Challenges in Forecasting Malicious Events from Incomplete Data

Nazgol Tavabi*
nazgolta@isi.edu
University of Southern California, Information Sciences Institute

Andrés Abeliuk*
aabeliuk@isi.edu
University of Southern California, Information Sciences Institute

Negar Mokhberian
nmokhber@isi.edu
University of Southern California, Information Sciences Institute

Jeremy Abramson
abramson@isi.edu
University of Southern California, Information Sciences Institute

Kristina Lerman
lerman@isi.edu
University of Southern California, Information Sciences Institute

ABSTRACT

The ability to accurately predict cyber-attacks would enable organizations to mitigate their growing threat and avert the financial losses and disruptions they cause. But how predictable are cyber-attacks? Researchers have attempted to combine external data – ranging from vulnerability disclosures to discussions on Twitter and the darkweb – with machine learning algorithms to learn indicators of impending cyber-attacks. However, *successful* cyber-attacks represent a tiny fraction of all *attempted* attacks: the vast majority are stopped, or filtered by the security appliances deployed at the target. As we show in this paper, the process of filtering reduces the predictability of cyber-attacks. The small number of attacks that do penetrate the target's defenses follow a different generative process compared to the whole data which is much harder to learn for predictive models. This could be caused by the fact that the resulting time series also depends on the filtering process in addition to all the different factors that the original time series depended on. We empirically quantify the loss of predictability due to filtering using real-world data from two organizations. Our work identifies the limits to forecasting cyber-attacks from highly filtered data.

CCS CONCEPTS

• **Security and privacy** → **Usability in security and privacy**; • **Computing methodologies** → *Machine learning approaches.*

KEYWORDS

predictability, cyber-attack, forecasting, time-series, permutation entropy

ACM Reference Format:
Nazgol Tavabi, Andrés Abeliuk, Negar Mokhberian, Jeremy Abramson, and Kristina Lerman. 2020. Challenges in Forecasting Malicious Events from Incomplete Data. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3366424.3385774

*Both authors contributed equally to this research.

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3385774

1 INTRODUCTION

Malicious behavior is increasingly common in online life. Social media platforms are racing to develop tools to detect—and in some cases anticipate—malicious behaviors in the form of manipulation, deception, misinformation, and cyberbullying. Cybercrime is one example of malicious behavior that has resulted in large financial losses, political and security risks. The 2016 hacking of the Democratic National Committee server was arguably a turning point in the 2016 US presidential elections. The leaks of potentially embarrassing emails upended the race and upset existing polls. The ability to anticipate cybercrime — and other security threats, such as violent protests in a country — would allow organizations to mitigate the risks associated with such disruptions. As an age-old saying goes: "to be forewarned is to be forearmed."

How predictable is malicious behavior? To narrow the scope of the question, we consider the predictability of cybercrime, since forecasting cybercrime shares many challenges of forecasting other types of malicious activities. In the case of cybercrime, conventional wisdom says indicators often exist that can be leveraged for detection and prediction. Successfully executing a cyber-attack requires preparation and planning: hackers carry out reconnaissance about the potential targets, identify its vulnerabilities, acquire relevant tools and exploits, etc. All these activities leave traces within the openly available data (as well as within the patterns of attacks themselves) that allow for forecasting new cyber-attacks.

An important consideration for an AI-based solution to cyber-attack prediction is that the successful attacks used to train machine learning models make up a *small fraction* of all attacks. In reality, the vast majority of attacks are stopped by the target's defenses: firewalls, domain blockers, spam filters, etc, as illustrated in Figure 1. Only a small fraction of attempted attacks reach the victim and are recorded as training data for machine learning algorithms. This also pertains to other types of malicious behaviors, which are at best only partially observed, as malicious actors attempt to obfuscate their behaviors. As we show in this paper, having access to only a subset of malicious events for use in training reduces the predictive utility of the data, and accuracy of the models learned from it. In this paper, we demonstrate this important problem in the context of cyber-attack forecasting.

This work makes the following two contributions: 1) We quantify the impact of incomplete observation due to filtering on the predictability of malicious events using real-world email data from

two distinct organizations and 2) We show that predictability decreases and prediction error grows when more malicious emails are filtered by the organization's defenses. Our work identifies an important challenge researchers have to consider when forecasting malicious activity, including cyber-attacks.

2 RELATED WORK

Researchers have used state-of-the-art machine learning methods to forecast cyber-attacks and extract predictive indicators from available data sources. Works such as [8, 12, 21, 24] have trained models to find associations between such indicators and successful attacks. Approaches using temporal forecasting include prediction of cyber breaches [23]. This type of analysis models the inter-arrival time of breach incidents inter-arrival as a stochastic process, described by an auto-regressive and Moving Average (ARMA). The authors also show a decrease between event inter-arrival times, indicating that cyber-attacks are becoming more frequent. A Bayesian framework for cyber-attack prediction is presented in [22]. The authors use attack graphs to represent and enumerate possible system vulnerabilities and attack paths. While this approach is promising, in practice it may not be actionable, as it requires an accurate picture of all system attack surfaces and vulnerabilities, which is unrealistic in a modern enterprise network. The authors in [3] propose a methodology to determine enterprise-level risk assessment against "untargeted" attacks, testing the framework on data accrued from a large financial institution. The work in [5] presents a Bayesian State Space Model (BSSM) model for forecasting cyber-attack events, and can do so with reasonable accuracy for non-bursty events up to one week out. This result is promising, however, the data set itself is exactly the sort of filtered view that this work addresses. The training data used in the study includes approximately seven years of weekly analyst-verified cyber incidents at a large US Department of Defense enterprise.

