Advancing SIEM Log Management Strategies through Vendor-Agnostic Measurement

GIAC (GCIH) Gold Certification

Author: Nate Street, nate.street@gmail.com
Advisor: Randy Marchany

Accepted: November 2, 2018

Abstract

The Security Information and Event Management (SIEM) system concept was introduced to the market decades ago, yet the initial promise of the tool remains largely unrealized for many organizations. Measuring the operational effectiveness of the SIEM continues to be a challenge due to the complexity of deployment, configuration of the numerous components, and the process of determining the logs the Incident Response team needs to perform a comprehensive security investigation. While security engineers may be able to deploy the necessary components according to the manual, they often lack the understanding of which logs provide valuable insight to answer the questions necessary to identify, contain, and eradicate malicious activity. Meanwhile, vendor recommendations often suggest sending all logs to the SIEM while simultaneously charging based on log ingestion rate. This creates a situation whereby vendors propose solutions to increase their bottom line budget versus the best interest of the customer. This paper explores a novel approach to quantifying the value of an individual log source sent to the SIEM. Through vendor-agnostic measurement, the algorithmic model utilized by the Log Quality Value (LQV) index enables security engineers and incident response teams to determine which logs provide the most value for security investigations. Two common attack patterns were assessed against the proof-of-concept tool, and a positive correlation was found between the LQV index and the critical logs used to investigate the attack. Future opportunities exist to evaluate the LQV algorithms against a more extensive dataset from live production environments and measure the tool effectiveness through periodically comparing the LQV index to logs used to detect security incidents. This research ultimately proposes a call-to-action for the security community to build more vendor-agnostic methods to independently measure the effectiveness of security products.

1. Introduction

Few security technologies exist that are as complex as the Security Information and Event Management (SIEM) tool. Once an organization purchases the product, security engineers are tasked with deploying a technology that will be utilized primarily by a separate Incident Response (IR) team to investigate potential security events or threats that occur on the organization's network. However, security engineers generally lack the Incident Response background required to fully understand the different types of log data that will be useful for the team during a potential security incident investigation.

Surprisingly, the very individuals within the organization who know how to perform log analysis in the context of a security investigation, i.e., members of the IR team, are typically not heavily involved during the deployment of the SIEM. For growing security programs, the organization might have yet to hire full-time and trained incident responders. This results in a situation in which an organization may have all the necessary hardware components installed successfully on the network, but may fail to gather the appropriate log data. The SIEM is thereby rendered useless or more often is the case, utilized for troubleshooting system and network connectivity problems versus alerting on malicious activity. This failure to gather requirements before implementing the project confirms Kumar's contention that "design problems based on poor requirements leads to design issues that are more difficult and expensive to resolve after project development" (Kumar 2006). The unfortunate outcome is a data breach that puts the organization in the headlines. To understand why so many organizations find themselves in this predicament, an appreciation of the many competing pressures that organizations labor under is necessary.

In the United States and abroad, governance and legislation surrounding data protection has grown in prevalence and severity requiring organizations to implement comprehensive security programs. Recently, the Securities and Exchange Commission (SEC) released interpretive Cybersecurity Disclosure Notices 33-10459 and 34-82746 stating “it is critical that public companies take all required actions to inform investors about material cybersecurity risks and incidents in a timely fashion, including those companies that are subject to material cybersecurity risks but may not have been the

target of a cyber-attack" (SEC, 2018). The impact of these directives is that regulators increase enquiry surrounding proper due diligence and handling of reported cyber incidents as well as the organizations historical treatment of security findings from audits and other assessments. The Federal Trade Commission (FTC) published the Privacy & Data Security Update detailing the enforcement actions and civil monetary penalties brought against U.S. companies in 2017 for failure to adequately protect consumers' personal information (FTC, 2017). This publication provides further evidence that the U.S. government is treating data protection seriously by invoking stiff financial penalties for negligent behavior.

In international markets, regulatory fines may result if the business fails to report a security incident in a timely fashion. In July 2014, the Monetary Authority of Singapore (MAS) Notice 644 went into effect which required data breaches to be disclosed "as soon as possible, but not later than 1 hour, upon the discovery of a relevant incident" (MAS Notice 164). While this notice primarily targets Financial Institutions (FI) that conduct business in Singapore, similar privacy regulations have been adopted in other Asia-Pacific countries including Hong Kong's Personal Data (Privacy) Ordinance (PDPO), Japan's Personal Information Protection Act (PIPA), Malaysia's Personal Data Protection Act 2010 (PDPA), and Taiwan's Personal Data Protection Act (PDPA) (Deloitte, 2017). Similarly, stronger cybersecurity-focused regulations which highlight data breach notification responsibility have already been introduced through the General Data Protection Regulation (GDPR). In May 2018, Facebook CEO, Mark Zuckerberg, encountered intense scrutiny while meeting with lawmakers at the European Parliament. For privacy violations, GDPR would enable international regulators to fine organizations up to 4% of their global revenue which would equal $1.6 billion for Facebook (Satariano & Schreur, 2018). Prior to this meeting with European Parliament, Zuckerberg stood before U.S. Congress and received harsh criticism by Chairmen Senator John Thune and other lawmakers, for handling of user privacy with relationship to the Cambridge Analytica scandal (Domonoske, 2018). Given the global shift towards stricter data protection laws and severe monetary penalties for violations, many organizations will need to invest heavily in security technologies that provide the appropriate coverage and ensure they are deployed correctly. The SIEM remains the industry de facto tool for

incident responders to monitor and detect potential security incidents which may require notification to federal and international governments.

In addition to regulatory pressures, organizations face internal challenges in the management of SIEM technology, which rely upon a growing number of diverse log sources generated by endpoints, network devices, cloud environments, and ultimately any organizational device that can potentially be compromised and utilized in an attack or fraudulent activity. Threat actors are continually inventing new attack variations while organizations are forced to aggressively adapt and implement countermeasures.

This research paper introduces a proof-of-concept tool that can assist incident response teams with advancing their SIEM deployment by measuring the value of log source data. The first section introduced the common SIEM engineering challenges and regulatory pressures faced by modern organizations. In Section 2, a brief history of SIEM technology, its structural operation, and functional implementation within the organization will be presented to provide further context and background for the problem. This section will also cover a literature review, and prior research conducted discussing the challenges of deploying SIEM technology and introduce the necessity to determine better ways to measure an effective deployment by evaluating log data. Section 3 will present the problem statement and research methodology used to evaluate the proposed tool. Next, Section 4 further explores the design of the artifact or tool followed by Section 5 which delivers the evaluation and results of the research. In Section 6, examples of practical applications of the tool in an organizational environment will be covered followed by future research opportunities and limitations in Section 7. Finally, the conclusion will summarize the results of the research and next steps.

2. Background – A Brief History of SIEM

The first generation of Security Information Event Management (SIEM) tools arose out of necessity. In the late 90s and early 2000s, organizational networks were flooded with a sea of alerts generated from Intrusion Detection Systems (IDS). To determine whether a potential security incident had occurred, system administrators would spend significant amounts of time checking a variety of systems on the network to determine malicious activity. By sending these events and alerts to centralized log

storage, a system administrator could reduce the amount of time required to detect and respond to these events. Centralizing events and alerts generated from different network devices and correlating potentially related events within a single application led to the invention of the SIEM.

While organizations today generate millions of events from systems, applications, and devices, only a small fraction may contain meaningful content necessary to make security-related decisions. To further narrow the scope, only a handful of events will have any relevance during a particular security investigation. Thus, the key purpose of the SIEM is to reduce this vast amount of generated event activity on the network into more manageable alerts. By using the SIEM, security analysts can investigate events that have a higher probability of being security-related which can drastically reduce the exposure time of the organization to malicious actors. The exposure time, or dwell time, can be defined by the following function:

$$\text{Exposure Time} = \text{Detection Time} + \text{Response Time}$$

The Security Operations Centers (SOC) mission is to reduce this exposure time as much as possible and to determine preventative controls to further improve the security posture of the organization. The SIEM is a critical security investigation tool for incident responders because of its ability to provide real-time alerting of potential indicators of compromise (IOCs) and detect violations of acceptable computer usage within the organization.

2.1. The SIEM – A Cyber Organism

While the details of SIEM architecture and design lie outside the scope of this paper, **Fig. 1** contains a simplified diagram that will assist with understanding how the tool operates and relies on input from components on the network.

