ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

Data Mining for Cyber Threat Intelligence: From Logs to Stories, Human-Centred

Ranjita Rout [2], Rakhi Jha [3], Tarun Kumar Behera [4], PRAVAT ROUTRAY[5]
[1, 2, 3, 4] Gandhi Institute for Education & Technology, Baniatangi, Khordha, Odisha
[5]NM Institute of Engineering & Technology,Bhubaneswar,Odisha
ranjitarout@giet.edu.in, rakhijha@giet.edu.in, tarunkumarbehera@giet.edu.in

ABSTRACT On the system, a typical medium-sized organisation logs between 10 and 500 million events every day. The specialised personnel only looks at fewer than 5% of threat signals, exposing a security gap that might be exploited by attackers. Cognitive overload is brought on by alert messages that provide insufficient information and are created in a way that is more user-friendly for machines than for people. This study proposes a paradigm for producing reports in normal language from security logs using unique storytelling approaches. The solution accommodates a range of reader preferences and skills by offering editable templates that are populated with information from both local and worldwide knowledge bases. A case study from an educational institution's Security Operations Center (SOC) is used for the validation. The resulting report outperforms the current method in terms of understanding (improved cognition) and thoroughness (enriched context). The examination shows the value of narrative in interpreting possible dangers in the context of cybersecurity.

INDEX TERMS Cybersecurity, storytelling, threat intelligence, human cognition, information extraction, knowledge Discovery.

I. INTRODUCTION

A. HIGH-VOLUME OF EVENTS ARE LOGGED, BUT NOT COMPREHENDED

Millions of activities and attempts are recorded on computer systems on a daily basis. As an example, a university of 3, 000 staff and 40, 000 students registers approx. 200 mln events every year. At the same time, only about 20% (or 40 mln) of the logs will be analysed by specialised security systems. To compare with the volume of events recorded, the cybersecurity team of such university consists of no more than 10 trained professionals.

Numerous algorithms have been proposed to automatically analyse the events and signal alerts for potential malicious activities [1]. There is a multitude of various types of monitoring systems in use that generate potential threat alerts. In order to appropriately respond to the suspected threat, the *synthesis* of currently *disintegrated* systems is required. Still, building the context around the potentially malicious alert is predominantly a manual task, which involves rich experience and knowledge regarding log files analysis [1]. Thus, comprehensive alert analysis has become a critical task in harmful events and fraudulent activities detection, their timely resolution, and future prevention [2].

Although monitoring systems are helpful in filtering through millions of logged events and generating security alerts, final human assessment is still part of the process. As such, thousands of potential security breaches received from different monitoring systems pose significant burden on cybersecurity team resources. Given the *machine-friendly* rather than *human-friendly* format of such alerts, as well as the substantial domain knowledge required, the interpretation of raised alerts is strictly limited to cybersecurity professionals.

The comprehensive and accurate alert assessment is also prone to the subjectivity aspect that forms an inherent part of any human evaluation process. Correct response then highly depend on long-standing experience of analysts from cyber threat management field. Dramatically increasing number of security alerts is currently outgrowing scarce and expensive cybersecurity resources.

B. KNOWLEDGE BEYOND EVENT LOGS IS REQUIRED FOR ANALYSIS

Despite the overwhelming volume of security alerts, only a fraction requires further investigation. Still, the time and

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

effort has to be dedicated by security analysts to confirm that the alert is indeed a false positive or a real incident. To properly assess the scale of the risk, the knowledge outside of the security logs is required. Local domain knowledge determines the risk of internal assets, and the potential risk of outsider is specified by Global domain knowledge. As an illustration, consider the examples below:

- **Local domain knowledge required:** A server of the organisation X is used for temporary storage and web testing, and is labelled as a *non-critical* host. Mostof the alerts from that server can be omitted unless a serious breach occurs. However, the server is located in the finance department for financial reporting and budget planning. Finance department usually holds crit- ical information. If an alert for a serious breach occurs for one of the servers in this department, other servers also can be at the potential cyber risk, warranting fur- ther investigation despite no explicit alert raised. Thus, the exceptional defence strategy should be adopted in advance following the complete knowledge obtained from an inside of the organisation.
- **Global domain knowledge required:** The organisationY with limited number of experienced cyber profes- sionals has to prioritise the crucial alerts over large volume of the remaining security breaches for prompt response. The selection is based on the prior knowl- edge and experience based on the repeated alerts from historical records. An appropriate response for the new attack requires an in-depth investigation of attacker's characteristics. However, the attacker may change its behaviour over the time for the repeated activities. The level of expert knowledge is usually not increasing at the same speed as the complexity of attacks in today's digital environment. As a result, a critical alert may not be given a required priority, leading to delayed response and potential escalation. Thus, knowledge obtained automat- ically from external sources is required to stay up to date with increasingly sophisticated and dynamically changing cyber attacks.

Both examples show that comprehensive alert analysis requires domain knowledge from Local *and* Global. If the complete knowledge cannot be modelled and integrated in alert analysis, either false alarms are triggered, or high-risk alerts are neglected.

C. EXPLAINABLE ARTIFICIAL INTELLIGENCE (AI)AND STORYTELLING

In prior research, the automated Cyber Situation Aware- ness (CSA) tools and models aiming to enhance the cog- nition of experts have been proposed [3]. As defined by Endsley: *''Situation awareness is the perception of the ele- ments of the environment within a volume of time and space,the comprehension of their meaning, and the projection of their status in the near future''.* As such, situational aware- ness system has been designed to compile, process, and fuse data from several different perspectives [4]. Yet, the existing

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

Cyber Situation Awareness systems have not been able to address the continuously evolving cybersecurity challenges completely [3]. Despite helpful, the security experts still have to digest vast volume of data and discover the hidden links and dependencies.

Storytelling is a method to assist and engage people to explore and interpret complex real-world problems. Accord-ing to Vink [5], telling stories in problem formulation phase merges synthesis and analysis, and makes abstract con- cepts more concrete. Storytelling can be used as a knowl- edge representation method to highlight the explicit and implicit information from log files, and convert it into a human-understandable format [6].