Another line of research is to use a variety of external data to improve predictions. This includes [14] and [16], which uses external data from Twitter discussions to automatically learn keywords associated with emerging cyber threats, such as malware or botnet names. Other works identify patterns within discussions of vulnerabilities on darkweb [4] or sentiment of posts in hacker forums [7], predictive of future cyber-attacks, or identify software vulnerabilities that are likely to get exploited [15, 20]. All of these approaches are useful in enhancing predictions of a specific domain based on external signals relevant to that domain, but none address the fundamental impact of data on the predictability of cyber-attacks.

3 LIMITS OF PREDICTABILITY

Our work builds upon our recent discovery of a fundamental limit to the predictability of partially observed dynamic systems. We developed a mathematical framework to characterize the loss of information in sampled time series [1]. The framework allows us to quantify how the predictability of dynamic systems is affected by temporal sampling under a variety of sampling conditions. In this work, we will extend our framework to quantify the impact of predictability loss on real-world forecasting tasks of interest

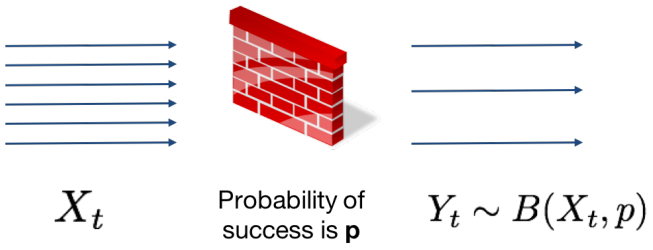

Figure 1: Sketch of the filtering framework. In the context of cyberattacks, X is the number of attempted attacks per day; Y is the number of successful attacks per day observed by the target. The probability of an attack being successful is p. The Binomial distribution $B(n, p)$ is used to model the time series Y.

in security applications. Specifically, we will measure how partial observation affects the accuracy of forecasting cyberattacks.

Consider a dynamic process generating events, (e.g.) cyberattacks against an organization or protests in a region. We represent the time series of this *ground truth* data as $X = [X_1, X_2, \ldots, X_T]$, each entry representing the number of events at time t. Observers of this process may not see all events. Twitter, for example, makes only a small fraction ($\leq 10\%$) of messages posted on its platform programmatically available, and an organization's defenses may stop the vast fraction of attempted attacks from reaching end users. We refer to the time series of the *observed* data as $Y = [Y_1, Y_2, \ldots, Y_T]$. Figure 1 depicts the cyber-attack prediction problem under our framework.

We model partial observability as a stochastic sampling process, where each event has some probability p to be observed, independent of other events. Therefore, we can represent the observed data as: $Y_t \sim B([X_t], p)$, where $B([X_t], p)$ represents the Binomial distribution with parameters $[X_t]$ for sample size and success probability p.

Under this framework, we derived the following theoretical result [1]:

> **Decay of auto-correlation of the observed signal.** The auto-correlation of the observed signal Y decays monotonically at lower sampling rates. (lower probability p)

From theory to application. In this paper, we aim to apply the theoretical framework – describing the effects of sampling on predictability – to real-world situations in the context of predicting cyber-attacks. We will explore whether the simplified assumptions of the theoretical model stand up to the test of empirical verification. Thus, our current study addresses two research questions:

- **RQ1** Our past work shows a linear (autocorrelation) loss of predictability due to sampling. How does sampling affect non-linear forecasting methods? Are these methods more/less robust than linear ones?
- **RQ2** For analytical purposes, partial observability is modeled as a random sampling, however, this is rarely the case in reality. Does incomplete data affect real-world cyber-attack forecasting scenarios?

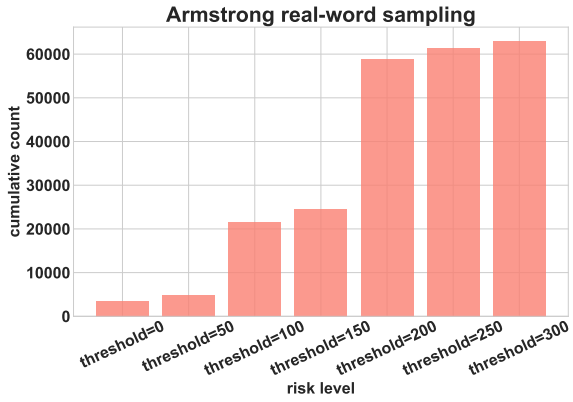

Figure 2: Number of threat messages in Armstrong data based on their risk score threshold

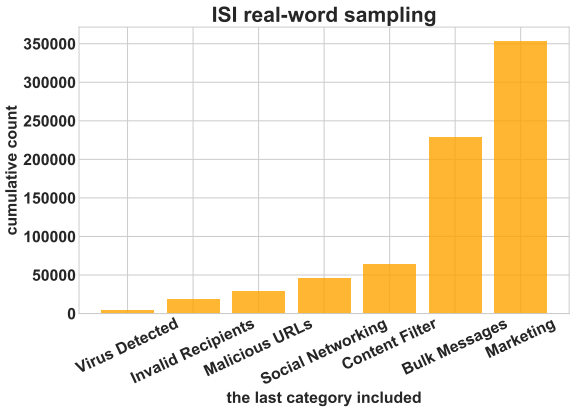

Figure 3: Cumulative number of messages in each category within ISI data.

