POSTER: Cyber Attack Prediction of Threats from Unconventional Resources (CAPTURE)*

Ahmet Okutan
Computer Engineering
Rochester Institute of Technology
Rochester, NY, USA
axoeec@rit.edu

Gordon Werner
Computer Engineering
Rochester Institute of Technology
Rochester, NY, USA
gxw9834@mail.rit.edu

Katie McConky
Industrial and Systems Engineering
Rochester Institute of Technology
Rochester, NY, USA
ktmeie@rit.edu

Shanchieh Jay Yang
Computer Engineering
Rochester Institute of Technology
Rochester, NY, USA
jay.yang@rit.edu

ABSTRACT

This paper outlines the design, implementation and evaluation of CAPTURE - a novel automated, continuously working cyber attack forecast system. It uses a broad range of unconventional signals from various public and private data sources and a set of signals forecasted via the Auto-Regressive Integrated Moving Average (ARIMA) model. While generating signals, auto cross correlation is used to find out the optimum signal aggregation and lead times. Generated signals are used to train a Bayesian classifier against the ground truth of each attack type. We show that it is possible to forecast future cyber incidents using CAPTURE and the consideration of the lead time could improve forecast performance.

CCS CONCEPTS

• **Security and privacy** → *Intrusion detection systems*;

KEYWORDS

Cyber-security, Unconventional signals, Bayesian Networks

1 INTRODUCTION

As computing and networking technologies are being embedded into our professional and personal activities, the impact of various and evolving cyber attacks continues to rise. This calls for an anticipatory capability to forecast potential cyber attacks before they happen [5, 6]. Such a capability requires examining beyond the traditional observables of malicious activities as they occur. This paper develops an automated, 24x7 system named CAPTURE that uses a broad range of unconventional signals from public data sources, including GDELT and Twitter, as well as reported cyber incidents, to forecast different types of cyber attacks. Signals are named unconventional as they are not necessarily specific to a target entity or any cyber attack, but might be indicative for potential future cyber incidents towards the entity.

Sometimes an attacker is angered by a news release and motivated to launch a cyber attack towards a target entity [1]. Once the attacker has the intent, it might take some time to have the opportunity to execute it. This time could be different for each attack type due to the reconnaissance needed in the exploration phase or the responsiveness of the attacker. CAPTURE defines the lead time (Lt) as the time elapsed between the most recent observation of a significant signal correlation and the execution of a cyber attack. On the other hand, using different aggregation periods (At) for a signal, *i. e.*, aggregating over the last day, week or month could affect its predictive power.

CAPTURE uses a novel and systematic methodology to determine appropriate Lt and At for each unconventional signal with respect to each attack type and each target entity. This paper applies ARIMA [8] to the reported binary cyber incident data for each attack type. Whether a specific type of cyber attack occurs each day is treated as a time series and CAPTURE considers the forecasted occurrence of each attack type and its associated probability as base signals to reflect the potential pattern of attack occurrences. The ARIMA-based signals along with other unconventional signals configured with various Lt and At parameters are used by CAPTURE to perform an ensemble forecasting of future cyber incidents. CAPTURE uses these signals along with the reported cyber incidents to train a Bayesian model for each attack type of each target entity. It continually monitors and aggregates the signal data based on the systematically selected Lt and At, and forecasts future cyber incidents.

2 SYSTEM ARCHITECTURE

Figure 1 shows the overall system architecture of CAPTURE. Consider a set of attack types $A = \{A_1, A_2, ...A_m\}$, a set of target entities $E = \{E_1, E_2, ...E_t\}$, a set of unconventional signals $U = \{U_1, U_2, ...U_n\}$, and a set of Time Series signals $V = \{V_1, V_2, ...V_{2m}\}$. Using binary observations B_i of each A_i as time series until day d, ARIMA is applied to forecast the number of cyber attacks that

*This research is 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 FA875016C0114.

