Annals of Telecommunications
https://doi.org/10.1007/s12243-023-00993-4

An approach for detecting multi-institution attacks

Saif Zabarah[1] · Omar Naman[1] · Mohammad A. Salahuddin[1] · Raouf Boutaba[1] · Samer Al-Kiswany[1,2]

Received: 10 May 2023 / Accepted: 19 September 2023

Abstract
We present Soteria, a data processing pipeline for detecting multi-institution attacks. Soteria uses a set of machine learning techniques to detect future attacks, predict their future targets, and rank attacks based on their predicted severity. Our evaluation with real data from Canada-wide academic institution networks shows that Soteria can predict future attacks with 95% recall rate, predict the next targets of an attack with 97% recall rate, and detect attacks in the first 20% of their life span. Soteria is deployed in production and is in use by tens of Canadian academic institutions that are part of the CANARIE IDS project.

Keywords Multi-institution attacks · Cybersecurity · Threat intelligence · Intrusion detection

1 Introduction

Multi-institution attacks look for vulnerabilities in large number of nodes located in multiple institutions. These attacks cause significant financial loses and information leaks. For instance, the loss caused by NotPetya attack exceeds $10 billion [1], and the WannaCry ransomware attack affected more than 200,000 computers in 150 countries [2] causing millions in damages.

Defending against multi-institution attacks is complicated because the target nodes are managed by tens of independent security teams. Detecting these attacks requires timely information sharing between institutions and analysis of potential threats. This is further complicated by the following.

First, the vulnerabilities an attacker can exploit change continuously making it harder to automate the defence mechanisms. Worst yet, it may take months until vulnerabilities are patched. For instance, the WannaCry ransomware attack targeted a vulnerability in old Windows versions, for which a patch had been released more than 2 months before the attack [2]. Second, the attacks often happen in a short period of time. For instance, our data set (Sect. 4) shows that attacks can initiate millions of connections in just 24 h. This short duration leaves little time for the cybersecurity personnel to detect, analyze, and deploy a defence mechanism. Third, large number of attacks happen at the same time. As the current process for analyzing the attacks involves cybersecurity personnel, the number of attacks that can be inspected at the same time is limited. This prolongs the attack detection time and increases the time window in which an attack can cause severe damage.

For academic institutions, defending against multi-institution attacks is harder because they operate large and constantly changing networks (e.g., research projects or students spawning new nodes and services), and they have smaller budgets and cybersecurity teams. This makes academic institutions a prime target for attacks. For instance, in Canada, cybercrime caused an average of $9.25 million in losses per academic institution in 2019 [3] and 46% of the Canadian institutions reported a cybersecurity incident in 2017, which was the second highest impacted sector in Canada [4].

The current defence technique against these attacks is inadequate. The main approach relies on sharing intelligence between cooperating institutions and using public databases

✉ Saif Zabarah
szabarah@uwaterloo.ca

Omar Naman
omar.naman@uwaterloo.ca

Mohammad A. Salahuddin
mohammad.salahuddin@uwaterloo.ca

Raouf Boutaba
rboutaba@uwaterloo.ca

Samer Al-Kiswany
alkiswany@uwaterloo.ca

[1] Computer Science, University of Waterloo, Waterloo, Ontario, Canada

[2] Acronis Research, Waterloo, Ontario, Canada

Published online: 09 October 2023

that list recent Indicators of Compromise (IoCs). This approach is slow to detect an attack, and the information shared is often limited due to regulatory and privacy policies.

We present Soteria, a novel data processing pipeline for detecting multi-institution attacks. Soteria overcomes the shortcomings of the current defence approaches. Soteria collects minimal information from cooperating institutions, mainly information about connections to an institution. It then uses a novel combination of machine learning (ML) techniques to detect current attacks and predict future attacks. Soteria also predicts the next targets of an attack and identifies the larger-scale attacks (i.e., the more severe attacks). These findings help focus each institution's limited resources on the most severe attacks that are targeting them now or in the near future.

Soteria is carefully designed to be able to scale to hundreds of institutions and detect attacks in a timely manner. Soteria uses graph analysis to extract features, linear regression to detect future attacks, and time-series analysis to predict the next targets of an attack. We use a bidirectional long short-term memory recurrent neural network with attention mechanism (ABiLSTM) to predict the future targets of an attack. Finally, to predict the severity of an attack, we capture static and dynamic features of the generated graphs and use normalization and reduction techniques to compute a severity indicator.

Through our study of the dataset and the exploration of different techniques, we offer a number of insights. We found that to accurately predict the next target of an attack, the used mechanism should capture (1) the relationships between institutions, as institutions with similar characteristics (e.g., institution size, services offered, and security posture) are usually targeted together; (2) the sequence of the attack; and (3) the level of activity of an attacker. One would expect that ML techniques that predict events that occur together, such as a co-occurrence matrix [5], to be efficient in detecting multi-institution attackers (MIAs). Surprisingly, based on our experiments, these techniques are not efficient; this is because a co-occurrence matrix does not capture the sequence or the level of activity of an attack. Techniques, such as a unidirectional LSTM which predicts a sequence of events, performed better, but did not achieve high accuracy in predicting the next target because attacks do not follow the same sequence in every attack incident.

It was interesting that a simple linear regression model over the right set of features is highly effective in detecting an attack, and hosts that contact more than one institution are, with high probability, initiating an attack. Surprisingly, using a short history of recent connections detects an attack faster than when using the data from the last 24 h.

Soteria has been deployed in production for the last year as part of the Canadian Network for the Advancement of Research, Industry and Education (CANARIE) Intrusion Detection System (IDS) program. CANARIE [6] is a Canada-wide backbone network connecting academic institutions. The CANARIE IDS is serving over 100 institutions in Canada. Over the last year, Soteria has identified numerous severe multi-institution attacks.

Our evaluation with real data from the CANARIE IDS participating institutions shows that external IPs communicating with more than one institution are 95% likely to be MIAs. Our evaluation also shows that Soteria detects future attacks with up to 95% recall rate and within the first 20% of the attack's lifetime. Finally, Soteria detects and notifies 97% of institutions that will be targeted in the future before the attacker initiates a connection to that institution.

This paper is an extension of our previous publication in [7]. In this paper, we detail the system design including a discussion on training and parameter tuning of the future target predictor and the reporting and dashboard mechanisms implemented. We present an evaluation of the Soteria runtime and showcase examples of common large-scale multi-institution attacks. Finally, we extend our survey to discuss current methods used by institutions to detect multi-institution attackers and discuss their limitations.

The rest of this paper is organized as follows: In Sect. 2, we survey related work. In Sect. 3, we detail the design of Soteria. In Sect. 4, we evaluate the accuracy of predicting a future attack and the future targets of an attack. We present our concluding remarks and plans for future work in Sect. 5.

2 Related work

Reconnaissance detection Reconnaissance attacks try to scan systems looking for vulnerabilities. Previous efforts on reconnaissance detection are limited to detecting port scans. For instance, Udhayan et al. [8] detect port scanning attempts by applying a set of heuristics on the connection timing and TCP header fields. Allen et al. [9] note that port scanning tools leave detectable features in the generated requests. For instance, they may contain invalid data or header information. Given the short list of scanning tools, they explore inspecting packets for special markers to identify port scanning attempts.

Heavy hitters detection Previous efforts attempted to detect large-scale attacks known as heavy-hitter attacks. Heavy-hitters communicate with an unusually large number of hosts. The main challenge in detecting heavy hitters is handling large amounts of data. Previous efforts [10, 11] resorted to filtering out low cardinality hosts and sampling. Yang et al. [12] summarize traffic measurements by using mergeable data structures and aggregating the summaries at the operator center. The merged summary is then used to detect heavy hitters.

