University of Capetown

Department of Computer Science
Masters Dissertation

Advancing Security Information and Event Management Frameworks in Managed Enterprises using GeoLocation

Author:
Herah Anwar Khan

Supervisor:
Dr. Andrew Hutchison

Dissertation presented to the Department of Computer Science at the University of Capetown in fulfiment of the requirements for the degree of
Master of Science

University of Cape Town

12 December 2014

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Plagiarism Declaration

I know the meaning of plagiarism and I declare that all the work done in this dissertation, save for that which is properly acknowledged, is my own.

Abstract

Security Information and Event Management (SIEM) technology supports security threat detection and response through real-time and historical analysis of security events from a range of data sources. Through the retrieval of mass feedback from many components and security systems within a computing environment, SIEMs are able to correlate and analyse events with a view to incident detection.

The hypothesis of this study is that existing Security Information and Event Management techniques and solutions can be complemented by location-based information provided by feeder systems. In addition, and associated with the introduction of location information, it is hypothesised that privacy-enforcing procedures on geolocation data in SIEMs and meta-systems alike are necessary and enforceable.

The method for the study was to augment a SIEM, established for the collection of events in an enterprise service management environment, with geo-location data. Through introducing the location dimension, it was possible to expand the correlation rules of the SIEM with location attributes and to see how this improved security confidence.

An important co-consideration is the effect on privacy, where location information of an individual or system is propagated to a SIEM. With a theoretical consideration of the current privacy directives and regulations (specifically as promulgated in the European Union), privacy supporting techniques are introduced to diminish the accuracy of the location information – while still enabling enhanced security analysis.

In the context of a European Union FP7 project relating to next generation SIEMs, the results of this work have been implemented based on systems, data, techniques and resilient features of the MASSIF project. In particular, AlienVault has been used as a platform for augmentation of a SIEM and an event set of several million events, collected over a three month period, have formed the basis for the implementation and experimentation.

A "brute-force attack" misuse case scenario was selected to highlight the benefits of geolocation information as an enhancement to SIEM detection (and false-positive prevention).

With respect to privacy, a privacy model is introduced for SIEM frameworks. This model utilises existing privacy legislation, that is most stringent in terms of privacy, as a basis.

An analysis of the implementation and testing is conducted, focusing equally on data security and privacy, that is, assessing location-based information in enhancing SIEM capability in

advanced security detection, and, determining if privacy-enforcing procedures on geolocation in SIEMs and other meta-systems are achievable and enforceable. Opportunities for geolocation enhancing various security techniques are considered, specifically for solving misuse cases identified as existing problems in enterprise environments.

In summary the research shows that additional security confidence and insight can be achieved through the augmentation of SIEM event information with geo-location information. Through the use of spatial cloaking it is also possible to incorporate location information without compromising individual privacy. Overall the research reveals that there are significant benefits for SIEMs to make use of geo-location in their analysis calculations, and that this can be effectively conducted in ways which are acceptable to privacy considerations when considered against prevailing privacy legislation and guidelines.

Acknowledgements

First and foremost, Allhumdullilah hazar dafa, Allhumdullilah (Praise be to Allah a thousand times).

I'd like to say thank you to Dr Andrew Hutchison (a.k.a the Travelling Professor), I have been very fortunate to have him as my supervisor throughout my studies. His years of experience are a culmination of ongoing excellence in industry and academia. The experiences I have gathered in presentations, demonstrations and developed writing skills will not be forgotten or let to rust. I hope to continue striving towards levels of high expectation.

I'd like to thank all members of the MASSIF project for the opportunity to work and spend time with them, it has been instrumental in my learning curve. Thank you in particular to Valerio Formicola of CINI, Italy, for all his help throughout, grazie!

Thank you to T-Systems South Africa, for providing the data for this research, and the great opportunity.

A token of gratitude to the department of Computer Science at the University of Capetown. Thank you for providing me a space in which to grow and reach the sunlight.

To my fellow peers of Lab 300, thank you for all the great laughs, critical anime and life talks, supportive environment and the hundreds of tea breaks.

To my gorgeous family, mom, dad, brother, sister, **thank you**. In all your different ways and impressions, you have helped make and shape this, one chip at a time.

My dear friends, far and near, thank you for forever helping me remember why I am here, if I forget my dreams you remind me, and often, that is all I need.

To Hei, Ichigo, Ulquiorra, Kamenashi Kazuya, Haruma Miura and dear Hayao Miyazaki - hontoni arigatou ne.

To everyone else, if you were there, thank you. Thank you from my heart.

Contents

List of Tables

List of Figures

Listings

Acronyms

APT Advanced Persistent Threat.