Given large volume of logs and alerts, the stories have to be generated automatically. Automated journalism is a recent accomplishment in story generation field [7]–[10]. The story-like technical and financial reports are produced based on the personal preference, which positively impacts their comprehension. The approach also outweighs the traditional 'human' journalists in both aspects, namely (i) faster reports generation, and (ii) lower propensity to errors [11]. Consid- ering its numerous benefits, the automated stories generation is yet to be explored in security log files interpretation.

Despite promising, the number of limitations of automated journalism in its current state have been identified, and are as follows:

1) Stories are based on basic template using the predefined format adapted to specific domain with high-quality of data required (flexibility limitation);
2) The mechanism for extension, additional information integration, and new knowledge contribution is cur- rently missing (contextual limitation).

In this paper, the automatic generation of storytelling reports at multiple levels of details (i.e. for expert and non- expert) provides a comprehensive view of the cyber situation (i.e. from local and global database) that fills the existing gap in the analysis of security log records. The model proposed, unlike current approaches that still (i) rely on security experts knowledge and expertise, or (ii) are limited in depth of the insight provided, allows to reveal the root causes of the prob-lem to facilitate the correct response to the potential threat. The novelty comes from the human-comprehensible format of the report, which proved successful in various applications (e.g. automatic journalist), yet it is still underutilised in cyber- security domain.

II. RELATED WORK

Much research has been dedicated to minimise human inter- actions in the process of log files analysis. The representation approaches of the analytical results can be categorised into four main categories, namely: 'Black box', 'Visual', 'Struc- tured', and 'Narrative'. In this section, the examples of the works falling within each group will be briefly introduced. Although the narrative approach has not been used for log files analysis, it will be discussed from different perspectives to determine its usability to cyber security domain.

A. 'BLACK BOX'

This representation group is named as 'Black box' since there is no explanation, or peering into its internal structure to justify *how* and *why* the analytical process works. The results usually are presented in the Boolean format to identify normal and abnormal (malicious) activities. For instance, abnormal activity was recognised by the application of var- ious machine learning techniques including Naive Bayes, K-Nearest Neighbours, and Support Vector Machines to high volumes of logs by Muggler *et al.* [2]. 'Black box' method is often considered as untrustworthy as there is insufficient reasoning about the situation and label assignment. As an example, a company simulates cyber attacks by a penetration test. Such activity should not be labelled as abnormal as an authorised person perform it. How can penetration test activ-ities be distinguished from real attacks given no explanation? Without situation awareness, the real attacks can be ignored by an expert.

B. VISUAL

A significant body of literature has already sought to involve human supervision in data analysis process by visualisation techniques utilisation [12]. Visualisation presentation facil- itates human cognition to improve potential issues identi- fication [13]. For example, a decision tree (as a level of analysis display) in the work by Xu *et al.* [14] was used to demonstrate how the system decides to assign a normal or abnormal label to a log record. It is based on limited and predefined criteria that does not offer the comprehensive view.

A graph is another presentation, which was used by Aharon *et al.* [15] to display system behaviour status. The graph shows different groups of log messages along with their labels (normal process or failure process) based on the clustering algorithm. Clustering similar messages on the graph is useful, though it does provide further explanation of *why* the particular messages belong to the one category.

Samii and Koh [16] considered more aspects of events by providing a search cability in an interactive query-based system. The information was displayed on an interactive visual interface from a high-level view to the original log files. Li *et al.* in [17] proposed a system to handle various types of events logs by providing a facile way of analysing. Statistic knowledge from logs was extracted and depicted on a dashboard. An Interactive dynamic query-based form has also been provided to support to explore more infor- mation about an event. An interactive visual interface and visual query-based

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

interactions are bound to specific graphi- cal features, which cannot fully support analysts to provide a comprehensive analytical report. For example, an expert cannot search all connections with'HTTP post method' if the HTTP method is not considered as a design feature in the interface or dashboard. By considering more design features, a high level of knowledge and specialist training are required to understand what should be searched through it, and what should be expected from the results.

Azodi *et al.* [18] attempted to address the issue by paths of attacks identification. Events correlated to an alert were discovered based on the regular expressions to obtain a bet- ter understanding of the progress of attack. A graph dis- played attack paths, and correlation between different attacks was shown by a link. Although the visual graph provides more design features and information about the connections between sources and destinations, the relevant details and explanations which are necessary for instant inference are missing. For instance, the graph shows a connection between server from our organisation to an external web site. However, it does not show what was the HTTP method used through this connection. Overall, the existing visualisation interfaces does not provide sufficient information to distinguish normal and malicious connections in order to assist an expert in cyber situation awareness.

C. STRUCTURED

Numerous studies have attempted to change the log struc- ture into a rich format to improve the understanding. Nimbalkar *et al.* [19] translated log files and added semantics keywords. The results are demonstrated in the semantic RDF linked data, which is a machine interpretable representation. Lack of concepts descriptions and their relations were poten- tial disadvantages of machine-readable formats for cyber ana- lysts. Furthermore, the representation format was particularly challenging for non-experts. In summary, RDF as a structured data format is highly machine-readable, but is not considered a good candidate for reporting and analysis by humans.

In cybersecurity area, information exchange formats were proposed to enhance knowledge of every single participant to address the lack of comprehensive analysis in the use of gathering all significant aspects [20]. Structured Threat Information eXpression (STIX) [21] and Incident Object Description and Exchange Format (IODEF) [22] are two of them. STIX is focused mostly on cyber threat intelligence from a holistic perspective, and IODEF is concentrated on attackers and defenders information. They are created for various purposes [23], and machine-readable format makes it extremely challenging to understand the components and the relations between them. The only human-readable exchange format is X-ARF [24]. However, the X-ARF is a basic format that can only exchange limited types of malicious alerts via an email. The email contains limited information such as alert description, alert category, initial information about the attack and attacker [25]. The exchange formats transfer alert messages to a new structure and add descriptions to enrich it. Therefore, the main aim of them is sharing the alert mes- sage, not interpreting the alert message and providing more evidence for improved understanding.

