10.48047/jocaaa.2024.33.08.71

A STOCHASTIC FRAMEWORK FOR MONITORING AND PREDICTING CYBER ATTACKS AND DATA BREACHES

Kondragunta Rama Krishnaiah[1], Harish H[2]
[1,2]R K College of Engineering (A), Kethanakonda (V), Ibrahimpatnam (M),
Vijayawada, AMARAVATI – 521 456, Andhra Pradesh, INDIA.
drkrk@rkce.ac.in, ORCID: 0000-0002-9069-766X
dr.hharish@rkce.ac.in, ORCID: 0000-0002-4572-1704

Abstract
With the increasing sophistication of cyber threats, traditional models based on static statistical distributions fall short in accurately characterizing the dynamics of cyber incidents. This study introduces a comprehensive stochastic modeling approach for analyzing and predicting hacking-related data breaches. Using a refined dataset comprising 600 incidents from 2005 to 2017, we apply time-series models including LACD1 and ARMA-GARCH, coupled with copula-based dependency modeling, to assess inter-arrival times and breach sizes. The results reveal statistically significant temporal and cross-variable dependencies, demonstrating that cyber incidents are better modeled through stochastic processes. These findings offer practical implications for cyber defense strategies and risk assessment frameworks, similar to how weather forecasting informs emergency preparedness.

Keywords: Cybersecurity, Data Breach, Stochastic Modeling, Time Series Forecasting, Risk Analytics

1. Introduction
Data breaches have emerged as a significant threat to organizations, with incidents ranging from unauthorized data access to large-scale information theft [1]. These breaches often occur through hacking, exploiting system vulnerabilities to extract confidential data. Understanding the dynamics of these events is crucial for enhancing cybersecurity defenses, formulating insurance strategies, and informing regulatory frameworks [2].

The economic and reputational impact of cyberattacks continues to rise, leading to growing investments in preventive and predictive technologies. However, the challenge remains in anticipating attack behavior with sufficient accuracy. Traditional models treat breach events as isolated, random occurrences, typically modeled by Poisson processes or static probability distributions [3][4]. While such approaches offer simplicity, they ignore the inherent temporal and structural complexities observed in real-world datasets.

This paper addresses these limitations by adopting a stochastic modeling framework. Specifically, we explore whether cyberattack occurrences and their respective breach sizes follow identifiable temporal patterns, and whether these characteristics exhibit autocorrelations or dependencies over time. Using insights from fields such as finance and seismology, where similar time-dependent behaviors are well-documented, we apply models capable of accounting for such complexities.

2. Related Work
Numerous prior works have examined the statistical properties of cyber breaches. Maillart and Sornette [3] analyzed personal identity theft incidents and demonstrated the existence of a heavy-tailed distribution. Edwards et al. [4] extended this by examining breach sizes and event

10.48047/jocaaa.2024.33.08.71

frequencies, suggesting the adequacy of log-normal distributions and negative binomial models, respectively.

Wheatley et al. [5] introduced Extreme Value Theory to study large-scale breaches, demonstrating that the distribution of breach sizes is more severe than previously thought. However, none of these models accounted for the temporal nature of breaches or considered joint dependencies between timing and severity.

Recent studies have begun to explore cyber risk dependencies. Böhme and Kataria [6] used copulas to model inter-organizational risk correlations. Herath and Herath, along with Mukhopadhyay et al. [7], developed actuarial frameworks integrating dependency measures into insurance models. Xu et al. [8] further applied vine copulas to evaluate the performance of early warning systems. Our research differentiates itself by applying these concepts specifically to malware-related hacking breaches using time-dependent modeling techniques.

3. Methodology

3.1 Data Collection

Our dataset, sourced from the Privacy Rights Clearinghouse [1], comprises 600 verified hacking-related data breaches reported in the U.S. between January 2005 and April 2017. We excluded breaches caused by internal errors, accidental disclosures, or non-malicious causes. Sectors affected include healthcare, finance, education, government, and nonprofits.

Each breach entry includes:

- Date of incident
- Breach size (number of records compromised)
- Industry classification

3.2 Data Preprocessing

Cyber incidents occurring on the same calendar day were often reported together, creating a potential bias. To mitigate this, we assigned randomized intra-day timestamps to each event, preserving their individual temporal identities. This was crucial for inter-arrival time analysis, ensuring fidelity in time-series modeling.