Our past work primarily focused on linear forecasting models, such as ARIMA, which we linked to linear measures of predictability, like autocorrelation. However, it is possible that due to nonlinear interactions, predictability may not be lost as quickly during sampling. To test this theory, in this work we investigate the possibility of mitigating the loss of information using non-linear relationships in data and metrics. We will identify such cases and compensate for the loss of linear predictability using *nonlinear forecasting models* and nonlinear measures of predictability, such as permutation entropy [6]. We will use neural networks in the prediction tasks and compare their performance to linear forecasting models, such as ARIMA. We will identify applications where linear and non-linear predictability measures diverge and explore whether non-linear models can compensate for the loss of information.

Second, the framework assumes that partial observability can be modeled as a Binomial sampling with a fixed probability of success. However, depending on the application, this may rarely be the case. For example, a Twitter sample obtained through their API may not be a good representation of the data due to non-homogeneous sampling [11]. In malicious email detection, multiple layers of spam filter techniques are combined for advanced protection [18]. In this work, we recreate as close as possible the sampling process over the attempted attacks by sequentially turning on these multiple layers of protection.

4 METHODS

4.1 Data

In this work, we explore email metadata sourced from two distinct enterprises. The first, denoted *ISI*, represents email metadata from the University of Southern California's Information Sciences Institute. The second is an anonymized applied science and technology company we denote as *Armstrong*. In both instances, we analyze the output from the specific enterprise's spam appliance, which does a number of classification and filtering steps in order to protect end-users from potentially malicious emails. We explain these data sets in more detail in the following section.

4.1.1 ISI. The ISI data set consists of daily summaries of all incoming mail activity from August 2018 through July 2019. These summaries contain count data for several threat classifications, as determined by the ISI spam appliance. The average number of total daily malicious emails is 2233 with a minimum of 0, maximum of 7163 and the standard deviation of 1733. The specific message classifications are as follows:

- *Invalid Recipients*: Messages dropped by a Lightweight Directory Access Protocol (LDAP) accept-query process
- *Virus Detected*: Messages with a malicious payload, as detected via the appliance virus scanner
- *Messages with Malicious URLs*: Messages containing malicious URLs, as determined via appliance blacklists and heuristics, in the message body or attachments
- *Stopped by Content Filter*: Messages having contents related to gaming, pornography, weapons, etc.
- *Marketing*: "Gray"-mail marketing messages sent by recognized marketing groups
- *Social Networking*: Messages from social networks, forums, etc.
- *Bulk*: Advertising and marketing messages sent by unrecognized sources.

Security appliances of this company classify unsolicited emails by categories and block any/all categories based on the organization's predefined policy.

4.1.2 Armstrong. This data was collected from a web security appliance deployed in the email infrastructure of the Armstrong organization. Armstrong data ranges from February 2018 to January 2019 and has on average 184 email messages per day. The minimum number of daily attacks is 0 and the maximum number is 1094 with the standard deviation is 163. This security appliance assigns 4 scores to each email, described below:

- *Impostor Score*: An aggregate score of rules and heuristics that compare (e.g.) message credentials and metadata to message contents in order to determine if the sender is who they purport to be