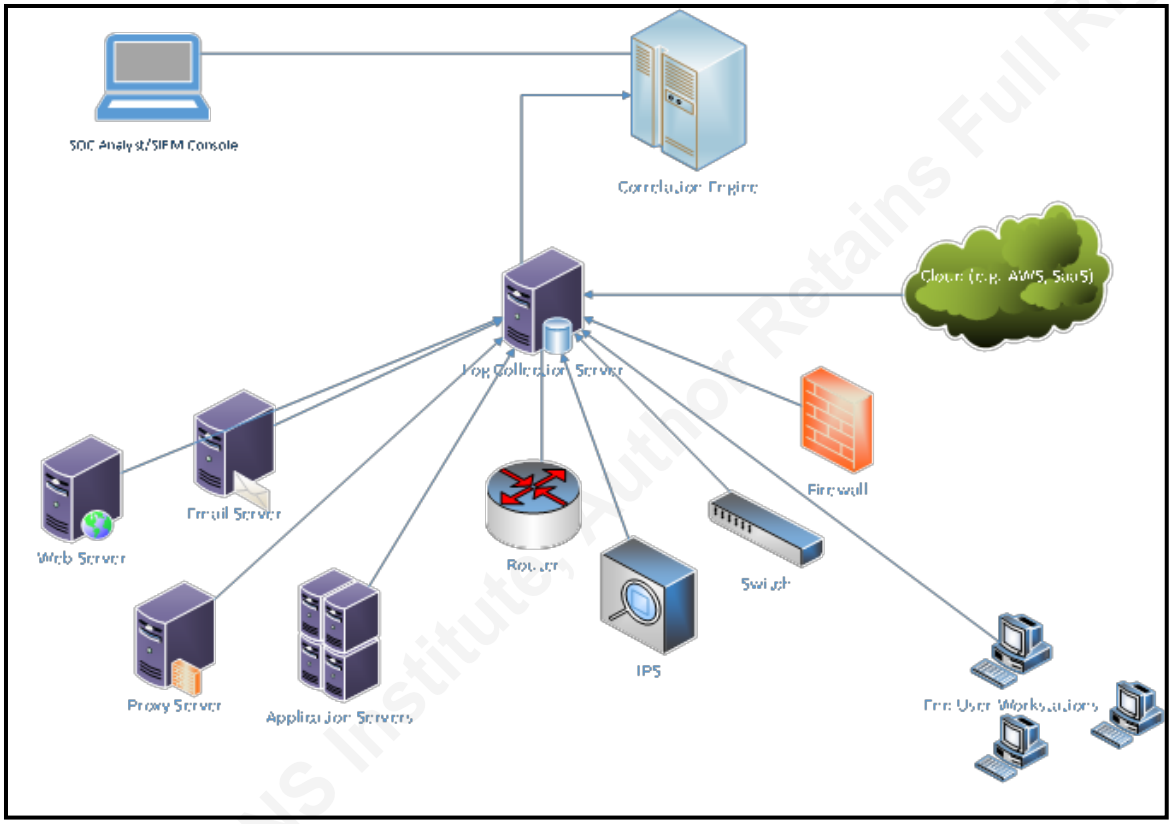

Figure 1 - High-level illustration of SIEM architecture

The devices above represent common components found within a modern SIEM deployment. The method of ingesting logs into the SIEM vary depending on the complexity of the infrastructure environment. Often the devices on the network send logs to a centralized log collection server in real time or a scheduled time interval. Other components of SIEM architecture include pulling logs from specified directories on the network, enriching data from external cloud-based feeds, direct database connectors, and Application Programming Interfaces (APIs) that query other systems in order to pull in data. In modern organizations, there may be hundreds of different types of devices on a single network, each with a separate log source. The Correlation Engine seen in **Fig. 1** is the critical component of the SIEM and is where the majority of the computational processing takes place. Once deployed, these devices perform a function similar to the human body's central nervous system whereby sensors transmit impulses from around the network to a centralized location. An incident responder will view alerts that indicate potentially malicious activity on a SIEM console and launch an investigation if needed.

Alerts, also known as “correlation rules” depending on the vendor, are created by SIEM content developers. These alerts often follow a Boolean logic such as the one seen in **Fig. 2**.

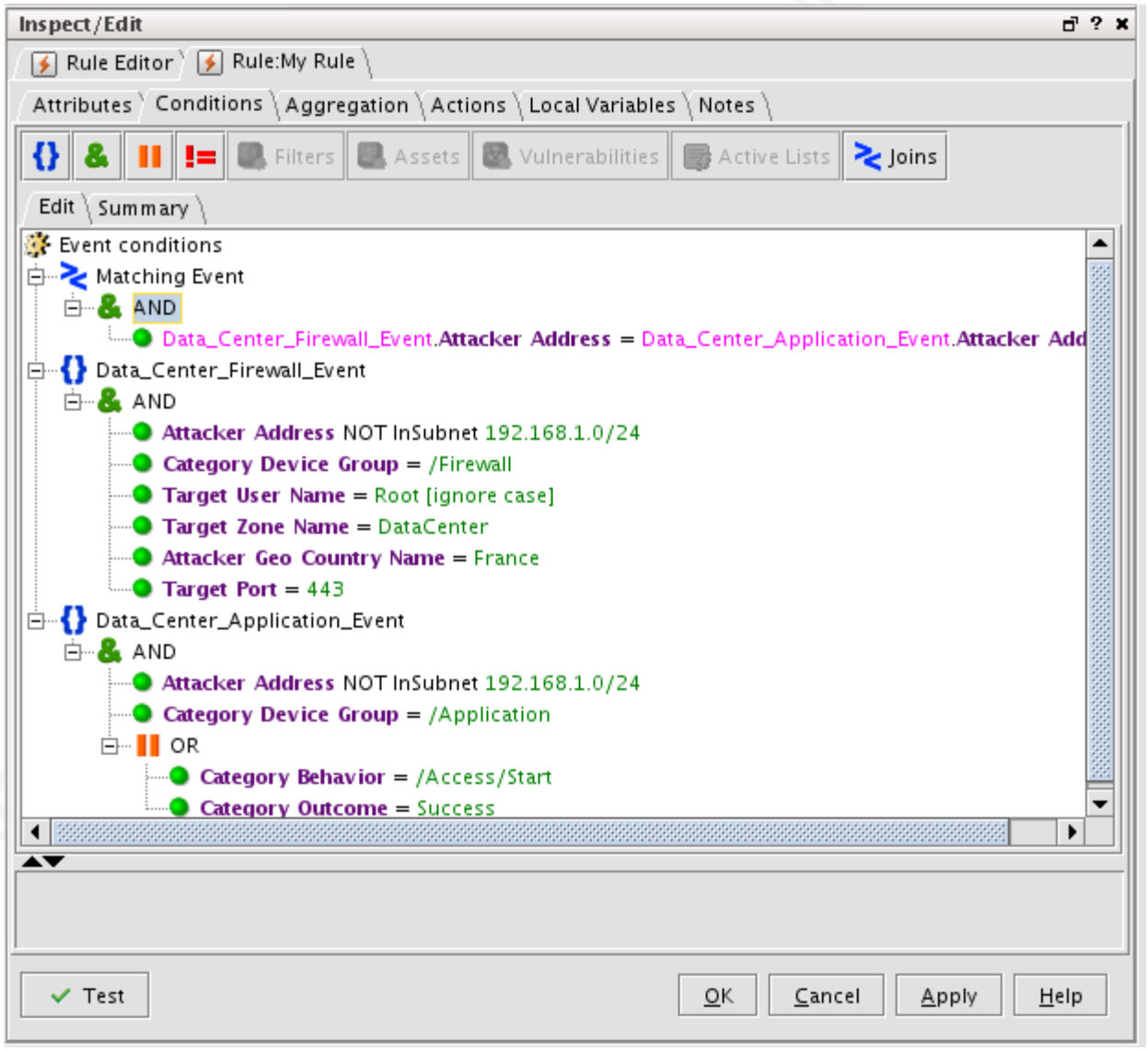

Figure 2 - Example of ArcSight correlation rule (Katoch, 2016)

The flow chart in **Fig. 3** depicts the entire process that occurs from the time an event occurs to when an alert is generated and finally reviewed by an incident responder.

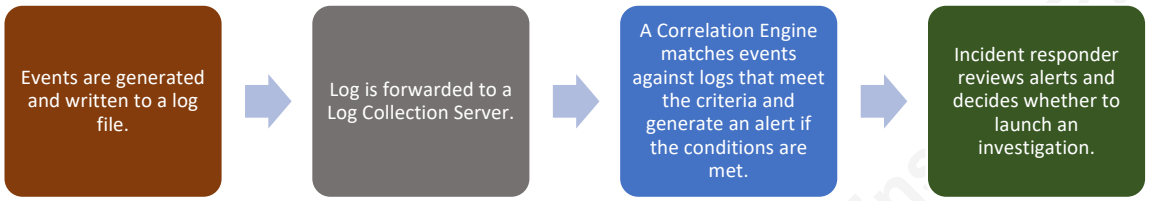

Figure 3 – Process demonstrating event generation to incident response.

As noted above, SIEM technology relies upon logs generated from various sources on the network. The SIEM serves as a single portal for the IR team to be alerted to both aggregated and correlated events that have been processed through the correlation engine. In order for teams to make high-confidence decisions, they must have the ability to review critical security logs that enable them to answer the 4Ws (i.e., Who, What, When, Where) and eventually "How" a security incident might have occurred.

For the SIEM to function as expected during the operate and maintain phase, several key roles and divisions of labor must be present in the organization. It is not uncommon for a staff of full-time security engineers to be solely responsible for keeping the SIEM operational and minimizing recovery time after an outage. During this operate and maintain phase, the incident responders, also known as security analysts, expect the security engineers to ensure the availability of the log source. Specifically, the devices on the network are expected to send logs to the SIEM in real-time or by a pre-determined schedule. A failure to receive a log as expected would lead to an escalation to the security engineers to determine the cause of the disruption. In practice, health monitoring of the SIEM is a joint effort between the security engineers and the incident responders. The latter are those who are most aware that a log is not being received while the security engineers will be the ones to determine the root cause and fix the problem.

Once the security engineers have deployed the SIEM, the IR team utilizes the tool to investigate anomalous and malicious activity on the network. The presence of high-quality logs facilitates this task. Without logs that contain meaningful data, the SIEM is ineffective in aiding the security analyst to make decisions with confidence. Instead, the analyst is forced to make decisions on incomplete data that can cost the business time and resources as well as cause analysts to miss actual security events. Furthermore, fear of being seen as incompetent may prevent incident responders from calling attention to their inability to conclude an investigation because the log data needed is not present within

the SIEM tool. Consider an expensive Mercedes vehicle that is being given to a junior driver. Upon receiving the Mercedes, the driver realizes that not all of the dashboard instruments, such as the speedometer, tire pressure alert, battery alerts, and engine warning light, are operating correctly. Although the dashboard is displaying information, the driver is unable to determine the source of the problem. This impedes the driver's ability to operate the vehicle. An outside observer sees erratic driving and deems the driver as inexperienced. Management decides to take another look and has the engineers perform a review. The engineers, having never driven a car, add more tools that provide information that seems relevant to them but is not valuable to the driver in effectively operating the vehicle. This scenario describes the relationship between the security engineer and the incident responder across many organizations that utilize SIEM technology. While competent team members may fill both engineering and incident response roles, the difficulty in clarifying the technical requirements is often a leading cause of failed SIEM implementations.