D. NARRATIVE

While narrative activity is a sense-making process rather than a finished product [26], a narrative explanation can be a good candidate in analysis facilitation. Currently, no efforts have been made in cybersecurity analysts support by using the narrative formats. Wu *et al.* [6] proposed a data-driven storytelling system for social connections improvement. The system transformed sensor data from IOT devices of elder's conditions for their loved ones in order to support a social connection between an alone elder and his/her family. Raw data was mapped to semantically meaningful variables through a GoalNet, and the dynamic storylines were gen- erated based on a set of curiosity rules. Wu *et al.* [6] only provided one level of explanation in their output results to attract the adult children's attention. Although the system could not explain the details of the elder's conditions and refer to a triggered sensor as evidence, they believed they reached their aims to captivate the adult children's atten- tion. A multi-level story from the alert message can be a novel approach to support the analytical process in cyber security domain. Simple concepts in sequential sentences can be organised to discern where the events are heading. It is easier for a human beings to identify correlations of events in the log files when they are modelled using storytelling design [11].

III. TERMINOLOGY

a: KNOWLEDGE BASE

In this paper, we use two main databases (Local and Global) to obtain contextual insight about an alert. In term of com- pleteness, internal sources and external sources are provided to enable sufficient level of comprehension.

Local knowledge base includes supplementary information that is internally processed, as well as the raw data collected from the security devices. Local knowledge base contains explicit knowledge about the situation of the event. The implicit knowledge is added to the knowledge base by prede- fined rules and procedures. Local Knowledge base contains (1) List of the internal servers and hosts with the associated information, including domain name, administrator, sever- ity (low, medium, high), location, and installed applications (2) Story templates, (3) Rules for analysis,(4) Regular expres- sions, and (4) List of keywords.

Global knowledge base contains supplementary infor- mation that is collected by external companies and researchers, and is processed internally. Global knowledge base is comprised of the following Information (1) Whois

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

Command [27], (2) Virus Total,[1] (3) Threat Miner,[2] (4) AlienVault,[3] (5) Snort Rules (6) Windows Defender Security Intelligence (WDSI),[4] and (7) Symantec.[5]

b: EVENT

In this paper, the event is a status of the action that is recorded in the log by the monitoring system.

c: ALERT

The alert is a generated message when abnormal event occurs. The security devices generate alert when observe that a part of an event specification matches their predefined patterns. The generated message (called an alert message in this paper), provides a short description for further analysis.

d: REPORT (SECURITY REPORT)

The report is a document that presents detailed information about the alert to assist the analysts to understand more about the abnormal events registered.

IV. LOG-DRIVEN STORYTELLING MODEL

The proposed model that consists of four individual layers and main procedures is illustrated in Figure 1. The details of each layer, i.e. primary purpose and associated steps are as follows:

A. PRE-PROCESSING LAYER

In this layer, alert message is parsed to extract the basic fields. The fields include Time, Date, Source Internet Protocol (SrcIP), Destination Internet Protocol (DesIP), as relevant to the alert. The alert generated by the Security Information and Event Management (SIEM) system[6] was used in the case study.

Since the selected fields are primary properties in each alert message, the proposed approach does not depend on the specific device. An alert record L can be represented as {Date, Time, SrcIP, DesIP, Message} from the alert's message by the monitoring systems.

- **Date and Time** values represent when the events are registered. These values can be different from alert Date and Time (as received after an event).
- **Source Internet Protocol (SrcIP)** value represents the address of the initiator of an event. In other words, who is the source of the connection (Subject or Object of an event).
- **Destination Internet Protocol (DesIP)** represents the objects of the events. In other words, DesIP is an address to which the connection has been made (Subject or Object of the event).
- **Message** value represents behaviors, which Subject conduct towards the Object. This value usually includes the classification group name for threat. Since this paper considers malware category, the value contains terms such as 'malware' or 'trojan'.

A collection of regular expressions is used to parse and tokenise the alert messages. The delimiters include '/', '?', '.', '=', '-', and '_'. The extraction parsers and tools before this layer are applied as pre-processing. The outputs produced will be further used in the Extraction layer.