BIOS Basic Input/Output System.

CBG Constraint-Based Geolocation.

CMS Central Management System.

CSV Comma Separated Value.

DdoS Distributed Denial of Service.

DPO Data Protection Officer.

EAM Event Aggregation Module.

EU European Union.

GET Generic Event Translator.

GIS Geographical Information Systems.

GPS Global Positioning System.

IAM Identity and Access Management.

IDS Intrusion Detection System.

IoT Internet of Things.

IP Internet Protocol.

IT Information Technology.

LDAP Lightweight Directory Access Protocol.

MASSIF MAnagement of Security and events in Service InFrastructures.

MESI Managed Enterprise Service Infrastructure.

MSSP Managed Security Service Providers.

NIST National Institute of Standards and Technology.

OSSIM Open Source SIeM.

OTP One Time Password.

PDU Protocol Data Unit.

POI Point of Interest.

PSA Predictive Security Analyser.

REB Resilient Event Bus.

RES Resilient Event Storage.

SIEM Security Information and Event Management.

SNMP Simple Network Management Protocol.

SSH Secure Shell.

TCP Transmission Control Protocol.

TSOM Tivoli Security Operations Manager.

TSSA T-Systems South Africa.

UCM Universal Content Management.

UDP User Datagram Protocol.

XML eXtensible Markup Language.

XSS Cross-site Scripting.

Chapter 1

Introduction

"There's a war out there, old friend. A world war. And it's not about who has got the most bullets. It's about who controls the information. What we see and hear, how we work, what we think...it's all about the information!" – *Sneakers(1992)*

1.1 Motivation

We knew for a long time this was coming - the great Information age. The influence of the internet in daily lives is a congealing reality of modern day. Security analysts are tasked with protecting information on networks facing a giant ocean of internet users, the count reaching 3 billion users[28] in 2014.

Advancing internet security tools to aid analysts in their efforts is a top priority within IT security research. The effectiveness of current procedures with fast-evolving technologies are re-assessed continually to stay a jump ahead of the user masses. In times where the concept of *Identity* still struggles to solidify itself on this mammoth world platform, strong security is cardinal.

To safeguard really large networks with many targetable users there are many security approaches available, one of these is the use of Security Information and Event Management (SIEM) technology. SIEM technology can be defined concisely as tools and processes for centralised real-time log event collection, integration and analysis in a distributed system. Using a centralised approach, SIEM tools can provide unified interfaces for disparate data, real-time correlation of different events in different parts of the system for early threat detection[6]. Administrators can utilise SIEMs to narrow down the large sets of security data generated from the multitudes of events to data of relevance, producing thorough and efficient reports[13].

SIEM is a differentiation in its particular capability to quickly sift through a sea of security and log data detecting behaviours in networks that indicate malfeasance.
According to industry research firm *Frost&Sullivan*, the world-wide market for SIEM systems are predicted to grow to $1.3 billion US Dollars in 2015, rising from $680 million in 2009[39].

Looking at SIEM technologies today, in order to maintain a strong hold their advancement

must be prioritised; questions on the current approach and methods of data analysis for security need to be assessed. Referring to the definition of SIEMs, the technology's primal focus is on the excavation of security event data from all possible sources for intelligent analysis. The potential of the data received by the SIEM directly influences the culmination potential of a SIEMs evaluatory abilities. The data that can be exploited from networks are now more richer in content and source, for until recent years the internet had not been incorporated with so many levels of devices. Remote sensing technology, for example, was generally known to have very limited capabilities for handling vector data proving unsuitable for network analysis and high quality plotting which are best done with vector format data[38]. Though at present, this is no longer the case. Current remote sensing technology provides compilation and analysis of data from the ground, atmosphere, and sky with links to Global Positioning System (GPS) data, Geographical Information Systems (GIS) data layers and functions, and other new modelling capabilities[51]. This data, often present in logs sent to SIEM collectors are rarely exploited for any security analysis procedures. Thus, the decided task, to evaluate these new data avenues in augmenting SIEM security analysis techniques, is a salient one.

The central theme of this study focuses on geolocation as an insufficiently exploited channel in these fields and determining the value it can bring to data management and security.
Especially today, in the wake of increasing terrorism concerns, governments and commercial companies around the world are looking to remote sensing and GIS to support security initiatives. When it comes to analysing data with physical relevance, using geolocation data offers a potential aid in monitoring and managing security operations.
Other possibilities of geolocation applicability is linked to the nature of the network being used. Should the criteria of physical location be an important aspect in the security and surveillance of the network, it becomes all the more useful.
A GIS itself can be seen as a versatile yet fundamental information management system. It offers strong ability of data transferring, data management and data analysis [33].
With the rising applications of geolocation for exploitation in all areas of technology it can prove to be more elemental with its incorporation. The study therefore centers on augmenting SIEM frameworks with the discovery of geolocation data exploitation techniques than can be capitalised on further than its basic use at present.