The quality of the log source must be determined by the security analysts who make up the IR team. Quality can be understood as the ability of the security analyst to make a high-confidence decision based on the data contained within the log. Currently, no standardized system of measurement exists to represent this quality. It is the focus of this paper to propose a proof-of-concept tool that will enable security analysts to move from intuitive, subjective assessments of logs towards one which quantitatively measures the log quality. This measurement will prove to be a useful communication tool between security engineers and security analysts in determining which logs should be captured for incident response.

2.2. Literature Review

Several publications from online blog posts, book publications, survey reports, and academic research papers have been released pertaining to SIEM deployment best practices. In 2012, Dr. Anton Chuvakin, a subject matter expert in SIEM technologies, wrote that in a perfect world: "nothing comes into your SIEM unless and until you know how it would be utilized" (Chuvakin 2012). Ideally, SIEM functional requirements are determined before the data is fed into the log management tool, where "goals drive security requirements, requirements drive use cases, use cases drive functionality, and

collection scope" (Chuvakin 2012). Chuvakin continues by writing that this ideal is "sadly, uncommon among the organizations deploying SIEM tools today" (Chuvakin 2012).

The "Ninth Log Management Survey Report" further underscores the challenges of log management using data polled from respondents working in a variety of industries including government, financial, health, and education. Most notably, the report states that "Organizations are having the most difficulty using their logs for the top reasons they collect logs, particularly in detecting APT-style malware, preventing incidents and tracking suspicious behavior" (Shenk 2014). The survey report proposes that the difficulty in normalizing log data across various devices and the need for setting up and maintaining the log management software and SIEM systems are possible reasons for these challenges.

Corporate stakeholders send their logs to storage for different functional reasons adding complexity to the situation. For example, in the financial industry, national and international regulations require the ability to show compliance through the retention of certain logs over a specific number of years. In "Successful SIEM and Log Management Strategies for Audit and Compliance," the author notes that "A single common denominator for all regulations requires that one[organization] log all events, and review them" (Swift 2010). Certain regulations such as PCI DSS v3.2.1 explicitly require one to "Track and Monitor all access to network resources and cardholder data." (Payment Card Industry (PCI) Security Standards Council, 2018) Other regulations are less explicit in indicating precisely what needs to be logged, monitored, and reviewed, leaving it open to the interpretation of auditors and regulators. By logging everything, although costly and inefficient, organizations can show compliance.

Security vendors are well aware of these nebulous legal requirements and offer their tool-specific guidelines for determining a successful implementation of their product. Unfortunately for the consumer, this typically leads to expensive solutions that entail the purchase of additional professional services or products that are offered to address the increasingly complex regulatory and threat environment. Because the industry lacks vendor-agnostic benchmark systems, standardized methods of determining

whether the SIEM is capturing information that will demonstrate regulatory compliance as well as provide appropriate security coverage do not exist.

The 2006 NIST publication "Guide to Computer Security Log Management" states that one of the challenges of log management is "balancing a limited amount of log management resources with an ever-increasing supply of log data" (NIST 2006). This statement remains true over a decade later. Logs from mobile devices, cloud service platforms, as well as next-generation security appliances, capture an incredible amount of data. The resources required to store these logs is only going to increase as organizations continue to grow. While the NIST publication proposes valuable insight into how an organization might manage their logs, it does not provide a method that organizations might use to help quantify the value of a given log source to an IR team.

Without standardized metrics, measuring the effectiveness of a security product, such as the SIEM, is arduous. The SIEM is a critical tool for the majority of SOCs and lacking metrics impacts the ability to measure the overall effectiveness of the Security Operation Center. Furthermore, the industry has yet to develop a standard definition for a Security Operations Center, nor has it determined appropriate metrics to measure effectiveness. In the SANS 2017 Security Operations Center Survey, the author confirms this lack "of clearly articulated metrics to express performance" and "selection of performance criteria that are valuable to the specific business needs and measure the effectiveness of the SOC's detection and remediation activities. Metrics are challenging, however, because there's not always a consensus on what makes good metrics within the SOC" (Crowley 2017). At best, the industry is gaining comfort with measuring the effectiveness of the SOC through the evaluation of the employees that use the technology (i.e., Incident Response team) versus accurately assessing whether the tools they have to work with are appropriate for them to do their jobs.

In 2018, Security Operations Centers still struggle with log management and governance throughout the lifecycle of the SIEM. Logs are the most fundamental component of an effective SIEM deployment, and by extension, the quality of the logs impact the organization's ability to prevent and detect potential security incidents. Feeding raw logs with no management strategy into the SIEM contributes to a low signal-to-noise ratio, thereby reducing the ability to determine malicious behavior in the

network. Therefore, a key control objective of successful log management is to send only high-quality logs to the SIEM. High-quality logs will significantly increase the effectiveness of the SIEM by providing the security analysts with the ability to develop better alerts resulting in a reduction in exposure time of malicious behavior within the organization.

The central question becomes, can the value of logs be measured? If so, how can this measurement be used to improve the operating effectiveness of the SIEM? Can the SOC determine if those logs were useful in incident response? What sort of governance lifecycle should a log go through to ensure its continued relevance to the IR team as well as to meeting regulatory requirements? This paper puts forth a tool that incident responders and Security Operations Centers can utilize to determine the relative value of a log source against the full inventory of logs fed into the SIEM. By developing a measurement system to evaluate logs at a point in time, security organizations can make better security and business decisions.

2.3. Previous Research and Other Related Work

In "A Methodology for Building a Log Management Infrastructure," Vasileios Anastopoulos provides a set of procedures tailored to security engineers deploying a log management system within an organization. Anastopoulos (2014) delineates useful engineering performance metrics primarily targeted at the real-time availability of the logs, i.e., the health monitoring aspect of keeping a log management system running within an organization. While this methodology is indeed useful for security engineers in designing the system, it does not address the problem that security analysts are confronted with evaluating the event data that these logs contain. Though this process is a critical component of deploying a SIEM solution, as noted above, security engineers often lack the necessary background training needed to fully appreciate the essential characteristics underpinning the value of a log during a security investigation.

The Open Web Application Security Project (OWASP) has published several articles that relate to logging and monitoring in a security environment. Particularly worthy of mention is their “Logging Cheat Sheet” (*Logging Cheat Sheet - OWASP*, May 2018) which is primarily targeted at helping application developers better understand what information or data should be sent to logs for security monitoring. The “Cheat

Sheet" also captures much of the information that is needed to produce a high-quality log that would be valuable for incident response purposes. If developers are incentivized to write code that logs potential application security incidents, policy violations, unusual application behavior as well as compliance monitoring deviations, this can drastically improve the detection capabilities necessary to deploy a SIEM solution successfully. Furthermore, application logs are distinctly different from the underlying host operating system (OS) logs which are frequently sent to the SIEM instead. This misconception often leads to organizational visibility gaps that are crucial for security analysts to perform an investigation.

In "SOC-CMM: Designing and Evaluating a Tool for Measurement of Capability Maturity in Security Operation Centers," the author, Rob van Os comments on the lack of academic research for evaluating SOCs that previous research is mostly based on: "some best-practices and whitepapers released by commercial companies. There have been attempts at defining a Capability Maturity Model to the SOC, and this is a step in the right direction. These are important indicators that research in this area is currently insufficient" (Van Os 2016). Security vendors engage with organizations to deliver practical solutions, but it is important to keep in mind that commercial entities are apt to release white papers that validate the purchase of their product. Hence, while these publications may provide valuable insight to the overarching security community, the research often exhibits a self-serving business interest.

Due to the lack of sufficient research surrounding the management of logs or measurement of SIEM maturity, an opportunity exists to create a measurement tool that organizations can use to quantify the value of a log to the IR team. This tool would have broader implications as organizations can utilize this information to make more cost-effective decisions when working with vendors.

3. Problem Statement and Research Methodology

The problem statement and subsequent research question is as follows: Given the difficulties and challenges in determining which logs are useful for Incident Responders in a security investigation, how can we quantify the value of logs sent to the SIEM? Is it

possible to derive a correlation between useful logs in a security investigation and a numerical value?

The research methodology adopted in this paper is the Design Science Research Method (DSRM) which "creates and evaluates IT artifacts intended to solve identified organizational problems" (Hevner 2004). According to Hevner, DSRM "involves a rigorous process to design artifacts to solve observed problems, to make research contributions, to evaluate the designs, and to communicate the results to appropriate audiences"(2004). The DSRM Process Model illustrates each phase of the methodology.

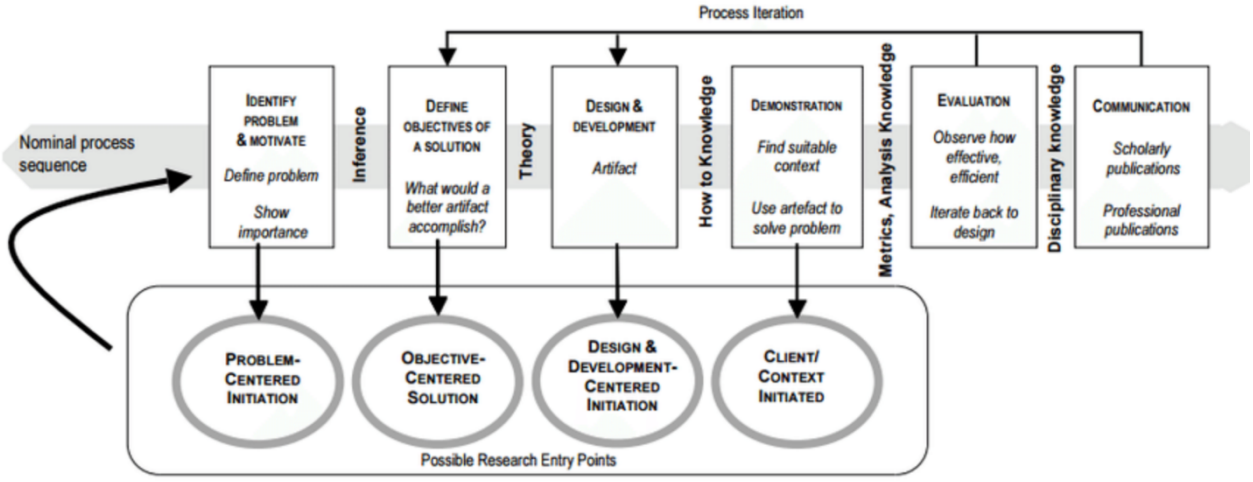

Figure 4 – DSRM Process Model (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007)

This research follows the DSRM model as displayed by Peffers et al. in **Fig. 4**.

