Correcting for Reporting Delays in Cyber Incidents

Seema Sangari* Dr. Eric Dallal†

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 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.

Key Words: Delay Distribution, Optimization, Reporting Delays

1. Introduction

The cyber security domain is evolving rapidly, with new attack vectors emerging regularly, and 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. 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. Marriott's major cyber incident occurred in 2014 but was only reported in 2018. Major cyber events become headlines in leading newspapers when reported publicly rather than when they happened. 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 the recent few years. Coleman et al. (2021) raised the concern that cyber incidents would remain undetected due to advanced threat techniques.

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, 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.

Terminology followed in this paper:
i: Incident
I: Set of all incidents

*School of Data Science and Analytics, Kennesaw State University, 3391 Town Point Dr. NW,Kennesaw, GA 30144

†AIR Worldwide, Verisk Cyber Solutions, Lafayette City Center, 2 Avenue de Lafayette, 2nd Floor, Boston, MA 02111

$delay, \delta_i$: Time between the incident date and the reporting date
age, A_i: Time between the incident date and the last incident reporting date in the data
f_Δ: Probability density function of the $delay$ distribution
F_Δ: Cumulative distribution function of the $delay$ distribution

2. Literature Review

Most of the literature on reporting delays addresses the medical space. We are not aware of any existing papers that propose a methodology for correcting cyber incident counts to account for reporting delays. However, Coleman et al. (2021) does examine both the distribution of the number of days to discover cyber incidents and the number of days to disclose them.

Harris (1987) described reporting delays as a statistical problem for the first time. Heisterkamp et al. (1988a,b, 1989) and Brookmeyer and Damiano (1989) made distributional assumptions and built linear/quadratic models whereas Rosenberg (1990) and Cheng and Ford (1991) suggested Poisson models. These are easy to implement but assume stationary reporting delays and don't capture trends. Morgan and Curran (1986), Downs et al. (1987, 1988)), Healy and Tillett (1988) and Heisterkamp et al. (1988a,b, 1989) fitted exponential, integrated logistic and log-linear models to capture trends. Gail and Brookmeyer (1988), Brookmeyer and Liao (1990), Kalbfleisch and Lawless (1991) and Esbjerg et al. (1999) applied conditional probabilities to capture trends but this resulted in over-fitting. Lawless (1994) 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.

Wang (1992) suggested maximum likelihood estimation (MLE) based non-parametric and semi-parametric approaches but with complete[1] data. Harris (2020) suggested correcting COVID cases with an expectation-maximization (EM) algorithm and trained the model with complete data to correct test data. White et al. (2009) and Weinberger et al. (2020) proposed a simpler method based on proportions but also require complete data to train.

Wang (1992), Keiding and Moeschberger (1992) and Midthune et al. (2005) applied truncated models to avoid random effects but require stable reporting delays.

Höhle and An Der Heiden (2014), Bastos et al. (2019) and Chitwood et al. (2020) suggested a Bayesian and hierarchical approach with Poisson and negative binomial distributions. It is easy to implement but makes distributional assumptions. Noufaily et al. (2015, 2016)) suggested a log-likelihood approach with a truncation model that is data driven but sensitive to the choice of three reporting time steps.

Bastos et al. (2019) suggested a chain-ladder approach but it is sensitive to outliers.

Jewell (1989), Zhao et al. (2009), Zhao and Zhou (2010), and Avanzi et al. (2016) investigated cyber claims data to account for reporting delays from a capital reserving perspective. This problem is different from the one being investigated, since reporting delays in claims are due only to detection delays.

As stated in Brookmeyer and Liao (1990), none of these approaches deal with delays longer than any previously observed.

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

3. Data

The data used is a proprietary data set constructed by merging multiple source cyber event sets together. The source cyber event sets contained differing fields that, at minimum, provided information as to the "what", "when" and "who" of an event. Concretely, the data sets included a description of the event, the name of the company to which the cyber event occurred (N.B.: aggregation events separately list each company known to be impacted), along with the date that the event occurred and the date that the event was reported.

Because company names frequently differ from one data set to another, the data sets could not be de-duplicated based on direct string matching. Instead, the event data sets were matched to a firmographic data set containing approximately 55 million businesses in the US. This was done via a previously developed matching algorithm that examined company name, industry classification (e.g., via NAICS codes), address information, and any other fields common to both the cyber event data set and the firmographic data set. Events in distinct cyber event data sets were identified as the same when:

1. They were matched to the same company in the firmographic data set; and
2. They were listed as having occurred within 1 week of each other.

