Data-Driven Field Mapping of Security Logs for Integrated Monitoring

Seungoh Choi, Yesol Kim, Jeong-Han Yun, Byung-Gil Min, Hyoung-Chun Kim

Chapter 13

DATA-DRIVEN FIELD MAPPING OF SECURITY LOGS FOR INTEGRATED MONITORING

Seungoh Choi, Yesol Kim, Jeong-Han Yun, Byung-Gil Min and Hyoung-Chun Kim

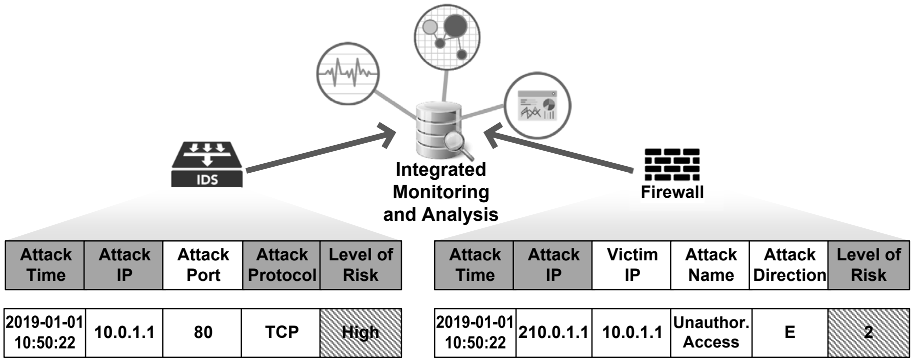
Abstract As industrial control system vulnerabilities and attacks increase, security controls must be applied to operational technologies. The growing demand for security threat monitoring and analysis techniques that integrate information from security logs has resulted in enterprise security management systems giving way to security information and event management systems. Nevertheless, it is vital to implement some form of pre-processing to collect, integrate and analyze security events efficiently. Operators still have to manually check entire security logs or write scripts or parsers that draw on domain knowledge, tasks that are time-consuming and error-prone.

To address these challenges, this chapter focuses on the data-driven mapping of security logs to support the integrated monitoring of operational technology systems. The characteristics of security logs from security appliances used in critical infrastructure assets are analyzed to create a tool that maps different security logs to field categories to support integrated system monitoring. The tool reduces the effort needed by operators to manually process security logs even when the logged data generated by security appliances has new or modified formats.

Keywords: Security, event logs, integrated system monitoring

1. Introduction

The vulnerabilities of industrial control systems used in critical infrastructure assets and the sophistication of attacks have increased significantly in recent years. In 2016, the U.S. Department of Homeland Security's ICS-CERT reported 257 new vulnerabilities in industrial control systems [9]. Meanwhile,



Attack Time	Attack IP	Attack Port	Attack Protocol	Level of Risk
2019-01-01 10:50:22	10.0.1.1	80	TCP	High

Attack Time	Attack IP	Victim IP	Attack Name	Attack Direction	Level of Risk
2019-01-01 10:50:22	210.0.1.1	10.0.1.1	Unauthor. Access	E	2

Figure 1. Challenges involved in integrating security log data of different formats.

new operating environments and wireless technologies used in industrial control systems are increasing their attack surfaces.

Security devices are being incorporated in operational technology environments to combat cyber threats to industrial control systems and the critical infrastructure assets they manage. The growing demand for security threat monitoring and analysis techniques that integrate information from security logs has resulted in enterprise security management systems giving way to security information and event management (SIEM) systems, whose security logs contain valuable information about the security status and traceability of the operating environments. The information, which is in structured or unstructured formats, helps detect anomalies and analyze the causes of security incidents, contributing to the development of appropriate countermeasures. Big data technologies are actively being applied to security log datasets to enhance attack detection, security incident investigations and mitigation techniques.

However, operators of critical infrastructure assets have great difficulty integrating the diverse formats of security data in multiple device logs to perform security monitoring and analyses. Figure 1 shows the challenges involved in integrating security log data of different formats. The fields in white are specific to security devices whereas the fields in grey are common to security devices. However, the levels of risk have different data types and semantics.

Although standards exist for presenting security event information from individual devices, the standards may not be supported by device manufacturers and/or the formats may differ considerably. Even in the case of the Common Event Format (CEF) [1], which is designed to support interoperability in security information and event management systems, different manufacturers use different extension fields. Manufacturers that use the Syslog standard often add new fields according to their needs. The formats and contents of security events also differ based on the security policies applied at field sites where the devices are configured and operated.

Additionally, even when security logs have data fields with the same semantics, the logs are difficult to integrate because the data is in different formats. For example, data in an attack severity field may be expressed numerically as 1, 5 and 10 in one log, but may expressed using the strings "caution," "severe" and "serious" in another log. Such correspondences cannot be solved simply by identifying the same expression type in order to analyze the fields correctly.