4. Artifact Design – Criticality and Value Algorithmic Model

It is the goal of this research to generate a usable artifact that assigns a numerical value to inherently subjective data. In this case, the subjective data being a log source, or machine data, that contains the system generated events an Incident Response team intuitively considers to be "good" or "bad." To measure usefulness or value, attributes of "known good" data need to be analyzed and broken into components which can be measured. In "How to Measure Anything – Finding the Value of 'Intangibles' in Business," Hubbard states that, when measuring intangible elements, the primary objective is to reduce uncertainty in decision-making (2014). As a result, even subjective

judgment can be useful in making decisions or as the author states "measurements can be fed directly into quantitative models so that optimal strategies are computed rather than guessed" (Hubbard 2014).

The tool developed in this study went through several iterations before arriving at a suitable version for research purposes. During the first round of tool design, an initial list of questions was developed after researching publicly available literary sources. This phase was intended to identify common thematic elements across the sources in order to construct questions that would capture the traits of a "good" log source. Google search queries included the following:

- “NIST Security Logging”
- “OWASP Logging”
- “Logging Best Practices”
- “SIEM Logging”
- “SIEM Best Practices”
- “Application Logging”
- “What logs do I send to my SIEM”

After examining the results and removing duplications, approximately eighteen sources were deemed relevant to the research question. Based on the similarities in the content of the different publications, there appeared to be a consensus in the industry on what constituted valuable log data. The blog post “Logging Best Practices” from a notable SIEM vendor, Splunk, captured many of the key themes (Splunk, 2018):

- Use timestamps for every event
- Create events that humans can read
- Use Unique identifiers (IDs)
- Log in a text format
- Identify the source
- Keep multi-line events to a minimum

With these considerations, approximately 20-25 questions were developed during this round, and an initial weight value for each question was assigned. The book publication “Logging and Log Management” contained a useful section on criteria for good logging

(Chuvakin, Phillips, & Schmidt, 2013, p.46) that helped drive additional attributes that would be valuable to security analysts. The questions then went through the second round of review.

During the second round of review, discussions were held with industry experts to obtain feedback on the initial list. Each question was carefully scrutinized to determine its primary objective as well as to ensure it was not overly broad in scope. For instance, the original question "Does the log contain Source IP Address and Destination IP Address?" was broken into two separate questions:

1) Does the log contain a Source IP Address?
2) Does the log contain a Destination IP Address?

The justification for this change being log sources may only contain the Source IP Address or the Destination IP Address. Moreover, the two pieces of data are differentially weighted with the higher value assigned to the Source IP Address since it would provide the IR team with a starting point for where the activity originated. Knowing the Source IP Address can be critical in determining the attack vector which can determine where to hunt for more IOCs during an investigation. Other changes based on these discussions included adding items that captured overall metadata information about the log source asset. An example of such a question is: Does/could any portion of the log contain PII? Other questions clarified whether the source of the log was a primary or secondary source of data. A primary source of log data would increase the overall value of that data as the absence of this log source within the SIEM would diminish the understanding or context of other logs (i.e., secondary sources).

In the third round of tool design and development, questions were finalized, and weighted values were determined. At this point, the next step in the DSRM process is to proceed with a tool evaluation against a dataset.

Several considerations influenced the design of the artifact. At concept inception, it was decided that incident responders would be the primary operators of the artifact or tool. As opposed to reading in raw machine data and outputting a value, the tool requires manual inputs from a knowledgeable worker that can validate the data. Despite advances in computing and machine learning, a human is still a necessary and critical component

in this process. For instance, a computer may be able to match a regular expression with an IP address; however, a human is still needed to confirm whether the IP address is a source or destination address. A human is also required to verify the integrity of the data contained in each log. Finally, as it relates to the usability of this tool, a human will be needed to execute the steps of the algorithmic model to ensure accuracy.

A second consideration in designing the tool was maintaining the clarity and simplicity of each question. It was essential that each question measured, to the greatest extent possible, a single aspect of the log data or field. In this way, the appropriate weighted value could be assigned to this piece of data.

A comprehensive set of questions was used to evaluate each log source. The goal of this process being to determine whether the log data is noise or signal, that is, whether or not it provides reliable information that can be utilized in a security investigation to reduce the exposure time.

The tool generates a **Log Quality Value** index for each log that is based on two variables. The first variable provides a measure of the device criticality of the log in determining a security incident or data breach. The second variable provides a more granular measure of the usefulness of individual event characteristics within the log to the Incident Responder. For instance, how does the log source answer the 4Ws? Does the log contain unique data fields that are not found in other log sources within the environment? Together, the pair of variables provides a more precise value of the log being sent to the SIEM within the organization. This can be expressed as follows:

$$\text{Log Quality Value} = \text{Criticality}_x \text{ , } \text{Value}_x$$

or

$$LQV = C_x, V_x$$

The next section provides greater detail as to the factors considered in the variable calculations.

4.1. Understanding Criticality

The **Criticality** (C_x) value is used to determine the importance or priority of the log in relationship to the other logs that exist in the environment. This value is calculated

through a set of questions that broadly describe the application or device generating the data in the log source. Based on the answers to these questions, the Criticality value will fall into only four types described below:

Criticality Type I – An application or device log that is classified as Type I may contain either or all of the following:

a) PII or sensitive data to which unauthorized exposure would result in financial loss, loss of life, or regulatory fines.

b) Events that are a Primary Source of key data for the Incident Response team. Loss of the log source would result in the inability to correlate or identify potentially malicious activity in the environment. Examples of this would include logs that contained who performed the action, what action occurred, when the action occurred, and where on the network it occurred.

Criticality Type II – An application or device log that is classified as Type II may contain contextual data that enriches Primary Source Data Events. Examples might be Threat Intelligence Data or Asset Inventory Information Management System.

Criticality Type III – An application or device log that contains events also contained in other log sources (e.g., Secondary Sources) to which the loss of such data would not result in the inability to correlate, detect, identify or prevent an investigation. Multiple log sources that contain redundant data decrease the Criticality of the log within the environment and in some instances should be considered for removal from the SIEM.

Criticality Type IV – An application or device log that is categorized as Type IV does not contain events that store or transmit PII/sensitive data. The logs do not contain events that have been used to confirm malicious activity in the past. Debug logs that provide no security value should be placed here. Logs that are required to be retained for regulatory or policy reasons will typically fall under this classification.

A **Criticality Type I** (C_1) is Most Critical. Whereas a Type IV or C_4 is Not Critical. **Fig. 5** below serves as a useful model for visualization purposes. The Criticality Type determines the quadrant the log will be assigned.

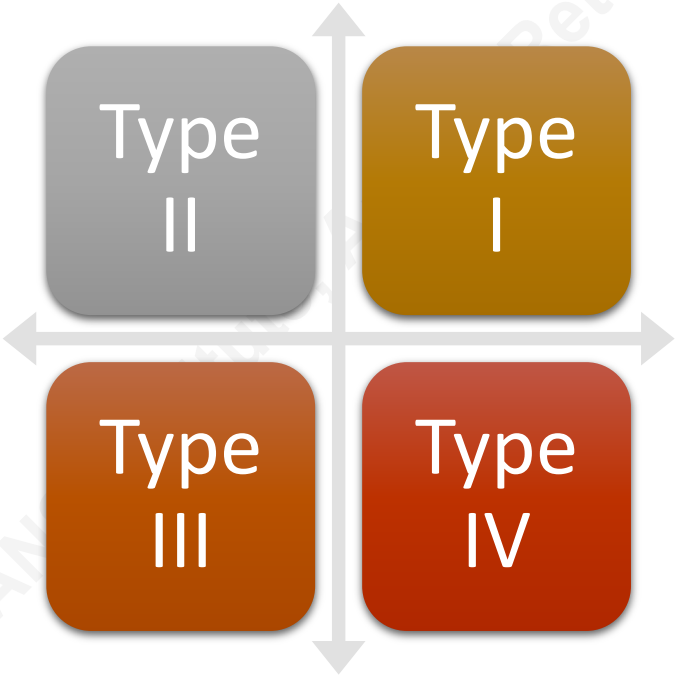

Figure 5 - Criticality quadrants for visualizing all inventoried logs

After an organization has inventoried all logs using the LQV index, this visualization method can help an organization understand the number of critical logs with relationship to each other. If the majority of the log sources fall into Type IV, this indicates that the SIEM has many logs that are not providing information that can be useful to an incident responder. While this helps to understand Criticality, the next section will dive into the algorithmic design used to calculate the Value variable.

4.2. Arriving at Value

The Value (V_x) algorithm is calculated by assigning an initial value of zero to V_x (e.g., V_x=0), performing a series of instructions, and returning the final output back to V_x. The V_x can be expressed as the following summation:

$$V_x = \sum_{n=1}^{29} \lambda_n(W_n)$$

This expands to:

$$V_x = \lambda_1(W_1) + \lambda_2(W_2) + ...\lambda_n(W_n)$$

Where lambda, λ, represents an abstract function taking the variable $W_{n,}$ a weighted value, as input, and returning an output. The abstract function is defined by the question and the response by the user will determine the final value. This method of defining the function is similarly used in modern programming languages such as C++ and Java while the use of λ is derived from Lambda Calculus. This abstraction allows for future research to involve more complex mathematics without changing the fundamental equation above.