CCS '17, October 30-November 3, 2017, Dallas, TX, USA

ACM ISBN 978-1-4503-4946-8/17/10.
https://doi.org/10.1145/3133956.3138834

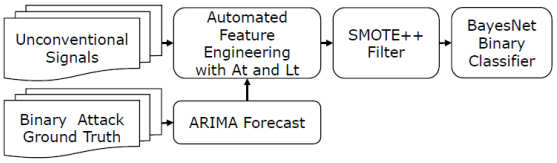

Figure 1: The brief overall architecture of CAPTURE.

will occur during the day $d + 1$. For each forecast a confidence value in the 95% confidence interval is generated. For each m attack types, forecasted daily counts and the associated confidences are used as signals to predict cyber attacks generating $2m$ Time Series signals in V. An instance of the CAPTURE architecture as shown in Figure 1 is created for each A_i and E_i combination. A key to the CAPTURE design is the ability to identify the optimal lead time and aggregation level for each signal. To perform this automated feature configuration a cross correlation is applied between the daily binary observations of B_i and each signal and the time lags for each signal are found in terms of days. Using the largest negative time lag as the lead time (Lt_i) and the difference between the largest negative lag and the smallest negative lag as aggregation time (At_i), the values of each signal are recalculated for each day. If there is no significant negative time lag, the signal is not used to predict attacks for A_i. If there is only one negative lag, the lag is used as At and Lt is set to 0 (See Figure 2). Repeating this process for each signal in U and V, a new set of signals say $Z = \{Z_1, Z_2, ...Z_k\}$ are generated for each attack type A_i where $k <= n + 2m$. A Bayesian network is a directed acyclic graph that is composed of k random variables and e edges that show the dependencies among these variables.

Let $Z = \{Z_1, Z_2, ...Z_k\}$ be k random variables (unconventional signals plus time series signals) with nominal or numeric values for a Bayesian network. The CAPTURE system trains a Bayesian classifier for each attack type A_i, using the set of aggregated signals in Z. Due to the nature of the cyber data, the data sets for some

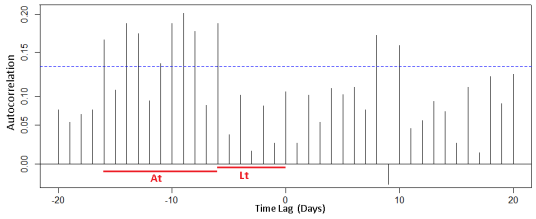

Figure 2: An example cross correlation that shows how the aggregation (At) and lead (Lt) times are found.

attack types are highly imbalanced. CAPTURE uses a novel filtering technique that is applied to the highly imbalanced data sets to make the data sets more balanced. The technique is named SMOTE++ and is built upon a previous technique called SMOTE [2]. It uses a combination of the majority under sampling, instance weighing, and minority over sampling (synthetic instance generation) techniques together. It uses k-means clustering starting with $k = 2$, and increments k until a cluster with minority instances is found. Then, it removes p percent of majority instances that are nearest to the center of the minority cluster in terms of the Euclidean distance where the parameter p is tuned. SMOTE++ improves the prediction performance of the CAPTURE system significantly on some highly skewed cyber data sets.

2.1 Signals Used

2.1.1 ***GDELT Signals.*** GDELT [3] provides the mentions of events across all of its source documents. Each event has an associated average tone and it can take a value between -100 (extremely negative) and +100 (extremely positive).

- **GDELT Event Mentions (GEM)**: The mentions of all events that have a negative average tone are counted.
- **GDELT Event Tone (GET)**: The negative event tones in GDELT are summed up for the requested date interval.
- **GDELT Event Articles (GEA)**: The number of GDELT documents containing one or more mentions of a negative event are counted.

In GDELT, a numeric score is assigned to each event to capture the potential impact of the event on the stability of the country where it occurred (Goldstein scale). Four instability signals are calculated based on the Goldstein scale, with the hypothesis that an increase in the instability may result in an increase in the cyber attack probability. All instability signals are calculated for a given date interval and the country where the target entity operates.

