Cyber Attacks Against Enterprise Networks: Characterization, Modeling and Forecasting

Zheyuan Sun[1], Maochao Xu[2], Kristin M. Schweitzer[3], Raymond M. Bateman[3], Alexander Kott[3], and Shouhuai Xu[4(✉)]

[1] Department of Computer Science, University of Texas at San Antonio, San Antonio, USA
[2] Department of Mathematics, Illinois State University, Normal, USA
[3] DEVCOM Army Research Laboratory, Adelphi, USA
[4] Department of Computer Science, University of Colorado Colorado Springs, Colorado Springs, USA
sxu@uccs.edu

Abstract. Cyber attacks are a major and routine threat to the modern society. This highlights the importance of forecasting (i.e., predicting) cyber attacks, just like weather forecasting in the real world. In this paper, we present a study on characterizing, modeling and forecasting the number of cyber attacks at an aggregate level by leveraging a high-quality, publicly-available dataset of cyber attacks against enterprise networks; the dataset is of high quality because more than 99% of the attacks were examined and confirmed by human analysts. We find that the attacks exhibit high volatilities and burstiness. These properties guide us to design statistical models to accurately forecast cyber attacksand draw useful insights.

Keywords: Cybersecurity data analytics · attack forecasting · attack prediction · burstiness · cyber threats · cybersecurity dynamics · statistical models

1 Introduction

The importance of forecasting cyber attacks (or attack events) is well appreciated because it can enable proactive cyber defense, similar to how weather forecasting helps us in planning our daily activities. For example, being ale to forecast cyber attacks against a network will give cyber defenders useful information in planning defenses [32]. Moreover, the forecasting capability allows the defender to dynamically adjust the allocation of defense resources [3,66,67], including both human analysts who need to examine the alerts triggered by defense tools [22] and sensor deployments when the prediction is geared toward specific type of attacks. Moreover, when the predicted number of attacks is high but the detected number of attacks is low, it hints that the defense may not be effective and/or the attacks may be new. Although there are studies on forecasting cyber attacks (e.g., [12,18,43,44,55,66,67]), these studies are limited primarily by the quality of the datasets they leverage because they are often collected from *non-enterprise networks* (e.g., cyber honeypots). This is not surprising because high-quality cyber attack data against production/enterprise networks is sensitive.

In this paper, we leverage a dataset which contains some aggregated information about cyber attacks against enterprise networks, rather than honeypots. The dataset

M. Yung et al. (Eds.): SciSec 2023, LNCS 14299, pp. 60–81, 2023.
https://doi.org/10.1007/978-3-031-45933-7_4

describes weekly-binned cyber attacks, but not the individual attacks. This dataset is of high quality in the sense that more than 99% of the attacks are confirmed by analysts.

Our Contributions. We make three contributions. First, we analyze three time series derived from the dataset: (i) the weekly-binned *number* of attacks, referred to as X_t; (ii) the weekly *average* attack report length, referred to as Y_t; and (iii) the weekly *total* attack report length, referred to as Z_t and derived from X_t and Y_t. Note that Y_t and Z_t are analyzed here for the first time. Second, we show that these time series exhibit high *volatilities* and *burstiness*, meaning that they cannot be modeled by simple stochastic process models (e.g., Poisson). Third, we show that these time series can be accurately modeled by an ARIMA+GARCH model, where ARIMA stands for "AutoRegressive Integrated Moving Average" and can model the dynamic mean, and GARCH stands for "Generalized AutoRegressive Conditional Heteroskedasticity" and can model the burstiness. Moreover, we show that the ARIMA+GARCH model can forecast the number of attacks one week ahead of time. These allow us to draw a number of useful insights, such as: (i) the average attack report length reflects attack sophistication but not attack newness; (ii) new attacks are not necessarily sophisticated; and (iii) burstiness in the average attack report length suggests that sophisticated attacks are seen often.

Related Work. The first study on forecasting cyber attacks based on real-world datasets may be [12], which leverages a dataset collected at a campus network from 6/14/2001 to 3/14/2007. The study *heuristically* uses the ARIMA model without giving statistical justification. Another study [18] forecasts distributed denial-of-service attack rate based on data collected by a network blackhole (not enterprise networks). These datasets were not publicly available. The dataset we analyze has been studied in [3], which was influenced by [66] but only considered the aforementioned time series X_t. Going beyond [3], we consider time series X_t, Y_t, Z_t, which allows us to draw more insights.

To our knowledge, the problem of forecasting cyber attacks was not systematically investigated until [66], which proposed a systematic *gray-box* framework for data-driven modeling and forecasting cyber attacks. The term "gray-box" means: first characterizing the statistical properties exhibited by the data, and then building models that can accommodate these properties to forecast cyber attacks. The gray-box nature explains why the resulting statistical models can predict, contrasting the *black-box* nature of machine learning, especially deep learning, models. The framework [66] has been extended to forecast cyber attacks with extreme values [43,67], investigate the effectiveness of cyber defense early-warning [55], forecast the distribution of multivariate cyber attack rates while accommodating the dependence between them [44], characterize the cyber threat posture [65], investigate the prediction upper bound of cyber attack rates [9], and forecast cyber data breach incidents [20,56]. The present study can be seen as an extension of the gray-box framework [66] to accommodate newly identified statistical properties exhibited by cyber attacks against enterprise networks (i.e., high volatilities and burstiness). Related to this statistical approach, deep learning has been applied to forecast this type of data [19] as well as causality-based forecasting [53].

At a higher level of abstraction, cyber threats forecasting, including the "grey-box" framework [66], belongs to cybersecurity data analytics, which is integral to Science of

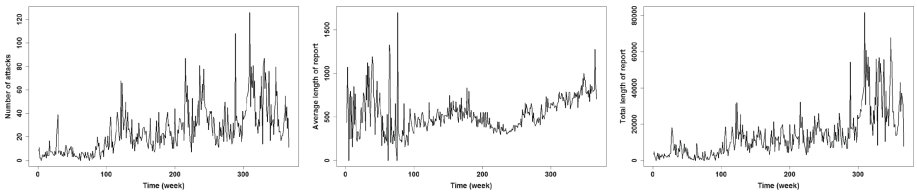
Cyber Security [64] and one pillar of the broader Cybersecurity Dynamics framework [58,59,61,62]. This broader framework is driven by cybersecurity metrics and quantification [6,10,15,38,42,63], including the quantification of attack and defense capabilities (e.g., cyber social engineering attacks [39–41,49]). Another pillar of the framework aims to model the evolution of the global cybersecurity state incurred by cyber attack-defense-use interactions, where "global" highlights the perspective of looking at a network (e.g., enterprise wide, nation wide, or even the entire cyberspace) as a whole. This pillar has led to many significant results (e.g., [7,8,23,25,26,34,35,57,60,68]).

Paper Outline. Section 2 describes the dataset and its cybersecurity implications. Section 3 characterizes the time series. Section 4 presents our model fitting and forecasting results. Section 5 concludes the paper. To improve the readability of the paper, Appendix A reviews the statistical knowledge that is used in this study.

2 Dataset and Its Cybersecurity Implications

2.1 Data Description

The dataset [2] is collected by a Network Defense Service Provider. It contains attack events against multiple enterprise networks managed by the Provider. These attack events are first detected by cyber defense tools and then more than 99% of them are manually examined and confirmed by human analysts (i.e., false positives are eliminated, while noting that the false-positive rate is not available). This means that at least 99% of the cyber attacks in the dataset are real attacks (i.e., true positives), while noting that the dataset may not contain all attacks (i.e., missing the false negatives). This does not invalidate the value of the present study on forecasting attacks based on true positives because the predictions can help defenders allocate resources to deal with detectable attacks. To our knowledge, this is the only publicly available dataset on cyber attacks against enterprise networks (rather than honeypots and network blackholes).



(a) X_t: Number of attacks (b) Y_t: Average report length (c) Z_t: Total report length

Fig. 1. Plots of time series X_t, Y_t, Z_t, where the x-axis represents time (unit: week).

The dataset contains 9,302 cyber attacks over a period of 366 weeks (or 7 years), but the precise data collection time is not given (except that it was after year 2000). For each attack, an *attack report* was written by a human analyst; unfortunately, no detailed information on attack reports is given. The dataset consists of two time series:

(i) the weekly-binned number of attacks, denoted by X_t for $t = 1, \ldots, 366$; (ii) the weekly-binned average length of attack reports, denoted by Y_t for $t = 1, \ldots, 366$. We propose deriving the weekly-binned *total* length of attack reports, denoted by Z_t, as follows: the weekly-binned total length of attack reports, denoted by $Z_t = X_t \cdot Y_t$ for $t = 1, \ldots, 366$. Figure 1a-1c respectively plot X_t, Y_t, and Z_t.