As mentioned, the response to the question determines the next instructions performed by the abstract function. An affirmative answer to a question will perform the operator against the current value contained in V_x. If a hypothetical log contains a Source IP Address and the Weighted Value for that question is +3, then the instruction is to add 3 to the current value contained within V_x. See below:

$V_x=V_x$ + Weighted Value

$V_x=0 + (+3)$

$V_x=3$

If a subsequent question had a Weighted Value equal to -2 and the answer to the question had been in the affirmative, then the instruction is to subtract 2 from $V_x=3$ and V_x would now equal 1. The user will perform that operator (e.g., subtraction) against the current value contained within V_x. The following calculations would be performed:

$V_x= V_x$ + Weighted Value

$V_x= 3 + (-2)$

$V_x=1$

The Value (V_x) is determined by the data fields contained within the log. A series of 29 questions were designed with the intent of identifying characteristics of that log

source that would make it valuable or less valuable to an Incident Response team. A non-exhaustive list of questions to provide a relevant example would be the following:

a) Does the log source contain WHERE (e.g., Source IP Address) the event occurred: **Weighted Value** = +3

b) Does the log source contain WHO or WHAT performed the action (e.g., Username, Hostname): **Weighted Value** = +2

c) Does the log source contain the ACTION that occurred? (e.g., Log On, Sign On, Virus Quarantined): **Weighted Value** = +3

d) Does the log source contain data that is difficult to parse and/or correlate with other events: **Weighted Value** = -2

For instances where the answers were not known, or the answers were not applicable to the type of log source, a default response is recommended. In this manner, all questions manually answered are considered in the result. The answers to the questions determine the next instruction to be performed to calculate the final value of V_x. Further opportunities for the development of the algorithm used to calculate V_x will be discussed in Section 7. A full listing of the questions developed and used for the research can be found in **Appendix D**. The next section will further describe the evaluation of the tool against a dataset as well as the results.

5. Evaluation and Results

During the third phase of DSRM, the design artifact, in this case, the Log Quality Value (LQV) index is evaluated against a simulated security environment. Publicly available log samples were identified to represent an environment with devices and applications found at most medium to large organizations. These devices or security products included log sources from the following:

- Anti-Virus
- Web Application Server
- Web Proxy
- Web Application Firewall
- Routers/Switches/Firewall
- File Integrity Monitoring

- Host Intrusion Detection System
- Vulnerability Scanner
- Mail Server (Exchange)

The publicly available Open Source Host Based Intrusion Detection System, OSSEC, provided a rich selection of sample log sources ("Log Samples — OSSEC 2.8.1 documentation," n.d.) that were used in the evaluation of the tool. These log samples contained generated log data as they would be configured in a real deployment. In many cases, the log samples were from open source products. To provide a more accurate representation of a real organizational environment, popular vendor product logs (e.g., Microsoft Exchange, Cisco IOS Firewall/Routers) were utilized as they were made available for evaluation. Other publicly available log samples were provided from the Security Repository available at secrepo.com ("SecRepo - Security Data Samples Repository," n.d.) or generated in a home lab environment. A “known bad” log source consisting of gibberish characters served as baseline data and received an index of C_4,V_{-6}. See **Appendix C** for additional log samples that were evaluated using the LQV index and subsequent results.

After the log sources were indexed using the LQV algorithm, an additional research step was performed to demonstrate how the results might reflect the value of the logs with relationship to real-world scenarios. As source material, the 2017 Verizon Data Breach Digest provided case study material. These scenarios are based on the Verizon Breach Investigations Report (VBIR) which contains information on recent cyber threats and trends impacting each major industry including financial, health, retail, and public sector amongst many more. Within each case study is a description of the event, mitigating activities, and security controls that would prevent or detect the attack and ultimately reduce the exposure time.

Security controls that detected the attack vectors should have a higher Log Quality Value if the sources have been configured to log the right information. After evaluation of the log source against the tool, the resulting values should mainly reside in the Criticality Type I (C_1) quadrant, and represent a relatively higher value V_x in our simulated environment. A low V_x score (e.g., $V_x<10$) indicates that the value of the log has room for improvement. This can mean that configurations exist in the device for

more granular logging or indicate that the log does not make a good candidate to send to the SIEM. This does not necessarily mean the device is not working as intended, only the value of sending the log to the SIEM for use by the Incident Response team is negligible.

Two case studies were chosen to evaluate the tool based on common attack patterns found in the real world. Each case study contains a description of the scenario, the evaluation of the tool against the key devices used in detection/prevention of the attack, and the findings.

Case Study 1 – Web Attack

Description: In CE-1 Website Defacement – the Hedley Kow, the data breach scenario involves the defacement of a web application (Verizon, 2017, p. 57). Since many organizations have an internet facing web portal that is used to engage with customers, web applications are a natural target for attackers. Improper configurations of the web application can invite path traversal attacks while unpatched vulnerable systems sitting on the public internet may be subject to exploits allowing unauthorized reverse shell access.

Evaluation: Several security controls can be implemented that will detect or prevent the risk of web attacks. The case study presented by Verizon, list several security products or tools which will be mapped to equivalent systems in the simulated environment under the column "Research Paper Simulated Environment Tool." **Table 1** below illustrates this mapping:

Verizon Identified Security Tool/Log	Detection Capability Usage	Research Paper Simulated Environment Tool	LQV
File Integrity Monitoring Tool	Detects file changes on the server	OSQuery	C_2, V_6
Web Server	Logs contain evidence of tampering or suspicious activity	Apache Web Server	C_3, V_{10}
Inbound and Outbound Network Connections	Detects web server connections to known suspicious IP addresses	Cisco IOS	C_1, V_{21}

Intrusion Detection System (IDS)	Detects network attacks that match known signatures. Acts as early warning or prevention	HIDS OSSEC	C_1, V_{11}
Web Application Firewall	Detects anomalous behavior and common web application attacks	Sophos WAF	C_1, V_{29}

Table 1 - Case Study 1 Mapping of Verizon Tools to Simulated Environment

Findings: The Log Quality Value index was effective in determining whether the tools that were deployed in the simulated environment contained logs of value to the Incident Response team during the investigation phase. Ideally, each of these log samples would be located in the C_1 quadrant and have a value towards the upper range of V_x scores within the environment. For example, a $V_x>20$ would indicate a higher value to the organization than a $V_x<10$. Log sources that demonstrate a higher LQV index should correspond with the tools used to detect the attack. A lower LQV index (i.e., C_3 or C_4 and $V_x<10$) would be an indicator to the organization that additional tuning or configuration should be applied to the detection devices or if further enhancements could not be made, then alternative vendor solutions should be considered by the organization.

An interesting observation can be derived from the findings in this particular case study. If not configured properly, otherwise forensically rich material might not be valuable to send to the SIEM to alert due to being hard to interpret by the SIEM. For instance, while File Integrity Monitoring may contain extremely valuable sources of information after an attack has occurred, the formats of the logs might be difficult to interpret or parse making this a problematic data source to correlate information. For the study, the File Integrity Monitoring tool produced logs in JSON format making specific events hard to parse which would make a correlation with other events more difficult within the SIEM. Logs that are difficult to parse is one criterion that will drive down the overall value (V_x) of the log to the Incident Response team.

Case Study 2 – Phishing Attack

Description: In MS-1: Crypto Malware – The Fetid Cheez, a network administrator downloaded malware via a malicious file attachment within a phishing email (Verizon, 2017, p. 77). This malware contained ransomware which subsequently exploited a vulnerability in an application that had yet to be patched. As a result, the ransomware encrypted files on a network share drive utilizing the network administrator's logged on user account to perform the operation.

Evaluation: Several security controls can be implemented to reduce the impact of malware introduced into the environment via phishing email. Once a malicious file attachment has been confirmed, mail exchange logs can be reviewed to identify the email recipients and remove the email from the mailbox during the eradication phase. Anti-Virus logs can be used to determine if any malicious activity has occurred on the end user workstation or file share. File Integrity Monitoring logs can be leveraged to check files that have been modified by infected users (e.g., the network administrator account) known to have received the phishing email. Finally, vulnerability scanning result logs may be useful in identifying additional unpatched systems within the organizational environment. According to the case study, the malware exploited an application vulnerability which allowed an attacker further access into the network via Command and Control (C2) servers.

The listed detection methods within the Verizon case study were mapped to equivalent devices in the simulated environment in **Table 2**.

Verizon Identified Security Tool/Log	Detection Capability Usage	Research Paper Simulated Environment Tool	LQV
Mail Exchange Server	Provides message transmission information such as the recipients of phishing emails	Mailbox (Exchange 2000)	C_1, V_{32}
AntiVirus	Detects malicious activity on endpoints	ZoneAlarm	C_1, V_{23}
File Integrity Monitoring	Detects file modifications and which users were involved with those changes	OSQuery	C_2, V_6

Inbound and Outbound Network Connections	Detects outbound C2 traffic if connection established with known suspicious IP	Cisco IOS	C_1, V_{21}
Vulnerability Scanning	Detects vulnerable systems if used proactively. Scan results fed into the SIEM can be valuable starting points for determining systems that might be at risk of exploit	OpenVAS	C_2, V_3

Table 2 - Case Study 2 Mapping of Verizon Tools to Simulated Environment

Findings: In Case Study 2, there was a stronger correlation between the Log Quality Values and the Verizon Identified Security Tools. In this scenario, 3 out of the 5 tools used showed a higher C_xV_x with a Criticality Type I and $V_x > 20$. This indicates that these log sources are providing valuable content to the Incident Response team. As mentioned in case study 1, the LQV for the File Integrity Monitoring tool showed a low V_x which can be interpreted as meaning that the security tool configuration (e.g., OpenVAS) in the simulated environment is not configured to log useful content. The Vulnerability Scanner used within the simulated environment generated a report that would not be easily parsed and fed into a SIEM for alerting purposes. Based on the LQV index of OpenVAS, the logs generated would deliver minimal benefit to the security analyst. Both the File Integrity Monitoring and Vulnerability Scanning applications had a Criticality Type II, which indicate that they primarily capture information that is contextual but would not be key sources used by an analyst during an investigation.