Furthermore, depending on the nature of the infrastructure assets, various security devices are employed to enhance security or monitor the operating environments; this requires the integration of diverse security logs. The problems are acerbated because it is often difficult to acquire data specifications from manufacturers. As a result, operators require pre-processing modules for the security logs or they have to apply manual efforts that draw on their knowledge and experience.

In order to overcome these problems and support integrated security monitoring, security logs were examined to identify the target fields needed for security analysis and context awareness. The results were used to create a tool that derives the characteristics of security logs and identifies fields that match the target fields. The tool enhances security monitoring by enabling the integration of security logs and new formats in security logs from newly added or replaced security devices while minimizing the manual effort required on the part of developers.

2. Related Work

RFC 4765 [6] published in 2007 by the Internet Research Task Force (IETF) specifies the Intrusion Detection Message Exchange Format (IDMEF) as an information exchange standard for operating and managing security devices and systems (e.g., intrusion detection systems (IDSs) and intrusion prevention systems (IPSs)). IDMEF specifies the heartbeats that pass status information between equipment and systems, and alerts that pass network attack detection information. In 2008, MITRE released the Common Event Expression (CEE) format [11] that expresses and standardizes log exchanges between systems and end users, log providers and security information and event management system vendors. CEE provides a common representation language through the profile, Common Log Syntax (CLS) and Common Log Transport (CLT) throughout event handling, including event structuring, event encoding/decoding and event transmission. In 2013, MITRE introduced the Structured Threat Information Expression (STIX) format for organizing and expressing cyber threat information. STIX has been adopted for the trusted automated exchange of indicator information about cyber threats in real time.

Other entities have developed and released common event specifications. Some of the important specifications are:

- **Common Event Format (CEF):** ArcSight [1] designed CEF for logging and audits, and for security information and event management. CEF is primarily used with Syslog and provides custom fields for scalability.

- **Log Event Extended Format (LEEF):** IBM [7] developed LEEF as a custom event format for its Security QRadar products.

- **Cisco Intrusion Detection Event Exchange (CIDEE):** The Cisco CIDEE format [5] extends the Security Device Event Exchange (SDEE) standard that provides specifications for the formats and protocols used to exchange events. CIDEE is a custom event format that is used by Cisco intrusion prevention systems to exchange intrusion information.

Commercial vendors of security appliances typically do not comply strictly with the standards for security log formats. There are some similar fields and formats, but the details are different for each vendor, product and version. Since security devices are not designed to interoperate with devices from other manufacturers, the formats of individual fields in the security information they generate are not disclosed.

Plaintext protocol reversing efforts have been conducted to extract information from communications protocols [4, 8]. Most research efforts have focused on open-text protocols such as SMB and HTTP. However, recent studies have attempted to obtain information about private communications protocols between command and control servers and bots in order to detect and respond to distributed denial-of-service (DDoS) attacks by botnets [2, 3, 10]. These studies concentrate on the field separation of communications data and analyzing the context or state from server-client conversations. As a result, this work cannot be applied to map different security log formats to a single field required for monitoring purposes.

3. Analysis of Field Characteristics

Four security appliances that are widely used in critical infrastructure assets were employed to analyze the characteristics of the fields in security logs. Ixia's Ixload, which can reproduce security violation situations, was used to enable the appliances to generate security logs.

Table 1 summarizes the attacks used in this research. They include 13 well-known flooding attacks and 6,740 vulnerabilities and malware attacks.

A total of 1,146,019 security logs were collected in the experimental environment over a ten-hour period. Table 2 shows a summary of the security logs. Note that the number of collected security logs differs from one security device to another due to differences in the types, numbers and detection methods of the security policies supported by the device vendors.

3.1 Target Fields in Security Logs

Analysis of the four types of security logs collected in the experimental environment confirmed that structural differences exist, e.g., for field numbers, types and contents. For security reasons, only limited information – not the detailed field structures – are described in this chapter.

Table 1. Summary of attacks.

Protocol	**Attack**
ARP	ARP flooding attack
ICMP	Fragmented ICMP message attack Ping of death attack Smurf attack
IGMP	Fragmented IGMP message attack
IP	Fragmented IP message attack Teardrop attack
TCP	Fragmented ACK flooding attack LAND attack SYN flooding attack Xmas tree attack
UDP	UDP flooding attack UDP fragment attack

Table 2. Summary of security logs.

Manufacturer	**Device**	**Security Logs**	
		Format	**Total (Proportion)**
A	IDS/IPS	I	117,850 (0.10)
B	IDS/IPS	II	672,624 (0.59)
C	IDS/IPS	III	52,801 (0.05)
C	Firewall	IV	302,744 (0.26)

Table 3. Target fields in security logs.