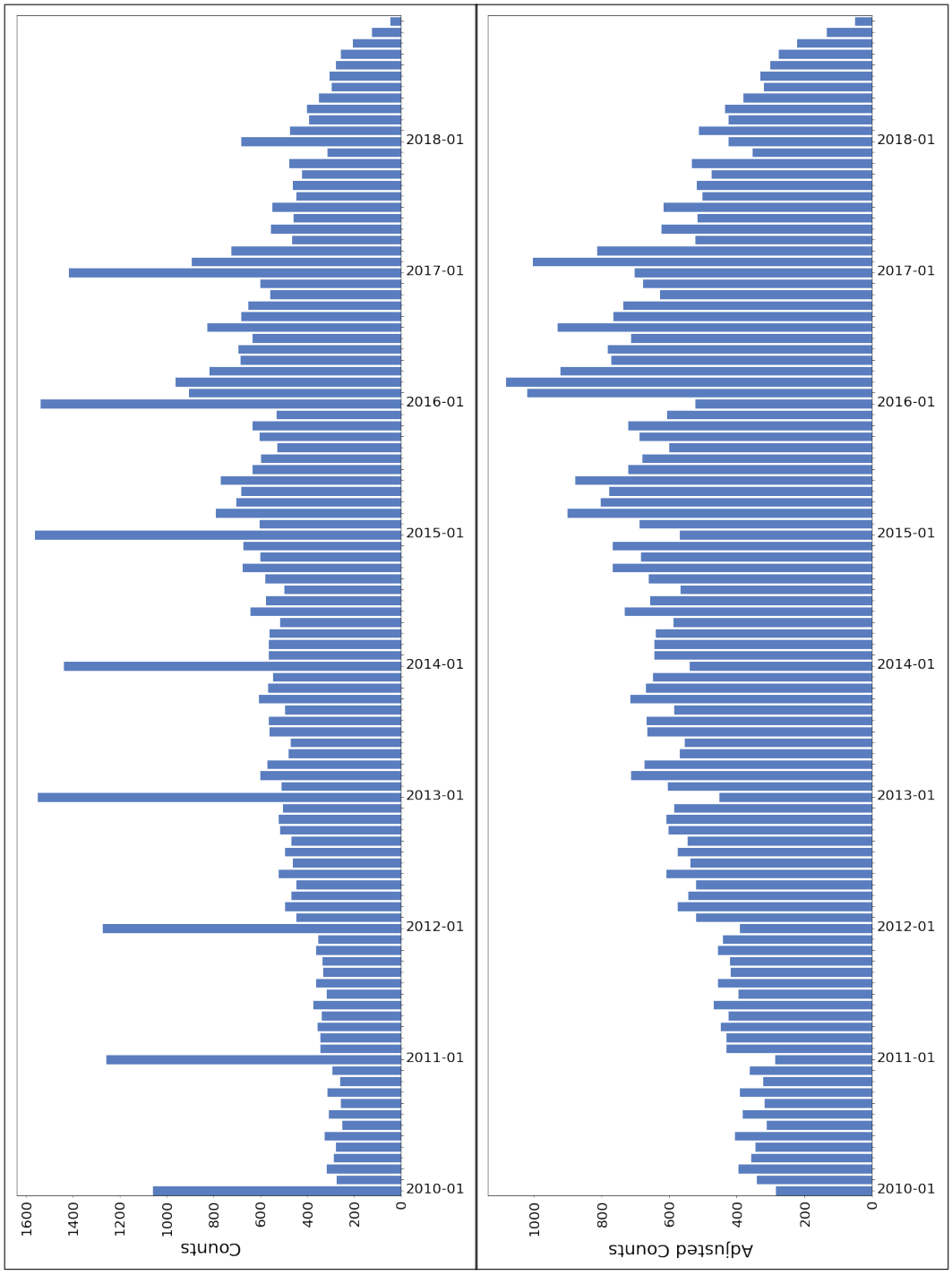
Limitations: Some events do not have an occurrence date listed, and were therefore excluded from this analysis. There are also large spikes in event counts listed as having occurred on January 1st in most years (Fig. 1a), which was assumed to be a default value when only the year of the event was known. These events were therefore re-distributed proportionally throughout the year as shown in Fig. 1b but excluded while developing the approach.

4. Proposed Approach

The proposed approach consists of estimating the reporting delay distribution f_Δ from the empirical data. With this distribution, a corrected count of events with age a can be determined by dividing the raw counts by $F_\Delta(a)$, as the latter expression represents the proportion of events that are reported within a delay, δ, of less than a or, equivalently, the proportion of events that are reported as of the present. There are four complications with estimating the delay distribution in the "obvious" way:

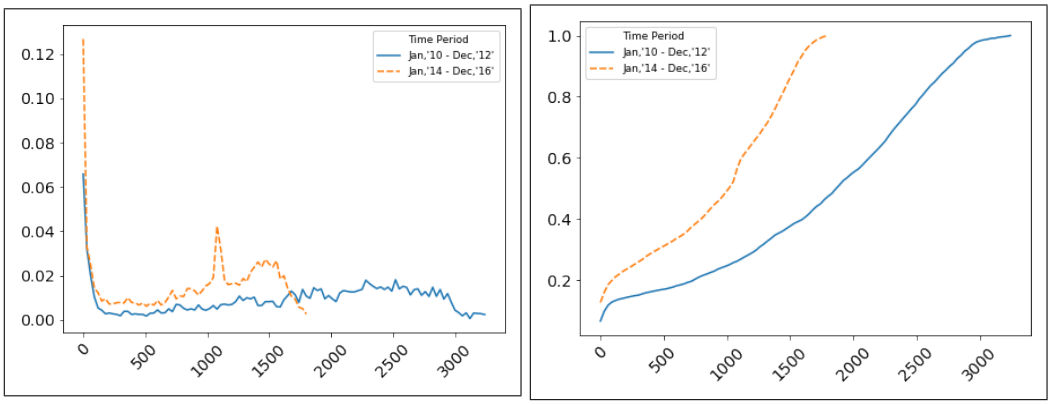
- The nature of reporting delays means that direct estimates from empirical data will be biased towards shorter delays, since recent events could only appear in the data set in the first place if the reporting delay is small.
- The estimates assume zero probability of any delay longer than the longest in the data set.
- The estimates for long reporting delays are based on few data points.
- The reporting delay distribution may not be stationary.