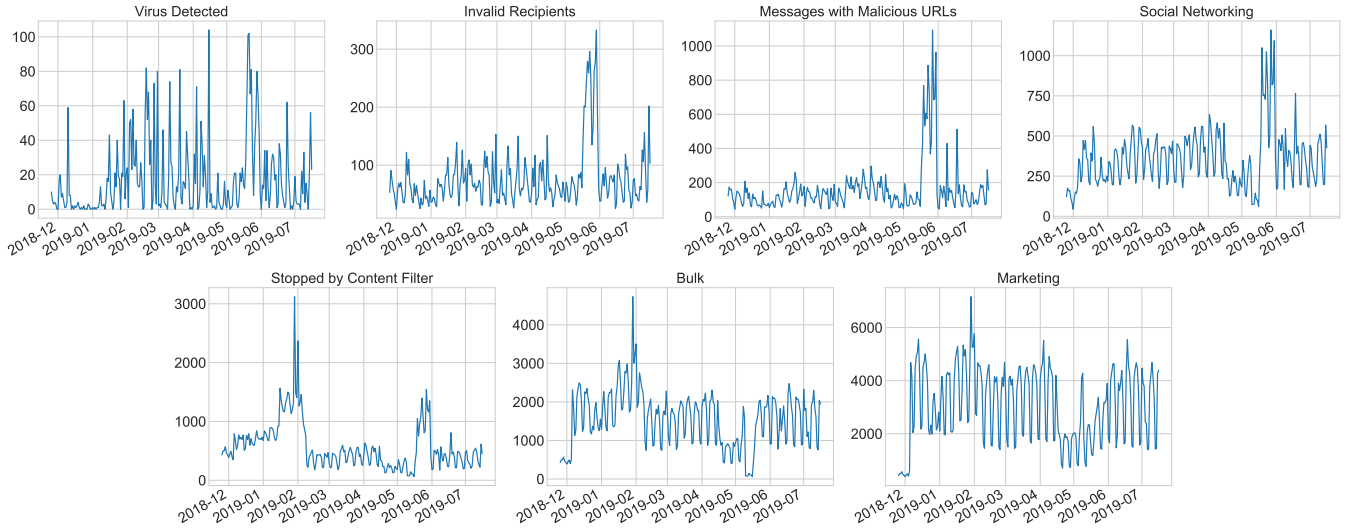

Figure 4: Cumulative time series of categories in ISI data used for real-world sampling. The top leftmost plot is the time series of messages from Virus Detected category only, we add other categories to sampled data one-by-one. The last plot (titled as Marketing) contains messages from all malicious categories.

- *Malware Score*: A determination of confidence that the message contains a malicious attachment or URL
- *Spam Score*: A score based on spam heuristics such as message keywords, metadata agreement, and other traditional methods
- *Phish Score*: An indication of how confident the email appliance is that the email is attempting to elicit information from the recipient maliciously

4.2 Sampling

To understand the effects of sampling on predictability, we used both random sampling and the more realistic real-world sampling.

4.2.1 Random Sampling. For random sampling, we filter events uniformly at random with some probability. This sampling can be modeled as a Bernoulli trial with parameter p. While changing the parameter from 0 to 1 we change the sampling rate. Also, since this process is stochastic for each probability p we sample the time series 50 times.

4.2.2 Real-world sampling. Real-world sampling for each dataset is described below.

ISI. To sample ISI data, we mimic the filtering done by the security appliances. We treat each category as a filter that prevents emails of that type to pass. The order of filters is based on the number of emails in each group: we filtered out the group with the highest number of emails ("Marketing") first, then "Bulk" and so on. We chose this ordering as it roughly aligns with our data; we consider the maliciousness of the message to be inversely proportional to the frequency such message types are observed (i.e. spam is common, therefore not as dangerous). In Figure 3, we can see that more serious attacks like messages with viruses occur less frequently in comparison to marketing or bulk messages. In Figure 4 we can see the daily time series of counts for each threat categories.

Armstrong. We sum over the aforementioned threat scores, and denote this the Risk score. This aggregate score has a potential range of 0 to 400 – although it only ranges from 0 to 300 in the data – is then used as a filter to sample emails. The distribution of the risk scores in the Armstrong data is shown in Figure 2. We filter the data by setting the risk score thresholds to $0, 50, \cdots 300$. For example, threshold 0 allows only email messages with a 0 risk score to pass through the filter, and threshold 300 allows all emails to pass.

4.3 Measuring Predictability

As measures of predictability we will use *auto-correlation*, *permutation entropy* and *prediction error*. The first two are model-free measures of predictability, while that latter requires model-based techniques such as ARIMA or RNN's (Recurrent Neural Network) to predict future values and compute their prediction error.

Auto-correlation captures how well a time series is correlated (using Pearson correlation) with its own time-lagged versions, and has been widely used in finance [10]. Auto-correlation is also linked to the performance of auto-regressive linear models such as ARIMA, with higher auto-correlation leading to better performing auto-regressive models. For each time series X, we find the lag τ, $1 \leq \tau \leq 7$, which has the highest auto-correlation. We then use the found lag to compute the auto-correlation of all the sampled time series Y.

Permutation Entropy [6] captures the complexity of a time series through statistics of its ordered sub-sequences (motifs), and

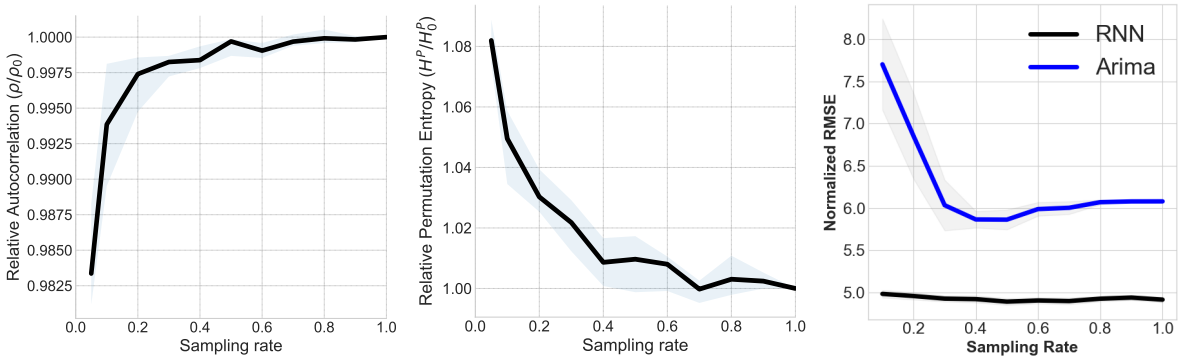

Figure 5: Decay of predictability of (randomly) sampled ISI data. The (a) auto-correlation decreases at low sampling rates and (b) permutation entropy increases; and (c) and error of model-bases techniques increases