Summary of Findings

Per the two Case Studies used to evaluate the tool, a number of conclusions can be reached. The key logs used in detecting the common attack pattern showed a positive correlation which mapped to items in the Criticality Type I quadrant and had a V_x greater than 20. Specifically, of the five Verizon identified tools used to detect the Web and Phishing attacks, three out of the five tools mapped to the most critical detection tools for an Incident Responder based on the LQV. Only one instance (i.e., HIDS in case study 1) did the V_x score show only an average value (V_x=11) with a corresponding Criticality of Type I. This finding can mean that the current IDS logging configurations that would be

available to the Incident Response team are not as valuable as other logs that can also be found in the C_1 quadrant. Another factor that might drive down the V_x score can be attributed to the difficulty in parsing the log. The ability to parse and view logging information outside of the tool generating the log is an important criterion used in the LQV algorithm. If the logging information is best viewed within the application and difficult to parse within the context of a SIEM, then perhaps that log should not be sent for indexing by the SIEM. This highlights a key goal of the Log Quality Value tool put forth in this research: to determine which logs contain content *of value* to the Incident Response team for use within a SIEM. These findings ultimately support that goal.

6. Practical Applications of the Log Quality Value index

This tool was designed to be used in conjunction with an overall Log Governance and Inventory strategy that organizations should implement when utilizing a SIEM. Given the tendency for logs to be sent and onboarded into the SIEM with no real oversight after initial setup, this tool would best work as a control objective to ensure high-quality logs are fed into the SIEM and retroactively used at organizations to assess the LQV of logs that are currently in place. **Appendix A** contains two sample processes that demonstrate how logs can be onboarded and periodically reviewed within an overarching Log Governance and Inventory Strategy. **Appendix B** contains additional use cases that demonstrate how the LQV index can be utilized by different teams within an organization to deliver higher quality logs for security usage.

7. Future Research Opportunities

At this stage, the Log Quality Value index remains a proof-of-concept. Further development and testing would be required to realize the full benefits of the tool to an organization. Given the research constraints of time and resources, opportunities exist to develop the tool in the following directions:

- Evaluating the tool against additional types of log sources (>50) which would provide greater research representation of logs that organizations

are sending to their SIEM instance. This increased population would help tool calibration.

- Evaluating the tool against multiple organizations within different industries which might identify trends in the usefulness of specific log sources in relation to other logs also commonly being used by those within that industry.
- Development of a robust Log Governance and Inventory strategy that is invoked upon onboarding of a log source and throughout the lifecycle of the log while it is being ingested into the SIEM. This strategy can incorporate the LQV to assess the value of the log periodically and make business decisions as to whether it remains a good candidate to send to the SIEM.
- Opportunities exist to have more customized questions for the type of device (e.g., Endpoint Detection and Response, Network Load Balancers, Threat Intelligence, Proxy, Firewall Logs) that is generating the log as some questions were not entirely appropriate for the device they were logging. Having more specific questions per type would be a better indicator of how valuable that log would be in the environment. For instance, if questions are asked that are not appropriate for the device, this will negatively skew the results.
- Using the 3-point vendor-agnostic measurement strategy described in Use Case 5 of **Appendix B**, a comprehensive open source guide could be created to provide organizations with a baseline for how they should configure their logs for that product. If the $M_{1,}$ or β, represents the highest security value the log source could be configured to log, Incident Responders would be able to evaluate the logs after deployment, determine the delta from the best logging configurations published in the guide, and communicate to Asset Owners which logging configurations need to be implemented in order to provide greater value to the SOC.
- For the purposes of the proof-of-concept used for this research, the calculations performed by the algorithm is basic arithmetic. The use of the

abstract function is to allow for more complex calculations in the future. This might include questions which consider previous user responses as a multiplier against the V_x or perform other averages on the value of V_x that will result in more precise calculations of the V_x. This would allow for more distinction between minimum and maximum LQVs enhancing the ability to determine valuable logs.

7.1. Research Limitations

Throughout the evaluation of the tool, several limitations and assumptions were necessary in order to complete testing. Here are a few that are worth noting:

- While significant effort was given towards testing the tool against production datasets that might exist in a real organizational environment, concern was expressed over the content of the logs and the resulting security inferences published in the paper even after anonymizing sources.
- In a real environment, security engineers might be able to answer some of the questions that the tool requires. Since this was a simulated environment, these answers were assumed to be "no" which subsequently resulted in a lower value.

8. Conclusion

The inspiration for this research evolved from challenges in assessing whether a SIEM solution had been deployed successfully. Through discussions with industry experts, it became clear that this problem is widespread and not limited to one particular organization or one vendor product. Furthermore, audits of this area typically focused on the performance of the incident response team, utilization of playbooks, and documentation versus an assessment of the primary tool used to detect and investigate malicious activity. This problem is further exacerbated by the lack of an industry standard in determining whether a SIEM has been deployed in a manner which protects the organization. The research described in this paper is an attempt to break down this complex system into measurable segments and evaluate those discrete segments of log data.

The Log Quality Value index is one method that Incident Response teams can use to evaluate and measure the usefulness of the logs being sent to their SIEM instance. This research showed how the LQV effectively addresses the research problem by quantifiably measuring the value of logs sent to the SIEM. As a result, the LQV helps to reduce the subjectivity of whether a log is providing valuable content in the context of a security investigation.

Incident responders play a critical role in identifying characteristics of a log that make the events beneficial to an investigation. They rely on application developers to write events to disk that are not cryptic in nature as well as indicate that potential fraudulent or unexpected behavior is occurring within the application or device. Industry pressure should be applied to application developers of all types of products to write logs that assist with security incidents and identify fraudulent behavior. It is not enough for developers to produce logs that enable them to debug problems in code and business logic. Additionally, organizations should strongly consider implementing an overall log governance and inventory strategy to aid in ensuring consistently valuable logs are being sent to the SIEM and periodically assessed to determine whether meaningful content continues to be derived from their consumption.

References

Anastopoulos, V. (2014). *A methodology for building a log management infrastructure* (Master's thesis). Retrieved from http://dione.lib.unipi.gr/xmlui/handle/unipi/6021

Chuvakin, A. (2012, September 24). On "Output-driven" SIEM [Web log post]. Retrieved from https://blogs.gartner.com/anton-chuvakin/2012/09/24/on-output-driven-siem/

Chuvakin, A. A., Phillips, C., & Schmidt, K. J. (2013). *Logging and log management: The authoritative guide to understanding the concepts surrounding logging and log management*. Waltham, MA: Syngress.

Crowley, C. (2017). *Future SOC: SANS 2017 Security Operations Center Survey*. Retrieved from SANS Institute website: https://www.sans.org/reading-room/whitepapers/analyst/future-soc-2017-security-operations-center-survey-37785

Deloitte. (2017). *Building trust across cultures: Privacy and data protection*. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Risk/gx-risk-building-trust-across-cultures.pdf

Domonoske, C. (2018, April 10). *Mark Zuckerberg Tells Senate: Election Security Is An 'Arms Race'*. National Public Radio. Retrieved from https://www.npr.org/sections/thetwo-way/2018/04/10/599808766/i-m-responsible-for-what-happens-at-facebook-mark-zuckerberg-will-tell-senate

Federal Trade Commission. (2017). *Privacy & Data Security Update: 2017*. Retrieved from https://www.ftc.gov/system/files/documents/reports/privacy-data-security-update-2017-overview-commissions-enforcement-policy-initiatives-consumer/privacy_and_data_security_update_2017.pdf

Hevner, A.R.; March, S.T.; and Park, J *Design research in information systems research*. MIS Quarterly, 28, 1 (2004), 75-105)

Hubbard, D. W. (2014). *How to measure anything: Finding the value of "intangibles" in business*. Hoboken, NJ: Wiley.

Katoch, P. (2016). *ArcSight Rule* [Example of ArcSight Join Rule]. Retrieved from https://katochcisco.blogspot.com/2016/12/understanding-siem-correlation-basics.html

Kent, K., & Souppaya, M. (2006). *Guide to Computer Security Log Management* (800-92). Retrieved from National Institute of Standards and Technology website: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-92.pdf

Nate Street, nate.street@gmail.com

Kumar, V. S. (2006). *Effective requirements management*. Paper presented at PMI® Global Congress 2006—EMEA, Madrid, Spain. Newtown Square, PA: Project Management Institute.

Log Samples — *OSSEC 2.8.1 documentation*. (n.d.). Retrieved August 1, 2018, from https://ossec-docs.readthedocs.io/en/latest/log_samples/

Monetary Authority of Singapore. (2013). *Notice on Technology Risk Management* (644). Retrieved from http://www.mas.gov.sg/~/media/MAS/Regulations%20and%20Financial%20Stability/Regulations%20Guidance%20and%20Licensing/Commercial%20Banks/Regulations%20Guidance%20and%20Licensing/Notices/Notice%20MAS%20644.pdf

OWASP. (2018). *Logging Cheat Sheet - OWASP*. Retrieved May 5, 2018, from https://www.owasp.org/index.php/Logging_Cheat_Sheet

Payment Card Industry Data Security Standard v3.2.1 (2018), Retrieved July 27, 2018 from the PCI Security Standards Council
https://www.pcisecuritystandards.org/security_standards/pci_dss_download_agreement.html

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). *A Design Science Research Methodology for Information Systems Research*. Journal Of Management Information Systems, 24(3), 45-77.