Intrusion detection systems IDSs monitors network traffic to detect malicious activities. Examples of IDSs are ZEEK [13], Snort [14], and Suricata [15]. IDS is primarily used for detecting attacker in a single institution, mostly based on network rules and policies. Soteria detects MIAs using ZEEK connection logs.

Current approach for detecting multi-institutional attacks The main approach for handling multi-institution attacks currently is through sharing attack intelligence [16]. Cyber threat intelligence takes on many forms [16].

- Strategic threat intelligence: the big picture of past, current, and future trends in the threat landscape.
- Tactical threat intelligence: techniques, tools, and tactics of the threat actors.
- Operational threat. intelligence: specifics about the nature and purpose of threats and actors.
- Technical threat intelligence: technical indicators about the campaigns.

The technical threat intelligence category is of concern to us as our goal is to provide live threat details.

Shared intelligence can either come from peer institutions that cooperate on detecting attacks or from public databases. A number of databases offer information about known attacks and provide a list of malicious IPs, such as AbuseIPDB [17] and VirusTotal [18]. Databases, such as CVE [19] and CWE [20], offer information about vulnerabilities in software.

Databases that provide lists of malicious IPs and domains are manyfold. Some are community-shared information, such as AbuseIPDB [17]. AbuseIPDB is a cybersecurity intelligence aggregation site that collects threat information from non-vetted sources (i.e., reliability of their information, is not confirmed). Others provide vetted and reliable information such as SolarWinds [21]. There are also sources that aggregate information from multiple databases, such as Virus Total [18].

Databases that list vulnerabilities discovered and the effected software versions are databases and sites that list known network software vulnerabilities, provide descriptions, and possible steps for remediation. For example, CVE [19] is a program to identify, define, and catalog publicly disclosed cybersecurity vulnerabilities. There are also DBs and sites, such as the Common Weakness Enumeration (CWE) [20], which offer a community-developed list of software and hardware weakness types.

Skopik et al. [22] surveyed technical, legal, regulatory, and organizational dimensions of threat intelligence sharing. They found that it is crucial to facilitate the dimensions for quick information distribution of threat information.

Threat intelligence sharing can be utilized in three ways:

- Manually: Cybersecurity specialists will receive alerts in an email or any other form of messaging. Then, they will decide to react to this information based on their knowledge of their network.
- Automated: In a completely automated fashion, security information and event management (SIEM), IDS, intrusion prevention system (IPS), or automated firewalls will receive threat feeds from reliable source and will immediately act on that information without user input. For example, malicious IPs will be blocked.
- Hybrid: Threat feeds will be fed to a SIEM, IDS, IPS, and other customized systems. These systems will use the fed data to search, correlate, and analyze, to provide alerting, visualization, and threat hunting capabilities, but action will be the responsibility of the cybersecurity personnel. For example, [23] proposes ECOSSIAN, a complex system that aggregates threat information from multiple National Security Operation Center (NSOC), which also aggregate threat information from Organization Security Opertation Center (OSOC).

Unfortunately, while helpful, these approaches do not adequately protect against multi-institution attacks. That is because the cycle starting from detection to information sharing is often slow, leading to delayed reaction to active attacks. Their networks are large and continuously changing, and there are a large number of attack attempts which overwhelms institution staff. They are also hesitant to share information [16] due to

- Privacy and liability concerns: Needing to share information that could be private.
- Nothing valuable to contribute: Organizations are missing a lot of the threats that may not be important to them but are important to others.
- Too much noise: There are so many malicious activities happening, it is not easy to report them all.
- Been hacked: The fear of sharing breach details more broadly than necessary.

Soteria aims to overcome the shortcomings of the previous techniques. It relies on minimal information shared by institutions to automatically detect multi-institution attacks. It also prioritizes information provided to security officers, by identifying the most severe attacks. Furthermore, it identifies the next potential targets for the attack. This information is used to notify the targeted institution before the attack reaches that institution. Finally, Soteria is privacy conscious and does not share institution-specific information.

3 System design

Soteria data processing pipeline is organized into five stages: (i) feature extraction, (ii) attack detection, (iii) severity estimation, (iv) next target detection, and (v) report generation. Figure 1 shows the Soteria system architecture.

Institutions periodically submit logs of the recent communication activities on their networks. The feature extraction step (Fig. 1) analyzes the data to extract a vector of features. The extracted features are leveraged by the attack detection step that uses an ML model to predict potential attacks. The ML model also outputs additional metrics to help with the next two steps. The severity estimation step uses the metrics calculated in the previous steps to estimate the severity of the predicted attacks. This step helps identify severe attacks. The next target prediction step uses deep learning to identify the next targets of the predicted attacks. Finally, Soteria combines the results of the attack detection, severity estimation, and next target prediction steps into user reports. The rest of this section details the design of each step.

3.1 Institutional data

Sharing connection and infrastructure information between institutions is complicated due to regulatory restrictions and privacy-related concerns. This is the case for the academic institutions participating in the CANARIE IDS program. The institutions periodically share connections logs collected by ZEEK. In addition to ZEEK connection logs, each institution identifies the IP addresses it owns. We present the details of the data set we use in Sect. 4.

Each row in a ZEEK connection log lists information about a connection. In our work, we use three fields for each connection:

- $id.orig\_h$: IP address of the host starting the connection.
- $id.resp\_h$: IP address of the host responding to the connection.
- ts: the time stamp when the connection occurred.

3.2 Feature extraction

To identify attacks on multiple institutions, we build a directed and weighted graph representing all the connections in the ZEEK connection logs. Each IP address represents a vertex. We add a directed edge from a source to a destination between two vertices that had one or more connections. Each edge has a weight. The weight is equal to the number of connections with the same direction between the two vertices. Vertices are labeled as internal vertex if they belong to an institution, or external vertex otherwise. Unfortunately, this approach for generating a graph creates enormous graphs that are challenging to analyze in a timely manner. For instance, for our data set, this approach resulted in an enormous graph with over 26.5 million vertices and 1.4 billion edges.

To reduce the graph size without losing information relevant to the attack, we do the following. First, we remove all edges representing a connection that is initiated by an internal vertex because these vertices are trusted. Second, we represent each institution by a single aggregate vertex and remove all its internal vertices. The aggregate vertex represents all the IP addresses belonging to an institution. We add a directed edge from an external vertex to an aggregate vertex if the external vertex has contacted any of the internal addresses of that institution. Each edge has two weights: the number of connections the external vertex initiated to any of the internal vertices, i.e., `conn_count`, and the count of unique internal vertices the external vertex is connected to, i.e., `vert_count`. These steps significantly reduce the graph size.

In the feature extraction step (Fig. 1), we compute the following for every external vertex:

- Outdegree (OD): The number of edges that begin from this specific vertex. For an external vertex, this equals the number of institutions it communicated with.
- Outdegree weighted by number of connections ($ODW(connection)$): The summation of all the `conn_count` weights of all the outgoing connections of an external vertex.
- Outdegree weighted by number of vertices ($ODW(ip)$): The summation of all the `vert_count` weights of all the connections of an external vertex.
- Adjacency list ($V(adj)$): The adjacency list associates each external vertex with the collection of its neighboring institution vertices.