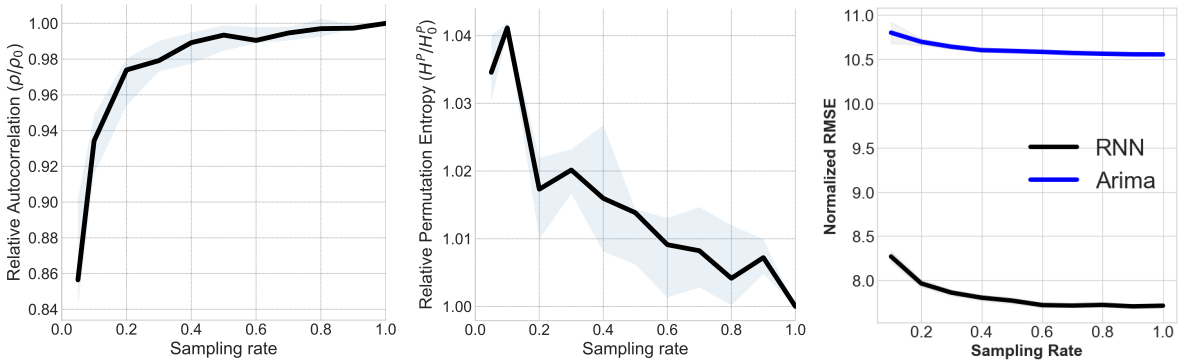

Figure 6: Decay of predictability of (randomly) sampled Armstrong data. The (a) auto-correlation decreases at low sampling rates and (b) permutation entropy increases; and (c) and error of model-bases techniques increases

it is used as a model-free, non-linear indicator of predictability, for example of infectious disease outbreaks [17] and human mobility [19]. Permutation entropy, can be interpreted as the entropy of all the $d!$ possible motifs of fixed size d present in a time series. The motifs represent ordinal patterns that measure the ordinal relation among successive time series values. As an example, if $x_1 = 3, x_2 = 6, x_3 = 1$, then the ordinal pattern of this subsequence $\{x_1, x_2, x_3\}$ is $\phi(x_1, x_2, x_3) = (312)$ because $x_3 \leq x_1 \leq x_2$. Lower permutation entropy is associated with better predictability.

4.4 Forecasting Models

For our model-based predictability measures, we use state-of-the-art forecasting models based on neural network architectures and autoregressive models. The forecasting task is as follows. Given a daily time series describing cyber-attack events, we predict new events occurring in the future. Finally, we quantify the prediction error with the Root Mean Square Error (RMSE). Before feeding the data into the model we normalize the time series using z-scores by negating the values by their mean and dividing them by their standard deviation. In this way, the mean of the time series becomes 0 and the standard deviation of 1. Since the data is already normalized, we do not need to normalize the RMSE and thus, we can compare the RMSE of the same data at different samplings rates.

4.4.1 Auto-regressive Models. For our linear model, we use Auto Regressive Integrated Moving Average (ARIMA) to predict the future points of the time series. The "AR" part of the name indicates using the lagged observed time series as a regressor for predicting future values. The "I" part shows that this model differences the raw observations (subtracts each point in the observation from a previous time-step) to make it a stationary time series. The "MA" part indicates that the model uses a linear combination of lags of the forecast errors as the regression error. An ARIMA model is specified as ARIMA(p,d,q) in which p is the number of autoregressive terms, d is the number of differences applied to make the time series stationary, and q is the order of moving average for the forecast errors. We performed a grid search to choose the set of parameters that minimizes the AIC (Akaike Information Criterion) value of goodness-of-fit.

4.4.2 Neural Networks. Neural network models have grown dramatically in popularity across many applications, including forecasting temporal phenomena. Neural network-based models can capture non-linearities by using multiple layers of non-linear activation functions. The caveat is that such models typically require

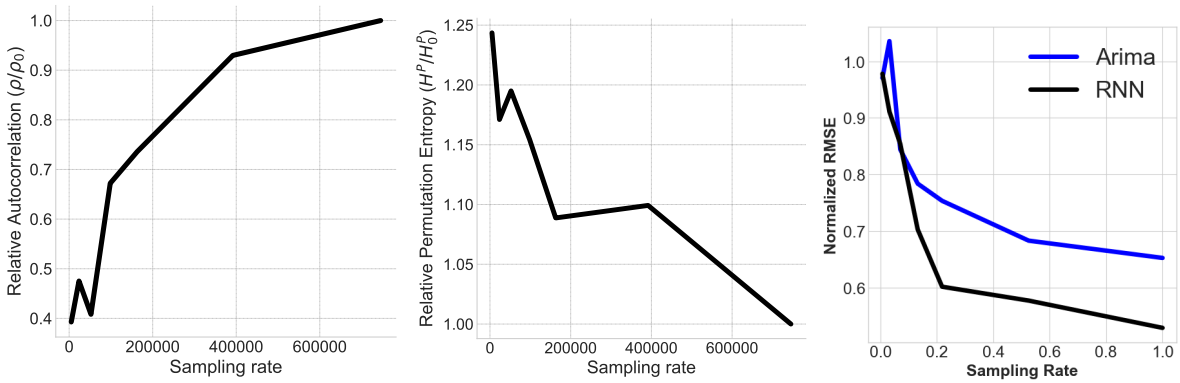

Figure 7: Decay of predictability of ISI data under real-world sampling. The (a) auto-correlation decreases at low sampling rates; (b) permutation entropy increases; and (c) and error of model-bases techniques increases.