Additional preprocessing steps included:

- Removal of missing or zero breach size entries
- Log transformation of breach size to normalize distribution
- Ordering events chronologically

3.3 Modeling Approach

Our modeling framework consists of three key components:

- **Inter-Arrival Time Modeling:** Implemented using the Log-Autoregressive Conditional Duration (LACD1) model, which captures autocorrelation and time-varying event intensities.
- **Breach Size Modeling:** Breach sizes were log-transformed and modeled using ARMA-GARCH to account for conditional heteroskedasticity (i.e., volatility clustering).
- **Joint Dependency Modeling:** Copulas were used to model the nonlinear dependency structure between event timing and size. Copula selection was based on Akaike Information Criterion (AIC) scores.

3.4 Statistical Validation

Model diagnostics included:

- **ACF and PACF Plots:** Used to visualize lagged correlations (Figure 2a-d).
- **Goodness-of-Fit Tests:** Kolmogorov-Smirnov and Ljung-Box tests validated the time-series models.

10.48047/jocaaa.2024.33.08.71

- **Dependence Tests:** Kendall's $\tau = 0.076$ and Spearman's $\rho = 0.115$ ($p < 0.05$), confirming the presence of non-trivial dependencies.

4. Results and Discussion

This section presents the key empirical findings of the study. It evaluates the effectiveness of the proposed stochastic models in capturing the dynamics of cyber breach incidents, including inter-arrival times and breach sizes. Visual analyses are provided to support the observed patterns, with statistical validation reinforcing the insights derived. Figures 1 and 2 illustrate critical aspects of these trends and dependencies in the dataset.

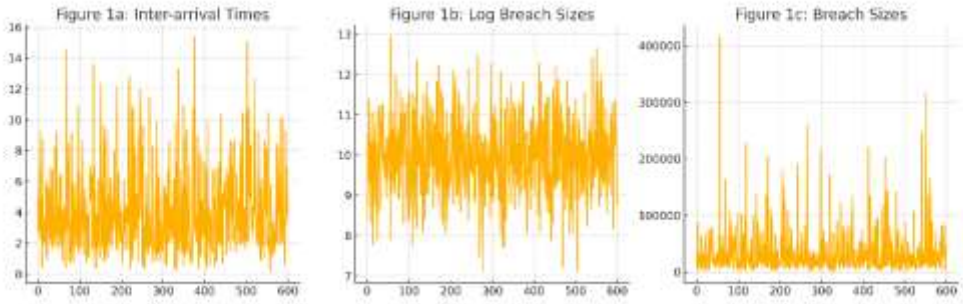

Figure 1: Temporal patterns of cyber breach incidents—(a) Inter-arrival times, (b) Log-transformed breach sizes, and (c) Original breach sizes.

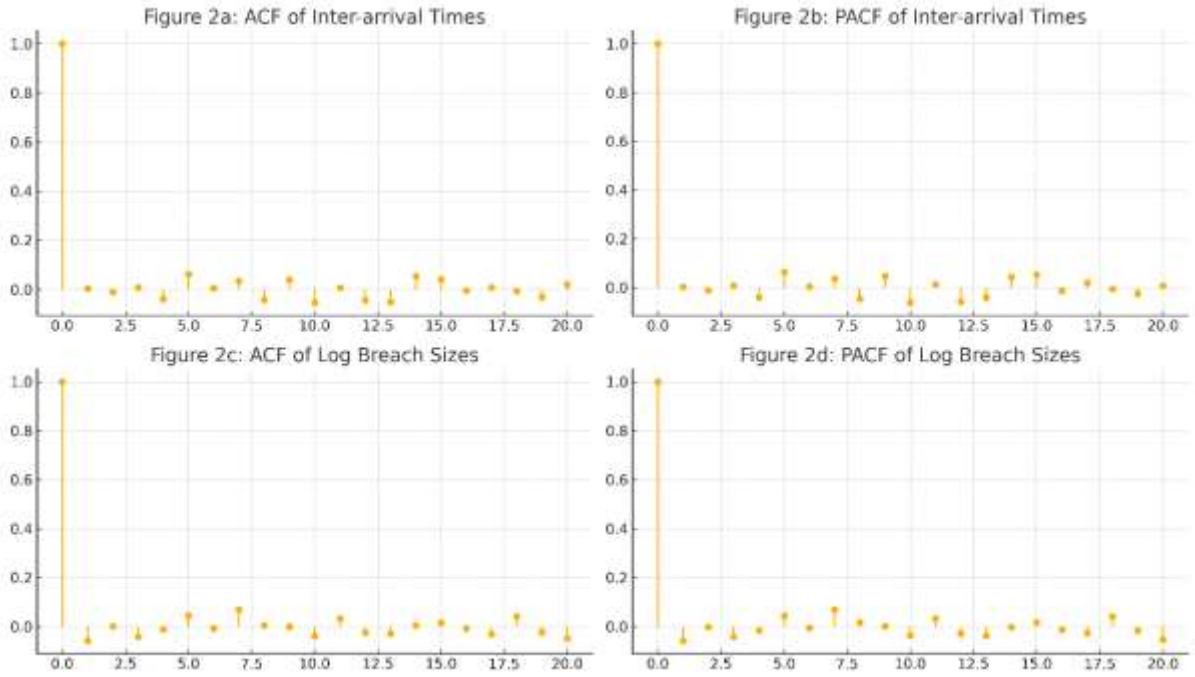

Figure 2: Autocorrelation structures in cyber breach data—(a) ACF and (b) PACF of inter-arrival times; (c) ACF and (d) PACF of log-transformed breach sizes.