Mathematically, the first problem is that, for incidents with age a, the delay δ_i can only be observed if $\delta_i <= a$. This problem can be resolved if one assumes that f_Δ is stationary, in which case the distribution can be estimated for delays, δ, larger than a from older events. This is the approach taken in Algorithm 1. To test for stationary, the algorithm was run on two year windows of the data set centered at different times. (Note that the two year windows include all events that *occur* within the window, irrespective of when they are reported.) This showed that the delay distribution is non-stationary, as can be seen in Fig. 2.



(a) With Default Date (b) Adjusted for Default Date

Figure 1: Cyber Event Counts until December 2018



(a) Probability Distribution (b) Cumulative Distribution

Figure 2: PDF and CDF of Delay Distribution generated for Dec.,'12 and Dec.,'16

To deal with this non-stationarity, the delay distribution was modeled via parametric distributions where parameters were estimated over monthly rolling two year windows. For each two year window, optimal parameters were determined based on matching cumulative distributions. Specifically, Algorithm 1 was run on each of the monthly two year windows to obtain an estimate of the delay distribution for that time window. For a given two year window and modeled distribution, an optimization algorithm was used to determine the parameters that gave the best match between:

- the delay distribution obtained by Algorithm 1; and
- the parametric distribution restricted to the domain where the maximum delay is less than the window's maximum reporting delay.

The optimal distribution parameters for each window were computed using a derivative free optimizer.

In what follows, the above two distributions will be referred to as the "debiased empirical delay distribution" and the "modeled delay distribution", respectively.

4.1 Generating the Debiased Empirical Delay Distribution

Inspired from Brookmeyer and Liao (1990), the proposed algorithm works on the limitation that the delay, δ, is not considered beyond age and hence only a distribution conditioned on the delay being less than or equal to age can be estimated. The algorithm corrects the incident counts with 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 been reported as the distribution itself is 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,

$$h_A(a) = |\{i \in \mathcal{I} : A_i = a\}| \quad (1)$$

$$h_\Delta(\delta) = |\{i \in \mathcal{I} : \Delta_i = \delta\}| \quad (2)$$

Then the equation to estimate the delay distribution is:

$$f_\Delta(\delta) = \frac{h_\Delta(\delta)}{\sum_{a=\delta}^{A_{\max}} h_A(a)/F_\Delta(a)} \tag{3}$$

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., where the age is larger than the delay under consideration), $\sum_{a=\delta}^{A_{\max}} h_A(a)/F_\Delta(a)$. Algorithm 1 shows the algorithmic implementation to generate the distribution.

Algorithm 1 Algorithm for computing the debiased empirical delay distribution

Input: The histograms, h_A and h_Δ, computed as in Eqs. (1) and (2), 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

4.2 Generating the Modeled Delay Distribution

As previously described, the modeled delay distribution is determined by an optimization that minimizes the difference in the cumulative distribution functions of the debiased empirical delay distribution and the modeled delay distribution, restricted to the domain $[0, \delta_{\max}]$ (N.B.: $\delta_{\max}$ is the maximum observed delay). The PDF and CDF plots shown in Fig. 2 suggest that a single distribution will not provide a good fit, as the delay distribution is bi-modal. Rather, a mixture of distributions would be required. Fig. 2 suggests a mix of exponential and normal distributions.

Mathematically,

$$F_\theta(\delta) = \alpha F_{Exp}(\delta, Scale) + (1 - \alpha) F_N(\delta, \mu, \sigma) \tag{4}$$
$$\text{where } F_{Exp} = \text{Exponential CDF with parameter, } Scale = 1/\lambda$$
$$F_N = \text{Normal CDF with parameters, } \mu \text{ and } \sigma$$

A possible interpretation of the bi-modal nature of the reporting delay distribution is that there are two distributions: one for events that are discovered almost immediately (modeled

by the exponential) and one for events which have a delay due to both discovery time and public disclosure time (modeled by the normal). The parameter α can therefore be interpreted as the proportion of events that are discovered right away by the organization.

Since the normal distribution is defined on $(-\infty, \infty)$ and reporting delays cannot be negative, the modeled delay distribution needs to be adjusted. The CDF in Eq. 4 truncated to $[0, \infty)$ (and re-normalized) can be expressed as

$$F_\theta(\delta) = \frac{\alpha(F_{Exp}(\delta, Scale)) + (1-\alpha)\overbrace{(F_N(\delta, \mu, \sigma) - F_N(0, \mu, \sigma)}^{\text{Truncated Normal Distribution until } \delta}}{\alpha + (1-\alpha)\underbrace{(1 - F_N(0, \mu, \sigma)}_{\text{Truncated Normal Distribution over } [0, \infty)}} \tag{5}$$

where $0 \leq \delta \leq \infty$

Since the debiased empirical delay distribution is only defined on $[0, \delta_{\max}]$, which is the domain on which it is compared to the modeled delay distribution, the truncation of the modeled distribution to this domain is also defined, denoted by F'_θ as below.

$$F'_\theta(\delta) = \frac{\alpha(F_{Exp}(\delta, Scale)) + (1-\alpha)\overbrace{(F_N(\delta, \mu, \sigma) - F_N(0, \mu, \sigma)}^{\text{Truncated Normal Distribution until } \delta}}{\alpha(F_{Exp}(\delta_{\max}, Scale)) + (1-\alpha)\underbrace{(F_N(\delta_{\max}, \mu, \sigma) - F_N(0, \mu, \sigma)}_{\text{Truncated Normal Distribution over } [0, \delta_{\max}]}} \tag{6}$$

where $0 \leq \delta \leq \delta_{\max}$

4.3 Defining the Optimization Function

Defining the optimization function was challenging due to two factors in particular:

- There are many combinations of parameters that give approximately the same distribution when restricted to the domain $[0, \delta_{\max}]$, but which differ substantially in how much of the total distribution's weight is contained in this domain.
- As two year windows closer to the present are considered, the quantity of data shrinks, resulting in increasingly unstable parameter estimates.