Satariano, A., & Schreur, M. (2018, May 22). Facebook's Mark Zuckerberg Gets an Earful From the E.U. *The New York Times*. Retrieved from https://www.nytimes.com/2018/05/22/technology/facebook-eu-parliament-mark-zuckerberg.html

Securities and Exchange Commission. (2018). *Commission Statement and Guidance on Public Company Cybersecurity Disclosures* (33-10459). Retrieved from https://www.sec.gov/rules/interp/2018/33-10459.pdf

SecRepo - *Security Data Samples Repository*. (n.d.). Retrieved August 1, 2018, from http://www.secrepo.com/

Shenk, J. (2014). *Ninth Log Management Survey Report* (9). Retrieved from SANS Institute website: https://www.sans.org/reading-room/whitepapers/analyst/ninth-log-management-survey-report-35497

Sophos WAF Log Sample. (n.d.). Retrieved August 1, 2018, from http://docs.sophos.com/nsg/sophos-firewall/v16058/Help/en-us/webhelp/onlinehelp/index.html#page/onlinehelp/WAFLogs.html

Splunk. (2018). Logging best practices | Splunk app intelligence [Web log post]. Retrieved from http://dev.splunk.com/view/logging/SP-CAAAFCK

Swift, D. (2010). *Successful SIEM and Log Management Strategies for Audit and Compliance*. Retrieved from https://www.sans.org/reading-room/whitepapers/auditing/successful-siem-log-management-strategies-audit-compliance-33528

Van Os, R. (2016). *SOC-CMM: Designing and Evaluating a Tool for Measurement of Capability Maturity in Security Operations Centers (Dissertation)*. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-59591

Verizon. (2017). *Data Breach Digest*. Retrieved from Verizon website: https://www.verizonenterprise.com/resources/reports/rp_data-breach-digest-2017-perspective-is-reality_xg_en.pdf

Appendix A – Log Governance and Inventory Strategy

Effective log management should be instantiated through a governance process where logs are onboarded and periodically reviewed. Whether the SIEM has already been deployed or in the planning stages, an inventory of all current logs should be indexed and documented within a Log Inventory Tracking Sheet. At minimum, the log inventory should contain a description of the device/application generating the log, the types of events contained in the log, and key event fields that analyst can utilize to search against or generate useful SIEM content. The process in **Fig. 1** suggests incorporating the LQV within the Log Inventory Tracking Sheet to determine the value of the log at the point in time at which the log is initially inventoried as well as periodically reviewed as suggested in **Fig. 2**. If possible, the raw log data source should be utilized against the LQV. It should be noted the SIEM itself can add enrichment data that enhances the logs such as lookup tables and correlations against other log sources (e.g., Threat Intelligence) that make the log more valuable. See **Appendix B**, Use Case 5 for a 3-point measurement strategy that might help quantify the additional value the SIEM is providing to the Incident Response team.

The Incident Response team should be the process owners of the log governance onboarding and periodic review as they are the primary stakeholders of the SIEM. A senior analyst that has a strong foundation in security investigations should be responsible for completing the review. Other key stakeholders include security engineers and Application/Asset Owners. The two diagrams below provide an example of how the IR team would engage the security engineers and asset owners during the log onboarding and periodic review process.

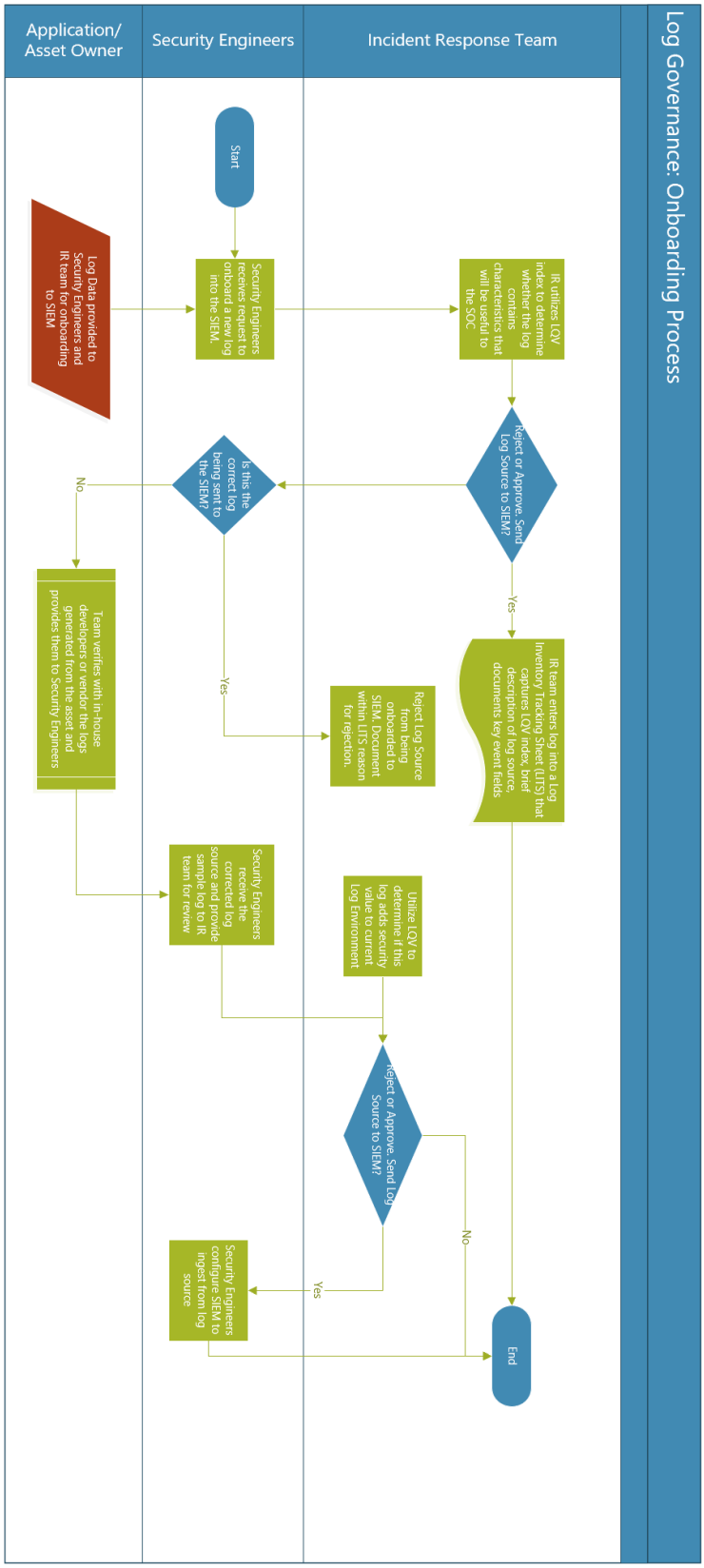

Figure 1 – Example of onboarding process of a log for consideration into the SIEM environment.

Nate Street, nate.street@gmail.com

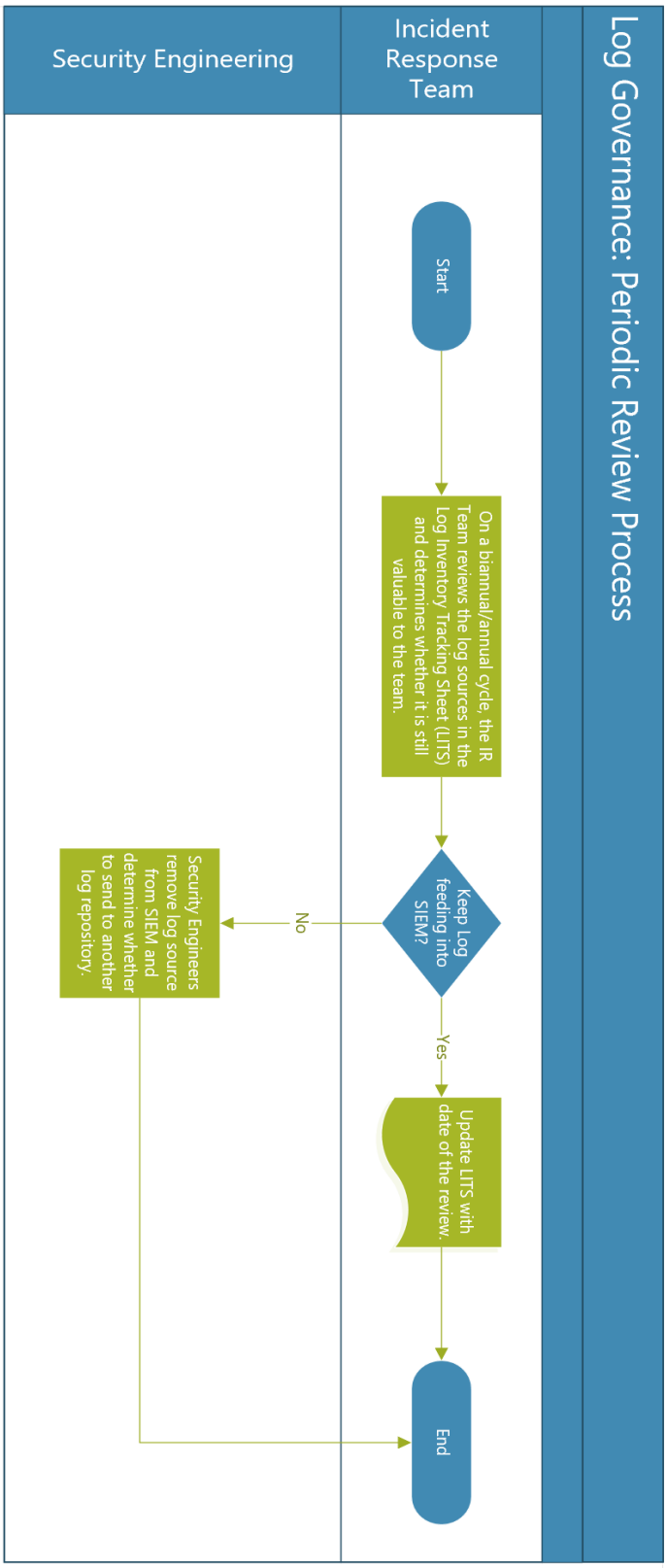

Figure 2 - Example of log periodic review process

Nate Street, nate.street@gmail.com

Appendix B – Practical Applications of the Log Quality Value index

Use Case 1: Incident Response team usage - As designed, the LQV is intended to be a tool primarily utilized by Incident Response teams to determine if a log source is or will provide valuable log data to the SIEM. With this in mind, after the log has been indexed using the LQV, further stratification can occur by determining which ranges qualify as being sent to the SIEM. For instance:

- If $V_x>20$, then send the log to SIEM
- If C_1 or C_2 and $V_x>10$, then send log to SIEM
- If C_3 or C_4 and $V_x<10$, then send to another log repository or do not log.