This research is focused on the missing exploitation gap between SIEMs as the security tools and the data they mine, specifically geolocation data, when advancing security to cope with the security challenges faced today. GIS applications were investigated in their use of geolocation that could be applied by SIEMs in security analysis.

However, the addition of new data realms brings along with it its own bone of contention. User geolocation data comes with significant privacy concerns in many countries, with efforts to creating legislative restrictions - understandably so - on the use and storage of this data. The second component of this thesis, focused on ensuring this additionally exploited data is 'defending the defender' by carrying out efforts towards ensuring user data privacy.

With focus equally on both security and privacy it is aimed to highlight the protection of user data as a fragile balance between enforcing and violating rights. The tools used to mine user data can be seen analogous to the firearm debate; the destructive or constructive potential hinged solely on the possessor, safe usage procedures need to be implicated leaving trust to

the mechanism rather, than on the bundle of contradictions that is man.

The following hypothesis is set forth addressing the research intentions with respect to SIEMs and geolocation data;

1.2 Hypothesis

- Location-based information enhances SIEM capability to perform advanced security detection.
- Privacy-enforcing procedures on geolocation in SIEMs and meta-systems alike are necessary and enforceable.

1.3 MASSIF

Recognising the importance of new generation SIEMs, the EU funded a large-scale FP7[1] project to investigate the topic surrounding advanced SIEMs. The project ran for the duration of September 2010 to October 2013. This research was concluded in the context of, and draws reference from, the scope and activities of this project.

The purpose of MASSIF was to achieve the following - an advanced SIEM Framework, defined as "a framework that aids the compilation and processing of security events from various layers down to the low level network components, physical sensing devices and business related applications through methods that are trustworthy, resilient and secure[40]".

Location-based approaches were not at the time a focus in the project so this work facilitated a contribution to this approach. The research was sufficiently independent to benefit from the use-case scenarios and 'testbed' of the MASSIF project but independent of the timeline or the deliverables of it.

1.3.1 Scenarios

The project, MAnagement of Security and events in Service InFrastructures (MASSIF), considered and investigated the topic of advancement in selected scenario environments. Using an approach focusing heavily on requirements, these environments are used as the foundation for extracting them; encouraging an advanced SIEM by providing solutions to specified needs rather than determining how to collectively satisfy potential users in loosely guided advancement aims. It makes use of these scenarios to provide design guidelines, using the scenario requirements as the challenges for research, adaptation, testing and evaluation.

[1] *See* http://www.massif-project.eu/

The use-case based approach is applied in four scenarios as seen in Figure 1.1. These selected industrial domains serve as a source of requirements to validate and demonstrate project results:

Figure 1.1: The Scenarios of MASSIF

1. *Olympic Games IT infrastructure.* The processing of huge amounts of generated data events in real time, deployed and managed by partner Atos.
2. *Mobile phone based money transfer service*, facing security events, especially for the "non-IT" and "service" events. This was handled by partner France Telecom.
3. *Critical Infrastructure*, such as a dam, using advancing concepts of SIEM to support their Information Technology (IT) system, managed by partner Epsilon.
4. *Managed IT outsource services* with a high degree of complexity in setting up SIEM systems for large distributed enterprises. This was managed by T-Systems SA.

This research work will be carried out within the Managed Enterprise Service Infrastructure scenario, sourcing environment data from scenario provider T-Systems International SA[2] (TSSA).

1.3.2 The Managed Enterprise Environment

TSSA is a typical example of a Managed Enterprise environment in industry with a large presence in the enterprise platform functioning as a global IT outsource service provider. Most often, the clients for TSSA are classified as very-large enterprises. The MASSIF project looked to contribute solving problems which current SIEM systems are facing to strengthen the resiliency and business continuity requirements of such large companies.
The main requirements of the managed enterprise service infrastructure (MESI) scenario were; (1) to improve the ability of service providers to identify and manage security attacks and; (2) increase security of enterprise information systems through the deployment of advanced security detection techniques.

[2]http://www.t-systems.co.za

Misuse Cases

The MESI scenario identified five misuse cases that apply to its environment and need to be addressed in the advancement of SIEMs to aid a better application of SIEM in this kind of distributed environment. These misuse cases identify the top five threats persistently encountered by security administrators within the live enterprise scenario.