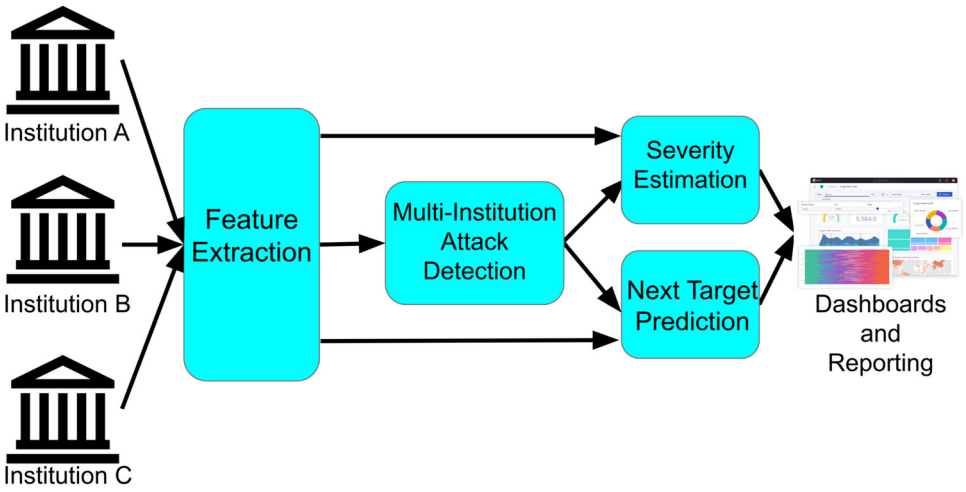
3.3 Tracking external IPs over time

Institutions periodically share their connection logs. To analyze the activities of external vertices over time, we discretize the logs. We divide the time into windows. Each window is l hours long. We analyze the connections of each window separately. For each window $W(t)$ at time t, we create a graph $G(t)$ and extract the four features as described in the previous section.

In Soteria, we track the features of each external vertex of the latest N windows. N and l are configurable, and we evaluate the efficiency of our approach while varying these two parameters in Sect. 4. Figure 2 shows an example of collecting logs from three institutions and dividing the time into three windows. A graph is built for each window.

We track the adjacency list ($V(adj)$) of an external IP starting from the window it becomes active even if this extends to more than the N latest windows. To avoid ana-

Fig. 1 Soteria pipeline



lyzing a previous time window graph, we compute the cumulative adjacency list $V(cumltv)$ at t for external vertex V_x as follows:

$$V_{xt}(cumltv) = \begin{cases} V_{xt}(adj) & if\ t = 0 \\ V_{x(t-1)}(cumltv) \cup V_{xt}(adj) & if\ t > 0, \end{cases} \tag{1}$$

where $t = 0$ is the first window V_x becomes active. $V_{xt}(adj)$ is the adjacency list of V_x at time t. $|V_{xt}(cumltv)|$ is the count of all institutions contacted by V_x since it began. For each external IP, we track OD, $ODW(connection)$, $ODW(ip)$, and $|V_{xt}(cumltv)|$ of the latest N windows in a $N \times 4$ matrix.

3.4 Detecting multi-institution attacks

We found that linear regression is effective in identifying vertices that will launch a multi-institution attack. Linear regression fits the data to a straight line that relates one independent variable t to a dependent variable Y.

$$F(t) = Y = B + tS, \tag{2}$$

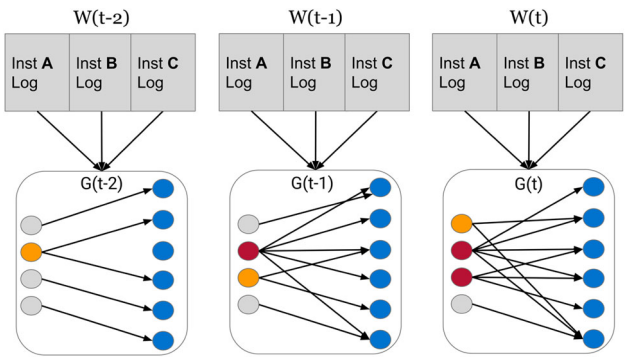


Fig. 2 Three graphs created over 3 consecutive time windows. Grey: benign external IP; orange: early detection of MIA; red: MIA

where t is the time window and Y is the feature of interest. $F(t)$ is the linear regression function that approximates Y. B is the value of $F(t)$ at $t = 0$, and S is the slope. S approximates the growth rate of a feature.

For each V_x, we keep track of OD, $ODW(connection)$, $ODW(ip)$, and $|V_{xt}(cumltv)|$ over N windows. For each V_x, we fit each feature to a linear regression line. In doing so, we can predict the future outcome of each feature. To fit the linear regression line, we minimize the residual sum of squares (RSS), a measure of the discrepancy between the data and the linear regression function.

$$RSS = \sum_{t=1}^{N} (Y(t) - F(t))^2 \tag{3}$$

We identify an IP as a MIA if it contacts $p(inst)$ institution or more. We set $p(inst)$ to 3 in our study. To predict if an IP V_x will become an MIA, we predict if V_x will contact more than $p(inst)$ of institutions in the current or a future time window. To identify a current or future MIA, we use

$$V_x\ is \begin{cases} MIA & if\ |V_{xt}(cumltv)| \geq p(inst)\ or \\ & F_{|V_x(cumltv)|}(t+n) + k \geq p(inst) \\ not\ MIA\ otherwise, \end{cases} \tag{4}$$

where $F_{|V_x(cumltv)|}(t+n)$ is the predicted number of institutions V_x will contact by the time window $t+n$. k is a constant that is used to tune the prediction.

In addition to predicting which node will become an MIA, we use linear regression to compute the slopes or the growth of the features for each external vertex. The growth of these features is used for the following steps in the Soteria pipeline.

- $V_{xt}(cumltv)$: The cumulative set of institutions V_x contacted throughout its lifetime till time t.

- $S_{|V_{xt}(cumltv)|}$: Growth of the cumulative number of institutions V_x connects to throughout its lifetime till time t.
- S_{ODx}: Growth of the number of institutions connected to V_x per time window.
- $S_{ODW(connection)_x}$: Growth of the number of outgoing connections of V_x per time window.
- $S_{ODW(ip)_x}$: Growth of the number of internal IPs communicated with per time window.

3.5 Severity estimation

The attack detection step may identify hundreds of potential MIAs. Tasking security analyst to analyze these potential attacks in a timely manner is a daunting task. The severity estimation step computes a severity indicator for each potential MIA and uses it to identify the most severe MIAs. This step helps the security analyst to prioritize analyzing attacks with high severity indicators.

The severity indicator uses all the static and growth features computed in the previous two steps of the pipeline. The severity estimation technique should have three properties. First, given the large number of external IPs, the severity indicator should be efficient to compute. Second, it should maintain the linearity of each feature, i.e., if feature X for IP1 is larger than X for IP2, this relation should be represented in the severity estimation mechanism. Third, it should tolerate highly skewed data.

We first considered normalizing each one of the seven features (three static and four growth) to the range [0, 1]. Adjacency list $V(adj)$ and cumulative adjacency list $V_{xt}(cumltv)$ are not included in the calculation, but the size of these lists is included. The classical normalization approach of a feature X is

$$X_{norm\_V} = \frac{X_V - X.\min}{X.\max - X.\min}, \tag{5}$$

where X_V is the value of a feature X for IP V and X_{norm_V} is the normalized value of X_V. $X.min$ is the smallest value for the feature among all IPs in the data set. $X.max$ is the largest value for X in the data set. While this approach is simple, it is not effective. This is because this approach does not handle highly skewed data well. For instance, the majority of attackers would create hundreds of connections, while an aggressive attacker may create hundreds of thousands of attacks which significantly skews $X.\max$ and stretches the bounds of normalization. This causes the majority of values to be placed on the lower end of the normalized range and makes it hard to differentiate between attacks since the normalized values are very close.