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

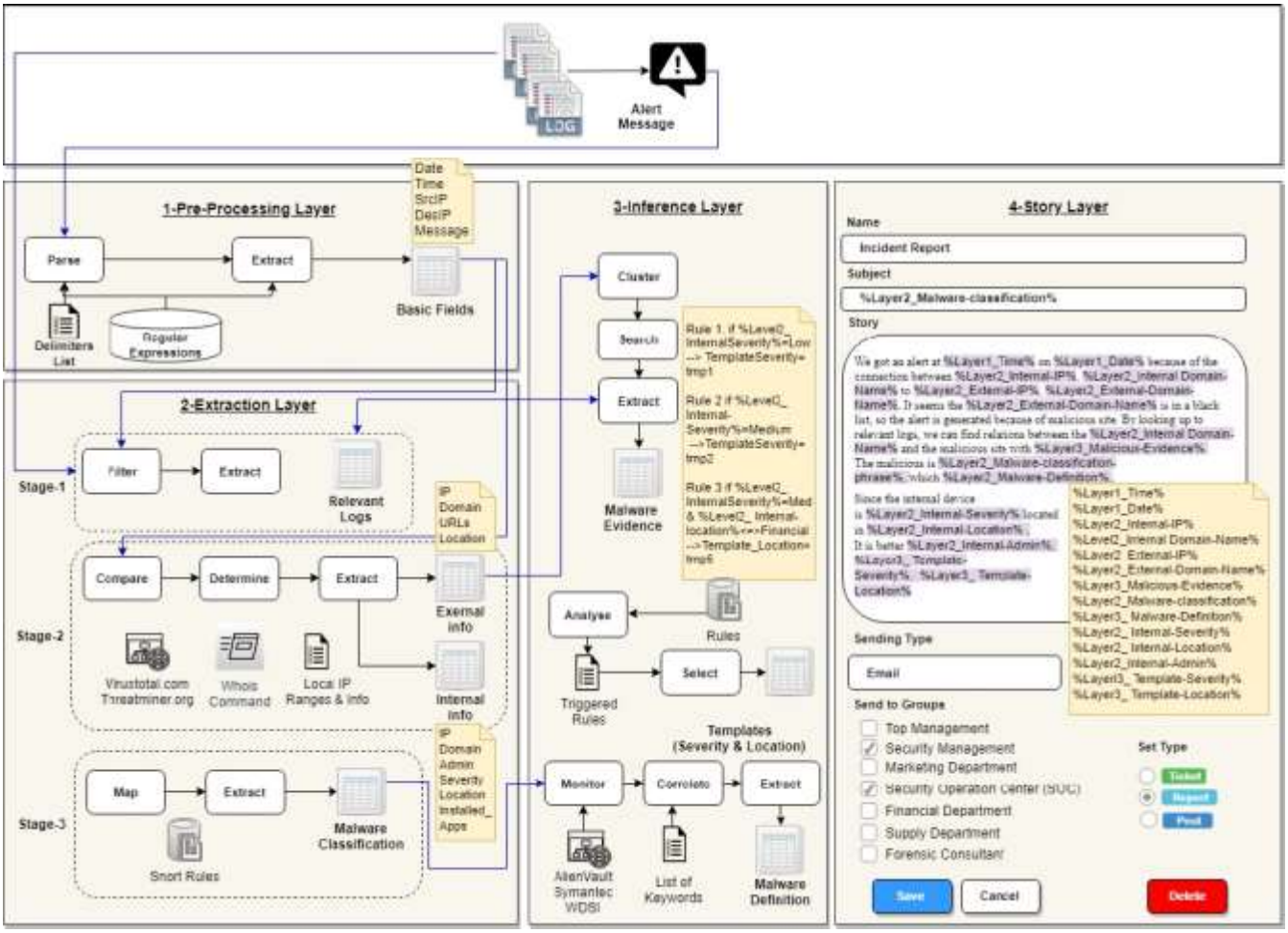


FIGURE 1. Overview of the Log-Driven Storytelling Model made of four layers (beige boxes) and operation procedures (white boxes, except the story layer). The story layer represents the final output with modification capability.

B. EXTRACTION LAYER

Although selection and retrieval of basic fields from an alert message is performed, the basic information about the alert, the relationships between basic fields and corresponding information allows to spot the potential logical links.

In this layer, the alert message is complemented with supplementary information to compensate for the lack of data, which leads to insufficient understanding [3]. As a result, full awareness about the alert situation from various heterogeneous sources such as different departments and owners can be achieved [28]. Associated information to the alert message is extracted from Local and Global knowledge base, which are mapped to the extracted basic fields in L (Extraction layer). The extraction layer consists of 3 main stages, which use different fields of L.

The 1st stage looks into the aggregated logs files that use Date and Time when the events were synchronised. Every single log record in log file has the Date and Time references. Events are sorted based on the time sequence. Date and Time of an event comes from the basic fields in L and log files, which are gathered log records from a variety of network devices. Binary search in terms of time is applied to retrieve events in a particular time interval. Since some logs are recorded based on the UTC, and others are recorded based on the local time, to cover all the related logs ∠ 1 day timespan is applied. The log file also provides the information about source and destination IP for the connection. Therefore, the corresponding connection between SrcIP and DesIP are found by tracing the entire particular interval. The output of this stage is a list of events that represent the connection between the source and destination that happened in the particular time interval.

The 2nd stage searches Local and Global knowledge base to find out about the IP address and Domain Information (which IP belongs to the organisation and which is from the outside, thus suspicious to be a source of infection). In this stage, Whois command identifies the names within a given registrar's registry. Therefore, the other registry out of organisation is used as *e*xternal. Furthermore, each organisation provides a list of IP address ranges based on their own network architecture. The matched IP address to this list is considered as *i*nternal. After determining the connection type: from Internal to External or from External to Internal, the corresponding information from Local and Global knowledge

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

base is extracted. The Global knowledge base contains set of information based on the online public repositories such as *''Virus Total''* [29] and *''Threat Miner''* [30]. Each set repre- sents the IP address, which is recorded in a black list, Domain names, Geography Location of the server, and URLs[7] that were repeated in previous infections (cause to be reported ina block list). Local knowledge base contains set of relevant information about internal hosts/servers (IP address, domain name, administrator, location, severity, and installed applica- tion). Although update of the Local and Global knowledge base is computationally expensive, it is a trade-off between automatic and complete information extraction, and the time and effort required for manual search.

The 3rd stage uses alert message to map to the Snort [31]Rules to extract the complete malware classification phrase. Snort is a lightweight network intrusion detection system thatuses rules to perform content pattern matching and detect a variety of malware. Snort Rules are open source and used in variety of security devices. By mapping the message field from L to the Snort malware rules, the complete phrase for the infection is extracted. While Snort and Snort Rules are usu- ally thought of as a list of independent - open source patternsto be tested in matching engines of security devices, the alertmessage usually contains Snort classification label, which defines the malware category [32]. In this paper, the approach is limited to security devices that lie at the core of Snort as a matching engine. Since Snort is popular Intrusion Detection System, this is not a severe limitation and variety of commer-cial and open-source devices worked with the Snort Rules.

C. INFERENCE LAYER

In this layer, information is analysed by using the artefact metadata and machine learning techniques to reconstruct the past events to answer three core questions about the actor (who), riskiness (what), and evidence (how) of the event in relevant logs. To understand who is the actor and what is the purpose of the action, the associated information to the malicious website has been extracted in the Extraction layer. There is still insufficient level of detail that would explain what is the aim of an action. Thus, the malware definition is automatically extracted from webpage articles that may carry the sentences related to malware explanation. To accomplishthis goal, we borrow the idea from [33] and use a scraperto monitor each website in the list of top security technical blogs to extract the associated supplementary information. It should be noted here that although the list of websites is limited, the approach is not restricted to them and the list can be customised. The examples of websites used in the case study are:

- AlienVault
- Symantec
- Windows Defender Security Intelligence (WDSI)

[7]Uniform Resource Locator (URL) form a part of the Uniform Resource Identifier (URI), and serves as a pointer to where the resources are located and the procedure to fetch them.

The scrapers perform the breadth-first crawling on eachwebsite to search for the *malware classification phrase* foundin the Extraction layer. Document Object Model (DOM) treesare generated for pages that are characterised by the sameHTML template. These pages contain relevant definitions asopposed to the ones with e.g. logins, subscriptions, advertise-ments - considered as non-relevant. All pages' DOM trees arecompared to identify the node with the combination of thetokenised phrase from the *malware classification phrase* 'is' text under the node with title 'summary', 'definition'or ' behaviour', starting with 'this malware', 'this virus', or'this trojan'. It is the way of providing further details aboutthe malware and clarify the aim of an action.