2.2 Attack Report Length, Attack Newness and Attack Sophistication

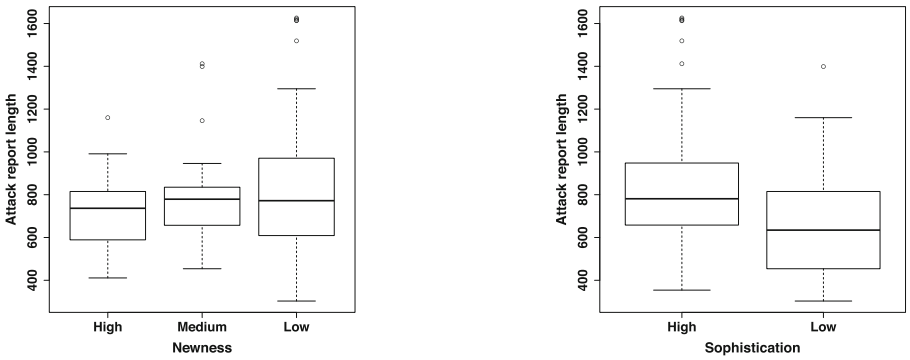
The Notions of Attack Newness and Attack Sophistication. We hypothesize that the attack report length may reflect (i) *attack newness*, meaning that a new attack that was not seen before would need to be documented in details, leading to a long report, and/or (ii) *attack sophistication*, meaning that a sophisticated attack would need to be documented in details, also leading to a long report. To (in)validate this, we are allowed to have access to a random sample of 100 attacks (rather than all the 9,302 attacks) in terms of the following three attributes: (i) attack report length, which is in the number of bytes but not the report itself; (ii) attack newness, which is subjectively labeled by a human analyst as "low", "medium" or "high"; (iii) attack sophistication, which is also subjectively labeled by a human analyst as "low", "medium" or "high". Among the 100 attacks, 14, 19, and 67 are respectively labeled as high-, medium-, and low-newness; whereas, 82, 15, and 3 are respectively labeled as high-, medium-, and low-sophistication. Since there are only 3 low-sophistication attacks (which are too few to make statistical sense), we combine medium- and low-sophistication attacks into one category, also called low-sophistication attacks in contrast to high-sophistication ones. This leads to 82 high-sophistication attacks and 18 low-sophistication attacks. Unfortunately, we are not authorized to share the data on these 100 attacks.

Given that newness and sophistication are subjectively labeled, we examine whether or not there is a dependence between the labels corresponding to them. For this purpose, we first use the Chi-square test, which is based on the contingency table [14] of the newness and sophistication labels. The Chi-square test result is 41.432 with a p-value $1.007e-09$, suggesting that there is a degree of dependence between these two notions. In order to characterize the dependence, we code the labels "low" as '0', "medium" as '1', and "high" as '2'; we then use the Kendall's rank correlation test [29] on these coded values. The estimated Kendall's tau is -0.5089 with a p-value 1.474e−07, suggesting that there is a negative dependence between attack newness and attack sophistication.

Insight 1. *A high-newness attack can have a low sophistication, namely that a new attack is not necessarily sophisticated.*

Relationship between Attack Report Length, Attack Newness, and Attack Sophistication. Figure 2a is the boxplot of the attack report length vs. attack newness based on the 100 samples mentioned above. We observe: (i) the boxplots of medium- and high-newness attacks are similar; (ii) the medians of attack report lengths of low-, medium-, and high-newness attacks are similar; (iii) there are very long attack reports (i.e., outliers) for some low-newness attacks; and (iv) the variability among attack report lengths of low-newness attacks is larger than that of high-newness and medium-newness attacks. These suggest that there is no significant differences in the attack report length

of varying attack newness, namely that the attack report length does not reflect attack newness. Figure 2b is the boxplot of the attack report length vs. attack sophistication. We observe that attack report lengths of low-sophistication attacks are substantially smaller than that of high-sophistication attacks. This suggests that attack report length indeed reflects the level of attack sophistication.



(a) Report length vs. newness

(b) Report length vs. sophistication

Fig. 2. Boxplots of attack report length vs. newness and report length vs. sophistication.

To formally confirm the intuitive findings mentioned above, we perform the following two statistical tests: ANOVA, which deals with the mean value of distributions [14], and Kruskal-Wallis, which is a non-parametric method dealing with distributions [29]. First, we use the ANOVA test to determine whether or not the mean attack report length is statistically the same in the three categories (i.e., high, medium, and low). For attack newness, the F statistic is 0.67 with a p-value 0.514, meaning that there is no statistical difference between the mean attack report length in the three categories of attack newness. For attack sophistication, the F statistic is 3.271 with a p-value 0.0736, where the small p-value indicates that there is a statistical difference between the mean attack report length in the two categories of attack sophistication. Second, we use the Kruskal-Wallis test to determine whether or not the attack report lengths in the two categories have the same distribution. For attack newness, the Kruskal-Wallis test statistic is 1.0182 with a p-value 0.601, which is consistent with the ANOVA test showing no statistical difference between the attack report length in the three categories of attack newness. For attack sophistication, the Kruskal-Wallis test statistic is 4.5789 with a p-value 0.0324, which confirms that there is a statistical difference between the attack report length of different attack sophistication. In summary, we draw:

Insight 2. *Attack report length reflects attack sophistication (i.e., less sophisticated attacks lead to shorter attack reports), but does not reflect attack newness (because a new attack is not necessarily sophisticated).*

3 Characterizing Time Series X_t, Y_t, Z_t

3.1 Basic Characteristics

From Fig. 1 we observe that Y_t for $t < 100$ can be large while the corresponding X_t is small, suggesting there are sophisticated attacks when $t < 100$. We observe that Z_t for $t \geq 300$ is large, which is caused by large X_t although the corresponding Y_t is not large. When Y_t is large, Z_t can still be small (e.g., for $t < 100$); when Y_t is not large, Z_t can still be large (e.g., for $x > 300$). This discrepancy between Y_t and Z_t justifies the importance of analyzing both Y_t and Z_t. In summary, we draw:

Insight 3. *The average attack report length and the total attack report length can exhibit different characteristics and can have different cybersecurity implications.*

From Fig. 1 we also observe that there exist large volatilities in all of the three time series. This phenomenon is confirmed by the basic statistics reported in Table 1, which shows that the variances are much larger than the corresponding mean values. This prompts that the classic Poisson process is not suitable for modeling these time series, and that we should investigate two statistical properties:

- Long-Range Dependence (LRD): This property, reviewed above, is important to characterize because it can guide the design of models to fit and forecast time series.
- Burstiness: This property, which is also reviewed above, is important because more analysts need to be allocated to cope with the bursts in attack events.

It is worth mentioning that *extreme values* [67] would be another property of interest, but the dataset contains too few data points to warrant an extreme value analysis.

Table 1. Basic statistics of X_t, Y_t, and Z_t sample values.

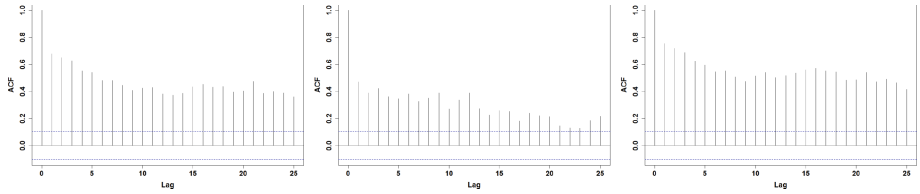
Time series	Min	Mean	Median	Var	Max
X_t	0	25.4153	21	394.0408	126
Y_t	0	533.7213	509	44617.46	1700
Z_t	0	14114.33	10884	170066803	81648

3.2 LRD Analysis

Figure 3a-3c plot the AutoCorrelation Function (ACF) of the three time series. We observe that in each case the ACF decays slowly, suggesting the presence of LRD. However, nonstationarity can also cause LRD [45,51], meaning that we need to study whether or not the slow decaying is indeed caused by the nonstationarity of the time series. For this purpose, we adopt the following widely-used hypothesis tests [37,47,51]:

- Unit root test: We use the Phillips-Perron test [46] to test if the slow decaying is caused by the *unit root* because it is nonparametric and robust against unspecified autocorrelation and heteroscedasticity. For time series R_t, we have:

$$H_0 : R_t \text{ has a unit root.} \qquad H_a : R_t \text{ is a stationary time series.}$$



(a) ACF of X_t (b) ACF of Y_t (c) ACF of Z_t

Fig. 3. Plots of the ACF (AutoCorrelation Functions) of time series X_t, Y_t, Z_t.