To overcome this shortcoming, we use a robust scaler, as shown below:

$$X_{norm\_V} = \begin{cases} 0 & X_V < X.Q_1 \\ 1 & X_V > X.Q_3 \\ \frac{X_V - X.Q_1}{X.Q_3 - X.Q_1} & otherwise, \end{cases} \tag{6}$$

where X_V is the value of a feature X for IP V and X_{norm_V} is the normalized value of X_V. The robust scaler finds the quartile for each feature X. $X.Q_1$ and $X.Q_3$ are the values of the first and third quartiles. The employeed Robust Scaler normalizes the skewed values less than the first and greater than the third quartiles to the values 0 and 1, respectively, then it normalizes the values between the first and third quartiles to the range [0, 1]. This approach effectively handles highly skewed data.

The normalized values of all features per external IP are added. To give an indicator with values in the range of [0, 1], the aggregated value is then normalized. This approach effectively computes the severity indicator, preserves the linearity of the features, and handles skewed values.

3.6 Predicting future targets

In the next step, we try to predict, for each MIA, which institution will be targeted next. Our first attempt to predict the future targets of each attack explored techniques to predict which institutions are often targeted together. We experiment with the co-occurrence matrix model which is successfully used in recommendation systems to predict the items that occur together [5]. Our results show (Sect. 4) that this approach is not effective in predicting future targets. This is because co-occurrence matrix can only capture the relationship between institutions and does not capture the sequence of the attack, neither does it capture the future possible growth of the attack.

We then explored using the Long Short Time Memory (LSTM) model [24]. LSTM is effective in predicting sequences of events. This approach captures the growth of an attack and uses it to predict the next targets. The LSTM approach achieved better results than the co-occurrence matrix model but did not achieve high accuracy. This is because attackers do not always attack institutions in the same order, and we need it to learn slight variations in these sequences. To overcome this challenge, we resorted to using bidirectional LSTM with an attention mechanism (ABiLSTM). ABiLSTM learns a sequence of events in both directions, forward and backward, to better predict targets despite variations in the order in which institutions are attacked. It also better captures the relationships between institutions in a specific window and across time windows.

3.6.1 Model design

Figure 3 shows the structure of the ABiLSTM model we use. The model has the following stages: input encoding, BiLSTM network, attention mechanism, and an output layer. The rest of this subsection details the design of each of these stages.

Input encoding We first encode *V(cumltv)* for each external IP using multi-hot encoding. For each time window, we create a bit map of size M, which is the total number of institutions. An index in the bitmap corresponds to an institution. A bit is set if the institution has ever been contacted by this IP address. Since we look into the last N time windows, the input to the model is an $M \times N$ array representing *V(cumltv)* in the last N time windows.

BiLSTM model The BiLSTM model uses an $M \times 2$ array of LSTM cells organized in M pairs. One LSTM in a pair learns the forward direction of a sequence, while the other learns the backward direction. The output of the two LSTM blocks is concatenated. The pairs are organized in a stack.

Each LSTM cell has three gates: input gate (i_t), output gate (o_t), and forget gate (f_t), where t is the window timestamp.

Equation 7 shows the input gate of a cell, while Eq. 8 generates a candidate vector. The combination of the input gate and the candidate vector controls the information stored in a cell at the current time window t.

$$i_t = \sigma(Z_i[h_{t-1}, x_t] + b_i), \tag{7}$$

$$\check{C}_t = \tanh(Z_c[h_{t-1}, h_t, x_t] + b_c), \tag{8}$$

where x_t represents the input of that cell at time t, h_t represents the output of the cell at time t and position m in the LSTM stack, h_{t-1} represents the outputs value one time step before the current time and $h_{t(m-1)}$ with the output value of the cell underneath it. Similarly, c_t and c_{t-1} represent the memory unit at time t and $t-1$. σ represents the sigmoid activation function, and $tanh$ represents the tangent function. Z_i and Z_c represent the weight matrices, and b_i and b_c are the bias values.

To decide whether to discard information from the previous time step and from the lower LSTM block, a forget gate is used as shown below:

$$f_t = \sigma(Z_f[h_{t-1}, h_t, x_t] + b_f). \tag{9}$$

The memory value for this time step is calculated using Eq. 10. Note that we use the memory value c_{t-1} from the previous time step.

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \check{C}_t \tag{10}$$

The output gate determines the value of the next hidden state. This state contains information on previous inputs. First, the output gate uses a sigmoid function to decide which portion of a cell state to return Eq. 10. We take the output of the output gate and perform the hadamard product ($\otimes$) with the output of the $tanh$ function of the memory value.

$$o_t = \sigma(Z_o[h_{t-1}, h_t, x_t] + b_o), \tag{11}$$

$$h_t = o_t \otimes \tanh(c_t), \tag{12}$$

where h_t is the hidden state of the cell which is shared with the next layer of the model and the next LSTM cell. The output of the BiLSTM model will be a concatenation of the outputs of both direction models, which will hence forth be denoted as h_t.

Attention layer The BiLSTM hidden layer outputs h_t through the activation function to obtain the correlation coefficient u_t using Eq. 13.

$$u_t = \tanh(Z_a h_t + b_a) \tag{13}$$

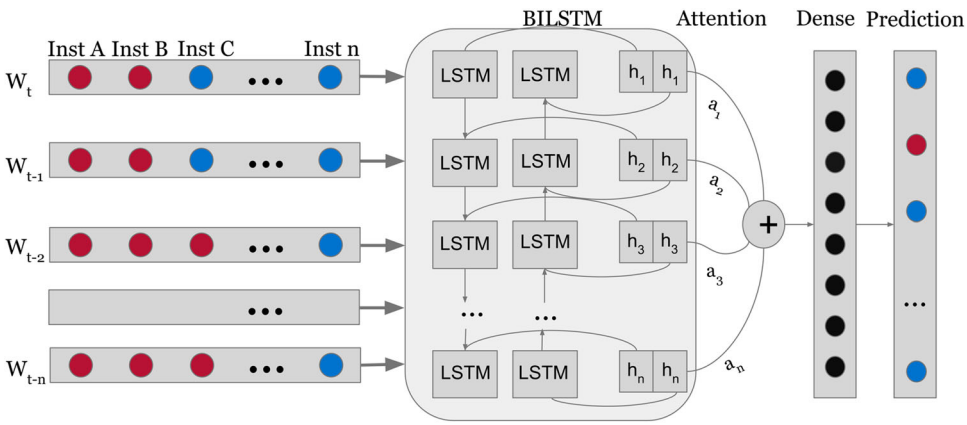


Fig. 3 Design of the model for next target prediction

Z_a represents the weight matrix, and b_a represents the bias values. First, we assign weights that demonstrate the importance of each output of the hidden layer by obtaining the weight coefficient a_t:

$$a_t = \frac{exp(u_t)}{\sum_{j=1}^{N} exp(u_j)} \tag{14}$$

. Then, we calculate the product of the weight coefficient and the output of the hidden layer to obtain the output vector v of the attention layer, as shown in Eq. 15.

$$v = \sum_{N} a_t h_t \tag{15}$$

Dense and output layer Finally, the prediction result is obtained through the output layer using a sigmoid function. The output layer contains as many neurons as there are institutions, each will produce an output for an institution. The model outputs the probability that each institution will be targeted. Because this is a classification problem, we need to convert probabilities into binary values. Therefore, we can simply round the probabilities into integers using a threshold. This threshold is tuned to provide better predictions. This gives us a multi-hot encoding vector, similar to the input matrix. We reverse the encoding on the input to get us a list of institutions that are most likely to be targeted next.