To obtain more information about the riskiness of an event, the information from potentially compromised internal server is applied to the list of rules to derive proofs. The proof isa sequence of the conclusions that demonstrates the risk of an event based on the internal information. A set of rules is used to infer valid conclusion, which defines the risk. The risk is based on the internal assets values in terms of locationand severity. For example, a server in financial department faces higher risk than other departments. The values of the severity and location from the Internal Info (Figure 1) can fire more than one rule. The output of the triggered rules are the template sentences about the risk, which will be selected to complete the story in the subsequent layer (Story Layer).

To provide the evidence of relevant events to the alert, the extracted information (URL, Downloaded files,and Com- munication files) from the external malicious website is searched among the relevant logs. This purpose is served by the application of k-Mean clustering on extracted URL to frame it as a classification problem. Input URLs are divided into disjoint subsets, then for each URL in each subset the distance to all the other URLs in the same subset is computed, and the URL that has the lowest sum of distances should be the centrist. To extract the max-length URL from each subset, the NLTK library, which offers an Ngrams function to iterate over values of N, is used. Then, the max-length URL from each subset, which presents the pattern of the URL,is searched among the relevant logs to extract the evidence. Repeated URLs are removed and the URL as a symptom is selected to enrich the report.

D. STORY LAYER

Story generation from analytically enriched data is the main contribution of this paper. It is much easier for human beings to

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

find the correlations between events in the log files if theyare modelled using storytelling design. A story can incorpo-rate different aspects of an event and can convey the *meaning* of an alert. Therefore both competence and comprehension are achieved by explaining the security alert in the storytellingdesign.

The story can be personalised based on the needs and preferences of the individual reader [11]. As Figure 1 shows, the intended audience can be selected in the 'Send to Group' section of the interface and the appropriate template based on their preference is shown in the Story section. The template is modifiable and can be customised based on the preference and internal policy. Each template contains set of variables (the yellow border) that are initialised through the previous layers. In this layer, the retrieved information and analytical results, which are automatically stored in the Local knowledge base are used to replace the variables in the story. Each variable contains it own original layer. For example, Date and Time are the variables that were extracted from the alert message in the Pre-processing Layer.

The riskiness of the event is explained in the separate templates based on the triggered rules, and are used to enrich the message with more internal recommendation. The results are the knowledge sets, and the relationships between them. In other words, the story is generated based on the template, and the relationships between retrieval information from previous layers. The generated story can be set as the 'Ticket' for future actions as a response to an incident, 'Report' for management, and 'Post' for broadcasting to increase an awareness about what has happened. Although storytelling design is template-based, the templates and rules are easily modifiable without an extensive technical expertise. The customisation can be achieved based on the organisational demands.

V. CASE STUDY

In order to validate the model proposed, the case study on real-world scenario was conducted. More specifically, the report generated by Log-Driven Storytelling Model was compared with the report generated by external vendor's tool - the Secureworks.[8]

Secureworks is the commercial cybersecurity analytical tool used by the SOC team at the educational institute. More specifically, Secureworks provides Incident Response Services for potential cyber threats detection among the monitored log files, and alert their clients by appropriate report generation. The vendor claims to combine human-machine analytical capability to assist in information security services. According to Secureworks, *''to ensure that even if our machine learning models occasionally encounter an issue, Human and Machine are Working Together''* [34]. Thus, the report generation still relies on human assistance to derive actionable cyber threat intelligence.

As for technical details, the machine side of Secureworks manages the logs from approx. 800 servers at the education institute, 2000 – 6000 MPS[9] (low - holiday period, high - semester period), and 600 – 700 *high-risk* incidents per year. The human side involves manual assistance and human-understandable report format generation about the incident registered (for the customer to understand their cybersecurity situation).

```
MALWARE-CNC Osx.Keylogger-Elite - 10.
233.62.247 -> 104.239.223.14 02/27/2019
5:05 PM
```

A. PRE-PROCESSING LAYER

The basic fields (i.e. *{(Date, Time, SrcIP, DesIP, Message)}*) were extracted from the alert using regular expressions presented in Table 1.

TABLE 1. Regular expressions used in the case study.

Type	RegEx
For message	*([A-Z][a-z]{0,*
For IPs	*(?:[0-9]{1,3}[0-9]{1,3*
For Date and Time	*(?:0?[1-9]|[0-5][0?=[012]|d:[0-5]|d[?:[ap]m)?|*

B. EXTRACTION LAYER

The information relevant to the basic fields were retrieved in the following stages.

The 1st stage: The relevant logs were identified based on Date and Time as well as source-destination connection. In order to ensure the coverage of maximum number of potentially relevant events, the timespan was set to 1 day before and 1 day after an event. Since Date and Time of an incident (based on the extracted basic fields) was 02/27/2019 5:05 PM, the timespan was set to the following: 02/26/2019 5:0 PM - 02/28/2019 5:0 PM (to allow all the devices to record their logs). In total, 644, 434, 681 logs were recorded by monitoring devices at the university throughout the time interval specified. After filtering based on both SrcIP and DesIP, the number of events was reduced to 12. This provides the final list of events that represent the connections that occurred between SrcIP and DesIP within the timespan specified.

The 2nd stage: The SrcIP was marked as Internal (by comparing with organisation IP addresses range), and the DesIP was marked as External (by applying Whois command and comparing with registry).

The retrieved information (i.e. IP, Domain, Admin, Severity, Location, Installed Application) about the internal server in the alert message including IP 10.233.62.247, and stored in the Local knowledge base is:

Internal Server = *{(10.233.62.247, Sev1.edu.au, Tommy Schart, IT-developer group, CoNsoleKit Microsoft Visual C++)*

The retrieved information (i.e. IP, Domain, URLs, Location) about the external server in the alert message including IP 104.239.223.14, and stored in the Global knowledge base is:

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

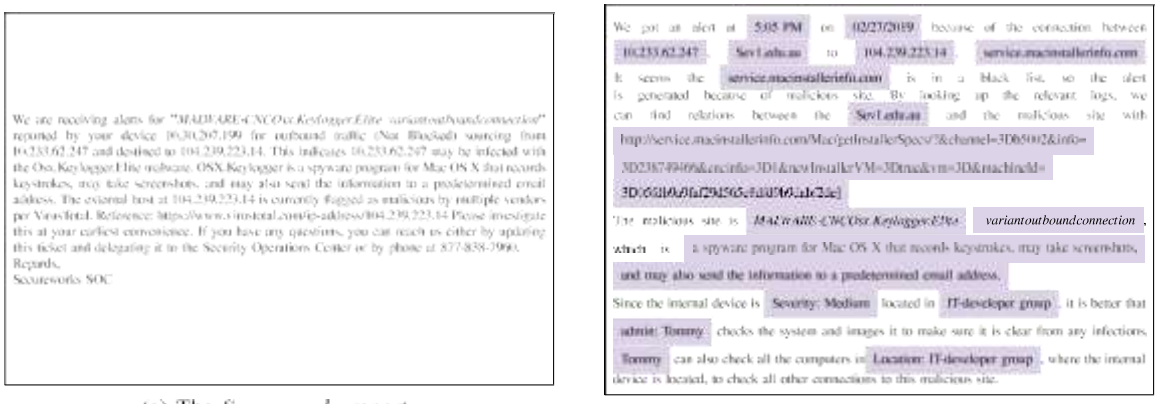
The example of the alert message produced by Secure- External Server = *{(104.239.223.14, service.*1
0

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

works is as follows:

macinstallerinfo.com, Urls, US)*

The 3rd stage: Since this paper focuses only on mal-ware, only Snort rules related to malware with the following



(a) The *Secureworks* report.

(b) The *Storytelling* report.

FIGURE 2. The reports generated in response to the security alert by both the default (a) and the proposed (b) solution.

titles were searched to identify the matched classification phrases: *snort3-malware-backdoor.rules,snort3-malware-cnc.rules,snort3-malware-other.rules, snort3-malware-tools.rules*. The matched Snort rule, which was mapped to the message part from the basic field, was as follows:

```
  alert tcp HOME_NET any ->
EXTERNAL_NET HTTP_PORTS (
msg:"MALWARE-CNC Osx.Keylogger.Elite
variant outbound connection";
flow:to_server,established; http_uri;
content:"/read-mip.php",fast_pattern,
nocase;metadata:impact_flag
red,policy balanced-ips drop,policy
security-ips drop; service:http;
reference:url,virustotal.com/en/file/
e23cae7189d6ca9c649afc22c638a45fd94f
19ef6b-585963164cca52c7b80f9b/analysis/;
classtype:trojan-activity; sid:41458;
rev:1; )
```

C. INFERENCE LAYER

The purpose of this layer is to answer the *what*, *who* and *why* questions about the incident. ‘‘MALWARE-CNC Osx.Keylogger.Elite variant outbound connection’’ was the malware classification phrase (according to: Extraction layer, 3rd stage). The definition of this malware was extracted from web articles in cybersecurity field stored in Global knowledge base. The extracted definition for the case study was compiled as follows: *malware classification phrase ’is’+ behaviour*. The definition was found under the ‘Behaviour’ node from the Symantec website,[11] and included:

> *‘‘OSX.Keylogger is a spyware program for Mac OS X that records keystrokes, may take screenshots, and may also send the information to a predetermined email address.’’*

Then, the malicious URLs were classified into five classes, each represented by the max-length URL. These were

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

searched among the 12 relevant logs to provide an evidencefor the incident. The URL that was matched in the relevant logs was randomly selected for use in the next layer. Since the infected server was not located in the financial department, and the severity was Medium, 2 rules based on Severity and Location were triggered, and the corresponding template foreach of them was selected.

D. STORY LAYER

The story based on the automatic retrieval of the variables from previous layers was generated in this layer. Complete template was contrasted against the report obtained from thecommercial tool.

The report produced by the proposed model (Figure 2a) was compiled fully automatically, while the Secureworks report (Figure 2b) required *both* machine processing and human assistance.

E. EVALUATION

Since formal evaluation of narrative format of both reports is qualitative in nature, the improvement in cyber threat man- agement proves a challenging task. In this paper, we focusedon the core questions to be answered in the report (i.e. actor(who), riskiness (what), and evidence (how)) as a basic for the proposed model evaluation. Thus, the following two cri- teria were defined: (1) Completeness, and (2) Comprehen- sion. In our case, the completeness refers to the amount of information required to obtain full comprehension about the situation. By assumption, the storytelling model due to its auto-fill function from various knowledge bases provide the complete information required to take action. On the other hand, the standard report (Secureworks in this case) entails manual search for missing information. To increase results reliability, the additional 10 alerts were investigated.

Since different types of alerts require different investiga- tion time, the random sample of 11 alerts in total messages was selected. An expert from the SOC team was involved in the empirical alert analysis consisting of filling the missing information from internal and external sources (similarly to

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

TABLE 2. Emprical Results.

Alert Messages	Completeness				Completeness Time				
	Definition	APPs in Definition	Malicious Site Info	Malware Evidence	Definition [s]	Relevant Logs [s]	Malicious Site Info [s]	Malware Evidence [s]	**Total Time [s]**
1. *Win.Trojan.Downloaderiani*	No	No	No	No	50	1200	33	201	**1484**
2. *Win.Trojan.Downloaderiani*	No	No	Yes*	No	50	1140	21	60	**1271**
3. *Gamarue.Andromeda.Trojan.PhoneHome*	No	No	No	No	75	1380	24	360	**1839**
4. *Agent.known.malicious.Virut*	No	Yes	No	No	68	1260	19	130	**1477**
5. *Agent.known.malicious.RookIE*	No	No	Yes*	No	154	1020	22	80	**1276**
6. *Win.Virut.Trojan*	Yes	No	No	No	0	1200	25	360	**1585**
8. *MALWARE/CNC/Osx.Keylogger.Elite*	Yes	Yes	No	No	0	1320	24	145	**1489**
9. *Agent.known.malicious.Virut*	No	Yes	No	No	68	1080	21	99	**1268**
10. *Win.Trojan.Downloaderiani*	No	No	No	No	60	1260	24	128	**1472**
11. *MALWARE/CNC/Osx.Keylogger.Elite*	Yes	Yes	No	No	0	1320	24	128	**1472**

the model proposed). The Secureworks reports for the alerts classified as malware (Potential Device Compromise) were obtained between 11/02.2019 and 28/02/2019 at the education institute. Table 2 shows the status of the knowledge required to complete the report ('Completeness' header). The expert manually retrieved the necessary information, and the extraction time was measured in seconds ('Completeness Time' header). The average extraction time across the 11 malware alerts was 1455.(36) s (approx. 25 mins). Thus, in total it took approx. 30 mins to answer the core questions about the actor, riskiness, and evidence (completeness = 25 mins comprehension 5 mins). As a result, the proposed model reduced the time to respond based on the full understanding of the situation by approx. 83% (25/30). In the storytelling model, given sufficient information on *what*, *who*, and *why* aspects, the time taken to obtain complete comprehension about an alert is approx. 5 mins (= 300 s). The time required for understanding is directly related to the degree of completeness (missing information has to be searched and extracted manually).