The misuse cases are listed with their corresponding ID (MC-X.Y.Z) given in context to the MASSIF project documentation; such that *MC* stands for Misuse Case, and *X.Y* links to the section listing the **final** approved misuse cases in the official Scenario Requirements Deliverable 2.1.1[41]. Finally, Z is the number allocated to the misuse case itself.

- MC-5.5.1 The Brute-Force Password Attack
- MC-5.5.2 Un-authorised Login to a computer system, network or application
- MC-5.5.3 SQL Injection
- MC-5.5.4 Session Hijacking through Cross-Site Scripting
- MC-5.5.5 Worm Propogation

Addressing these challenges put forth by the environment aligns with the objectives of this study. Within the scope of research there are two industry SIEMs that are investigated in MASSIF, namely, Alienvault Open Source SIeM (OSSIM)[3] and 6Cure[4].
Alienvault's OSSIM is a leading innovative SIEM product and provides an open source version available to users all over the world. Gartner[5] identifies OSSIM as one of the visionaries in the SIEM market, proving salient in the influence of effective SIEMS. Figure 1.2 shows the Gartner Magic Quadrant for SIEM frameworks in 2013.
This SIEM solution was additionally incorporated into research investigations and facilitated part of the testbed experiments of enhancing SIEMs with geolocation. TSSA provided the environment data in which they manage security services for their clients as source data used in the experiment.

1.3.3 Context: Involved Partners

The project running for three years till the end of 2013, set out to demonstrate the advancement of SIEMs with a research consortium[6] consisting of SIEM product providers;

- AlienVault OSSIM
- 6Cure

scientific research partners,

- Télécom SudParis (IT), France.

[3]http://www.alienvault.com
[4]http://www.6cure.com
[5]Technology research firm Gartner, is the leading global provider of independent and objective technology related research and advice. *See* http://www.gartner.com
[6] *See* http://www.massif-project.eu/partners

Figure 1.2: Gartner Magic Quadrant for SIEM Frameworks

- Fraunhofer Institute for Secure Information Technology (Fraunhofer SIT), Germany.
- Institution of the Russian Academy of Sciences St.Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), Russia.
- Consorzio Interuniversitario Nazionale per l'Informatica (CINI), Italy.
- Distributed Systems Laboratory, Universidad Politcnica de Madrid (UPM), Spain
- Fundaao da Faculdade de Ciencias da Universidade de Lisboa (FFCUL), Portugal.

and industry use case providers,

- Atos Research and Innovation (ATOS), Spain.
- Orange Labs - France Telecom (FT), France.
- T-Systems South Africa (TSSA), South Africa.
- Epsilon srl, Italy.

as the contributors to the project.

1.4 Methodology

The methodology applied comprised of an investigative research approach using the MASSIF platform for experimental application. The primary research objective concerned advancing SIEMs with geolocation and was carried out in two main steps;

- An investigative study into security methods capitalising on geolocation data in order to assess geolocation credibility in the field of security; and conducting research into the need for privacy enforcement of systems carrying personally identifiable data. For the latter, a privacy model for meta-systems such as SIEMs is introduced based on privacy legislation.
- The use of location-based information was investigated in Managed Enterprises, making use of the final stages highlighted in Figure 1.3, in MASSIFs use-case based approach.

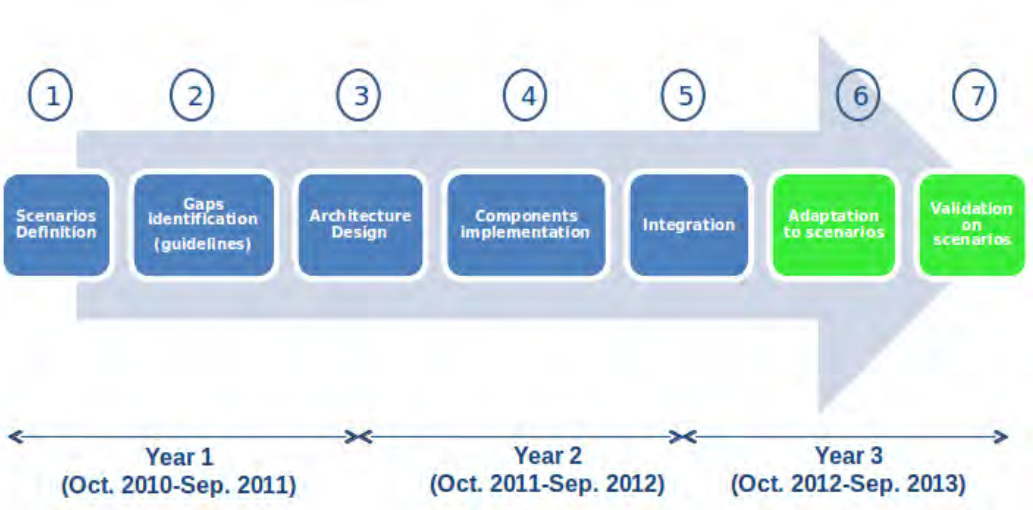

Figure 1.3: The MASSIF Approach[22]

The study focused on the experimental tool integrations in the final stages, 6 and 7, where the adaptation and validation of scenarios with the MASSIF tools are performed. The primary result of the tool adaptation and validation was to determine the application of tools in each scenario and their contribution to each context.