Manufacturer	**Device**	**Security Log Fields**		
		Format	**Number**	**Targets**
A	IDS/IPS	I	13	10
B	IDS/IPS	II	34	17
C	IDS/IPS	III	27	12
C	Firewall	IV	18	8

Table 3 shows the target fields in the security logs. Note that the IDS/IPS from A has the fewest fields (13) whereas the IDS/IPS from B has the most fields (34), more than 2.6 times more fields than A. The IDS/IPS from C has more fields (27) because its manufacturer uses various field structures in its devices according to their models and functions. Furthermore, even if each security log has the same field name, the field type or content may be different. For example, the Protocol field is represented differently by case-sensitive strings or numbers, such as “TCP” or “Tcp” or 6, depending on the manufacturer.

It is not necessary to use all the fields because other fields in a log may contain the associated information. For example, when ID-Rule and Name-Rule have the same meaning, ID-Rule is the key value that uniquely distinguishes the security policy whereas Name-Rule is an annotation used by administrators for easy recognition.

There may be unnecessary fields in terms of the semantics when performing integrated security monitoring. For example, Type-CategoryAttack is mainly used to classify detection results. However, it does not precisely classify and identify the attack type because the classification is too broad. In the case of Length-RawPacket, there are some difficulties in deriving a security threat by only examining the packet length.

Therefore, original field structure analyses were performed for three IDS/IPS devices and one firewall from the perspective of security monitoring and analysis. This resulted in the exclusion of three (minimum) to 17 fields (maximum). A field that was not included in all the security logs was excluded, but it was retained if it was deemed necessary for security monitoring and analysis.

3.2 Field Categories in Security Logs

In order to categorize the target fields listed above, the meanings of the 47 target fields in the security logs were analyzed. The target fields could be represented using 17 field-category-consolidated fields. Table 4 shows the categories of fields included by the manufacturers along with their contents. Seven categories of fields were included in all the security logs – Time-Sent, IP-Attacker, IP-Victim, Port-Attacker, Port-Victim, Type-AttackProtocol and Type-Action. The other categories of fields were included in some of the security logs.

The analysis also confirmed that the field categories depended on the types of security devices. The field categories of the security logs generated by IDS/IPS devices mainly deal with attack-related information such as the attack name, type and direction. On the other hand, certain categories of log fields were common regardless of the types of security devices. For example, ID-Rule was generated by IDS/IPS devices for the signature-based detection function. In the case of the firewall with an access-control-list-based security policy, ID-Rule was used even in the deny rules.

3.3 Syntax of Field Categories

The data types and main features of the fields in the security logs were analyzed in order to map field information such as field name and field meaning based on the field categories.

First, the field data types were analyzed and classified as String and Number as shown in Table 5. The String type is divided into Word (single length of text that does not contain spaces) and Sentence (collection of words separated

Table 4. Categories of fields in the security logs.

Field Category		**Security Log Format**				**Information**
Major	**Minor**	**I**	**II**	**III**	**IV**	
Time	Sent	✓	✓	✓	✓	Time of sent log
	Attack	–	✓	–	–	Time of attack start
	AttackEnd	–	✓	–	–	Time of attack end
IP	Detector	–	✓	–	–	IP address of device that detected attack
	Attacker	✓	✓	✓	✓	IP address of attacker
	Victim	✓	✓	✓	✓	IP address of victim
Port	Attacker	✓	✓	✓	✓	Port number of attacker
	Victim	✓	✓	✓	✓	Port number of victim
Name	Machine	–	✓	–	–	Name of device that detected attack
	Attack	✓	✓	✓	-	Name of detected attack
Type	Attack	–	✓	✓	–	Type of detected attack
	AttackDirection	-	✓	✓	–	Type of detected attack
	AttackProtocol	✓	✓	✓	✓	Type of transport protocol
	Action	✓	✓	✓	✓	Type of action against detected attack
Level	Risk	✓	✓	✓	–	Level of severity of detected attack
Count	TotalAttack	–	✓	✓	–	Total number of detected attacks
ID	Rule	✓	✓	–	✓	Rule ID that detected attack

by spaces). The Keyword type is a subtype of the Word type when the text is unique. In addition, special subtypes such as Time and IP are included for the String type. The Number type has the subtypes Constant (fixed numerical value) and Variable (variable numerical values).

Second, the field data was analyzed based on the field categories. The analysis yielded the data types shown in Table 6. To enhance understanding, each field category is arranged according to its type. The analysis confirmed that the field categories and types cannot be matched uniquely due to the different data formats in the security logs produced by the appliances.

Table 5. Field type categories.

Type	Subtype	Context
String	Word	Single text
	Keyword	Single unique text
	Sentence	Multiple text
	Time	Timestamp
	IP	IP address
Number	Constant	Single fixed numerical value
	Variable	Variable numerical values

Table 6. Mapping between field categories and types.