We also investigated the scenario where the 11 alerts occurred in a consecutive manner (busy period). To avoid potential damage and further escalation, the alerts should be addressed immediately. Time to respond to all alerts was set as a cumulative sum of Completeness Comprehension Time of each consecutive alert. Since the alerts are processed sequentially, the total response time builds up. Table 3 demonstrates the cumulative delay time to respond to an alert in the case of 11 consecutive alerts received in a day. Given the scenario, the proposed model has the potential to reduce the response time by approx. 17000 s (approx. 6 days) in comparison with the report derived in a semi-manual manner by the SOC team (existing approach). Please note that human limitation and environment limitation were not considered in the experiment.

TABLE 3. Empirical evaluation of the consecutive alerts.

Alert No	Total Time by SOC		Total Time by LDSM	
	Completeness	Comprehension	Completeness	Comprehension
1	1484	300	0	300
2	2755	600	0	600
3	4594	900	0	900
4	6671	1200	0	1200
5	7347	1500	0	1500
6	8766	1800	0	1800
7	10308	2100	0	2100
8	11797	2400	0	2400
9	13065	2700	0	2700
10	14537	3000	0	3000
11	16009	3300	0	3300
	19309 [s]		3300 [s]	

VI. DISCUSSION

A. STUDY CONTRIBUTIONS

The improvement from human-computer interaction perspective in security alerts handling will be discussed using the main two criteria: (1) Completeness, and (2) Comprehension.

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

Completeness: The information in the Secureworks report was insufficient for prompt inference, and the SOC member had to manually gather the complementary data from different sources. For instance, the information about the risk severity *(medium)* as well as the internal location of the device *(IT-developer group)* were missing. Also, the action recommendation *(check the system and images)* and person designation *(admin Tommy)* proved beneficial for timely and coordinated response. The utilisation of Local and Global knowledge bases aimed to provide the rich and comprehensive context around the incident. The template was filled using both internal information as well as the external sources. While the proposed model extracted the relevant knowledge automatically, the Secureworks report still required human involvement in the process. Also, the interpretation of cybersecurity is heavily reliant on analytical experience and knowledge (where and how to search for relevant information?), which puts strain on already scarce cybersecurity resources.

Comprehension: Narrative technique application in cyber risk management domain was aimed to reduce the cognitive load imposed on cybersecurity analysts while processing the large number of logs. The reports generated in storytelling manner proved more human-readable, facilitated comprehension, and effectively allowed for faster response to potential threat (time factor is found crucial in cybersecurity domain). Also, human-friendly format of the report contributed towards wider audience engagement into cyber situation awareness (currently restricted to security professionals). As an example, the user of the infected device can receive the storytelling report and obtain an insight into the cyber situation instantly, thus preventing further problem escalation. The narrative format assists understanding despite lack of expertise in cyber security domain. Finally, the capability to provide the reports at different level of details automatically enabled to cater for various information needs and intended aim (i.e. low-level for Security Operation Centre, high-level for Top Management).

Summary: By comparison between the generated story and the Secureworks report, the following can be inferred:

- The storytelling report is generated fully automatically, reducing the burden on cybersecurity resources;
- The implicit knowledge (what happened and why?), which analysts have to investigate manually, is included in the generated story;
- The log files with private information that cannot be sent to the third party for further processing are protected.

In terms of the current limitations, in this paper we only focused on malware taxonomy for approach demonstration. Still, the model can be easily adapted to other types of incidents by providing the complementary sources in Local and Global knowledge base. Also, since the enriched report for a security alert in a story design is not available, we were not able to perform the direct comparison with the proposed storytelling model. Thus, the impact of the narrative format has been assumed to be beneficial for cognitive workload reduction based on empirical observation at SOC team at the university.

In terms of future directions, the proposed solution can be extended beyond the educational sector. Cyber threats are currently commonplace across organisations. The overall benefits of narrative style would contribute staff comprehension, regardless the industry. Also, the additional validation metrics (readability score, user survey, time-to-respond, etc.) on larger-scale data could be provided to further confirm the benefits of the approach.

VII. CONCLUSION

The report generated by the proposed model proved to be more complete and more comprehensible for the SOC team in comparison with the Secureworks report. As a result, the cognitive effort in information digestion and understanding was significantly reduced. Also, due to the human-friendly format, a wide range of staff with different levels of expertise was able to be involved in cyber risk management process.

REFERENCES

[1] I. Dickson. (2017). *Text Classification of Network Intrusion Alerts to Enhance Cyber Situation Awareness and Automate Alert Triage*. [Online]. Available: https://www.dst.defence.gov.au/publication/

[2] M. Muggler, R. Eshwarappa, and E. C. Cankaya, ''Cybersecurity management through logging analytics,'' in *Proc. Adv. Intell. Syst. Comput.*, Jul. 2017, pp. 3–15.

[3] M. Albanese, H. Cam, and S. Jajodia, ''Automated cyber situation awareness tools and models for improving analyst performance,'' in *Proc. Adv. Inf. Secur. Cybersecur. Syst. Hum. Cognition Augmentation*, Sep. 2014, pp. 47–60.

[4] P. Jayathilake, N. R. Weeraddana, and H. K. E. P. Hettiarachchi, ''Automatic detection of multi-line templates in software log files,'' in *Proc. 17th Int. Conf. Adv. ICT Emerg. Regions (ICTer)*, Sep. 2017, pp. 1–8.