For this research, using these stages, we contributed and performed experimentations with location information in the Managed Enterprise scenario for the security augmentation.

The conclusions drawn from our investigations in terms of location-based security and meta-system privacy were applied to the managed enterprise scenario testbed and achieved through some selected tools of MASSIF and an Alienvault OSSIM SIEM solution.

1.4.1 Research Approach

Using the conclusions drawn from the investigative study and learning from the MASSIF project, an existing SIEM Framework was evaluated to determine if GIS methods can be applied to the security challenges discovered and can be seen as apt and applicable solutions to address them. Once applicability was determined, further study was carried out to explore the integration of GIS into an existing SIEM. Using a feasible integration, an integrated prototype was created and tested. The privacy implications of geolocation data and its use were

addressed, investigating its adherence in the prototype experiment.

The identified security issues in enterprise are applied to SIEMs and methods of addressing them were discussed. The success of the privacy and security solutions determined, if carried forward across to SIEMs, can augment SIEMs in the predicted future and strategise the longevity of these frameworks as security giants of technology.

1.5 Thesis Outline

This thesis is divided into two main sections, SIEM Security and SIEM Privacy. It encapsulates the efforts of applying geolocation information to strengthen these two areas in SIEM technologies. The thesis continues henceforth with Chapter 2, a background study on SIEMs and the position of geolocation today. In chapter 3, SIEMs and Security are discussed, paticularly the application of security augmented with geolocation. Chapter 4 investigates SIEMs and Privacy, how concerns of privacy can be addressed in SIEMs with personal data like geolocation. This follows the experimental design and testbed architecture in Chapter 5. In the implementation, Chapter 6, the technical procedures and strategy utilised in the prototype developed as the proof of concept are documented. The prototype results are confirmed through test executions and inspection of the applied test data in Chapter 7. Chapter 8 evaluates the results of the implementation in the context of determining overall SIEM advantage in terms of security, efficiency, and adaptability. The methods used are also evaluated and interpreted in their approach conclusions. Finally, in Chapter 9, a summary of findings is provided and an affirmation of the goals of this research is presented, closing with recommendations for future work.