Field Category		Time	IP	Word	Keyword	Sentence	Constant	Variable
Major	Minor							
Time	Sent	✓	–	–	–	–	–	–
Time	Attack	✓	–	–	–	–	–	–
Time	AttackEnd	✓	–	–	–	–	–	–
IP	Detector	–	✓	–	–	–	–	–
IP	Attacker	–	✓	–	–	–	–	–
IP	Victim	–	✓	–	–	–	–	–
Port	Attacker	–	–	–	–	–	–	✓
Port	Victim	–	–	–	–	–	–	✓
Count	TotalAttack	–	–	–	–	–	–	✓
Type	AttackProtocol	–	–	✓	–	–	–	✓
Level	Risk	–	–	✓	–	–	–	✓
Type	Action	–	–	✓	–	–	–	✓
ID	Rule	–	–	✓	–	–	–	✓
Name	Attack	–	–	✓	–	✓	–	–
Type	Attack	–	–	✓	–	✓	–	–
Type	AttackDirection	–	–	✓	–	–	–	–
Name	Machine	–	–	–	✓	–	–	–

3.4 Semantics of Field Categories

The data characteristics are prominent in the case of a field category that maps to the Number type. To clarify the semantics, features were extracted from predefined information such as the communications protocol. The fields Port-Attacker and Port-Victim use numbers in the range 1 to 65,536 corresponding to two bytes of storage. On the other hand, Type-AttackProtocol has values from 0 to 255 because its values are represented by one byte in the IP headers.

Next, a situation was considered where a security event was generated in the operational environment as a result of an attack. Count-TotalAttack is always greater than zero because a security event occurs during an attack. Also, the

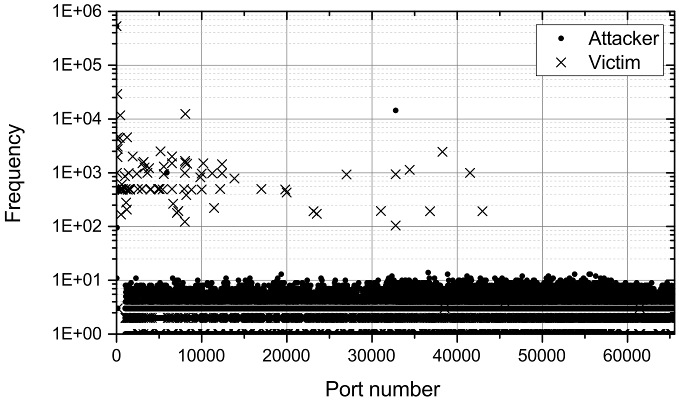


Figure 2. Field characteristics in security log format II.

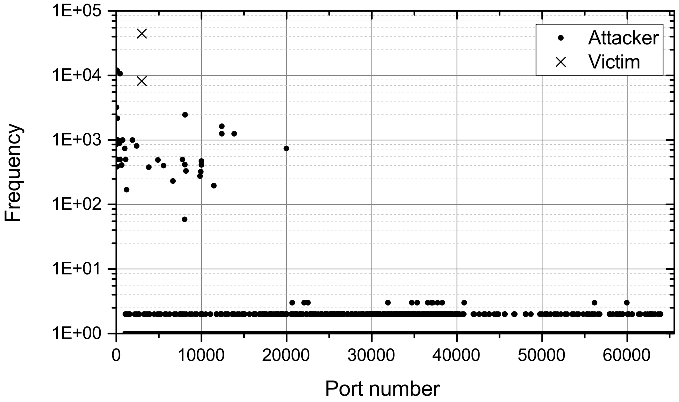


Figure 3. Field characteristics in security log format III.

number of attacks detected per unit time varies, so various values (including zero) could result.

Finally, characteristics were extracted from statistical patterns. The fields Port-Attacker, Port-Victim and Count-TotalAttack, which are mapped only to the Variable type, have different characteristics in terms of distributions of values (e.g., variance, skewness and kurtosis). Figures 2 and 3 show the significant differences that exist in the value distributions.

Tables 7 and 8 summarize the characteristics of the field categories.

4. Mapping Security Logs to Field Categories

This section demonstrates how security logs are mapped to field categories via data-driven analysis of the security logs.

Table 7. Characteristics of field categories.