3.6.2 Prediction model training and parameter tuning

We implemented the model on Keras with a Tensorflow backend [25]. For experiments, we use the dataset indicated in Sect. 4.1 which is split into three groups: 65% as the training set, 15% for validation, and 20% for testing. We use the training set to train the model, then use the validation set to tune the hyper parameters. We also attempted to use dropout layer with rates varying from 0.05 to 0.2, but we found that it degrades the models performance. We also used the binary cross entropy-based loss function and the gradient descent algorithm with Adaptive Moment Estimation [26] to learn the model parameters. The default parameters of Adaptive Moment Estimation are learning rate=0.001, beta 1=0.9, beta 2=0.999, and epsilon=1e-08. We perform a full cycle of training, validating and provide prediction in each pipeline run and store the weights for the next cycle.

3.7 Dashboards and reporting

The last step of the pipeline creates a dashboard and reports to present the findings to the security analysts. All the static and growth features as well as the severity metric are presented to the institutions.

Table 1 Table of severity score classification

Severity score	Severity label
< 0.25	Very low severity
≤ 0.25 and < 0.5	Low severity
≥ 0.5 and < 0.75	Medium severity
≥ 0.75	High severity

The severity metric is used to order the external IPs. The display of the list of MIAs is customized for each institution. The list of MIAs is split into two categories: a list of MIAs already targeting an institution and a list of predicted MIAs that will potentially attack an institution in the near future. In order for the institutions to comprehend the meaning of the severity score, we classify different ranges into labels. Table 1 demonstrates the different classifications and their ranges.

Also in the dashboard, we provide a possible target rate for each external threat that has not reached the corresponding institution. This value is produced using the raw numerical outputs of the model. These raw numbers are normalized using the simple normalization (Eq. 5). In order for the institutions to comprehend the meaning of the target rates, we classify different ranges into labels. Table 2 demonstrates the different classifications and their ranges.

Figure 4 demonstrates the dashboard as seen by the participating institutions. This is the main dashboard for Soteria, which contains analysis extracted from our pipeline and others. We will only explain the analysis that pertains to Soteria's pipeline. As per the numbered labels on the figure:

1. As we indicated earlier, ZEEK logs are supplied by the tens of participants; however, some institutions are not supplying their logs in time. This affects model prediction. Therefore, we decided to provide an indicator that describes how available and timely is their data to the Soteria pipeline. In this case, this sample institution has been providing their logs in a timely manner.
2. Pie chart of threats that have been detected by Soteria. This is used to gauge what are the portions of MIAs in each category.

Table 2 Table of target rate classification

Possible target rate	Label
< 0.25	Very low possibility
≤ 0.25 and < 0.5	Low possibility
≥ 0.5 and < 0.75	Possible
≥ 0.75	High possibility

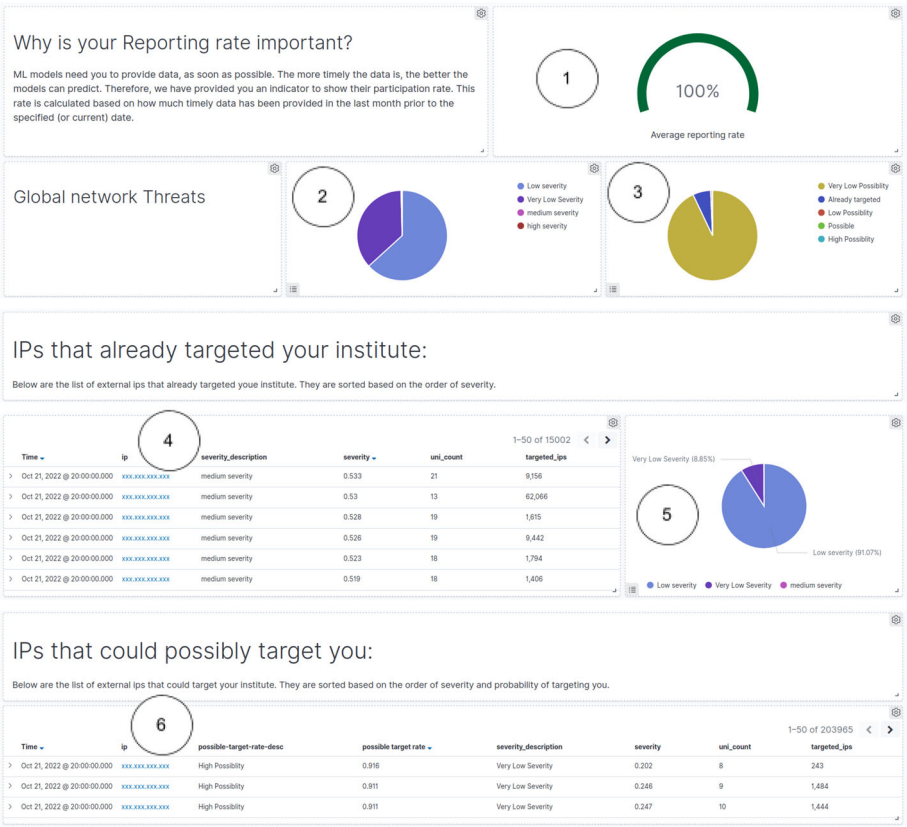


Fig. 4 Soteria example main dashboard as by participating institutions

3. Pie chart of how possible threats that have been detected by Soteria will reach this institution.
4. List of MIAs detected by Soteria that have reached this institution. They are sorted based on the severity score. Each IP is clickable to another dashboard that details information on that IP.
5. Pie chart of the severity of threats detected by Soteria that have reached this institution.
6. List of MIA detected by Soteria that have not reached this institution. They are sorted based on the severity score and possibility of reaching the institution. Each IP is clickable to another dashboard that details information on that IP.

4 Evaluation

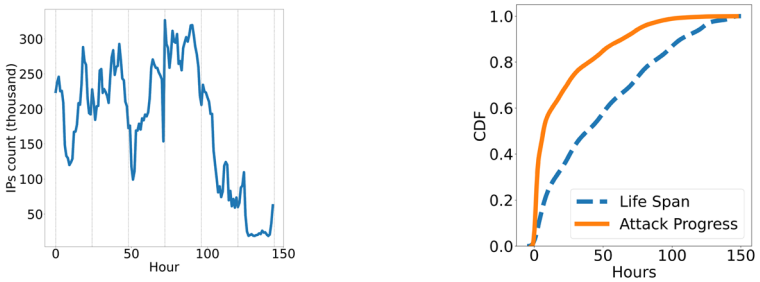
We evaluate the accuracy of Soteria in detecting future attacks, its accuracy in identifying the next targets, and the impact different configurations have on the detection performance.

4.1 Setup

Dataset details Our dataset includes the ZEEK connection logs from 52 Canadian institutions participating in the CANARIE IDS program. We use the data collected over 6 days between the 25th and 30th of January 2022.The dataset consists of over 15.5 billion connections. There were over 12 million unique external IP addresses initiating a connection to any of the institutions during the 6 days. Out of the 12 million, 2.7 million of them initiated connections to multiple institutions. The number of connections from external IP addresses per day fluctuates across the 6 days with noticeable drops during the weekends. The drop could be interpreted as a result of the institutions being less active during the weekends. Figure 5a shows the number of connections per hour.

Each external IP is tracked from the moment it starts the first connection to an institution until it stops communicating with any institution. For external IPs that communicate with multiple institutions, Fig. 5b shows the life span and attack progress of external IPs. Life span is the total time the external IP was active. It is the time period between its first connection and last connection that IP made in the data set. The life span of 70% of the external IPs is less than 3 days. Fifty percent of the external IPs had a life span of less than 1 day. For each external IP, we extract the full list of institutions it contacts. The attack progress line in Fig. 5b shows the progress an external IP makes in contacting institutions in this list. The figure shows that an attacker contacts 70% of its target list within the first 24 h.