[5] J. Vink. (2010). *Storytelling|Design Research Techniques*. [Online]. Available: http://designresearchtechniques.com/casestudies/storytelling/

[6] Q. Wu, Z. Shen, C. Leungy, H. Zhang, Ailiya, Y. Cai, and C. Miao, ''Internet of Things based data driven storytelling for supporting social connections,'' in *Proc. IEEE Int. Conf. Green Comput. Commun. IEEE Internet Things IEEE Cyber, Phys. Social Comput.*, Aug. 2013, pp. 383–390.

[7] D. P. Barrett, S. A. Bronikowski, H. Yu, and J. M. Siskind, ''Robot language learning, generation, and comprehension,'' *CoRR*, vol. abs/1508.06161, 2015. [Online]. Available: http://arxiv.org/abs/1508.06161

[8] D. Kim and J. Lee, ''Designing an algorithm-driven text generation system for personalized and interactive news reading,'' *Int. J. Hum. Comput. Interact.*, vol. 35, no. 2, pp. 109–122, Jan. 2019.

[9] C. Matt, ''The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority,'' *Digit. J.*, vol. 3, pp. 416–433, May 2015.

[10] N. Lee, K. Kim, and T. Yoon, ''Implementation of robot journalism by programming custombot using tokenization and custom tagging,'' in *Proc. 19th Int. Conf. Adv. Commun. Technol. (ICACT)*, 2017, pp. 566–570.

[11] A. Graefe, *Guide to Automated Journalism*. New York, NY, USA: Tow Center for Digital Journalism, 2016.

[12] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins, ''#FluxFlow: Visual analysis of anomalous information spreading on social media,'' *IEEE Trans. Visual. Comput. Graph.*, vol. 20, no. 12, pp. 1773–1782, Dec. 2014.

[13] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, ''The role of uncertainty, awareness, and trust in visual analytics,'' *IEEE Trans. Visual. Comput. Graph.*, vol. 22, no. 1, pp. 240–249, Jan. 2016.

[14] W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan, ''Detecting large-scale system problems by mining console logs,'' in *Proc. ACM SIGOPS 22nd Symp. Operating Syst. Princ. (SOSP)*, 2009, pp. 117–132.

[15] M. Aharon, G. Barash, I. Cohen, and E. Mordechai, ''One graph is worth a thousand logs: Uncovering hidden structures in massive system event logs,'' in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2009, pp. 227–243.

ISSN: 2278-4632
Vol-12 Issue-07 No.03 July 2022

[16] A. Samii and W. Koh. *Interactive Visualization for System Log Analysis*. Accessed: Sep. 6, 2019. [Online]. Available: https://pdfs.semanticscholar.org/880c/99fafd7a0c051139fc95d01ddbca327553f5.pdf
[17] T. Li, J. Wu, L. Xue, D. Bao, Y. Jiang, C. Zeng, B. Xia, Z. Liu, W. Zhou, X. Zhu, W. Wang, and L. Zhang, ''FLAP: An end-to-end event log analysis platform for system management,'' in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2017, pp. 1547–1556.
[18] A. Azodi, F. Cheng, and C. Meinel, ''Towards better attack path visualizations based on deep normalization of host/network IDS alerts,'' in *Proc. IEEE 30th Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, Mar. 2016, pp. 1064–1071.
[19] P. Nimbalkar, V. Mulwad, N. Puranik, A. Joshi, and T. Finin, ''Semantic interpretation of structured log files,'' in *Proc. IEEE 17th Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2016, pp. 549–555.
[20] F. Menges and G. Pernul, ''A comparative analysis of incident reporting formats,'' *Comput. Security*, vol. 73, pp. 87–101, Mar. 2018.
[21] (2017). *Structured Threat Information Expression STIXTM*. [Online]. Available: https://docs.google.com/document/d/1IcA5KhglNdyX3tO17bBluC5nqSf70M5qgK9%nuAoYJgw/
[22] R. Cover. (2008). *Incident Object Description and Exchange Format (IODEF)*. Accessed: Sep. 6, 2019. [Online]. Available: http://xml.coverpages.org/iodef.html
[23] J. Komárková, M. Husák, M. Laštovička, and D. Tovarňák, ''CRUSOE: Data model for cyber situational awareness,'' in *Proc. 13th Int. Conf. Availability, Rel. Secur.*, 2018, p. 36.
[24] (2008). *X-ARF Network Abuse Reporting 2.0*. Accessed: Sep. 6, 2019. [Online]. Available: http://xarf.org/
[25] P. Pawlinski, P. Jaroszewski, J. Urbanowicz, P. Jacewicz, P. Zielony, P. Kijewski, and K. Gorzelak, ''Standards and tools for exchange and processing of actionable information,'' in *Proc. Eur. Union Agency for Netw. Inf. Secur.*, Heraklion, Greece, 2014.
[26] M. Hyvärinen, ''Analyzing narratives and story-telling,'' *The SAGE handbook social Res. methods*, pp. 447–460, 2008.
[27] J. Gargano and K. Weiss. *Windows Defender Security Intelligence*. Accessed: Sep. 6, 2019. [Online]. Available: https://www.microsoft.com/en-us/wdsi/threats/malware-encyclopedia-description
[28] K. Jana, M. Husak, M. Lastovicka, and D. Tovarnak, ''Crusoe: Data model for cyber situational awareness,'' in *Proc. 13th Int. Conf. Availability, Rel. Security*. Hamburg, Germany, Aug. 2018.
[29] V. Total. *Virustotal-Free Online Virus, Malware and URL Scanner*. Accessed: Sep. 6, 2019. [Online]. Available: https://www.virustotal.com/en
[30] ThreatMiner. *Data Mining for Threat Intelligence*. Accessed: Sep. 6, 2019. [Online]. Available: https://www.threatminer.org/
[31] M. Roesch, ''Snort: Lightweight intrusion detection for networks,'' *Lisa*, vol. 99, pp. 229–238, Jun. 1999.
[32] S. Egorov and G. Savchuk, ''SNORTRAN: An optimizing compiler for snort rules,'' *Fidelis Secur. Syst.*, pp. 1–8, 2002.
[33] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, ''Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence,'' in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2016, pp. 755–766.
[34] D. Stevenson. (Apr. 2018). *Foresee: Human and Machine Learning Working Together*. https://www.secureworks.com/blog/foresee-human-and-machine-learning-wor%king-together