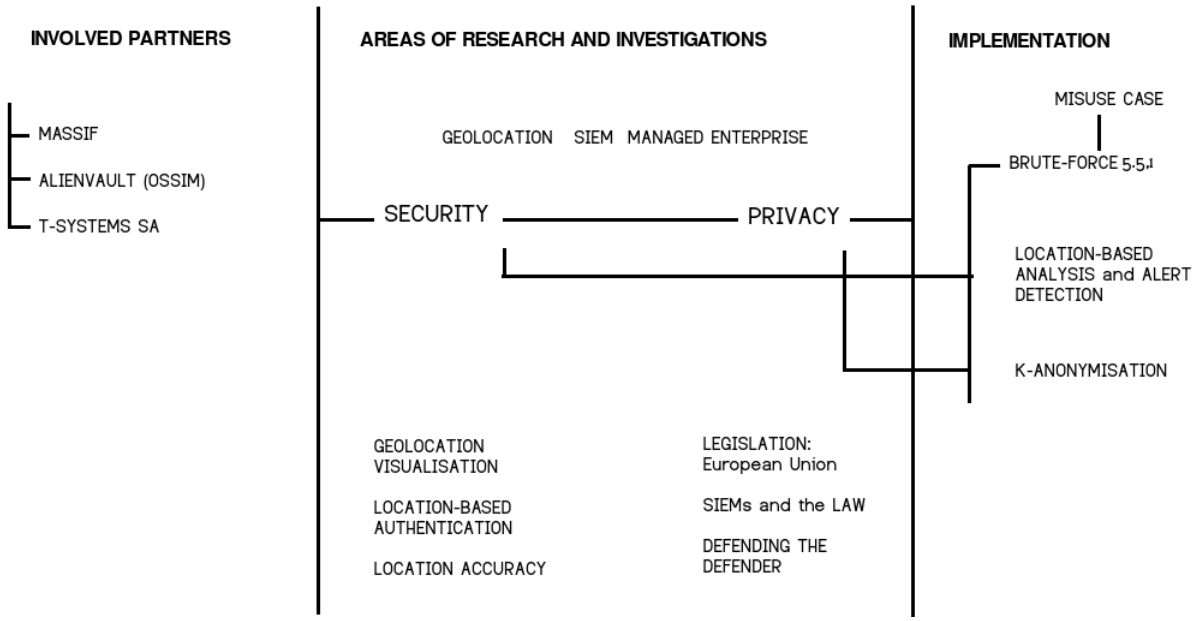

Figure 1.4: Thesis Mind Map

Chapter 2

Background

The following chapter concerns the definition of SIEM technology, defining components of the SIEM, from the low-level collection to the interface at high level. Common security analysis techniques used by a SIEM framework are discussed to provide insight in the provisions and expectations of a typical SIEM. The methods of security data management and exploitation for analytics are primary aspects of concern. Geographic co-ordinate data and it's application in various security-enforcing systems through GIS are investigated. The advantages of geolocation (geographic location) data are discussed in the context of SIEM augmentation in areas of security and privacy. The two areas are then investigated in detail in the chapters following.

2.1 SIEM Overview

Security information and event management (SIEM) is an approach to security management with the focal intention of providing a holistic view of IT security within an organisation. The term is a combination of two functions, Security Information Management(SIM) and Security Event Management(SEM).

2.1.1 SIEM Components

A typical SIEM framework can be seen as nothing more than a large security management layer set across existing systems and security controls. It unifies security data through low level devices to high level application sources, acting like a cosmic security lens. The underlying tools and process structures required to support this data conglomeration, however, requires a considerably high level of technical expertise. The four major SIEM functions on the centralised data can be seen as log consolidation, threat correlation, incident management and reporting[57].

The framework has two main components, as mentioned earlier, the SIM component and the SEM component. The SIM provides log management, the collecting and logging of all supporting regulatory compliance data as well as internal threat management, resource access monitoring and all other security-relevant information. The SEM concerns itself with event correlation, intrustion detection and incident response[6]. Both of these components are viewed together as a combined concept - SIEM.

The SIM component receives data from multiple sources - intrusion detection systems, syslog tools, firewalls, vulnerability scanners and any sensor agents deployed to retrieve security relevant information. This data is then collected, combined and centrally stored in logs, typically secured with encryption procedures. The SEM is tasked to use these amalgamated logs for data mining activities such as correlation of events, real-time monitoring, tracking data assets and scanning for vulnerabilities.
The SEM typically has highly developed reporting capabilities, administrators and managers of the system can be interested in variant statistics and data or may need independent reports for different compliance requirements[6].

The combined tools provide the following minimum functionalities as a SIEM:

- Log collection and drill down analysis capabilities
- Event, file integrity and user activity monitoring
- Event normalisation and aggregation
- Correlation rules, security policies
- Real-time monitoring
- Incident response, alerts, messages
- Statistical analysis and feedback

Further functionalities that can be provided from application to interface level to be used by analyst and security administrators vary on the distribution of SIEM that is chosen. SIEM tools have been widely developed in industry with many SIEM solutions available, some of which are present in the field of open source. Depending on their various approaches each have a specific overall goal to offer, aiming to meet user expectations in their best practices, with solutions often providing more than less.

2.1.2 SIEM Architecture

The data that is collected from various sources by the SIEM is aggregated - combined into a single stream, then normalised - translated into standardised format[3]. This is done to reduce duplicates and accelerate the subsequent analysis. It is then correlated between data sources and analysed against rules defined by vendors, users or correlation algorithms created by security analysts. This process aids the provision of real-time incident/event reporting and alerting of events that need handling. The resulting data is stored in a secure manner with methods such as encryption, to ensure data integrity, which allows them to use it as evidence in investigations or for meeting compliance requirements. The SIEM also provides maintenance and authoring of correlation rules, often supporting rules that can execute based on arbitrary conditions as well as anomalous behaviour. An example of such a rule is a negative condition rule, where the absence of an event over a period of time executes a rule, such as a backup process that misses a scheduled routine[3]. Figure 2.1 depicts this flow of information processes throughout the various areas within the SIEM machine.