Field Category		Type	Characteristics
Major	Minor		
Time	Sent	Time	Highest priority among fields of the Time type
Time	Attack	Time	Corresponds to earlier time among fields of the Time type, except for the Time-Sent field
Time	AttackEnd	Time	Corresponds to later time among fields of the Time type, except for the Time-Sent field
IP	Detector	IP	Value of the IP type field is one (unique); predefined IP address resolution is required
IP	Attacker	IP	Does not include a specific string (.1, .255); predefined IP address resolution is required
IP	Victim	IP	IP fields except for IP-Detector and IP-Attacker; predefined IP address resolution is required
Port	Attacker	Variable	Fields with range 0–65,536; occurrence distribution is forward
Port	Victim	Variable	Fields with range 0–65,536; occurrence distribution is backward
Count	TotalAttack	Variable	Fields with 1–max range; kurtosis is high
Type	AttackProtocol	Word	Fewer word values and higher frequencies of occurrence; predefined protocol name verification is required
		Variable	Fields with range 0–255; fewer numbers of values and higher frequencies of occurrence

4.1 Overview

The data-driven mapping of security logs to field categories involves three phases:

- **Phase 1: Field Preparation:** During this phase, the security log that is the subject of the field mapping is received as input. The security log is parsed to remove delimiters and produce individual fields.

- **Phase 2: Field Analysis:** During this phase, the type of each field in the security log is classified. The classification results are mapped to the data characteristics. Details about the fields are presented in Section 3.2.

Table 8. Characteristics of field categories (continued).

Field Category		Type	Characteristics
Major	Minor		
Level	Risk	Word	Distribution is biased
		Variable	Fields with range 1–5 (10); distribution is biased
Type	Action	Word	Distribution is biased; predefined information verification is required
		Variable	Number of field values is low; distribution is biased
ID	Rule	Word	Meaningless text; no predefined information
		Variable	Fields with range 65,536–max; lowest priority among fields of the Variable type
Name	Attack	Word	Predefined attack name verification is required
		Sentence	Highest priority among fields of the Sentence type
Type	Attack	Word	Low priority among fields of the Word type
		Sentence	Longest text among fields of the Sentence type
Name	Machine	Keyword	Highest priority among fields of the Keyword type
Type	AttackDirection	Word	Predefined string (E,I) verification is required

- **Phase 3: Field Mapping:** During this final phase, a field category is identified by combining the mapped type and data characteristics. The output is a candidate field category for each field.

4.2 Phase 1: Field Preparation

During the field preparation phase, the raw security log is processed as an input for the subsequent field analysis phase (Figure 4). Since raw security logs have different formats depending on the manufacturer and device model, they have to be grouped into the same format. The grouped security logs are separated into fields based on delimiters. The data is organized in a structure (e.g., matrix or data frame) that simplifies the analysis based on the field type that is conducted in the next phase.

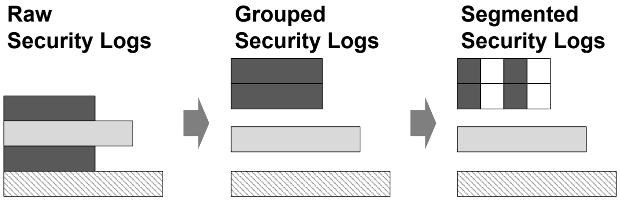


Figure 4. Field preparation using security logs.

Table 9. Results of mapping security logs to field categories.

Security Log Format	**Total Fields**	**Correctly Mapping to Field Category**		
		1st Candidate	**2nd Candidate**	**Total (%)**
I	10	9	0	9 (90)
II	17	16	1	17 (100)
III	12	9	3	12 (100)
IV	8	6	1	7 (87.5)
Total	47	40	5	45 (95.74)

4.3 Phase 2: Field Analysis

During the field analysis phase, the field type is first analyzed based on the field data. Following this, the field characteristics are analyzed to assign the field characteristics based on Tables 7 and 8.

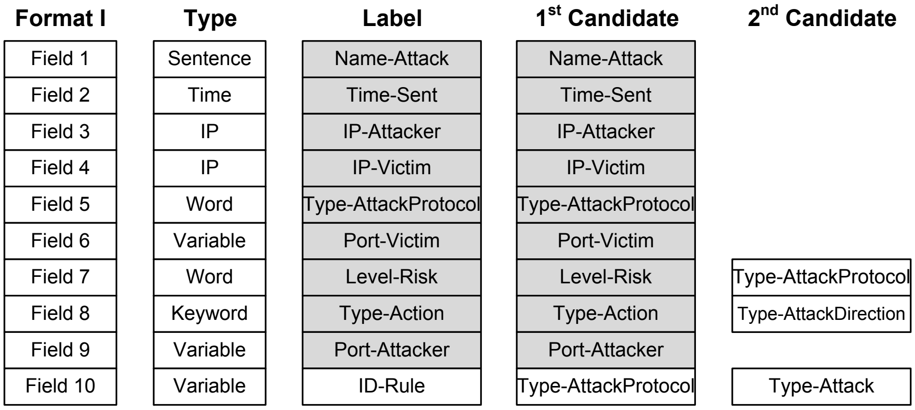
After the field data type analysis is complete, field category candidates are identified and mapped according to the individual field types. However, as described above, there could be multiple candidates for a given field. For this reason, the field data characteristics also have to be analyzed.