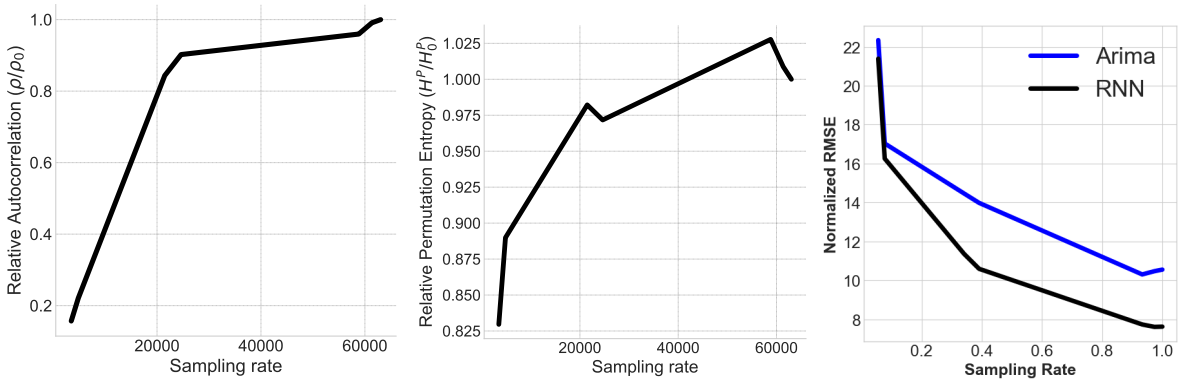

Figure 8: Decay of predictability of Armstrong data under real-world sampling for three out of the four measures. The (a) auto-correlation decreases at low sampling rates; (b) permutation entropy decreases; and (c) and error of model-bases techniques increases.

a large amount of training data to accurately estimate the model parameters. Many variants of recurrent neural network units have been proposed including Long Short Term Memory (LSTM)[9] which has been widely used for time series prediction [13] and even more specifically for cyber attack prediction [2]. Each LSTM unit contains cell state plus input, output and forget gates. This architecture is designed to remember long term dependencies and forget irrelevant information which can be very useful for time series prediction. We use a simple one layer LSTM plus activation function for training the model. Since we had limited amount of data more complex models failed due to overfitting.

We measure predictability with auto-correlation, permutation entropy and prediction error. For prediction error, we use ARIMA as a representative for linear models and LSTM as a representative for non-linear models. We use RMSE (Root Mean Squared Error) to measure prediction error.

In all of the experiments below, for each model, data set, and sampling rate, we first trained the models with the first month of the historical data. Then, we performed next day predictions on the rest of the data set. The models were retrained with the whole historical data every week of predictions. ARIMA generally uses all the historical data as input to predict, however, for the RNNs we only used the 7 previous days as input.

5 RESULTS

In this section, we show our findings surrounding the effect of sampling/filtering on the predictability of cyberattacks. We use the two data sets described in previous sections, ISI and Armstrong. For each data set, we use both random sampling and real-world sampling to compare their effects on predictability. See Methods Section 4.2 for details on the different sampling approaches.

5.1 Random Sampling

We first examine the loss of predictability due to sampling under the random sampling scenario, where the email events have equal, but variable probability to be filtered by the security appliance. Figures 5 and 6 show that the predictability decays at lower sampling rates for both our data sets. This effect holds true no matter the measure of predictability used. Whether it is a model-free measure (such

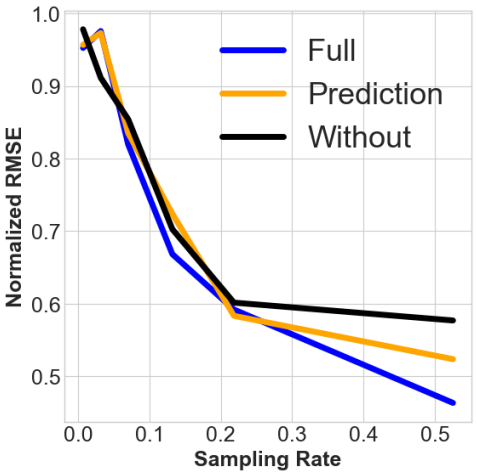

Figure 9: Using raw (unfiltered) data as external signal to predict ISI data (with real-world sampling) Full: raw data as external signal, Prediction: predicted raw data as external signal, without: without external signal