Fig. 5 Life span of external IPs



(a) Number of connections per hour during the six days.

(b) CDF of the lifespan of an External IP and the attack progress.

Data preprocessing To simulate a stream of updates from institutions, we split the connections into time windows. Given that a large percentage of attacks complete within 24 h, we vary the window size l from 1, 3, 6, and 12 h. Each window contains all the connections for that given time period. To simulate a real workload, we split the data and then feed the data to the pipeline for analysis. When we train the model, we use multiple windows as input. Unless otherwise specified, we use the current windows and 2 previous windows as input.

4.2 Labeling multi-institution attackers

Unfortunately, the attackers in our dataset are not labeled. We attempted to utilize public databases to label our data, but we found them unreliable as there is a high degree of false positives (i.e., an IP is recycled and is no longer malicious such as the case with malicious actors using cloud services) and false negatives (i.e., threat has not been reported by anyone). To overcome this shortcoming in our dataset, we select 387 random samples of external IP addresses that contacted 3 or more institutions. We manually inspect the logs and interactions with each one of the IP addresses and identify MIAs.

We found 369 real multi-institution attacks and 18 benign IP addresses. This indicates that 95% of external IP addresses that contact multiple institutions are malicious actors with a 95% confidence interval and a margin of error of 5%. In this paper, if an external IP address contacts more than 3 institutions, we label it as a MIA.

Metrics We use the following two metrics in our evaluation: recall, which gauges how many of the true positive we have detected, and false alarm, which gauges how many of the true negatives we have misclassified as positives.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{16}$$

$$FalseAlarm = \frac{FalsePositives}{TrueNegatives + FalsePositives} \tag{17}$$

Testbed We conduct our experiments using a 17-node cluster. Sixteen nodes are used to run an Open Distro Elastic Search cluster for data ingestion and static feature extraction. These nodes have an Intel(R) Xeon(R) Silver 4208 CPU with 32 cores, 188GB of RAM, and 48TB of storage space. One node is used to run the rest of Soteria pipeline. The node has an Intel(R) Xeon(R) Gold 5120 CPU with 56 cores, 376 GB of RAM, and a NVIDIA Tesla P40 GPU.

4.3 Detection of future multi-institution attacker

Using our dataset, we measure the accuracy of our MIA detection step. For this evaluation, we use a window size l of 3 h and use a history N of 3 previous windows, and we try to predict if an IP address will become a MIA in the next 24 h.

Figure 6 shows the recall and false alarm of our attack detection step. The figure shows that our linear-regression-based technique detected more than 95% of attacks with lower than 15% false alarms. All false alarms have a severity level of less than 25%, which are presented last in the list of threats to search.

4.4 Predicting the next target

We use our dataset to measure the accuracy of predicting the next target. We use a window size l of 3 h and use a sequence

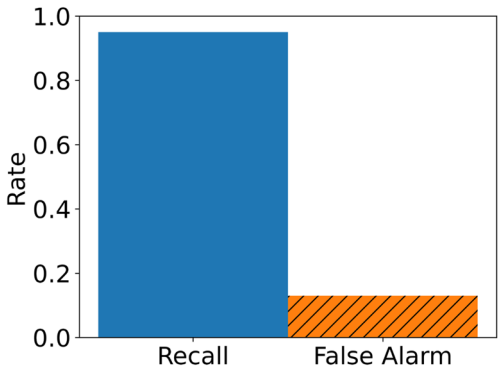


Fig. 6 Recall and false alarm of detecting MIA

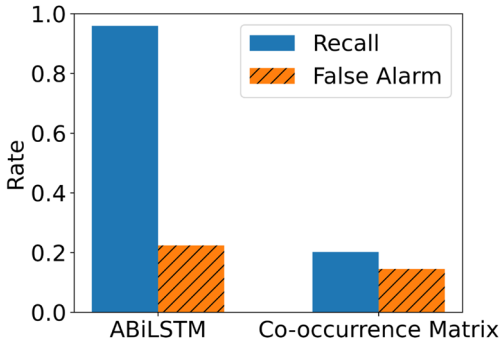


Fig. 7 Recall and false alarm of detecting next targets

N of 3 windows. Figure 7 shows the recall and false alarm of the next target prediction step. We compare the performance of using ABiLSTM and co-occurrence matrix.

Figure 7 shows that our approach with ABiLSTM achieves 4.7 times higher recall rate. ABiLSTM achieves 95% recall rate with 20% false alarm rate, while the co-occurrence matrix achieves only 20% recall rate with 15% false alarm rate. This asserts our previous discussion that BiLSTM's supersedes co-occurence matrix due to its ability to learn data sequences and future growth of attacker.

4.5 Effect of window size

In this section, we evaluate the effect of window size l on the accuracy and speed of detection. We fix the number of windows N to 3 and vary the window size l between 1, 3, 6, and 12 h. Figure 8a shows the recall and false alarm rate for identifying future attacks. The results show a slight variation in the recall rate with smaller window sizes having better recall rates. Due to the shorter lifetime of the attackers and their rapid attacking rates, smaller windows are able to capture this type of behavior. For instance, window size of 1 achieves 97% recall rate compared to 92% recall rate for window size of 12. There is no significant change in the false alarm rate. Figure 8b and c evaluate the effect of window size on the accuracy of predicting the next target. We evaluate both ABiLSTM and co-occurrence matrix. The figures show that changing the windows size does not significantly change the recall or the false alarm rate of ABiLSTM. For co-occurrence matrix, changing the window size changes the recall rate with the best being with a window size of 1, achieving 26% and the worst being 20% with a window size of 3. The results show that under all window sizes, ABiLSTM achieves 3.5 to 7 times higher recall rate without a significant increase in false alarm rate.

4.6 Effect of the number of windows

In the previous section, we kept the number of windows N fixed but varied the window size l. This results in each configuration processing a variable size of history. The number of connections in 3 windows of a 1 h window size is much smaller than 3 windows of 12 h window size. In this section, we set the look-back time to 24 h, regardless of the window size. We use 24 windows with 1-h long windows, 8 windows with a 3-h window size, 4 windows with a 6-h window size, and 2 windows with a 12-h window size.

Figure 9a shows the recall and false alarm rate for identifying future attacks. The results do not show a noticeable change in the recall or the false alarm rate with different window sizes.

Figure 9b and c evaluate the effect of window size on the accuracy of predicting the next target using both ABiLSTM and co-occurrence matrix. Similar to the previous results, the figures show that changing the window size while using the history size does not bring significant change to the recall or false alarm of these techniques. Interestingly, there is no noticeable change in results for both MIA detection and path prediction between fixing the number N to 3 or fixing the look-back time to 24 h.