– Nonstationary test: We use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [33] to test if the slow decaying is caused by *non-stationarity*.

$$H_0 : R_t \text{ is stationary.} \qquad H_a : R_t \text{ is nonstationary.}$$

The p-values of the Phillips-Perron test for X_t, Y_t, and Z_t are all small (< 0.01), suggesting no evidence of unit root in any of the three time series. The p-values of the KPSS test are all small (< 0.01), suggesting that the three time series are nonstationary. Therefore, the LRD observed in the time series is indeed caused by nonstationarity.

Insight 4. *Cyber attack processes X_t, Y_t and Z_t are not Poisson but exhibit nonstationarity, which should be leveraged to design gray-box forecasting models.*

3.3 Burstiness Analysis

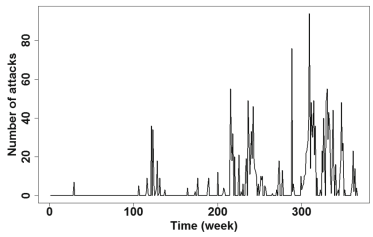
Burstiness is characterized via inter-event time. Since X_t, Y_t and Z_t are regular time series (i.e., their inter-event times are fixed), we pre-process them to obtain irregular time series of larger values, namely creating new time series by cutting off the small values of X_t, Y_t and Z_t to respectively obtain time series X'_t, Y'_t and Z'_t. We here focus on obtaining X'_t from X_t, while noting that Y'_t and Z'_t are obtained in the same fashion. Consider X_t for $1 \le t \le 366$, we first sort them as $X_{t_1} \le X_{t_2} \le \ldots \le X_{t_{366}}$. Then, we select a threshold ζ and omit any $X_t \le X_{t_\zeta}$ because we only consider the larger values. That is, we will analyze the time series X'_t for $t = 1, \ldots, 366$, where

$$X'_t = \begin{cases} X_t - X_{t_\zeta} & \text{if } X_t > X_{t_\zeta} \\ 0 & \text{otherwise.} \end{cases}$$

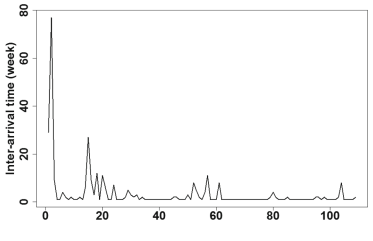
There are two guiding principles for selection threshold ζ: (i) X'_t, Y'_t and Z'_t should have sufficiently many non-zero values for building statistically significant models; and (ii) X'_t, Y'_t and Z'_t should have roughly the same number of non-zero values, which assures that they are equally significant in a statistical sense. Based on X'_t, we can define *inter-arrival time* between two consecutive, non-zero weekly-binned number of attacks as follows. We say two non-zero values X'_t and $X'_{t'}$ are *consecutive* if there does not exist t^* such that $t < t^* < t'$ and $X'_{t^*} > 0$. Then, we define the *inter-arrival time* between the two consecutive, non-zero values X'_t and $X'_{t'}$ as $t' - t$.

By examining the dataset, we set $\zeta = 257$ for X_t to obtain $366 - \zeta = 109$ non-zero values for X'_t, where $X_{t_{257}} = X_{319} = 32$; we set $\zeta = 256$ for Y_t to obtain 110 non-zero values for Y'_t, where $Y_{t_{256}} = Y_{297} = 610$; we set $\zeta = 256$ for Z_t to obtain 110 non-zero values for Z'_t, where $Z_{t_{256}} = Z_{173} = 16,524$.

Burstiness Analysis of X'_t. Figure 4a plots time series X'_t. We observe that more attacks are detected in the later weeks perhaps because the networks grow over time. Figure 4b plots the inter-arrival time between two consecutive, non-zero values. We observe many small inter-arrival times (i.e., 1 week), indicating that many attacks are waged during consecutive weeks and that bursts are exhibited. Corresponding to Fig. 4b and according to Eq. (4), the burstiness measure is $B = 0.424$, meaning that attacks are bursty.



(a) Large number of attacks X'_t (y-axis) vs. time t (x-axis)



(b) Inter-arrival times in X'_t (y-axis) vs. sequence # of intervals (x-axis)

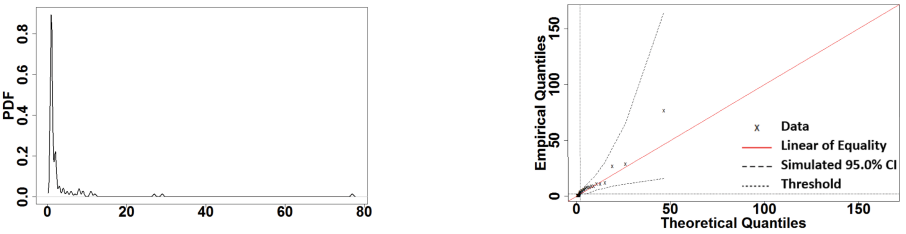
Fig. 4. Plots of X'_t and inter-arrival times between two consecutive, non-zero weekly-binned number of attacks.

To compute the more delicate burstiness measure δ given in Eq. (5), we need to fit the distribution of the inter-arrival times. Figure 5a plots the empirical density of inter-arrival times. We observe that the density is asymmetric with a long tail, meaning that distributions like normal, exponential, and weibull cannot fit inter-arrival times and that we should fit with a mixed distribution. We propose modeling the tail via GPD given in Eq. (7), and the other part by a lognormal distribution. The mixture density function is

$$f(x) = \begin{cases} \frac{1}{x\sigma_1\sqrt{2\pi}} e^{-\frac{(\ln x - \mu_1)^2}{2\sigma^2}}, & \text{if } x \le u, \\ \frac{1}{k}\left(1 + \xi z\right)^{-\frac{1}{\xi}-1}, & \text{if } x \ge u, \end{cases} \tag{1}$$

where σ_1 and μ_1 are respectively the standard deviation and the mean of lognormal distribution, ξ is the shape parameter of GPD, $z = (x - u)/\sigma$, and $k = 1/(\sigma \cdot \xi)$ is the scale parameter. Figure 5b shows the qq-plot of the inter-arrival times. We observe that the proposed mixed distribution fits the inter-arrival times well except for one data point, while all of the data points are within the simulated 95% confidence interval.

Table 2 describes the estimated parameters of the fitted distribution and their standard deviations. We observe that threshold u for the mixture distribution is 2.001, meaning that the tail proportion for the GPD fitting is around 21% and that the proposed mixture distribution fits the data well. Based on the fitted mixture distribution, we use Eq. (5) to derive the burstiness parameter $\delta = 0.556$, which suggests the presence of



(a) Empirical density (b) qq-plot (with threshold u)

Fig. 5. Empirical density and qq-plot of inter-arrival time (CI stands for "confidence interval")

burstiness. The memory parameter defined in Eq. (6) is $\theta = 0.381$, suggesting the existence of positive memory in inter-arrival times, namely that short (long) inter-arrival times are often followed by short (long) inter-arrival times.

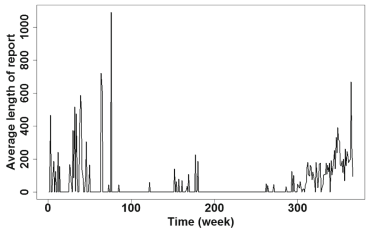
Table 2. Estimated parameters and standard deviations

	μ_1	σ_1	ξ	k	u
Parameter	0.301	0.409	0.387	5.670	2.001
Standard deviation	0.045	0.045	0.379	4.481	0.001

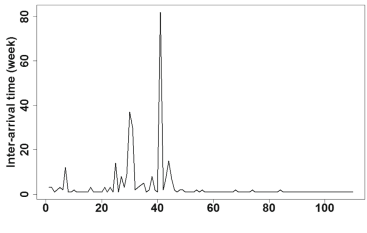
Burstiness Analysis of Y_t'. Figure 6a plots time series of large, weekly-binned average attack report lengths. We observe that larger average lengths are mainly exhibited when $t < 100$ and when $t > 300$. This further confirms the non-uniformality in the average attack report length and therefore attack sophistication. Figure 6b plots the inter-arrival times between two consecutive, non-zero Y_t' values, which are similarly defined as in the case of X_t'. We observe that most inter-arrival times between two large, average attack report lengths are very small (i.e., 1 week). Corresponding to Fig. 6b and according to Eq. (4), the burstiness measure B is $B = 0.462$, which suggests that the inter-arrival times between large, average attack report lengths exhibit the burstiness property.