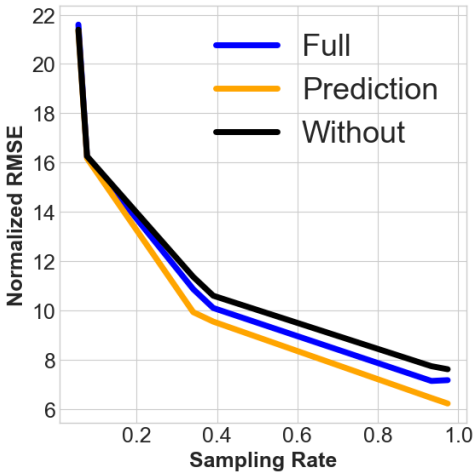

Figure 10: Using raw (unfiltered) data as external signal to predict Armstrong data (with real-world sampling) Full: raw data as external signal, Prediction: predicted raw data as external signal, without: without external signal

as auto-correlation or permutation entropy) or using our model-based measures (prediction error). Research question RQ1 asked whether non-linear methods are more or less robust to the effects of sampling. Comparing the error of the ARIMA vs RNN predictions in both enterprises we highlight two main observations:

- The RNN (non-linear) method clearly outperforms the linear based ARIMA predictions, at all sampling rates, in both enterprises.
- For ISI data (Figure 5), the accuracy of the predictions deteriorates more sharply and abruptly at lower sampling rates (around 40%) compared to the RNN. Whereas for Armstrong (Figure 6), both methods exhibit similar and milder behavior as a function of the sampling rate.

5.2 Real-World Sampling

Under the real-world sampling scenario, predictability also decays as more of the unsolicited emails are removed by the security appliances. Figures 7 and 8 show the loss of predictability under real-world sampling scenarios. Note that the curves look similar to those for random sampling, with the exception of Armstrong's decreasing permutation entropy trend. Yet the effects are much more pronounced, which is evident in the fast incline of error for both ARIMA and RNN methods alike. Addressing question RQ1, we highlight that,

- Real-world sampling affects similarly to both linear and non-linear methods. Specifically, for low sampling rates, both methods have a similar high error.

Regarding question RQ2, we compare real-world sampling vs random sampling.

- Our results suggest that in real-world scenarios, partial observability has a higher impact on predictability.

Together, both results suggest that the practical implications of incomplete data when predicting cyber-attacks not only are significant, but are stronger than what the current theory indicates. We have empirically validated that our models and theories do assert an important limit in predictability. Yet, as our results show, further steps and future work are needed to better quantify, mitigate and predict the effects of partial data in real-world situations.

5.3 Effects of External Signals

Results presented above show how predicting cyber-attacks become increasingly difficult as we consider more dangerous threats (which are fewer in number). At the same time predicting attempted cyber-attacks, which are of less interest, is considerably easier. We can take advantage of this fact and use the prediction of the attempted cyber-attacks as an external signal. We test this theory using the RNN model, since it consistently gave better results than ARIMA. We first forecast the external signal for the next day ($\hat{x}_{t+1}$), then feed 7 historical days of both target data and external data plus $\hat{x}_{t+1}$, to the model. We ran this experiment at different sampling rates (except for the raw data).

Figures 10 and 9 show results for Armstrong and ISI data, both with real-world sampling. In both plots, "Prediction" presents the experiment described above. To compare the results, we also plot predictions without external signals, the same results presented in Figures 8C and 7C, as "Without". The "Full" line represents the error of forecasts using the unfiltered data as an external signal. We expected "Full" and "Without" to be upper bounds and lower bounds for "Prediction" respectively. This is the case in Figure 9; however, in Figure 10 "Prediction" outperforms "Full". This could be because of the high permutation entropy of unfiltered Armstrong which also causes an unexpected trend in Figure 8B. Figures 9 and 10 also show that the raw signal is a better external signal when we are trying to predict signals with similar sampling rates. This

suggests that signals with higher yet similar sampling rates could be the best candidates for external signals.

6 CONCLUSION

Artificial intelligence and machine learning have generated much excitement in recent years with their ability to identify elusive patterns in large volumes of data. The hope for the cybersafety community is that AI systems can be trained to recognize precursors of malicious events, such as cyber-attacks, in the voluminous data streams from open online sources. Our work identifies inherent obstacles to realizing this vision. Importantly, the performance of AI systems is only as good as the training data they are provided. However, if training data represents only a partial observation of the processes of interest, the performance of the predictive model necessarily degrades. In the cybersecurity setting, this means that AI systems trained on filtered data – e.g., successful cyber-attacks or events passing through an organization's firewall – will not be able to accurately predict future cyber-attacks. Using data from two organizations, we demonstrated the loss of predictability as more and more of the data was filtered, for example, by the organization's security appliances.

Our work identifies potentially fruitful avenues for future research. Figures 8 and 7 show that non-linear models such as RNN do have an advantage over linear models, however, this gap decreases for lower sampling rates. Suggesting that in practice, the full advantage of these methods is not fully exploited. Future work will be devoted to further increase this gap, investigating the possibility of mitigating the loss of information using new non-linear, neural network based methods. Second, our framework is very flexible and general, and can be applied to a myriad of applications beyond cybersecurity, for example, predicting armed conflict, protest, and political unrest around the world.