The priority and frequency of data are assigned to each field of the String type (e.g., Word, Sentence, Keyword, IP and Time). Because the Variable type has multiple numerical values, the minimum and maximum, variance and skewness of the values are as recorded as characteristics.

4.4 Phase 3: Field Mapping

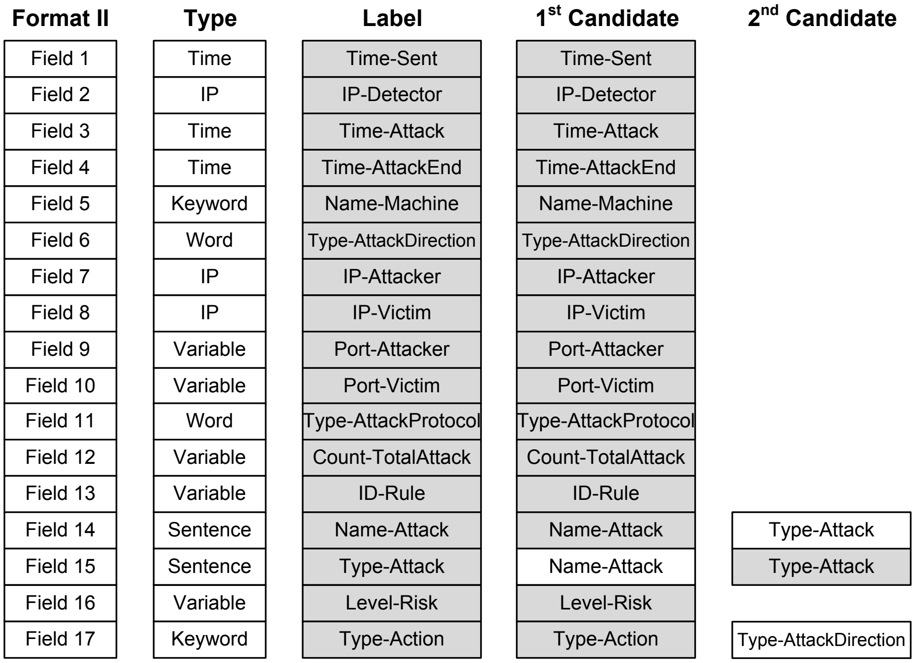
During the field mapping phase, the final type of each field is considered according to the priority of the candidate field category by analyzing the field type and characteristics provided by the field analysis phase. Upon applying the proposed method to the four security logs considered in this work, the field types were mapped as shown in Figures 5 through 8. The mapped candidates are presented in order of priority according to the data characteristics.

Table 9 summarizes the results of mapping security logs to field categories. As seen in the table, when the correct field categories were mapped to the first



Format I	Type	Label	1st Candidate	2nd Candidate
Field 1	Sentence	Name-Attack	Name-Attack	
Field 2	Time	Time-Sent	Time-Sent	
Field 3	IP	IP-Attacker	IP-Attacker	
Field 4	IP	IP-Victim	IP-Victim	
Field 5	Word	Type-AttackProtocol	Type-AttackProtocol	
Field 6	Variable	Port-Victim	Port-Victim	
Field 7	Word	Level-Risk	Level-Risk	Type-AttackProtocol
Field 8	Keyword	Type-Action	Type-Action	Type-AttackDirection
Field 9	Variable	Port-Attacker	Port-Attacker	
Field 10	Variable	ID-Rule	Type-AttackProtocol	Type-Attack

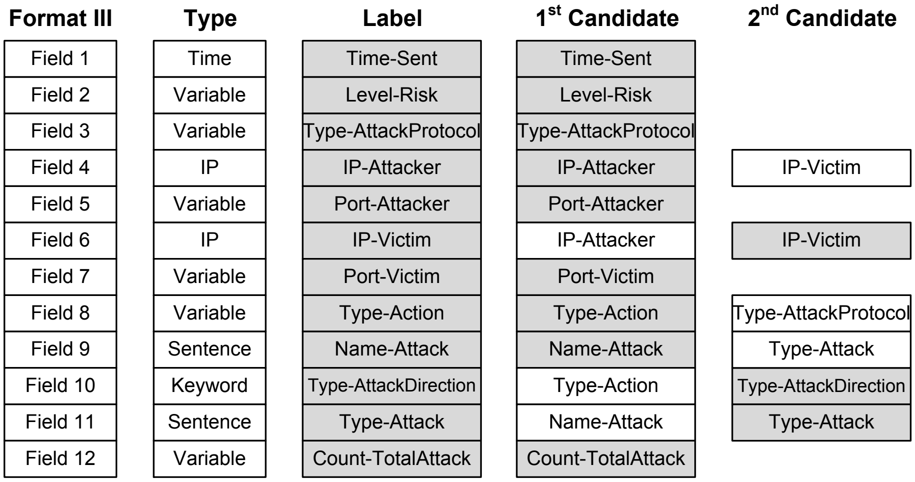
Figure 5. Results of mapping field categories in the security log with format I.