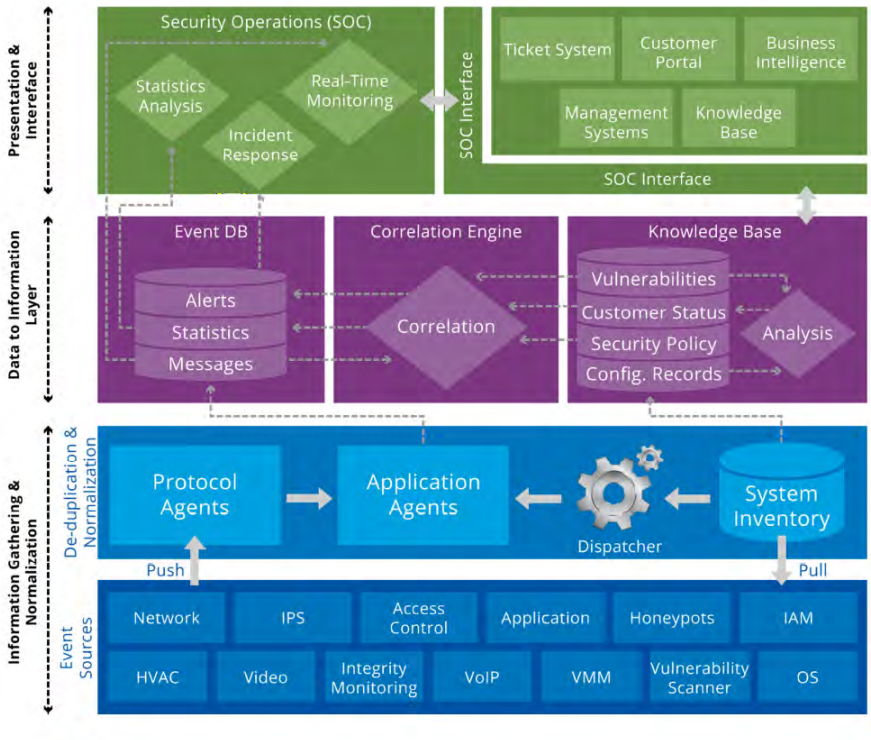

Figure 2.1: SIEM Architecture Reference[3]

Information gathering & Normalisation

- Event Sources.
  SIEM collects data from the network (including servers like Domain Name Servers (DNS), web servers, Active Directory (AD) servers, database servers and resources such as switches, routers, hosts, databases, application logs, vulnerability management systems, firewalls, anti-malware systems, honey-pots, and Intrusion Detection Systems. Other less traditional sources of data include Identity and Access Management (IAM) systems, occupancy sensors or the lighting systems they control, Closed Circuit Television (CCTV) systems, fire control systems, power management systems, and any other system that can provide applicable data that aids in defining a security event in the enterprise. A SIEM system can basically take almost any data that can be digitized and, through a proper rule-set, create a useful output[3].

- De-duplication & Normalisation
  Data can either be sourced directly into inventory from sources or through the process of agent collectors. The two important factors that need to be considered before correlation are de-duplication and normalisation. These procedures are the first stage in efforts to remove the clutter of noisy data and retain the usefulness. De-duplication removes repetition and simliarly normalisation standardises the data and removes redundancies. The data is then dispatched to the application agents for ensuing correlation procedures.

Data to Information Layer

- Event Database.
  A SIEM establishes an early warning system to take helpful preventative actions, providing alerts, messages or generating relevant statistics for investigation. This type of information is generated from the preceding analysis and correlation and are stored in the event database to be sent through to the user interface.

- Correlation Engine.
  Functionalities such as event correlation are very powerful procedures that do the magic of going through tons of event network data and spotting the interesting, odd bits that analysts can precede investigations with. Event correlation is more powerful than the entire event aggregation. An example of the magic in correlating data, is in the case of a large number of messages from different nodes in a distributed system being directed towards the same destination node, this sort of pattern is indicative of a distributed denial attack launched by the system[6].

- Knowledge Base.
  As part of the implementation of a SIEM solution, tuning is performed to reduce false positive alerts so that the device is giving relevant information to each specific environment[3]. With the use of important context such as security policies and customer status we can avoid the threshold for false positives in alarms and statistic predictions. Analysing current vulnerabilities requires a fundamental application of knowledge when mining security events for malicious activity.

Presentation and Interface

- Security Operations Center (SOC).
  The Security Operations Center is the area that delivers the mined security data in formats that provide the utmost usefulness to a user/system on the receiving end. The data is given in statistical formats, such as reports, graphs. Alerts generated provide response procedures, and of course the data itself is delivered as it is received for real-time analysis and monitoring.

- SOC Interface.
  At the interface, the output can be delivered to response systems, third parties, system analysts, ticketing systems or maintenance systems. It all depends on the the system in which the SIEM is placed and how it prefers to perform its next step of defence after it has received the incriminating evidence and indications. Business Intelligence is a very important part of applying a SIEM correctly to a system. As the needs of a business change, in turn so do the vulnerabilities and operations which requires different rules and data feeds to achieve the business objectives of the SIEM system[3].