- **Goldstein Event Count (GGC)**: The number of events associated with a negative Goldstein score.
- **Goldstein Score Average (GGA)**: The average of the Goldstein scores.
- **Goldstein Score Less Than Zero (GLZ)**: The average of scores that are less than zero.
- **Goldstein Score Less Than Minus Five (GLF)**: The average of scores that are less than minus five.

2.1.2 ***Twitter Events (TEC).*** The number of significant malicious events are counted based on the data from a previous study by Ritter *et al.* [7].

2.1.3 ***Level of Mentions of Entities (LME).*** Increased discussion of a target entity may indicate an increased surveillance towards the entity and could lead to increase in the likelihood of cyber attack towards it. To generate LME, the number of mentions of the target entity and its related keywords are counted in Twitter.

2.1.4 ***Sentiment Signals.*** According to Baumeister *et al.* a threat actor could be motivated to launch an attack based on anger [1]. Indications of outrage towards a target entity could be predictive for cyber incidents towards that entity. Two sentiment scores (named Affect (AFF) and Intensity (INT)) are used that are calculated by a private company in the industry for each of the three target entities, *i. e.*, anonymized K5 company, Defense, and Banking sectors. Six sentiment signals are used (two for each of the three targets) represented as *K5_AFF*, *K5_INT*, *DEF_AFF*, *DEF_INT*, *BANK_AFF*, and *BANK_INT*.

2.1.5 ***Time Series Signals.*** Applying ARIMA to the previously observed binary ground truth data of a given attack type, we forecast whether a cyber attack will occur at a future date. Using ARIMA forecasting, two values are generated for the future date,

i. e., the forecasted number of attacks (F) and the confidence associated with the forecast (C). These two values are generated daily for each attack type and used as signals for each trained Bayesian prediction model.

3 EVALUATION RESULTS

The CAPTURE system is evaluated based on the binary ground truth data of an anonymized company aliased as K5. The signals defined in 2.1 are used as inputs for the Bayesian classifier which is trained against the K5 ground truth on a daily basis. The signals until a date d are used with a Bayesian classifier to forecast the cyber attacks for the date $d + 1$. A separate Bayesian classifier is trained for each attack type defined for K5, *i. e.*, Malware (MW), Scan, Defacement (Def), Malicious Email (ME), Malicious URL (MU), and Denial of Service (DoS). The signals described in 2.1 are generated daily for all dates between July 1 2016 and Jan 1 2017. For each defined attack type, an auto cross correlation (CCR) analysis is carried out between each signal and the binary ground truth of the attack type. Using the CCR results, statistically significant time lags between the signal and the ground truth are determined. This process is repeated for each signal and attack type, and the optimum At and Lt parameters shown in Table 1 are found for each signal and attack type pair. We observe that different signals might have quite different aggregation and lead times for different attack types. The signals defined in 2.1 are recalculated based on the found At and Lt for each attack type. The recalculated signals are then used to train a Bayesian classifier for each attack type using its corresponding binary ground truth. The area under the ROC curve (AUC) gets better, when the true positive rate is high and false positive rate is low. Therefore, AUC is used to check the performance of each Bayesian model for each attack type. FilteredClassifier in Weka [4] is used to apply SMOTE++ and train a Bayesian classifier for each attack type with 10×10 folds cross validation. When the aggregated signals (Signals with At and Lt) are used, the AUC values shown in the first row of Table 2 are found for each attack type. To check the significance of using Lt, a new set of signals (Signals with At) were calculated based on the At found with CCR and $Lt = 0$. With these signals, the AUC values of the Bayesian models for each attack type are shown in the second row of Table 2. We apply a t-test with a p-value of 0.05 and observe that the differences in the AUC values are significant for the Scan, Def and DoS attack types. We observe that using a lead time for the aggregated signals could improve the performance for some attack types, but believe that further research with different data sets is needed to generalize our findings.

Table 1: *At-Lt* of signals found with cross correlation for each attack type. TS denotes Time Series signals (TS_MW_F and TS_MW_C denote the forecasted count and confidence respectively for Malware attacks).

Signal	MW	Scan	Def	MU	ME	DoS
GEM		11-1		14-3		
GET		12-1		12-5		
GEA		11-1		14-3		
TEC		5-16	20-1	5-14		
LME		2-4			14-1	20-1
GGC	20-1	16-1		6-1	6-13	9-8
GGA	20-1	16-1		6-1	6-13	9-8
GLF	20-1	16-1		6-1	6-13	9-8
GLZ	20-1	16-1		6-1	6-13	9-8
DEF_AFF		19-1			20-1	12-1
DEF_INT		6-6	13-4		9-1	14-1
BANK_AFF	5-1	11-1	7-5	4-1	7-14	17-1
BANK_INT		5-5	11-5		8-1	
K5_AFF			11-6	4-1	5-2	1-1
K5_INT	7-1	6-4	13-4		3-7	8-12
TS_MW_F	15-16	7-5	9-1			16-5
TS_MW_C	8-4	15-6	17-4		8-11	17-3
TS_SCAN_F					20-1	17-4
TS_SCAN_C		4-1	4-1	5-1	17-4	15-6
TS_DEF_F		16-1	9-1		12-8	11-10
TS_DEF_C		16-1	15-1			8-13
TS_MU_F		2-1	4-15		18-3	16-3
TS_MU_C	20-1	17-1	12-7	7-1	17-4	12-3
TS_ME_F		1-1	20-1		1-1	13-7
TS_ME_C	3-1	3-1		6-1		20-1
TS_DOS_F			18-1		8-11	3-4
TS_DOS_C			12-8			6-15

Table 2: The AUC values found for each attack type for two cases.

AUC	MW	Scan	Def	MU	ME	DoS
Signals with At and Lt	0.51	0.67	0.69	0.56	0.64	0.82
Signals with At	0.51	0.61	0.66	0.57	0.62	0.78

4 CONCLUSION

The evaluations with the ground truth data of the anonymized company K5 illustrates the capability of CAPTURE. The results show that unconventional signals that are not directly related to a target entity, could be used to forecast cyber attacks towards it. Furthermore, cross correlation analysis is useful in determining the aggregation and lead times of the signals and using a lead time could improve the forecast performance.

REFERENCES

[1] Roy F. Baumeister, Kathleen D. Vohs, C. Nathan DeWall, and Liqing Zhang. 2007. How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, Rather Than Direct Causation. *Personality and Social Psychology Review* 11, 2 (2007), 167–203.

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal Of Artificial Intelligence Research* 16, 1 (June 2002), 321–357.

[3] GDELT. 2017. The GDELT Project. (2017). http://www.gdeltproject.org/ [Online; accessed 6-February-2017].

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11, 1 (Nov. 2009), 10–18.

[5] B. Munkhdorj and S. Yuji. 2017. Cyber attack prediction using social data analysis. 23 (01 2017), 109–135.

[6] A. Okutan, S. Yang, and K. McConky. 2017. Predicting Cyber Attacks With Bayesian Networks Using Unconventional Signals. In *Proceedings of the 12th Annual Conference on Cyber and Information Security Research (CISRC '17)*. ACM.

[7] A. Ritter, E. Wright, W. Casey, and T. Mitchell. 2015. Weakly Supervised Extraction of Computer Security Events from Twitter. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. Geneva, Switzerland, 896–905.

[8] G. Werner, S. Yang, and K. McConky. 2017. Time Series Forecasting of Cyber Attack Intensity. In *Proceedings of the 12th Annual Conference on Cyber and Information Security Research (CISRC '17)*. ACM.