4.1 Inter-Arrival Times

10.48047/jocaaa.2024.33.08.71

Inter-arrival times showed significant deviation from the memoryless exponential distribution. Figure 1a illustrates the non-uniform distribution of these durations. ACF and PACF (Figures 2a and 2b) show persistent autocorrelations, refuting the suitability of Poisson assumptions.

4.2 Breach Sizes

As depicted in Figure 1b, log-transformed breach sizes follow a more normalized pattern but still exhibit volatility clustering. Figure 1c shows several extremely large breaches. ACF and PACF plots (Figures 2c and 2d) further confirm autocorrelation, indicating suitability for ARMA-GARCH modeling.

4.3 Correlation Between Variables

Our copula-based analysis revealed statistically significant positive dependence between breach frequency and severity. Large breaches were more likely to follow long dormant periods, suggesting strategic attacker behavior, possibly due to the need to develop or acquire more sophisticated attack tools (Figure 2).

4.4 Predictive Performance

Using a rolling forecast approach, we predicted Value-at-Risk (VaR) for both inter-arrival times and breach sizes. Results showed:

- At $\alpha = 0.90$: 28 predicted violations vs. 31 actual
- At $\alpha = 0.95$: 14 predicted vs. 20 actual

These results indicate the model's conservativeness in predicting extreme events. While large deviations still present forecasting challenges, overall predictive fidelity remains robust.

Table 1. Summary Statistics

Metric	Mean	Std Dev	Max	Min
Inter-Arrival Time	3.95 days	2.92	17.5	0.15
Breach Size (log)	10.05	1.02	14.23	7.08

4.5 Practical Implications

The models proposed enable various stakeholders to make informed, data-driven decisions:

- **Cyber Insurance Providers:** Can improve actuarial models and dynamic premium pricing.
- **Security Analysts:** Gain better tools for incident forecasting and threat mitigation.
- **Policy Makers:** Obtain empirical support for designing adaptive regulatory frameworks.

Moreover, this predictive capability allows for dynamic defense postures. For example, if the model forecasts an impending high-severity event, organizations can proactively allocate resources, enhance monitoring, and activate contingency protocols (as guided by Figure 1 and 2).

5. Conclusion

This study proposes a stochastic modeling approach for analyzing the occurrence and severity of cyber hacking breaches. By applying time-series models and copula theory, we uncover patterns and dependencies that traditional methods overlook. Our analysis confirms that cyber breaches are temporally and structurally interdependent phenomena.

Key contributions include:

- Empirical validation of autocorrelation in breach data (Figure 2)
- Novel joint modeling of inter-arrival times and breach magnitudes (Figures 1 and 2)
- Practical frameworks for predictive cyber defense and risk forecasting

10.48047/jocaaa.2024.33.08.71

Future work will address the accurate modeling of outlier events and explore the integration of additional covariates such as geopolitical risk, sector-specific vulnerability, and attacker sophistication.

References

[1] Privacy Rights Clearinghouse. Chronology of Data Breaches. 2017.
[2] Eling, M. & Schnell, W. (2016). What do we know about cyber risk and cyber risk insurance? *Journal of Risk Finance*, 17(5), 474–491.
[3] Maillart, T., & Sornette, D. (2010). Heavy-tailed distribution of cyber-risks. *Eur. Phys. J. B*, 75(3), 357–364.
[4] Edwards, B., Hofmeyr, S., & Forrest, S. (2016). Hype and heavy tails: A closer look at data breaches. *J. Cybersecurity*, 2(1), 3–14.
[5] Wheatley, S., Maillart, T., & Sornette, D. (2016). The extreme risk of personal data breaches. *Eur. Phys. J. B*, 89(1), 7.
[6] Böhme, R., & Kataria, G. (2006). Models and measures for correlation in cyber-insurance. *Workshop Econ. Inf. Secur*.
[7] Mukhopadhyay, A., et al. (2013). Cyber-risk decision models: To insure it or not? *Decision Support Systems*, 56, 11–26.
[8] Xu, M., Hua, L., & Xu, S. (2017). A vine copula model for predicting the effectiveness of cyber defense early-warning. *Technometrics*, 59(4), 508–520.