To obtain the more delicate burstiness measure δ given by Eq. (5), we fit the distribution of inter-arrival times. Figure 7a plots the empirical density of inter-arrival times of large average attack report lengths, which also shows asymmetry with a long tail. This also suggests us to use the mixture distribution in Eq. (1) to fit inter-arrival times in Y_t'. Figure 7b shows the qq-plot, indicating the mixed distribution fits inter-arrival times very well and all of the data points are within the simulated 95% confidence interval.

Table 3 summarizes the estimated parameters of the fitted distributions and their standard deviations. We observe that the threshold u for the mixture distribution is 4.999, which means that the tail proportion for the GPD fitting is around 11%. Having fitted the distribution, we compute the burstiness parameter δ according to Eq. (5) to

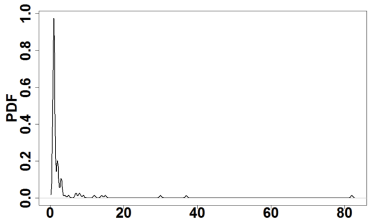


(a) Large average report length Y_t' (y-axis) vs. time (x-axis)

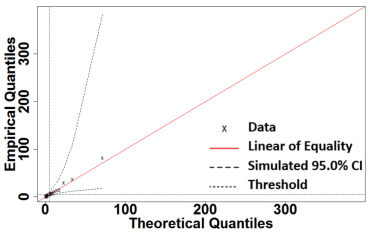


(b) Inter-arrival times in Y_t' (y-axis) vs. sequence # of intervals (x-axis)

Fig. 6. Plots of Y_t' and inter-arrival time between two consecutive, non-zero attack report lengths.



(a) Empirical density plot



(b) qq-plot of inter-arrival times

Fig. 7. Empirical density and qq-plot of inter-arrival time (CI stands for confidence interval).

Table 3. Estimated parameters and standard deviations

	μ_1	σ_1	ξ	k	u
Parameter	0.211	0.385	0.594	7.049	4.999
Standard deviation	0.040	0.031	0.487	3.938	0.001

have $\delta = 0.862$, which suggests the presence of strong burstiness. The memory parameter θ as defined in Eq. (6) is $\theta = 0.125$, which suggests positive memory, namely that short (long) inter-arrival times are often followed by short (long) inter-arrival times.

Burstiness Analysis of Z_t'. According to Eq.(4), burstiness of Z_t' is $B = 0.425$. When we proceed to derive burstiness measure δ, which requires to fitting the distribution of inter-arrival times of two consecutive, non-zero total report length in Z_t', we did *not* find any accurate fitting of the distribution, despite we tried the normal, weibull, gamma, GPD, and several mixed models. This means burstiness measure δ as defined in Eq. (5) is not always practical because fitting distributions may not be feasible especially when the distribution involves many parameters. Nevertheless, the other burstiness measure defined in Eq. (6) is $\theta = 0.392$, implying positive memory in inter-arrival times, namely that long (short) inter-arrival times are often followed by short (long) inter-arrival times.

Insight 5. *The detection or discovery of cyber attacks is bursty. Longer attack reports, which indicate more sophisticated attacks, are also bursty.*

Discussion. It would be ideal to pin down the root cause of burstiness. To this end, Harang and Kott [28] offer one hypothesis: burstiness is related to a threshold of analyst knowledge. It is conjectured [30] that the common element of various bursty processes is a threshold mechanism, namely that events occur infrequently until some domain-specific quantity accumulates to a threshold value, at which point events "burst out" at a high frequency. It is interesting to note that cyber analysts recalled episodes that multiple attacks are detected after the arrival of a crucial piece of new information about a previously unknown attack behavior or characteristic [28]. This new information enables analysts to recognize a particular type of attacks that until then was difficult or impossible to detect. At that point, analysts are able to rapidly detect a number of pre-existing attacks within a short period of time (a "burst"). These newly detected attacks actually represent false negatives before the arrival of a crucial piece of information; unfortunately, the dataset does not describe which attacks are in this category.

4 Modeling and Forecasting X_t, Y_t, and Z_t

For this purpose, we divide each time series into an in-sample part (for fitting) and an out-of-sample part (for forecasting). We use the first 266 samples (e.g., X_t for $1 \leq t \leq 266$) for in-sample fitting, and use the rest 100 samples (i.e., X_t for $267 \leq t \leq 366$) for out-of-sample forecasting. We first need to test if there is correlation between X_t and Y_t, while noting that there is correlation between X_t and Z_t and between Y_t and Z_t because $Z_t = X_t \times Y_t$. This is important because correlation, if existing, can be leveraged to achieve more accurate fitting and/or forecasting, as shown in other settings [44,55]. Examining Fig. 1a and 1b leads to the following *Hypothesis: there is a negative relation between the number of attacks and the length of reports.* One possible mechanism underpinning this hypothesis is that with many attacks arriving, there is less time to write long reports but, conversely, with few attacks arriving, there is more time to write long reports. Since we have showed the existence of temporal correlation within each individual time series X_t and Y_t, we need to eliminate this temporal correlation so as not to interfere with the evaluation of the correlation between X_t and Y_t. For this purpose, we measure the correlation between X_t and Y_t via the correlation between their respective standardized model residuals. By using the Pearson's correlation test to their residuals, we obtain a p-value of 0.117; by using the Kendall's rank correlation test, we obtain a p-value 0.5084. These p-values suggest no correlation between X_t and Y_t.

4.1 Model Fitting X_t, Y_t and Z_t for $1 \leq t \leq 266$

Motivated by the presence of volatilities and burstiness in X_t as shown above, we propose using the ARIMA+GARCH model reviewed above to fit X_t because ARIMA can accommodate the mean part and GARCH can accommodate high volatilities or burstiness. To be flexible, we allow the orders of p' and q' in the mean part of

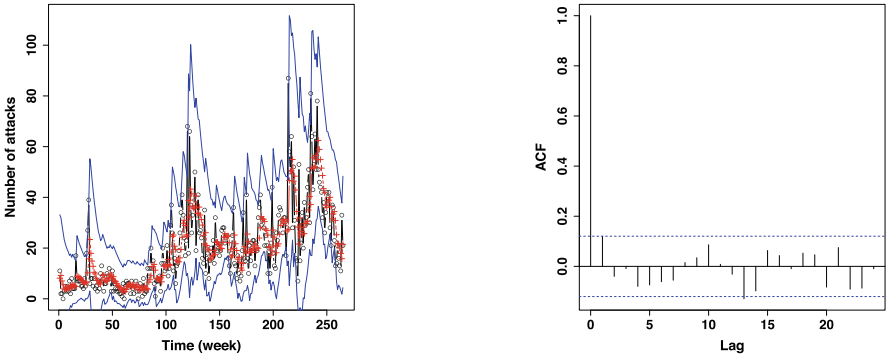
ARIMA(p', d, q') to vary between 0 and 5, and we use the Akaike Information Criterion (AIC) criterion to select the orders. For the GARCH part, we fix the order as GARCH$(1, 1)$ because it is adequate to accommodate volatilities in the residuals, while recalling that higher-order GARCH models are not necessarily better than GARCH$(1, 1)$ [27].

Table 4 summarizes the selected model ARIMA$(0, 1, 1)$+IGARCH$(1, 1)$ with the skewed Student-T distribution innovations and estimated parameters. We observe that the estimated coefficients are all significant (i.e., statistical significance at level .05) except for ω which is the intercept of GARCH model.

Table 4. Estimated parameters (EST.) and their standard deviations (SD) of the ARIMA$(0, 1, 1)$+IGARCH$(1, 1)$ fitting of X_t.

	ψ_1	ω	α_1	β_1	ξ	ν
EST.	-0.67	1.88	0.20	0.80	1.53	5.20
SD	0.05	1.18	0.05	–	0.13	1.22

Figure 8a plots the ARIMA$(0, 1, 1)$+IGARCH$(1, 1)$ fitting of X_t, where red-colored crosses represent fitted values, black-colored empty circles represent the observed values, and blue-colored lines are the fitted 95% confidence intervals. We observe that the overall fitting is good except for a few very large data points, and that the fitted confidence intervals can accommodate the variability of X_t. Figure 8b plots the ACF of the model residuals. We observe that the residuals fall in between the two blue-colored dashed lines (indicating the 95% confidence limits) with probability 95%, meaning that none of the correlation is significant. Therefore, ARIMA$(0, 1, 1)$+IGARCH$(1, 1)$ can adequately fit the temporal dependence between the X_t's with varying t. By using the Ljung-Box test on the standardized residuals, we obtain a p-value of 0.09, which implies that ARIMA$(0, 1, 1)$+IGARCH$(1, 1)$ can accurately fit X_t.