In general, we see slight improvement in recall with smaller number of windows and with smaller window sizes, which we attribute to the quick nature of these attacks. The performance gap between the smaller and larger windows is not large and that is because our comparison so far compares performance of predicting future attacks. This

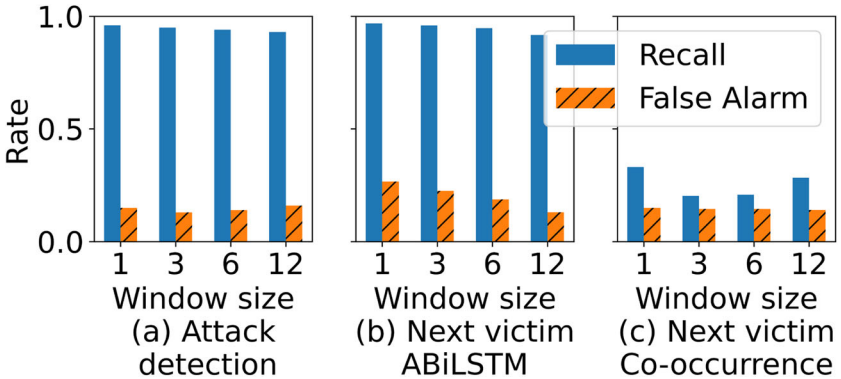


Fig. 8 Performance using three look-back windows

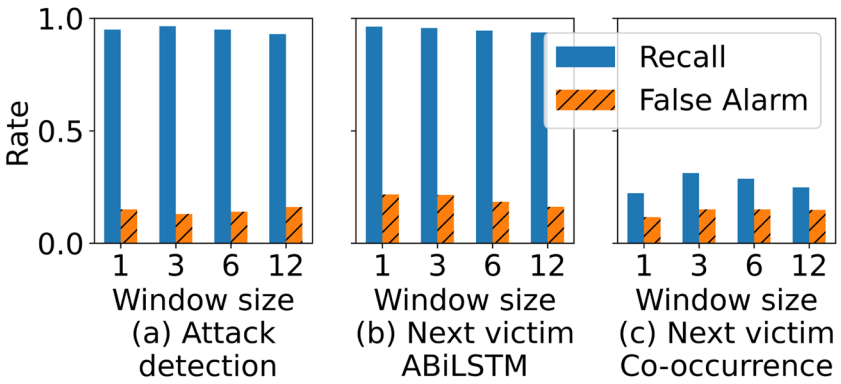


Fig. 9 Performance while looking back for 24 h

comparison does not highlight that larger windows are unable to capture attack progress as well as smaller windows. We evaluate the utility of different configurations in the following subsections.

4.7 Speed of attack detection

We evaluate how early our technique can detect an attack. We analyze the dataset and identify for each MIA the complete list of institutions it will attack. Figure 10 shows a box plot of the percentage of the MIA life at which Soteria detects the attack. We compare two configurations: window size of 3 with a fixed number of 3 windows, and a windows size of 3 with a total look-back period of 24 h. Figure 10 shows that using smaller window sizes allows for predicting the attack earlier, with a window size of 1 h detecting the attacks before 20–40% of its life span compared to 75–85% with a window size of 12. Surprisingly, using a fixed number of windows achieves better results. With a window size of 1, using 3 windows, the attack is detected at around 20% of its life span, while when using 24 h, the attack is detected when it is around 40% of its life span on average. This is because smaller windows help to detect an attack earlier, and a smaller number of windows speeds up the detection step. Under all window sizes, using a fixed number of windows performs better on average.

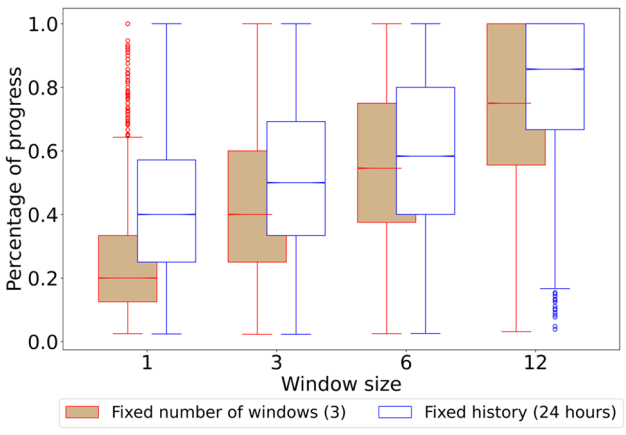


Fig. 10 The attack detection speed. The box plot shows when an attack is detected during its life span

4.8 Soteria execution time

We evaluated the execution time of Soteria to consider the difference in run time between the different window size and number of windows combinations. For each combination, we captured the time to run each component in seconds. In Fig. 11, the solid bar is the fixed H and the stripped bar is the fixed look-back time. Fetching, aggregation, and static feature extraction takes 95% of running time. Both window size and number of windows positively correlate to running time, but H is a much bigger factor. The amount of resources dedicated to run the pipeline is an important factor, and due to the congestion on the production Elastic Search cluster, the $h = 1$ and $h = 3$ could not be produced in time before the start of the next window.

We investigated the source of the low performance in the fetch and feature extraction phases. Our investigation shows that the Elastic Search cluster is congested with the sheer amount of high priority jobs that search through terabytes of data daily. To mitigate this issue, we are currently expanding our Elastic Search cluster by adding new nodes, swapping

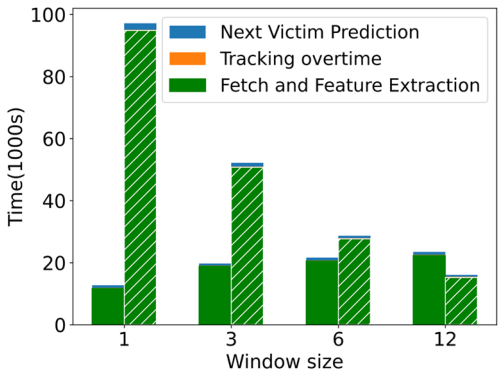


Fig. 11 Soteria run time for each of the multiple window size and number combinations

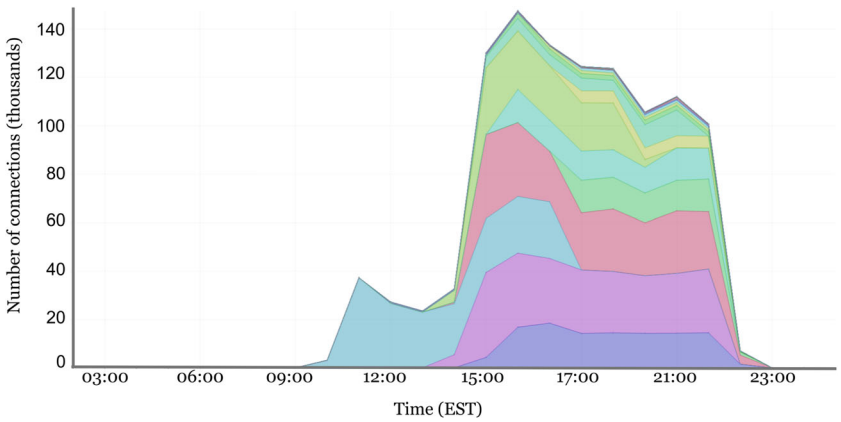


Fig. 12 Activity of a high scale external short-lived MIA. This stacked area graph represents the number of connections targeting each institution in a 24 h time frame

hard drives with solid state drives, and switching files systems from zettabyte file system (ZFS), which is not optimal for Elastic Search deployment, to a fourth extended file system (EXT4).

4.9 Example MIAs

In this section, we demonstrate two examples of the most common types of large-scale MIAs. Figures 12 and 13 are stacked area figures counting the number of connections each MIA has initiated and to which institution. Each color represents a different institution. The connections are counted utilizing an hourly windowing. Figure 12 shows the most dominant type of large-scale MIA. It started communications quickly and grow rapidly. It reached the peak of its activity in approximately 6 h, in which it has most connections, and targeted all the institutions at that point. It has a relatively short life span of 14 h. Upon further investigation, on the type of vulnerability it was searching, we have found that it targeted multiple application types, including webapps, ssh, and others if possible.

Figure 12 shows the second most dominant type of large-scale MIA. This type is constantly, actively monitoring for vulnerabilities. It has an almost cyclical pattern, initiating connections quickly then dying off. It continually targets the same set of institutions on a consistent basis. This MIA seems to be dedicated to its task. Upon further investigation, on the type of vulnerability it was searching, we found that it targets a large range of ports.