While the actual numerical value used to determine ranges will differ per the organization, these ranges can be adjusted as the Security Operations Team collects more historical data. Logs that are frequently utilized during security incidents should correlate with a higher LQV index. Alternatively, if there is an increase in SIEM alerts being triggered from log sources categorized with a Criticality Type III or IV, this may indicate a root cause analysis should be performed to understand the problem.

Use Case 2: Audit usage – Use of the LQV can enable Incident Response teams to meet the Critical Security Control #6 objective of "Maintenance, Monitoring, and Analysis of Audit Logs." Auditors may want to determine if the SIEM is receiving the right logs in order to mitigate the risk of failure to detect and prevent malicious activity. By providing as evidence the LQV index of each log source currently being fed into the SIEM, Incident Response teams can show how the control is mitigating the risk by ensuring that high-quality logs are being sent to the SIEM and periodically reviewed through a log governance process.

Use Case 3: Application Developer team usage – As more organizations move towards DevSecOps, the LQV can be utilized by developers during application development to understand how the logs being written to disc will enable an Incident Response team to

determine if potentially malicious, anomalous, or fraudulent activity is occurring within the application. Developers can provide samples of logs generated by the Application, and the IR team can provide the LQV index back as a reference point.

Use Case 4: Security engineer usage – Security engineers are typically responsible for deploying, maintaining, and operating security tools as well as configuring settings that will determine what is logged to disc. As part of the review, the LQV can be utilized during the deployment process to determine whether the log makes a good candidate for log collection by the SIEM. Engineers should work closely with the Incident Response team to ensure the tool generating the log is configured to log the right data before deployment.

Use Case 5: Three-point vendor-agnostic measurement – A primary value of the SIEM is the ability to enrich raw log data to provide better context for security events which enables faster analysis. Therefore, further research could be performed using a three-point measurement(M) approach:

M_0 – Initial LQV at raw log data source, default log configuration after installation of product.

M_1 – LQV measurement after maximum log configuration. Maximum being defined as all log configurations that contain data field attributes that enable security investigations.

M_2 – LQV measurement within the SIEM with data enrichment configured (e.g. Outbound DNS logs paired with Threat Intelligence Blacklist).

M_0 and M_1 can be evaluated per vendor software/hardware product creating a Minimum LQV, or α, and Maximum LQV or β. The M_2, or $\mathfrak{S}$, measurement would be evaluated after SIEM deployment and data enrichment log sources have been ingested by the SIEM. With [α, β] calculated independent of the vendor with the $\mathfrak{S}$ calculated after ingestion to the SIEM, then the $\Delta LQV = \mathfrak{S} - \beta$ would equal the value of the data enrichment the SIEM is providing to the IR team.

Appendix C – Log Sources

The log sources in **Table 1** contain a variety of tools and products along with their calculated C_xV_x index that might be found in a fictional company. The graph in **Fig. 3** is one method a company could use to visualize the complete log inventory and notice any trends or patterns over time.

Log Source	Product	C_xV_x Index
Cisco Secure ACS (AAA)	Cisco ACS	C_2, V_{11}
Anti-Virus	ZoneAlarm	C_1, V_{23}
Web Application Server	Apache	C_3, V_{10}
Web Proxy	Squid	C_4, V_4
Host IDS	OSSEC HIDS	C_1, V_{11}
File Integrity Monitoring	OSQuery	C_2, V_6
Routers/Switches	Cisco IOS	C_1, V_{21}
Web Application Firewall	Sophos WAF	C_1, V_{29}
Operating System	Windows 10	C_2, V_{10}
Operating System	Mac OS X	C_4, V_{12}
Threat Feed	Alienvault IP Reputation	C_2, V_{14}
VMWare	ESX	C_1, V_{14}
Remote Access (Linux)	SSHD Logs	C_1, V_{16}
Mail Server	Microsoft Exchange Logs	C_1, V_{32}
Windows Firewall	Windows Firewall	C_1, V_{22}
Baseline – Gibberish Log (Minimum LQV)	N/A	C_4, V_{-6}
Fictional Log Source (Maximum LQV)	N/A	C_1, V_{41}

Table 1 - LQVs of inventoried log sources for a fictional company

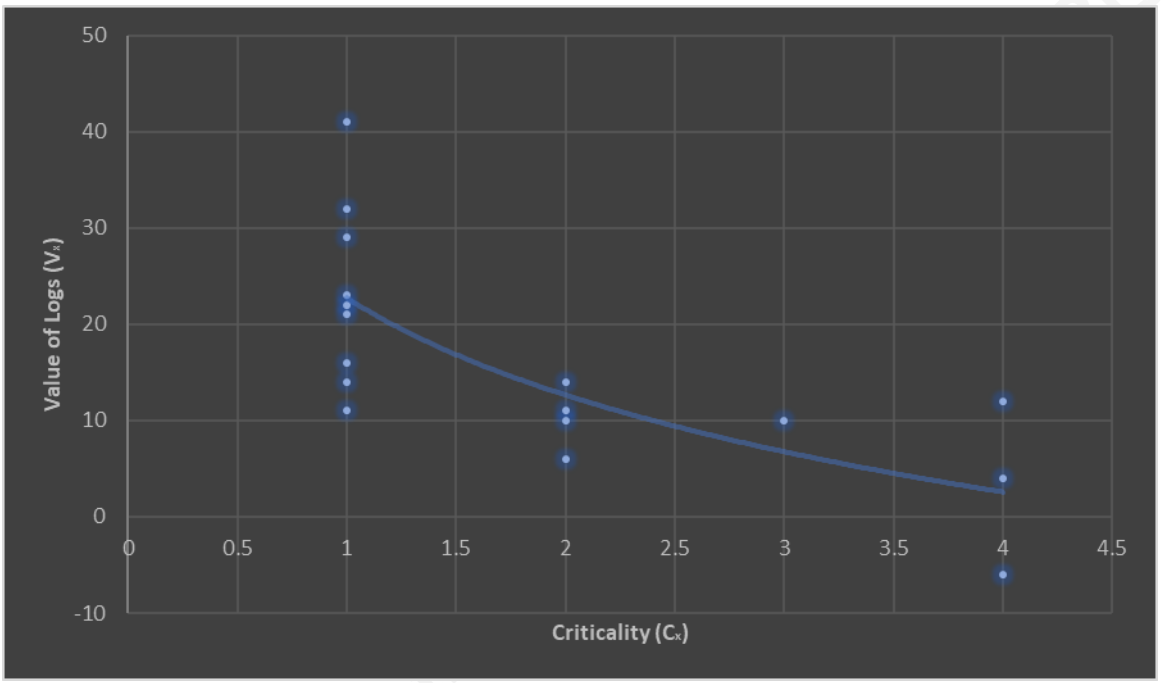

Figure 3 – A graph of LQVs of all inventoried log sources for a fictional company.

Nate Street, nate.street@gmail.com

Appendix D – Log Evaluation Characteristics

The full list of questions and associated weight values used in the research to evaluate the log sources are contained in **Table 2** below.

Evaluation Characteristics	Weighted Value
Do the events contain a single field which clearly describes what happened?	+1
Are the events time stamped?	+3
Are the time stamped events synchronized with an NTP server?	+2
Do the events contain a field which uniquely identifies the device where it happened? (e.g., Hostname, Server Name)	+1
Do the events have a field which contains the Source IP address?	+3
Do the events have a field which contains the Destination IP address?	+2
Do the events contain Source Port?	+1
Do the events contain Destination Port?	+1
Are the event's Source IP or Destination IPs NAT'd requiring an additional log to map the NAT'd IP to the actual device or application where the event occurred?	-2
Do the events contain information on user(s) involved? (i.e., Username)	+2
Do the events contain what data, account, host or information has been affected? What was the target of this action? Do the events identify what has changed?	+2
Can the context of the event be understood independently without correlation with another log/source?	+3
Is the confidence level in the accuracy of the event high?	+2
Can these logs be paired with other logs to enhance context of an action that occurred? (i.e., correlation candidate)	+2
Does the format of the log negatively impact the ability to understand when the action occurred (e.g. XML, "obfuscation", JSON)	-2
Is the log a report (e.g. Nessus Vulnerability Scanning Results) that is not easily parsed?	-2
Is this log a PRIMARY source of data in the organizational environment?	+3
Is it known whether this device would or could be used to determine that an attack occurred?	+1
Does this log source have the potential to capture events that require urgent response without any additional correlation?	+3
Do these events contain a severity value that is relevant to the organizational environment?	+2
Does the log source contain a default severity value data field assigned by the vendor that are not appropriate for the organizational environment?	-1
Historically (within 2-3 years), has this log source been utilized to detect or prevent a potentially malicious activity?	+3
Does another application/device on the network capture similar information?	-2

Nate Street, nate.street@gmail.com

Would this log source capture whether data/PII left the environment? Exfiltration?	+3
Is this log source only applicable to one type of platform despite more types being in the environment?	-2
Does this log source cover all platforms in the environment (e.g. Win, Linux, Mac)	+3
If data is masked in the field due to being sensitive (e.g. PII, Passwords, etc.), is the event no longer actionable?	-2
Does the log contain multi-line events?	-2

Table 2 - Characteristics used to evaluate log sources in research