(a) ARIMA$(0, 1, 1)$+GARCH$(1, 1)$ (b) Plot of ACF of model residuals

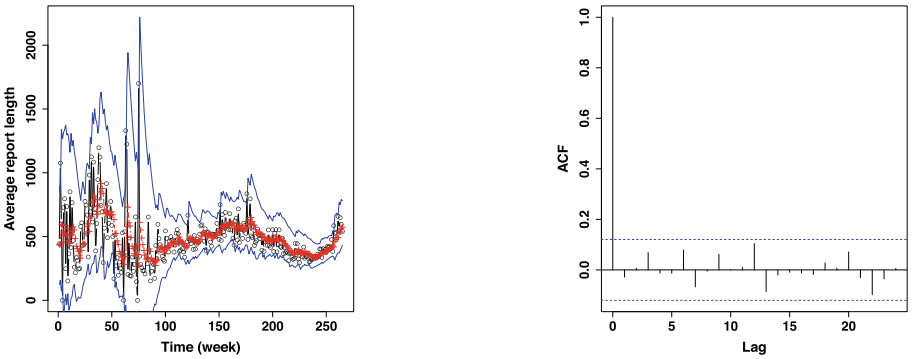
Fig. 8. ARIMA$(0, 1, 1)$+GARCH$(1, 1)$ fitting X_t for $1 \leq t \leq 266$.

Using the same model fitting method as in fitting X_t, the selected model is also ARIMA$(0,1,1)$+IGARCH$(1,1)$ with the skewed Student-T distribution innovations and estimated parameters summarized in Table 5. We again observe that all the coefficients for the are significant except for ω.

Figure 9 plots the ARIMA$(0,1,1)$+IGARCH$(1,1)$ fitting of Y_t, where red-colored crosses represent the fitted values, black-colored empty circles represent the observed values, and blue-colored lines represent the fitted 95% confidence intervals. We observe that the fitting is good and the fitted confidence intervals can accommodate the variability in Y_t. Figure 9b plots the ACF of the model residuals. We observe that none of the correlation is significant, meaning the selected ARIMA$(0,1,1)$+IGARCH$(1,1)$ model can adequately fit the temporal dependence in Y_t. By applying the weighted Ljung-Box test to the standardized residuals, we obtain a p-value of 0.91, which implies that ARIMA$(1,0,1)$+IGARCH$(1,1)$ can accurately fit Y_t.

Table 5. Estimated parameters (EST.) and their standard deviations (SD) of the ARIMA$(0,1,1)$+IGARCH$(1,1)$ fitting of Y_t.

	ψ_1	ω	α_1	β_1	ξ	ν
EST.	-0.76	237.00	0.20	0.80	1.42	5.81
SD	0.37	176.21	0.05	–	0.13	1.72



(a) ARIMA$(0,1,1)$+IGARCH$(1,1)$ (b) ACF of model residuals

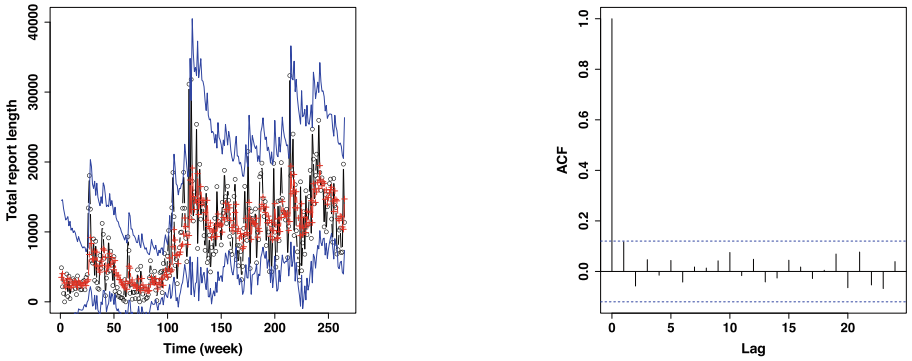
Fig. 9. ARIMA$(0,1,1)$+IGARCH$(1,1)$ fitting Y_t.

Similarly, the selected model for fitting Z_t is ARIMA$(1,1,1)$+GARCH$(1,1)$ with the skewed Student-T distribution innovations and estimated parameters summarized in Table 6. The coefficient for AR, ϕ_1, is significant at the 0.1 level. The coefficients for GARCH$(1,1)$ are significant, and $\alpha_1 + \beta_1 = 0.99$, which explains the burstiness.

Table 6. Estimated parameters (EST.) and standard deviations (SD) for the ARIMA$(1,1,1)$+GARCH$(1,1)$ fitting of Z_t

	ϕ_1	ψ_1	ω	α_1	β_1	ξ	ν
EST.	0.27	-0.90	31343	0.08	0.91	1.59	6.14
SD	0.07	0.05	187540	0.04	0.06	0.23	1.86

Figure 10 plots the ARIMA$(1,1,1)$+GARCH$(1,1)$ fitting result, where red-colored crosses represent the fitted values, black-colored empty circles represent the observed values, and blue-colored lines represent the fitted 95% confidence intervals. We observe that the overall fitting is good except for a few large samples, and the fitted confidence intervals can accommodate the variability in Z_t. Figure 10b plots the ACF of the model residuals, and shows that there is no significant correlation because they are within the blue-colored lines (i.e., the 95% confidence intervals). That it, the selected ARIMA$(1,1,1)$+GARCH$(1,1)$ model can adequately fit Z_t. By using the weighted Ljung-Box test to the standardized residuals, we obtain a p-value of 0.38, meaning ARIMA$(1,1,1)$+GARCH$(1,1)$ can accurately fit Z_t.



(a) ARIMA$(1,1,1)$+GARCH$(1,1)$ (b) ACF of model residuals

Fig. 10. ARIMA$(1,1,1)$+GARCH$(1,1)$ fitting Z_t.

Insight 6. *The three time series X_t, Y_t, and Z_t can all be accurately fitted using the ARIMA+GARCH model because they exhibit the burstiness property (which can be accommodated by the GARCH part).*

4.2 Forecasting X_t, Y_t and Z_t for $267 \leq t \leq 366$

Now we use the fitted models of X_t, Y_t and Z_t for $1 \leq t \leq 266$ to forecast X_t, Y_t and Z_t for $267 \leq t \leq 366$, respectively. To provide more information, we propose

forecasting distributions of X_t, Y_t and Z_t for $267 \leq t \leq 366$. To highlight ideas, we focus on forecasting X_t, but the idea is equally applicable to Y_t and Z_t. Recall the fitted model $\Psi_{t-1} = \{X_{t-1}, X_{t-2}, \ldots, X_1\}$. Let $\{f_t(X|\Psi_{t-1})\}_{t=1}^{\infty}$ be a sequence of one-step ahead (i.e., one-week ahead in this paper) density forecasting conditioned on the information available at time $t-1$, The cumulative density forecasts are given by

$$F_t(X_t) = \int_{-\infty}^{X_t} f_t(u|\Psi_{t-1})du. \tag{2}$$

If the model is accurate, the sequence of probability integral transforms $\{F_t(X_t)|t = 1, 2, \ldots\}$ are independent and identically random variables $U(0, 1)$ [1]. That is, the sequence of transforms being $U(0, 1)$ means we cannot reject the model being accurate. To evaluating accuracy of the forecasted probability density, we use two metrics:

- Density accuracy: Berkowitz [4] developed a formal test for evaluating the performance of density forecasting. The basic idea is to transform $\{F_t(X_t)|t = 1, 2, \ldots\}$ to the standard normal distribution $N(0, 1)$ by using the normal quantile function, and then test the normality of transformed data via the Lagrange Multiplier test [52]. Passing the test means the forecasted density is accurate.
- VaR violation: For a random variable X_t, the Value-at-Risk (VaR) at level α $(0 < \alpha < 1)$ is defined as [36]: $\mathrm{VaR}_\alpha(t) = \inf\{l : P(X_t \leq l) \geq \alpha\}$. For example, $\mathrm{VaR}_{.95}(t)$ means that there is only a 5% probability that the observed value is greater than the forecasted $\mathrm{VaR}_{.95}(t)$, which leads to a *violation* and indicates inaccurate forecasting. In order to evaluate the accuracy of the forecasted VaR values, we propose using the following three popular tests [11,17]:
  - The unconditional coverage test, denoted by LR_{uc}: It evaluates whether or not the fraction of violations is significantly different from the model's violations. If so, the forecasting is inaccurate; otherwise, we cannot reject that the forecasting is accurate.
  - The conditional coverage test, denoted by LR_{cc}: It is a joint likelihood ratio test for the independence of violations and unconditional coverage. Passing this test means that the violations are independent and the coverage is accurate, and hence we cannot reject that the forecasting is accurate.
  - The dynamic quantile test, denoted by (DQ): It is based on the sequence of 'hit' variables and it measures whether the present VaR violations are uncorrelated to the past violations or not. Passing this test means they are uncorrelated, and hence we cannot reject that the forecasting is accurate.

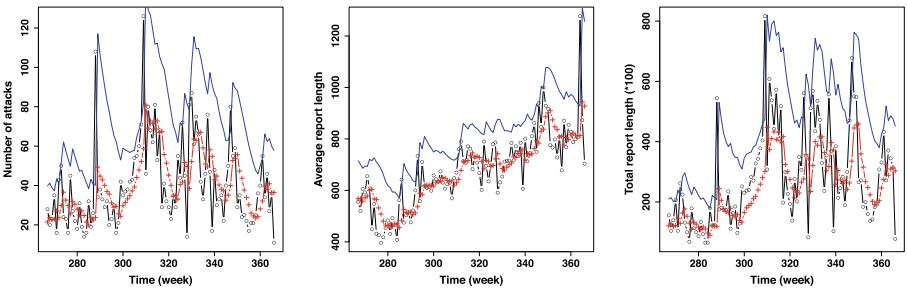
Forecasting Algorithm. Algorithm 1 describes the *rolling* algorithm for forecasting X_t, and can be adapted to forecast Y_t and Z_t by replacing X_t with Y_t or Z_t and by replacing $\hat{X}_t$ with $\hat{Y}_t$ or $\hat{Z}_t$, respectively.

Forecasting Results. Figure 11 plots the forecasting results of X_t, Y_t, and Z_t for $267 \leq t \leq 366$. We observe that the forecasted values are close to the observed ones, and that the forecasted $\mathrm{VaR}_{.95}$ can describe the violations well. We use rigorous statistical tests to validate the observation mentioned above. Table 7 summarizes the results. For X_t, the p-value of the Berkowitz test is 0.1925, indicating the forecasted density functions

Algorithm 1. Algorithm for forecasting X_t

Input: Historical time series dataset $\{(t, X_t)|t = 1, \ldots, 366\}$
Output: The predicted $\{(t, \hat{X}_t)|t = 267, \ldots, 366\}$

for $t = 266, \ldots, 365$ **do**
  Fit the historical data $\{(s, X_s)|s = 1, \ldots, t\}$ with the selected ARIMA(p', d, q')+GARCH$(1, 1)$ model
  Use the fitted model to forecast X_{t+1}, denoted the forecasted value by $\hat{X}_{t+1}$
end for
Return $\{(t, \hat{X}_t)|t = 267, \ldots, 366\}$



(a) Forecasted $\hat{X}_t$ (b) Forecasted $\hat{Y}_t$ (c) Forecasted $\hat{Z}_t$

Fig. 11. Plots of forecasting results based on ARIMA(p', d, q')+GARCH$(1, 1)$ models, where the red-colored $+$'s represent forecasted values, black-colored empty circles are the observed values, and blue-colored lines are the forecasted VaR$_{.95}$ (Color figure online).

are accurate. For the VaR violation tests at the $\alpha = .95$ level, all of the p-values are large. This also suggests that the ARIMA(p', d, q')+GARCH$(1, 1)$ models have a good accuracy performance. For Y_t, the p-value of the Berkowitz test is 0.5657, indicating the forecasted density functions are accurate. The VaR violation tests at the $\alpha = .95$ level all lead to large p-values, indicating the ARIMA(p', d, q')+GARCH$(1, 1)$ models have a good forecasting accuracy. For Z_t, the p-value of the Berkowitz test is 0.1550, meaning the forecasted density functions are accurate. The VaR violation tests at the $\alpha = .95$ level all lead to large p-values, meaning the ARIMA(p', d, q')+GARCH$(1, 1)$ models have a good forecasting accuracy. Summarizing the discussion, we draw:

Table 7. Statistical tests for forecast accuracy of ARIMA(p', d, q')+GARCH$(1, 1)$ models.

	Density accuracy	VaR violation metrics		
	Berkowitz	LR$_{uc}$	LR$_{cc}$	DQ
X_t	0.1925	0.3855	0.4027	0.2725
Y_t	0.5657	0.6560	0.6147	1
Z_t	0.1550	1	0.7664	0.999

Insight 7. *The distribution of the number of attacks, of the average report length, and of the total report length can be accurately forecasted, by using models that can accommodate the statistical properties exhibited by the data (i.e., burstiness in this case).*

Insight 7 is valuable because being able to predict report length, which reflects the sophistication of incoming attacks, provides a means to proactively allocate defense resources (e.g., human experts) to achieve more effective defense.

5 Conclusion and Discussion

We have presented an empirical study on a real-world high quality cyber attack dataset, leading to useful insights, such as: burstiness is commonly exhibited by cyber attacks; new attacks can be relatively simple (rather than sophisticated); attack report length reflects attack sophistication (but not attack newness); the detection of cyber attacks and the detection of sophisticated attacks are both bursty; ARIMA+GARCH model can fit the number of cyber attacks well; the distribution of the number of attacks against enterprise networks can be predicted accurately.

The study is limited by the dataset. First, the dataset does not provide information about individual attacks. Should the available dataset contain information about cyber attacks (e.g., types of attacks), deeper analysis can be conducted. Second, it is an outstanding open problem to precisely identify the cause of bursts. Third, the dataset does not provide any false negative information. Fourth, Insights 1 and 2 are drawn based on 100 random samples rather than all the 9,302 samples because we are only given the privilege to request for 100 random samples.

Acknowledgment. This work was supported in part by ARL Grant #W911NF-17-2-0127, NSF Grants #2122631 and #2115134, and Colorado State Bill 18-086.

A Statistical Preliminaries

Long Range Dependence (LRD). Intuitively, LRD means a stochastic process exhibits persistent temporal correlations, namely its autocorrelation decays slowly. Formally, a stationary time series $\{\mathsf{X}_i, i \geq 1\}$ is said to possess LRD [50,54] if its autocorrelation function $\rho(h)$, which is defined below, has the following property:

$$\rho(h) = \mathrm{Cor}(\mathsf{X}_i, \mathsf{X}_{i+h}) \sim h^{-\beta} L(h), \quad h \to \infty, \tag{3}$$

for $0 < \beta < 1$, where $\mathrm{Cor}(\cdot, \cdot)$ is the correlation function and $L(\cdot)$ is a *slowly varying* function satisfying $\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1$ for $t > 0$ [16]. The degree of LRD is expressed by the Hurst parameter (H), which is related to the parameter β in Eq. (3) as $\beta = 2-2H$. For LRD, we have $1/2 < H < 1$ and the degree of LRD increases as $H \to 1$.

Burstiness. Intuitively, burstiness indicates an abnormally large number of events within a short period of time when compared with the Poisson and regular stochastic processes. Unlike LRD, burstiness has no universally accepted definition. The simplest

definition of burstiness is perhaps the *coefficient of variation*, namely $r = \sigma/\mu$, where σ and μ are respectively the standard deviation and the mean of inter-event times [24,31]. Since r can be an arbitrary number, a refined burstiness definition is [31]:

$$B = \frac{\sigma - \mu}{\sigma + \mu} = \frac{r - 1}{r + 1}, \tag{4}$$

where $B \in [-1, 1]$, $B = -1$ corresponds to the regular time series (i.e., $r = 0$), $B = 0$ corresponds to the Poisson process, and $B \to 1$ indicates bursty time series [31].

By observing that burstiness can be rooted in two deviations from the Poisson process (i.e., the distribution and memory of the inter-event time), burstiness has also been defined as a vector (δ, θ), where $\delta, \theta \in [-1, 1]$ [24]. The first element δ reflects the distribution of inter-event time τ and is defined as

$$\delta = \frac{\text{sign}(\sigma - \mu)}{2} \int_0^\infty |\Pr(\tau) - \Pr_p(\tau)| d\tau, \tag{5}$$

where "sign" is the *sign* function, σ is the standard deviation, μ is the mean of inter-event time τ, $\Pr(\cdot)$ is the density function of τ, and $\Pr_p(\cdot)$ is the exponential distribution. Note that δ measures the difference between $\Pr(\cdot)$ and $\Pr_p(\cdot)$, $\delta = -1$ indicates the regular time series, $\delta = 0$ indicates the Poisson process, and $\delta \to 1$ indicates bursty time series. The second element θ reflects the memory of inter-event times, namely the correlation coefficient of consecutive inter-event times

$$\theta = \frac{1}{N} \sum_{i=1}^{N-1} \frac{(T_i - \mu_1)(T_{i+1} - \mu_2)}{\sigma_1 \times \sigma_2}, \tag{6}$$

where μ_1 (μ_2) and σ_1 (σ_2) are respectively the sample mean and the standard deviation of inter-event times $T_1, \ldots, T_{N-1}$ ($T_2, \ldots, T_N$), and N is the sample size. Note that θ is the memory coefficient describing the correlation of consecutive inter-event times, where $\theta > 0$ means positive memory, namely short (long) inter-event times are often followed by short (long) ones, and $\theta < 0$ means negative memory, namely short (long) inter-event times are often followed by long (short) ones.

Generalized Pareto Distribution (GPD). To characterize burstiness in inter-arrival times, we can first transform a regular time series into an irregular time series as follows. Given a sequence of independent and identically distributed (iid) observations $\mathsf{X}_1, \ldots, \mathsf{X}_n$, the excesses $\mathsf{X}_i - \ell$ of some suitably large threshold ℓ can be modeled by, under certain mild conditions, the *generalized Pareto distribution* (GPD) [16,48]. For characterizing burstiness, we choose the ℓ such that 30% of the excess values will be investigated, which is in contrast to the study of extreme values that would only consider the largest 1% samples. The survival function of the GPD is

$$\bar{G}_{\xi,\sigma_1,\ell}(x) = 1 - G_{\xi,\sigma_1,\ell} = \begin{cases} \left(1 + \xi \frac{x-\ell}{\sigma_1}\right)^{-1/\xi}, & \xi \neq 0, \\ \exp\left\{-\frac{x-\ell}{\sigma_1}\right\}, & \xi = 0, \end{cases} \tag{7}$$

where $x \geq \ell$ if $\xi \in \mathbb{R}^+$, $x \in [\ell, \ell - \sigma_1/\xi]$ if $\xi \in \mathbb{R}^-$, and ξ and σ_1 are respectively called the *shape* and *scale* parameters.

ARIMA and GARCH Models. ARIMA (AutoRegressive Integrated Moving Average) and GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) are widely-used time series models [13]. Intuitively, ARIMA can model the mean of a time series, and GARCH can model the high volatility of a time series. Formally, let $\phi(x) = 1 - \sum_{j=1}^{p'} \phi_j x^j$, $\psi(x) = 1 + \sum_{j=1}^{q'} \psi_j x^j$, and ϵ_t be independent and identical normal random variables with mean 0 and variance σ_ϵ^2. A time series $\{\mathsf{X}_t\}$ is said to be a ARIMA(p', d, q') process if $\phi(B)(1-B)^d(\mathsf{X}_t - \mu) = \psi(B)\epsilon_t$, where d is the number of difference, B is the back shift operator, and μ is the mean. A time series $\{\mathsf{Y}_t\}$ is called a GARCH process [5] if $\mathsf{Y}_t = \sigma_t \epsilon_t$, where ϵ_t (also called *innovation*) is the standard white noise. For the standard GARCH model, we have $\sigma_t^2 = w + \sum_{j=1}^{q'} \alpha_j \epsilon_{t-j}^2 + \sum_{j=1}^{p'} \beta_j \sigma_{t-j}^2$. A restricted version of the GARCH model is called the integrated GARCH (IGARCH) by requiring $\sum_{j=1}^{q'} \alpha_j + \sum_{j=1}^{p'} \beta_j = 1$. To accommodate more general classes of noise, we will use the skewed Student-T distribution, whose density can be written as [21]:

$$g(z; \boldsymbol{\vartheta}_j) = \frac{2}{\xi + \xi^{-1}} \left[t_\nu(\xi z) I(z < 0) + t_\nu v \left(\xi^{-1} z \right) \mathrm{I}(z \geq 0) \right],$$

where $\mathrm{I}(\cdot)$ is the indicator function, $\vartheta_j = (\xi_j, \nu_j)$, $\xi > 0$ is the skewness parameter, $\nu > 0$ is the shape parameter, and

$$t_\nu(z) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left[1 + z^2/\nu \right]^{-(\nu+1)/2}.$$

Akaike's Information Criterion (AIC). When fitting time series, we need criteria for model section. AIC is a widely used criterion [13,36,45] for balancing the goodness-of-fit of a model and its complexity such that a smaller AIC value indicates a better model. Formally, $\mathrm{AIC} = -2\log(\mathrm{MLE}) + 2k$, where MLE measures the goodness-of-fit of a model and k is the number of model parameters (indicating model complexity).

References

1. Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P.: Modeling and forecasting realized volatility. Econometrica **71**(2), 579–625 (2003)
2. Bakdash, J., et al.: Dataset associated with 'malware in the future? forecasting analyst detection of cyber events' (2019). https://osf.io/hjffm/
3. Bakdash, J.Z., et al.: Malware in the future? Forecasting of analyst detection of cyber events. J. Cybersecurity **4**(1) (2018)
4. Berkowitz, J.: Testing density forecasts, with applications to risk management. J. Bus. Econ. Stat. **19**(4), 465–474 (2001)
5. Bollerslev, T., Russell, J., Watson, M.W.: Volatility and Time Series Econometrics: Essays in Honor of Robert Engle. Oxford University Press, Oxford (2010)
6. Charlton, J., Du, P., Xu, S.: A new method for inferring ground-truth labels and malware detector effectiveness metrics. In: Lu, W., Sun, K., Yung, M., Liu, F. (eds.) SciSec 2021. LNCS, vol. 13005, pp. 77–92. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-89137-4_6

7. Chen, H., Cho, J., Xu, S.: Quantifying the security effectiveness of firewalls and DMZs. In: Proceedings of the HoTSoS 2018, pp. 9:1–9:11 (2018)
8. Chen, H., Cam, H., Xu, S.: Quantifying cybersecurity effectiveness of dynamic network diversity. IEEE Trans. Dependable Secur. Comput. (2021). https://doi.org/10.1109/TDSC.2021.3107514
9. Chen, Y., Huang, Z., Xu, S., Lai, Y.: Spatiotemporal patterns and predictability of cyberattacks. PLoS One **10**(5), e0124472 (2015)
10. Cho, J.H., Xu, S., Hurley, P.M., Mackay, M., Benjamin, T., Beaumont, M.: STRAM: measuring the trustworthiness of computer-based systems. ACM Comput. Surv. **51**(6), 128:1–128:47 (2019)
11. Christoffersen, P.F.: Evaluating interval forecasts. International Economic Review, pp. 841–862 (1998)
12. Condon, E., He, A., Cukier, M.: Analysis of computer security incident data using time series models. In: International Symposium on Software Reliability Engineering, pp. 77–86 (2008)
13. Cryer, J.D., Chan, K.S.: Time Series Analysis With Applications in R. Springer, New York (2008). https://doi.org/10.1007/978-0-387-75959-3
14. Devore, J.L., Berk, K.N., Carlton, M.A.: Modern Mathematical Statistics with Applications. STS, Springer, Cham (2021). https://doi.org/10.1007/978-3-030-55156-8
15. Du, P., Sun, Z., Chen, H., Cho, J.H., Xu, S.: Statistical estimation of malware detection metrics in the absence of ground truth. IEEE T-IFS **13**(12), 2965–2980 (2018)
16. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events. AM, vol. 33. Springer, Heidelberg (1997). https://doi.org/10.1007/978-3-642-33483-2
17. Engle, R.F., Manganelli, S.: CAViaR: conditional autoregressive value at risk by regression quantiles. J. Bus. Econ. Stat. **22**(4), 367–381 (2004)
18. Fachkha, C., Bou-Harb, E., Debbabi, M.: Towards a forecasting model for distributed denial of service activities. In: 2013 IEEE 12th International Symposium on Network Computing and Applications, pp. 110–117 (2013)
19. Fang, X., Xu, M., Xu, S., Zhao, P.: A deep learning framework for predicting cyber attacks rates. EURASIP J. Inf. Secur. **2019**, 5 (2019)
20. Fang, Z., Xu, M., Xu, S., Hu, T.: A framework for predicting data breach risk: leveraging dependence to cope with sparsity. IEEE T-IFS **16**, 2186–2201 (2021)
21. Fernandez, C., Steel, M.F.J.: On Bayesian modeling of fat tails and skewness. J. Am. Stat. Assoc. **93**(441), 359–371 (1998)
22. Ganesan, R., Jajodia, S., Cam, H.: Optimal scheduling of cybersecurity analysts for minimizing risk. ACM Trans. Intell. Syst. Technol. **8**(4), 52:1–52:32 (2017)
23. Garcia-Lebron, R., Myers, D.J., Xu, S., Sun, J.: Node diversification in complex networks by decentralized colouring. J. Complex Netw. **7**(4), 554–563 (2019)
24. Goh, K.I., Barabási, A.L.: Burstiness and memory in complex systems. EPL (Europhys. Lett.) **81**(4), 48002 (2008)
25. Han, Y., Lu, W., Xu, S.: Characterizing the power of moving target defense via cyber epidemic dynamics. In: HotSoS, pp. 1–12 (2014)
26. Han, Y., Lu, W., Xu, S.: Preventive and reactive cyber defense dynamics with ergodic time-dependent parameters is globally attractive. IEEE TNSE **8**(3), 2517–2532 (2021)
27. Hansen, P.R., Lunde, A.: A forecast comparison of volatility models: does anything beat a GARCH (1, 1)? J. Appl. Economet. **20**(7), 873–889 (2005)
28. Harang, R., Kott, A.: Burstiness of intrusion detection process: empirical evidence and a modeling approach. IEEE Trans. Inf. Forensics Secur. **12**(10), 2348–2359 (2017)
29. Hollander, M., Wolfe, D.A., Chicken, E.: Nonparametric Statistical Methods. vol. 751. Wiley, Hoboken (2013)
30. Karsai, M., Kaski, K., Barabási, A.L., Kertész, J.: Universal features of correlated bursty behaviour. Sci. Rep. **2**, 1–7 (2012)