5 Conclusion

We present Soteria a data processing pipeline for detecting multi-institution attacks. Soteria can detect current and future multi-institution attacks, rate the severity of the attacks, and predict its future targets. Our evaluation shows that Soteria is able to identify future attacks and identify their targets with high accuracy. Soteria is currently deployed in production as part of the CANARIE IDS program.

In our future work, we plan to explore three main directions. We filter out two types of external IPs, and we do not detect threats coming from them, due to the large number of false positives they generate. The first are external IPs that are shared with large number of users, such as a Tor node. It is difficult to differentiate between the different users using

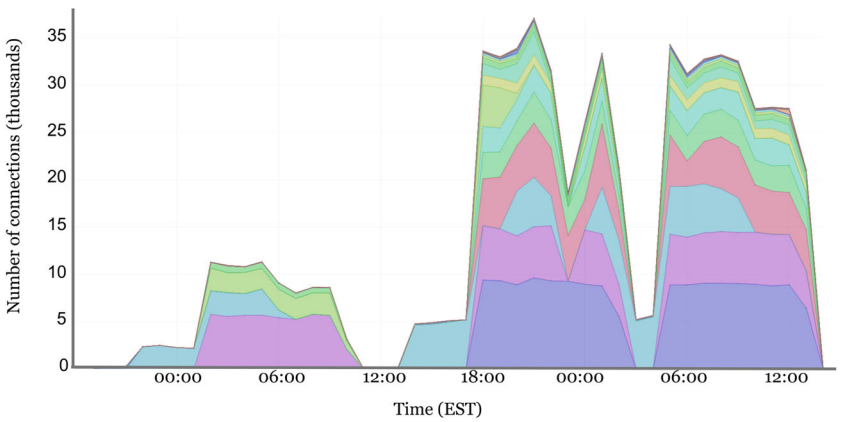


Fig. 13 Activity of a high scale external long-lived MIA. This stacked area graph represents the number of connections targeting each institution in a 48 h time frame

the same exit node, and it would be incorrect to monitor the node like any other external IP. We need to differentiate users of shared IPs; one way is by clustering connections. The second are IPs that are filtered as they are registered DNS servers that only initiate DNS connections. These are not guaranteed to be safe and could be MIAs; therefore, we need to find a proper detection mechanism in this regard.

The second issue is that some threat actors distribute their attacks by using multiple IPs to launch their attack. Although we can monitor the attackers individual IPs, we can not learn the whole picture of the attacker. A research direction is to cluster IPs together to identify the attacker. These clustered IPs can be identified as a single attacker and feed to Soteria as we did with a single IP giving the full scope of the attacker for a more accurate analysis.

A third direction would be to work on correlating Soteria with the results of other threat analysis tools and threat feeds. Soteria analysis does a good job at detecting MIAs, but does not provide information on the attack type. Therefore, correlating results with other tools can provide additional information into the kind of threat. Furthermore, this correlation can provide additional information to improve severity ratings of these threats.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Government Accountability Office (2021). Cyber Insurance-Insurers and policyholders face challenges in an evolving market, from https://www.gao.gov/assets/gao-21-477.pdf. Accessed Jan 2023
2. Akbanov M, Vassilakis V (2019) WannaCry ransomware: analysis of infection, persistence, recovery prevention and propagation mechanisms. J Telecommun Inf Tech 1:113–124
3. Accenture Security (2021). Ninth Annual cost of cybercrime study, from https://www.digitalmarketingcommunity.com/researches/ninth-annual-cost-of-cybercrime-research-2019. Accessed Jan 2023
4. Bilodeau H, Lari M, Uhrbach M (2019) Cyber security and cybercrime challenges of Canadian businesses in 2017, from https://www150.statcan.gc.ca/n1/pub/85-002-x/2019001/article/00006-eng.htm. Accessed Jan 2023
5. Dunning T, Friedman E (2014) In: Practical Machine Learning: Innovations in Recommendation. O'Reilly
6. CANARIE (2022). Canarie.ca, from https://www.canarie.ca/. Accessed Jan 2023
7. Zabarah S, Naman O, Salahuddin MA, Boutaba R, Al-Kiswany S (2023) Soteria: an approach for detecting multi-institution attacks. In: 2023 26th Conference on innovation in clouds, internet and networks and workshops (ICIN), pp 113–120. https://doi.org/10.1109/ICIN56760.2023.10073491
8. Udhayan J, Prabu M, Krishnan V, Anitha R (2009) Reconnaissance scan detection heuristics to disrupt the preattack information gathering. In: International conference on network and service security
9. Allen WH, Marin GA, Rivera LA (2005) Automated detection of malicious reconnaissance to enhance network security. Proceedings. IEEE SoutheastCon 2005:450–454. https://doi.org/10.1109/SECON.2005.1423286
10. Cao J, Jin Y, Chen A, Bu T, Zhang Z-L (2009) Identifying high cardinality internet hosts. In: IEEE INFOCOM 2009. https://doi.org/10.1109/INFCOM.2009.5061990
11. Kamiyama N, Mori T, Kawahara R (2007) Simple and adaptive identification of superspreaders by flow sampling. In: IEEE INFOCOM. https://doi.org/10.1109/INFCOM.2007.305
12. Liu Y, Chen W, Guan Y (2016) Identifying high-cardinality hosts from network-wide traffic measurements. IEEE Trans Dependable and Secure Comput 13(5):547–558. https://doi.org/10.1109/TDSC.2015.2423675
13. The Zeek Project (2022). conn.log - Book of ZEEK, from https://docs.zeek.org/en/master/logs/conn.html. Accessed Jan 2023
14. Cisco: networking, cloud, and cybersecurity solutions (2022). Snort, from https://www.snort.org. Accessed Jan 2023
15. The Open Information Security Foundation (OISF) (2022). Suricata, from https://www.suricata.io/. Accessed Jan 2023
16. Feng B (2021) Threat intelligence sharing: what kind of intelligence to share? Concordia, from https://www.concordia-h2020.eu/blog-post/threat-intelligence-sharing/. Accessed Jan 2023
17. Marathon Studios Inc (2016). AbuseIPDB - IP address abuse reports, from https://www.abuseipdb.com/. Accessed Jan 2023
18. Hispasec Sistemas (2004). virustotal.com, from https://www.virustotal.com/. Accessed Jan 2023
19. The MITRE Corporation (1999). CVE - common vulnerabilities and exposures, from https://cve.mitre.org/. Accessed Jan 2023
20. The MITRE Corporation (2006). CWE - common weakness enumeration, from https://cwe.mitre.org/. Accessed Jan 2023
21. Solarwinds (2023). Intrusion Detection Software, from https://www.solarwinds.com/security-event-manager/use-cases/intrusion-detection-software. Accessed Jan 2023
22. Skopik F, Settanni G, Fiedler R (2016) A problem shared is a problem halved: a survey on the dimensions of collective cyber defense through security information sharing. Comput Secur 60:154–176. https://doi.org/10.1016/j.cose.2016.04.003
23. Settanni G, Skopik F, Shovgenya Y, Fiedler R, Carolan M, Conroy D, Boettinger K, Gall M, Brost G, Ponchel C, Haustein M, Kaufmann H, Theuerkauf K, Olli P (2017) A collaborative cyber incident management system for European interconnected critical infrastructures. J Inf Secur Appl 34:166–182. https://doi.org/10.1016/j.jisa.2016.05.005
24. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735
25. Chollet F et al (2015) Keras. https://keras.io
26. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization