10):1336–1344. 13, 19, 36, 62

[133] Weirich, P. (2015). Separability. In *Models of Decision Making*, chapter 1, pages 23–50. Cambridge University Press, first edition. 84

[134] White, L. F., Wallinga, J., Finelli, L., Reed, C., Riley, S., Lipsitch, M., and Pagano, M. (2009). Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza and other Respiratory Viruses*, 3(6):267–276. 13, 19, 36, 62

[135] Wood, J. S., Donnell, E. T., and Fariss, C. J. (2016). A method to account for and estimate underreporting in crash frequency research. *Accident Analysis and Prevention*, 95:57–66. 20, 22, 24, 26, 82

[136] World Economic Forum (2021). The Global Risks Report 2021: 16th Edition. Technical report. 1

[137] Zhao, X. B. and Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance: Mathematics and Economics*, 46(2):290–299. 16, 36

[138] Zhao, X. B., Zhou, X., and Wang, J. L. (2009). Semiparametric model for prediction of individual claim loss reserving. *Insurance: Mathematics and Economics*, 45(1):1–8. 16, 36