The optimization function used is shown in Eq. 7 below. The first term[2] $\| \log_{10} F'_\theta - \log_{10} F_\Delta \|^2$ reduces the CDF difference between the debiased empirical delay distribution and the modeled delay distribution over the domain $[0, \delta_{\max}]$. The purpose of applying $\log_{10}$ weights is to place more emphasis on a good fit for the initial months[3]. As mentioned above, there are many different combinations of parameters that would result in comparable errors in the first optimization term, but which nevertheless differ substantially over the domain $[0, \infty)$. This problem is relatively minor when the range $[0, \delta_{\max}]$ contains the bulk of the distribution, which it does when $\delta_{\max}$ is substantially greater than the second peak of the debiased empirical delay distribution. But as two year windows closer to the present are considered, this ceases to be the case. In order to avoid this problem, the optimization function must consider modeled delay distribution values beyond $\delta_{\max}$. The second term $\| \log_{10} S_\theta - \log_{10} S_{\theta'} \|^2$ penalizes large differences between the CDFs of consecutive

[2] F'_θ computed as defined in Eq. 6

[3] The rationale behind $\log_{10}$, is to obtain CDF values close from the point of the ratio between the two distributions (modeled and debiased empirical), not in terms of absolute difference - a $\log_{10}$ CDF difference between 0.03 and 0.06 would be a factor of two difference in the correction whereas a difference between 0.93 and 0.96 would be much smaller despite the fact that the absolute difference is 0.03 in both cases.

modeled distributions beyond $\delta_{\max}$. In order to further minimize parameter instability for recent two year windows, another set of weights is assigned to the first two terms, which effectively diminishes the importance of a good fit with the empirical data and increases the importance of parameter stability as the amount of available data diminishes.

The third and fourth terms are penalization terms - $F_N^2(0, \mu, \sigma)$ term penalizes negative delays introduced by the normal distribution (defined over $(-\infty, +\infty)$) whereas $S_\theta^2(10Y)$ term penalizes delays beyond 10 years.

Mathematically, the optimization function is defined as

$$\begin{aligned}\theta_{Opt} = \underset{\theta=(\alpha,Scale,\mu,\sigma)}{argmin} & \frac{\delta_{\max}}{\delta_{Fix}} \underbrace{\| \log_{10} F'_\theta - \log_{10} F_\Delta \|^2}_{\delta \in [0, \delta_{\max}]} \\ & + \left(1 - \frac{\delta_{\max}}{\delta_{Fix}}\right) \underbrace{\| \log_{10} S_{\theta'} - \log_{10} S_\theta \|^2}_{\delta \in (\delta_{\max}, \delta_{Fix}]} \\ & + \underbrace{F_N^2(0, \mu, \sigma)}_{\delta < 0} + \underbrace{S_\theta^2(10Y)}_{\delta > 10Years}\end{aligned} \tag{7}$$

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

F'_θ is as defined in Eq. 6 whereas S_θ is the survival function, defined as the complement of F_θ of Eq. 5:

$$S_\theta = 1 - F_\theta \tag{8}$$

Finally, θ' refers to the previous two year window's optimal parameters. Since no previous parameters are available for the first two year window, the second term is taken to be zero for that window.

$$\| \log_{10} S_{\theta'} - \log_{10} S_\theta \|^2 = 0 \tag{9}$$

θ' refers to optimal parameters at previous step.

The covariance matrix adaptation evolution strategy (CMA-ES) was applied to compute the modeled distribution parameters. It is a derivative free optimization algorithm, a type typically used when derivatives are difficult or costly to compute (Hansen (2006, 2016, 2019)).

4.4 Computing the Corrected Counts

Once the modeled delay distributions (one for each two year window) have been obtained by optimization, event counts can be corrected based on the cumulative distribution function of the full modeled distribution (i.e., defined over $[0, \infty)$) defined by Eq. 5.

Specifically,

$$\text{Corrected Count} = \frac{\text{Reported Counts for month, 'm'}}{F_\theta(a)} \tag{10}$$

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

From each modeled distribution, the correction factor is computed at only one point, age, of the modeled distribution for the given month's correction.

5. Results

Fig. 3 shows two examples of plots comparing PDFs of the fitted parametric modeled distribution, its truncation to the domain $[0, \delta_{\max}]$ — where δ_{max} is computed for each window individually, and the debiased empirical delay distribution. Fig. 3a shows this comparison for the two year window starting from July 2012 until June 2016 and Fig. 3b shows this comparison for the most recent window starting from January, 2017 until December 2018.

5.1 Parameters and Interpretation

Fig. 4 shows the parameter plots of the delay distribution generated for each monthly two year rolling window.

The alpha plot (Fig. 4a) suggests that organizations discover 8-18% of cyber events right away ($8\% \leq \alpha \leq 18\%$).

The scale plot (Fig. 4b) 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. 4c) and standard deviation, σ, (Fig. 4d) 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 $\pm 10\%$ range.

5.2 Corrections and Validation

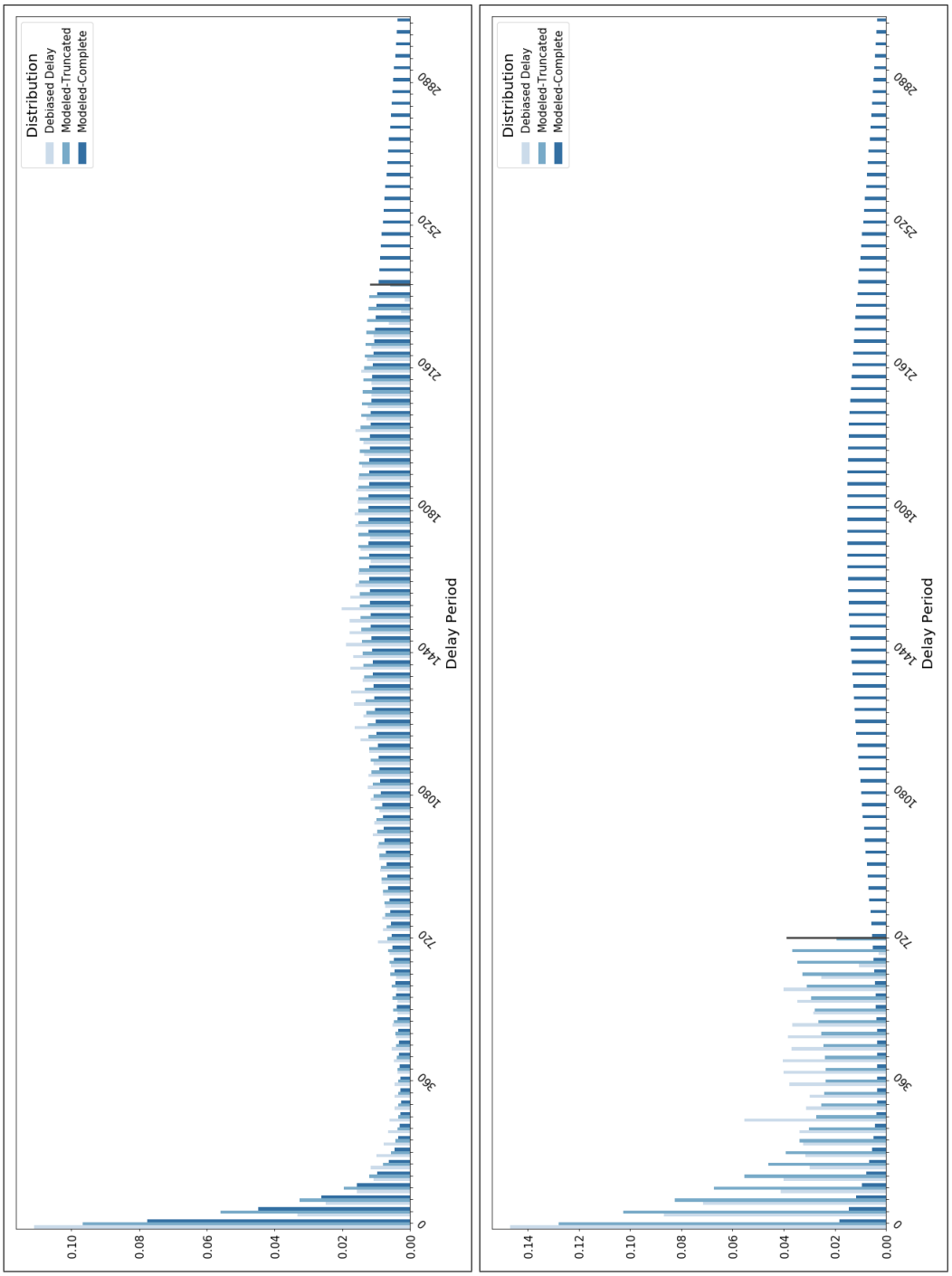
Fig. 5 shows the corrected incident counts based on the proposed methodology. Figs. 5a and 5b show the corrected counts for the events reported by Dec. 2017 and by Dec. 2018, respectively. Although the corrections follow similar trends in both, the correction factors vary substantially.

To validate the proposed algorithm, the counts reported until December 2017 (2018) were corrected for a year ahead and compared against the counts reported as of December 2018 (2019). The year ahead correction factor is computed as

$$\text{Year ahead } F_\theta(a, a + 1\,Year) = \frac{F_\theta(a)}{F_\theta(a + 1\,Year)} \tag{11}$$

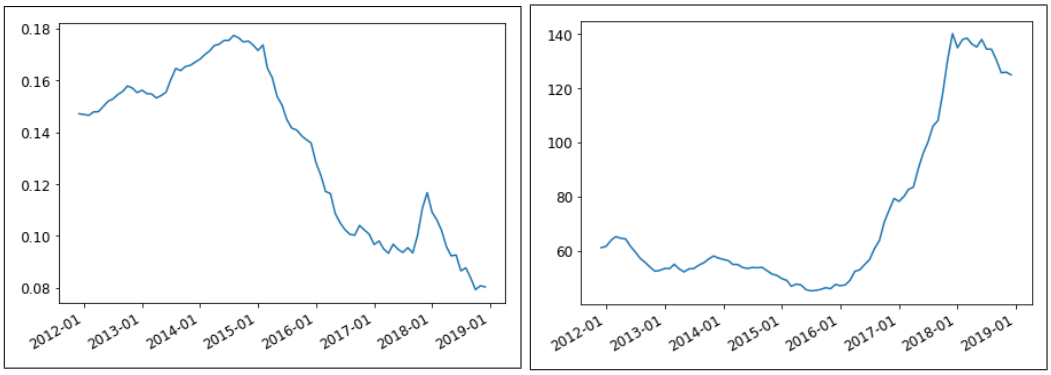
where a is the age of the event counts being corrected.

Whereas the 2017 year ahead corrections (Fig. 5a) 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. 5b) generally overestimate the 2019 counts, except for the most recent months, which show close agreement. As stated in section 4.3, 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. 5a or Fig. 5b, 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 counts is therefore much more in line with reports from insurers and other organizations that release reports on cyber risk.



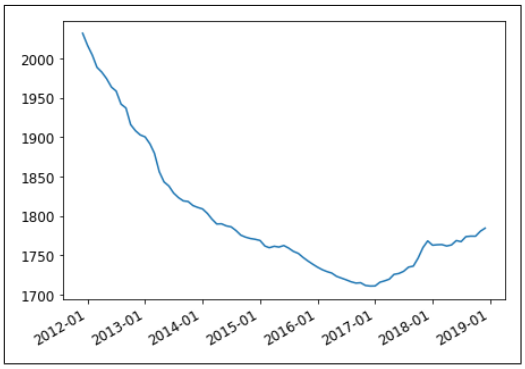
(a) From July 2012 to June, 2014 (b) From January, 2017 to December, 2018

Figure 3: Comparing PDFs of Debiased Delay Distribution with Parametric Modeled Distribution

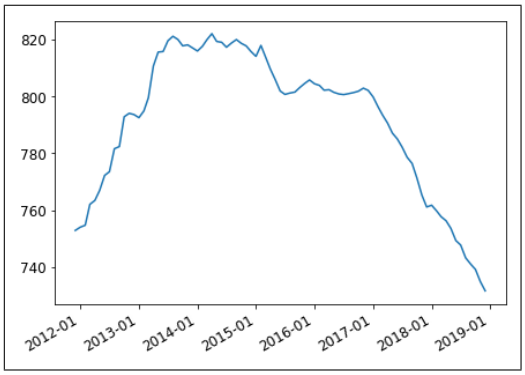


(a) Alpha Parameter

(b) Scale Parameter



(c) Mean Parameter



(d) Sigma Parameter

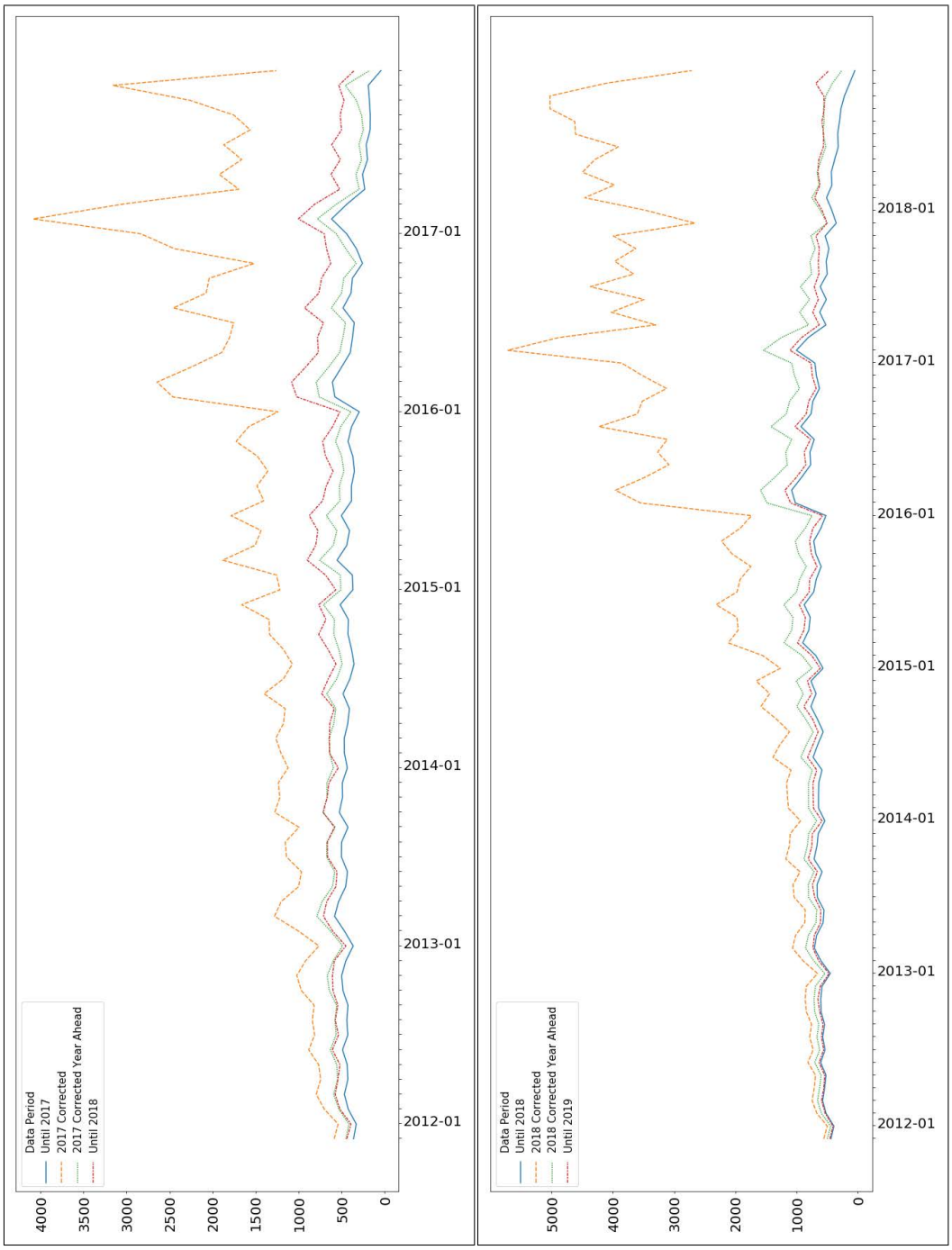
Figure 4: Plots of Modeled Distribution Parameters based on Empirical Debiased Delay Distribution

6. Conclusion

This work examined the long known problem of reporting delays in historical cyber events databases and proposed an algorithm to correct for 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.

The general approach can be applied to any form of data with reporting delays including other long tailed insurance claims (such as liability), and COVID-19 pandemic data. The current study corrects the overall counts for US cyber events, and the industry specific correction of cyber event counts is a future direction of this work.



(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. 10
201X Corrected Year Ahead - "Counts until 201X" corrected based on Eq. 11

Figure 5: Validation Plots

Acknowledgment

The study was done in collaboration with AIR Worldwide using their proprietary cyber data. The authors would like to thank Scott Stransky for his comments and suggestions that helped improve the approach.

REFERENCES

Benjamin Avanzi, Bernard Wong, and Xinda Yang. A micro-level claim count model with overdispersion and reporting delays. *Insurance: Mathematics and Economics*, 71:1–14, 2016. ISSN 01676687. doi: 10.1016/j.insmatheco.2016.07.002.

Leonardo S. Bastos, Theodoros Economou, Marcelo F.C. Gomes, Daniel A.M. Villela, Flavio C. Coelho, Oswaldo G. Cruz, Oliver Stoner, Trevor Bailey, and Claudia T. Codeço. A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine*, 38(22):4363–4377, 2019. ISSN 10970258. doi: 10.1002/sim.8303.

Ron Brookmeyer and Anne Damiano. Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine*, 8(1):23–34, 1989. ISSN 10970258. doi: 10.1002/sim.4780080105.

Ron Brookmeyer and Mitchell H. Gail. Minimum Size of the Acquired Immunodeficiency Syndrome (Aids) Epidemic in the United States. *The Lancet*, 328(8519):1320–1322, 1986. ISSN 01406736. doi: 10.1016/S0140-6736(86)91444-3.

Ron Brookmeyer and Jiangang Liao. The analysis of delays in disease reporting: Methods and results for the acquired immunodeficiency syndrome. *American Journal of Epidemiology*, 132(2):355–365, 1990. ISSN 00029262. doi: 10.1093/oxfordjournals.aje.a115665.

Fenny F.K. Cheng and Wesley L. Ford. Adjustment of aids surveillance data for reporting delay to the editor:, 1991. ISSN 10779450.

Melanie H Chitwood, Marcus Russi, Kenneth Gunasekera, Joshua Havumaki, Virginia E Pitzer, Joshua L Warren, Daniel Weinberger, Ted Cohen, and Nicolas A Menzies. Bayesian nowcasting with adjustment for delayed and incomplete reporting to estimate COVID-19 infections in the United States. *medRxiv*, page 2020.06.17.20133983, 2020. doi: 10.1101/2020.06.17.20133983. URL `http://medrxiv.org/content/early/2020/06/20/2020.06.17.20133983.abstract`.

Derryck Coleman, Madeleine Conley, and Nicole Hallas. Trends in Cybersecurity Breaches. Technical report, Audit Analytics, Massachusetts, USA, 2021. URL `https://go.auditanalytics.com/cybersecurityreport`.

A. M. Downs, R. A. Ancelle, H. J.C. Jager, and J. B. Brunet. AIDS in Europe: Current trends and short-term predictions estimated from surveillance data, January 1981-June 1986. *Aids*, 1(1):53–57, 1987. ISSN 02699370.

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

Sille Esbjerg, Niels Keiding, and Nils Koch-Henriksen. Reporting delay and corrected incidence of multiple sclerosis. *Statistics in Medicine*, 1999. ISSN 02776715. doi: 10.1002/(SICI)1097-0258(19990715)18:13⟨1691::AID-SIM160⟩3.0.CO;2-D.

Mitchell H. Gail and Ron Brookmeyer. Methods for projecting course of acquired immunodeficiency syndrome epidemic, 1988. ISSN 00278874.

N Hansen. 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, 2006.

Nikolaus Hansen. The CMA Evolution Strategy: A Tutorial. *Computing Research Repository*, apr 2016. URL `http://arxiv.org/abs/1604.00772`.

Nikolaus Hansen. CMA - Python Package, 2019. URL `https://pypi.org/project/cma/`.

Jeffrey E Harris. Delay in Reporting Acquired Immune Deficiency Syndrome (AIDS). *National Bureau of Economic Research Working Paper Series*, No. 2278, 1987. URL https://www.nber.org/papers/w2278.

Jeffrey E Harris. Overcoming Reporting Delays Is Critical to Timely Epidemic Monitoring: The Case of COVID-19 in New York City. *medRxiv*, page 2020.08.02.20159418, 2020. doi: 10.1101/2020.08.02.20159418. URL https://doi.org/10.1101/2020.08.02.20159418.

M. J. R. Healy and H. E. Tillett. Short-Term Extrapolation of the AIDS Epidemic. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 151(1):50, 1988. ISSN 09641998. doi: 10.2307/2982184.

S. H. Heisterkamp, J. C. Jager, A. M. Downs, and J. A.M. Van Druten. The use of Genstat in the estimation of expected numbers of AIDS cases adjusted for reporting delays. In *Fifth Genstat Conference*, pages 4–18, 1988a.

S. H. Heisterkamp, J. C. Jager, A. M. Downs, J. A.M. Van Druten, and E. J. Ruitenberg. 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, 1988b.

S. H. Heisterkamp, J. C. Jager, E. J. Ruitenberg, J. A.M. Van Druten, and A. M. Downs. Correcting reported aids incidence: A statistical approach. *Statistics in Medicine*, 8(8):963–976, 1989. ISSN 10970258. doi: 10.1002/sim.4780080807.

Michael Höhle and Matthias An Der Heiden. Bayesian nowcasting during the STEC O104: H4 outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002, 2014. ISSN 15410420. doi: 10.1111/biom.12194.

WS S. Jewell. Predicting ibnyr events and delays. *ASTIN Bulletin*, 19(I):25–56, 1989. ISSN 17831350.

J D Kalbfleisch and J F Lawless. Regression Models for Right Truncated Data With Applications To Aids Incubation Times and Reporting Lags. *Statistica Sinica*, 1(1):19–32, 1991.

Niels Keiding and Melvin Moeschberger. Independent Delayed Entry. In *Survival Analysis: State of the Art*, pages 309–326. Springer Netherlands, 1992. doi: 10.1007/978-94-015-7983-4_18.

J. F. Lawless. 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, 1994. ISSN 03195724. doi: 10.2307/3315820.

Douglas N. Midthune, Michael P. Fay, Limin X. Clegg, and Eric J. Feuer. Modeling reporting delays and reporting corrections in cancer registry data. *Journal of the American Statistical Association*, 100(469):61–70, 2005. ISSN 01621459. doi: 10.1198/016214504000001899.

W. M. Morgan and J. W. Curran. Acquired immunodeficiency syndrome: Current and future trends. *Public Health Reports*, 101(5):459–465, 1986. ISSN 00333549.

Angela Noufaily, Yonas Ghebremichael-Weldeselassie, Doyo Gragn Enki, Paul Garthwaite, Nick Andrews, André Charlett, and Paddy Farrington. 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, 2015. ISSN 1467985X. doi: 10.1111/rssa.12055.

Angela Noufaily, Paddy Farrington, Paul Garthwaite, Doyo Gragn Enki, Nick Andrews, and Andre Charlett. Detection of Infectious Disease Outbreaks From Laboratory Data With Reporting Delays. *Journal of the American Statistical Association*, 111(514):488–499, 2016. ISSN 1537274X. doi: 10.1080/01621459.2015.1119047.

Philip S. Rosenberg. A simple correction of AIDS surveillance data for reporting delays. *Journal of Acquired Immune Deficiency Syndromes*, 3(1):49–54, 1990. ISSN 10779450.

Mei-Cheng Wang. The Analysis of Retrospectively Ascertained Data in the Presence of Reporting Delays. *Journal of the American Statistical Association*, 87(418):397, 1992. ISSN 01621459. doi: 10.2307/2290270.

Daniel M. Weinberger, Jenny Chen, Ted Cohen, Forrest W. Crawford, Farzad Mostashari, Don Olson, Virginia E. Pitzer, Nicholas G. Reich, Marcus Russi, Lone Simonsen, Anne Watkins, and Cecile Viboud. Estimation of Excess Deaths Associated with the COVID-19 Pandemic in the United States, March to May 2020. *JAMA Internal Medicine*, 06520(May):E1–E9, 2020. ISSN 21686114. doi: 10.1001/jamainternmed.2020.3391. URL https://sci-hub.tw/downloads-ii/2020-07-02/3c/10.1001@jamainternmed.2020.3391.pdf.

Laura Forsberg White, Jacco Wallinga, Lyn Finelli, Carrie Reed, Steven Riley, Marc Lipsitch, and Marcello Pagano. 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, nov 2009. ISSN 17502640. doi: 10.1111/j.1750-2659.2009.00106.x.

Xiao Bing Zhao and Xian Zhou. Applying copula models to individual claim loss reserving methods. *Insurance: Mathematics and Economics*, 46(2):290–299, 2010. ISSN 01676687. doi: 10.1016/j.insmatheco.2009.11.001.

Xiao Bing Zhao, Xian Zhou, and Jing Long Wang. Semiparametric model for prediction of individual claim loss reserving. *Insurance: Mathematics and Economics*, 45(1):1–8, 2009. ISSN 01676687. doi: 10.1016/j.insmatheco.2009.02.009.