Format II	Type	Label	1st Candidate	2nd Candidate
Field 1	Time	Time-Sent	Time-Sent	
Field 2	IP	IP-Detector	IP-Detector	
Field 3	Time	Time-Attack	Time-Attack	
Field 4	Time	Time-AttackEnd	Time-AttackEnd	
Field 5	Keyword	Name-Machine	Name-Machine	
Field 6	Word	Type-AttackDirection	Type-AttackDirection	
Field 7	IP	IP-Attacker	IP-Attacker	
Field 8	IP	IP-Victim	IP-Victim	
Field 9	Variable	Port-Attacker	Port-Attacker	
Field 10	Variable	Port-Victim	Port-Victim	
Field 11	Word	Type-AttackProtocol	Type-AttackProtocol	
Field 12	Variable	Count-TotalAttack	Count-TotalAttack	
Field 13	Variable	ID-Rule	ID-Rule	
Field 14	Sentence	Name-Attack	Name-Attack	Type-Attack
Field 15	Sentence	Type-Attack	Name-Attack	Type-Attack
Field 16	Variable	Level-Risk	Level-Risk	
Field 17	Keyword	Type-Action	Type-Action	Type-AttackDirection

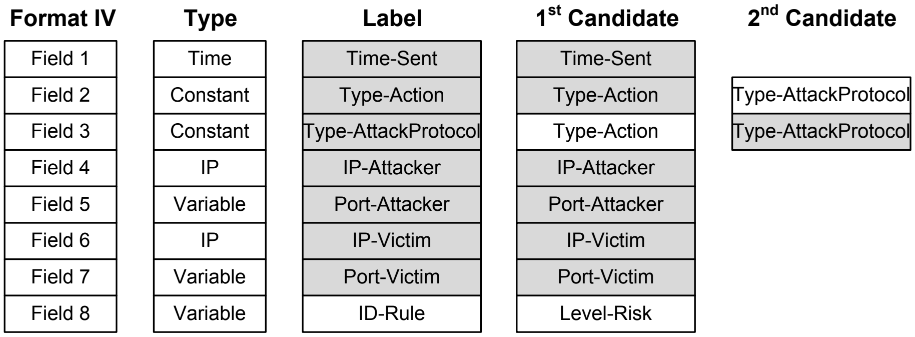
Figure 6. Results of mapping field categories in the security log with format II.

candidates, 40 field categories correspond to approximately 85.11% of the total 47 fields. When the ranges of the choices are extended to the second candidates, 45 field categories correspond to 95.74% of the total 47 fields. Note that ID-Rule, Type-Action and Level-Risk are generally not found in the security logs



Format III	Type	Label	1^{st} Candidate	2^{nd} Candidate
Field 1	Time	Time-Sent	Time-Sent	
Field 2	Variable	Level-Risk	Level-Risk	
Field 3	Variable	Type-AttackProtocol	Type-AttackProtocol	
Field 4	IP	IP-Attacker	IP-Attacker	IP-Victim
Field 5	Variable	Port-Attacker	Port-Attacker	
Field 6	IP	IP-Victim	IP-Attacker	IP-Victim
Field 7	Variable	Port-Victim	Port-Victim	
Field 8	Variable	Type-Action	Type-Action	Type-AttackProtocol
Field 9	Sentence	Name-Attack	Name-Attack	Type-Attack
Field 10	Keyword	Type-AttackDirection	Type-Action	Type-AttackDirection
Field 11	Sentence	Type-Attack	Name-Attack	Type-Attack
Field 12	Variable	Count-TotalAttack	Count-TotalAttack	

Figure 7. Results of mapping field categories in the security log with format III.



Format IV	Type	Label	1^{st} Candidate	2^{nd} Candidate
Field 1	Time	Time-Sent	Time-Sent	
Field 2	Constant	Type-Action	Type-Action	Type-AttackProtocol
Field 3	Constant	Type-AttackProtocol	Type-Action	Type-AttackProtocol
Field 4	IP	IP-Attacker	IP-Attacker	
Field 5	Variable	Port-Attacker	Port-Attacker	
Field 6	IP	IP-Victim	IP-Victim	
Field 7	Variable	Port-Victim	Port-Victim	
Field 8	Variable	ID-Rule	Level-Risk	

Figure 8. Results of mapping field categories in the security log with format IV.

because the fields have the same semantics but different types. The limitations are discussed in the following section.