ACKNOWLEDGMENTS

This work was supported by the Office of the *Director of National Intelligence* (ODNI) and the *Intelligence Advanced Research Projects Activity* (IARPA) via the *Air Force Research Laboratory* (AFRL) contract number FA8750-16-C- 0112, and by the *Defense Advanced Research Projects Agency* (DARPA), contract number W911NF-17-C-0094. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, DARPA, or the U.S. Government.

REFERENCES

[1] Andrés Abeliuk, Zhishen Huang, Emilio Ferrara, and Kristina Lerman. 2019. Predictability limit of partially observed systems. *arXiv preprint arXiv:2001.06547* (2019).

[2] Santosh Aditham, Nagarajan Ranganathan, and Srinivas Katkoori. 2017. LSTM-based memory profiling for predicting data attacks in distributed big data systems. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 1259–1267.

[3] Luca Allodi and Fabio Massacci. 2017. Security Events and Vulnerability Data for Cybersecurity Risk Estimation. *Risk Analysis* 37, 8 (Aug 2017), 1606–1627. https://doi.org/10.1111/risa.12864

[4] Mohammed Almukaynizi, Ericsson Marin, Eric Nunes, Paulo Shakarian, Gerardo I Simari, Dipsy Kapoor, and Timothy Siedlecki. 2018. DARKMENTION: A Deployed System to Predict Enterprise-Targeted External Cyberattacks. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 31–36.

[5] Jonathan Z Bakdash, Steve Hutchinson, Erin G Zaroukian, Laura R Marusich, Saravanan Thirumuruganathan, Charmaine Sample, Blaine Hoffman, and Gautam Das. 2018. Malware in the future? Forecasting of analyst detection of cyber events. *Journal of Cybersecurity* 4, 1 (2018), tyy007.

[6] Christoph Bandt and Bernd Pompe. 2002. Permutation Entropy: A Natural Complexity Measure for Time Series. *Phys. Rev. Lett.* 88 (Apr 2002), 174102. Issue 17.

[7] Ashok Deb, Kristina Lerman, and Emilio Ferrara. 2018. Predicting Cyber-Events by Leveraging Hacker Sentiment. *Information* 9, 11 (2018), 280.

[8] Palash Goyal, KSM Hossain, Ashok Deb, Nazgol Tavabi, Nathan Bartley, Andrés Abeliuk, Emilio Ferrara, and Kristina Lerman. 2018. Discovering signals from web sources to predict cyber attacks. *arXiv preprint arXiv:1806.03342* (2018).

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[10] Kian-Ping Lim, Weiwei Luo, and Jae H Kim. 2013. Are US stock index returns predictable? Evidence from automatic autocorrelation-based tests. *Applied Economics* 45, 8 (2013), 953–962.

[11] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose.. In *ICWSM*.

[12] Ahmet Okutan, Shanchieh Jay Yang, and Katie McConky. 2017. Predicting cyber attacks with bayesian networks using unconventional signals. In *Proceedings of the 12th Annual Conference on Cyber and Information Security Research*. ACM, 13.

[13] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971* (2017).

[14] Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. 2015. Weakly Supervised Extraction of Computer Security Events from Twitter. *Proceedings of the 24th International Conference on World Wide Web - WWW '15* (2015).

[15] Carl Sabottke, Octavian Suciu, and Tudor Dumitras. 2015. Vulnerability disclosure in the age of social media: exploiting twitter for predicting real-world exploits. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*. 1041–1056.

[16] Anna Sapienza, Sindhu Kiranmai Ernala, Alessandro Bessi, Kristina Lerman, and Emilio Ferrara. 2018. DISCOVER. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (2018). https://doi.org/10.1145/3184558.3191528

[17] Samuel V Scarpino and Giovanni Petri. 2019. On the predictability of infectious disease outbreaks. *Nature communications* 10, 1 (2019), 898.

[18] Vivek Shandilya, Fahad Polash, and Sajjan Shiva. 2014. A Multi-LAYER ARCHITECTURE FOR SPAM-DETECTION SYSTEM. *Computer Science & Information Technology* (2014), 193–200.

[19] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[20] Nazgol Tavabi, Palash Goyal, Mohammed Almukaynizi, Paulo Shakarian, and Kristina Lerman. 2018. Darkembed: Exploit prediction with neural language models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[21] Gordon Werner, Shanchieh Yang, and Katie McConky. 2017. Time series forecasting of cyber attack intensity. In *Proceedings of the 12th Annual Conference on cyber and information security research*. ACM, 18.

[22] Jinyu Wu, Lihua Yin, and Yunchuan Guo. 2012. Cyber Attacks Prediction Model Based on Bayesian Network. *2012 IEEE 18th International Conference on Parallel and Distributed Systems* (Dec 2012). https://doi.org/10.1109/icpads.2012.117

[23] M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu. 2018. Modeling and Predicting Cyber Hacking Breaches. *IEEE Transactions on Information Forensics and Security* 13, 11 (Nov 2018), 2856–2871.

[24] Zhenxin Zhan, Maochao Xu, and Shouhuai Xu. 2015. Predicting cyber attack rates with extreme values. *IEEE Transactions on Information Forensics and Security* 10, 8 (2015), 1666–1677.