31. Kim, E.K., Jo, H.H.: Measuring burstiness for finite event sequences. Phys. Rev. E **94**(3), 032311 (2016)
32. Kott, A., Arnold, C.: The promises and challenges of continuous monitoring and risk scoring. IEEE Secur. Priv. **11**(1), 90–93 (2013)
33. Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y., et al.: Testing the null hypothesis of stationarity against the alternative of a unit root. J. Econometrics **54**(1–3), 159–178 (1992)
34. Li, X., Parker, P., Xu, S.: A stochastic model for quantitative security analyses of networked systems. IEEE TDSC **8**(1), 28–43 (2011)
35. Lin, Z., Lu, W., Xu, S.: Unified preventive and reactive cyber defense dynamics is still globally convergent. IEEE/ACM ToN **27**(3), 1098–1111 (2019)
36. McNeil, A.J., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques, and Tools. Princeton University Press, Princeton (2010)
37. Mikosch, T., Starica, C.: Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. Rev. Econ. Stat. **86**(1), 378–390 (2004)
38. Mireles, J.D., Ficke, E., Cho, J., Hurley, P., Xu, S.: Metrics towards measuring cyber agility. IEEE Trans. Inf. Forensics Secur. **14**(12), 3217–3232 (2019)
39. Montañez Rodriguez, R., Longtchi, T., Gwartney, K., Ear, E., Azari, D.P., Kelley, C.P., Xu, S.: Quantifying psychological sophistication of malicious emails. In: Yung, M., et al. (eds.) SciSec 2023, LNCS, vol. 14299, pp. 319–331. Springer, Cham (2023)
40. Montañez, R., Atyabi, A., Xu, S.: Book chapter in "cybersecurity and cognitive science", chap. social engineering attacks and defenses in the physical world vs. cyberspace: a contrast study. Elsevier, pp. 3–41 (2022)
41. Montañez, R., Golob, E., Xu, S.: Human cognition through the lens of social engineering cyberattacks. Front. Psychol. **11**, 1755 (2020)
42. Pendleton, M., Garcia-Lebron, R., Cho, J.H., Xu, S.: A survey on systems security metrics. ACM Comput. Surv. **49**(4), 62:1–62:35 (2016)
43. Peng, C., Xu, M., Xu, S., Hu, T.: Modeling and predicting extreme cyber attack rates via marked point processes. J. Appl. Stat. **44**(14), 2534–2563 (2017)
44. Peng, C., Xu, M., Xu, S., Hu, T.: Modeling multivariate cybersecurity risks. J. Appl. Stat. **45**(15), 2718–2740 (2018)
45. Peter, B., Richard, D.: Introduction to Time Series and Forecasting. Springer, New York (2002). https://doi.org/10.1007/b97391
46. Phillips, P.C., Perron, P.: Testing for a unit root in time series regression. Biometrika **75**(2), 335–346 (1988)
47. Qu, Z.: A test against spurious long memory. J. Bus. Econ. Stat. **29**(3), 423–438 (2011)
48. Resnick, S.: Heavy-Tail Phenomena: Probabilistic and Statistical Modeling. Springer, New York (2007). https://doi.org/10.1007/978-0-387-45024-7
49. Rodriguez, R.M., Xu, S.: Cyber social engineering kill chain. In: Proceedings of International Conference on Science of Cyber Security (SciSec 2022), pp. 487–504 (2022)
50. Samorodnitsky, G.: Long range dependence. Founda. Trends Stoch. Syst. **1**(3), 163–257 (2006)
51. Shao, X.: A simple test of changes in mean in the possible presence of long-range dependence. J. Time Ser. Anal. **32**(6), 598–606 (2011)
52. Silvey, S.D.: The Lagrangian multiplier test. Ann. Math. Stat. **30**(2), 389–407 (1959)
53. Trieu-Do, V., Garcia-Lebron, R., Xu, M., Xu, S., Feng, Y.: Characterizing and leveraging granger causality in cybersecurity: framework and case study. EAI Endorsed Trans. Secur. Safety **7**(25), e4 (2020)
54. Willinger, W., Taqqu, M.S., Leland, W.E., Wilson, V.: Self-similarity in high-speed packet traffic: analysis and modeling of ethernet traffic measurements. Stat. Sci. **10**, 67–85 (1995)
55. Xu, M., Hua, L., Xu, S.: A vine copula model for predicting the effectiveness of cyber defense early-warning. Technometrics **59**(4), 508–520 (2017)

56. Xu, M., Schweitzer, K.M., Bateman, R.M., Xu, S.: Modeling and predicting cyber hacking breaches. IEEE T-IFS **13**(11), 2856–2871 (2018)
57. Xu, M., Xu, S.: An extended stochastic model for quantitative security analysis of networked systems. Internet Math. **8**(3), 288–320 (2012)
58. Xu, S.: Emergent behavior in cybersecurity. In: Proceedings of the HotSoS 2014, pp. 13:1–13:2 (2014)
59. Xu, S.: The cybersecurity dynamics way of thinking and landscape (invited paper). In: ACM Workshop on Moving Target Defense (2020)
60. Xu, S., Lu, W., Zhan, Z.: A stochastic model of multivirus dynamics. IEEE Trans. Dependable Secur. Comput. **9**(1), 30–45 (2012)
61. Xu, S.: Cybersecurity dynamics. In: Proceedings of the Symposium on the Science of Security (HotSoS 14), pp. 14:1–14:2 (2014)
62. Xu, S.: Cybersecurity dynamics: a foundation for the science of cybersecurity. In: Wang, C., Lu, Z. (eds.) Proactive and Dynamic Network Defense. Advances in Information Security, vol. 74, pp. 1–31. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10597-6_1
63. Xu, S.: SARR: a cybersecurity metrics and quantification framework (keynote). In: Lu, W., Sun, K., Yung, M., Liu, F. (eds.) SciSec 2021. LNCS, vol. 13005, pp. 3–17. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-89137-4_1
64. Xu, S., Yung, M., Wang, J.: Seeking foundations for the science of cyber security. Inf. Syst. Front. **23**(2), 263–267 (2021)
65. Zhan, Z., Xu, M., Xu, S.: A characterization of cybersecurity posture from network telescope data. In: Proceedings of the InTrust, pp. 105–126 (2014)
66. Zhan, Z., Xu, M., Xu, S.: Characterizing honeypot-captured cyber attacks: Statistical framework and case study. IEEE T-IFS **8**(11), 1775–1789 (2013)
67. Zhan, Z., Xu, M., Xu, S.: Predicting cyber attack rates with extreme values. IEEE Trans. Inf. Forensics Secur. **10**(8), 1666–1677 (2015)
68. Zheng, R., Lu, W., Xu, S.: Preventive and reactive cyber defense dynamics is globally stable. IEEE TNSE **5**(2), 156–170 (2018)