5. Discussion

The three principal discussion points are:

- **Dictionary for Semantics:** The semantics of the same fields in the security logs must be reconciled. Fields may be semantically equivalent based on predefined information that is commonly used, such as standards and specifications, but there may be differences in the field categories. For example, in the case of the Type-AttackProtocol field, "HTTP" in the String type and 80 in the Number type have to be considered as having the same meaning. However, the semantics of fields can be different regardless of the field types according to the predefined information from the manufacturer. For example, in the case of the Type-Action field, even if the field type is Number and the value is 1, then the field meaning can be changed by the "Deny" or "Allow" characteristics, depending on the predefined information. In the case of the Level-Risk field, a value of 1 for the Number type may correspond to "Low" or "High" depending on the predefined information.

- **Correlated Analysis of Fields:** *A priori* information and the analysis results can support field inference. The analysis of security logs produced by the IDS/IPS device created by manufacturer A confirms that security events are configured in a key-value manner. In other words, since the key is already known, it is possible to derive the characteristics of the value corresponding to the key and to apply it to infer the fields in the same or other security logs with similar characteristics. The fields can be more accurately inferred using security log fields that are related to each other. For example, it is possible to apply association analysis between an IP (address) field type and a Variable field type such as Port-Number with the range 1 to 65,536 for more precise classification of an attacker or a victim.

- **Manual Field Mapping Process:** This research is a preliminary attempt to support integrated monitoring of critical infrastructure assets because only four major security appliances were considered. In a real-world environment, monitoring personnel must handle all the formats in the security logs maintained in critical infrastructure assets. This is a highly manual process that relies on domain knowledge and experience. Although the proposed approach has involved some manual analysis, it is still a useful first step to removing dependencies and providing useful information that can reduce operator error.

6. Conclusions

The data-driven mapping of security logs can support the integrated monitoring of operational technology systems in the critical infrastructure. The characteristics of security logs from security appliances used in critical infrastructure assets have been analyzed to create a tool that maps different security logs to field categories based on their field types and characteristics. This enables events in multiple security logs to be integrated automatically. Moreover, it reduces the effort on the part of operators to manually process security logs for integrated security monitoring when the logged data generated by existing or new security appliances have diverse formats. Future research will focus on improving the field mapping tool by considering a variety of security appliances and critical infrastructure assets and applications.

References

[1] ArcSight, Common Event Format, Revision 15, ArcSight Technical Note, Cupertino, California, 2009.

[2] J. Caballero, P. Poosankam, C. Kreibich and D. Song, Dispatcher: Enabling active botnet infiltration using automatic protocol reverse-engineering, *Proceedings of the Sixteenth ACM Conference on Computer and Communications Security*, pp. 621–364, 2009.

[3] J. Caballero and D. Song, Automatic protocol reverse-engineering: Message format extraction and field semantics inference, *Computer Networks*, vol. 57(2), pp. 451–474, 2013.

[4] J. Caballero, H. Yin, Z. Liang and D. Song, Polyglot: Automatic extraction of protocol message format using dynamic binary analysis, *Proceedings of the Fourteenth ACM Conference on Computer and Communications Security*, pp. 317–329, 2007.

[5] Cisco Systems, Cisco Intrusion Detection Event Exchange (CIDEE) Specification, San Jose, California (`www.cisco.com/c/en/us/td/docs/security/ips/specs/CIDEE_Specification.html`), 2009.

[6] H. Debar, D. Curry and B. Feinstein, The Intrusion Detection Message Exchange Format (IDMEF), RFC 4765, 2007.

[7] International Business Machines, IBM QRadar: Log Event Extension Format (LEEF), Version 2, Armonk, New York (`www.ibm.com/support/knowledgecenter/SS42VS_DSM/b_Leef_format_guide.pdf`), 2016.

[8] H. Li, B. Zhang, B. Shuai, J. Wang and C. Tang, Automatic protocol feature word construction based on machine learning, *Proceedings of the IEEE International Conference on Progress in Informatics and Computing*, pp. 93–97, 2015.

[9] National Cybersecurity and Communications Integration Center, ICS-CERT – Year in Review, Department of Homeland Security, Washington, DC (`ics-cert.us-cert.gov/Year-Review-2016`), 2016.

[10] A. Sood, R. Enbody and R. Bansal, Dissecting SpyEye – Understanding the design of third generation botnets, *Computer Networks*, vol. 57(2), pp. 436–450, 2013.

[11] The CEE Board, Common Event Expression, MITRE, McLean, Virginia (`cee.mitre.org/docs/Common_Event_Expression_White_Paper_June_2008.pdf`), 2008.

[12] Z. Wang, X. Jiang, W. Cui, X. Wang and M. Grace, ReFormat: Automatic reverse engineering of encrypted messages, *Proceedings of the Fourteenth European Conference on Research in Computer Security*, pp. 200–215, 2009.