2.1.3 SIEM Benefit Summary

A SIEM framework can help gain centralised visibility, leverage the value of existing investments and prepare for potential threats that could compromise their business-critical information assets[3]. Overall, SIEM is an approach that can provide powerful insights with the strength of conglomeration and intelligent data mining. Of course, the effectiveness of a

SIEM is stringent on the rules and policies put in place and its deployment correctly applied to the relevant environment. This requires skilled analysts and know-how if it is to effectively extrapolate the benefits of such a giant management service. Using a SOC approach it helps an organisation focus on data patterns from all over the network, predict the networks behaviour in cases of possible security incidents. Advances in correlation and analysis of events is a saviour in large data volume situations which is a matter that is all the more increasing in the times of the Information age. With the evolution to an " Internet of things " more devices are IP enabled and providing data that can be digitized and applied in security analysis, without the use of correlative analysis techniques data interpretation would be an astronomical task[26].

SIEM will continue to play a vital role as a strong security handler of large data, with its mining, integrity and quick response real-time monitoring abilities.
In the next section some of the main analysis tools used in SIEMs are discussed and how they approach data to identify anomalies.

2.2 SIEM Security Analysis Techniques

In this section, elemental techniques applied in SIEMs are described and the various ingenuities that are administered to sniff out malicious user intention. These tools focus on specific qualities of the data to mine intelligent patterns that can trace irregularities, it can also be determined if geolocation data can benefit within these mining techniques.

2.2.1 Event Normalisation

A SIEM framework performs the task of centrally combining all security data for unified analysis, normalisation is the process that makes this possible. The process of taking raw data input and extracting only relevant fields is called normalisation, for example, when collecting data from various sources, the range and formats in which the data is received varies extensively. These events are converted into a standardised format using a generic schema, not only does this help combining data but it also aids better comparison of events. The main purpose of the event normalisation technique is to create this common event format which all source data can be translated into, enabling the unified comparison and collection of information across the entire network.

2.2.2 Event Correlation

Security events need to be analysed from as many sources as possible to aid threat assessment for the appropriate response, the greater the number of sources the better the potential situational awareness can be invoked. From this, enters samaritan methods such as event correlation. Event correlation is a technique for analysing a large number of events and indicating the few events that are really important in that mass of information.

Correlation is the process of transforming a sequence of events matching certain conditions to output an event. The output event is an indication of a triggered alarm to indicate the identification of a pattern. These patterns are event sequences corresponding to suspicious

behaviour such as cyber attacks or an unauthorised login. Listed below are just a few of the many attacks that correlation rules can detect through their pattern-based analysis:

- Web service attacks (e.g. SQL injections, cross site scripting, etc.)
- Bruteforce authentication attacks across protocols (e.g. Secure Shell (SSH), Lightweight Directory Access (LDAP))
- Policy violations (e.g. the use of torrents, anonymous proxies)
- Distributed denial of service attacks (DDoS)

The events generated from the correlation process are sent to the server as an event from the SIEM sensors, and is stored in the SIEM databases to keep a record of the triggering event incident. The real-time correlation of events greatly aids security administrators in filtering through the extensive loads of security events entering the SIEM framework and highlight data of security relevance, indicating activity with high potential of malfeasance.

2.2.3 Process Mining

Some advanced tools used for the evaluation of security events make use of process mining. This concerns the extraction of knowledge of a business process from sources such as process execution logs. This alternate perspective of mining aims to gain insight from varied views such as process flow control and performance[50].

In such cases, the tool analyses the known-control flow of the event-driven processes involved against any deviations from required security properties for that environment. Deviations can be found from anomalies caused by attacker interactions, problems incurred with measurements(e.g loss of events) or an evolution in the process specification[27].

Figure 2.2 is an example of event processing performed by MASSIFs process mining tool[27], the Predictive Security Analyser (PSA); to use this technique, there a few steps that need to be followed. First, certain security requirements are identified that the event-driven processes need to adhere to, sort of like the business rules of a process flow.

These rules are specifications of the required security properties the monitored process should adhere to. Process specifications can be determined from process discovery tools e.g. Petri Net specifications can converted by ProM, a downloadable process mining tool[1] that caters for many modeling languages.

After the requirements have been determined a process model is created that will encompass the incoming events and the events themselves need to be made available in real-time through a collecting process.

If a critical state is incurred (violation of business rules) the tool provides a process-oriented visualisation of the problem, providing a method of situational awareness of process states and alarm generation methods that can be directed to reporting into the central response center of a SIEM.

[1] http://www.promtools.org/prom6/

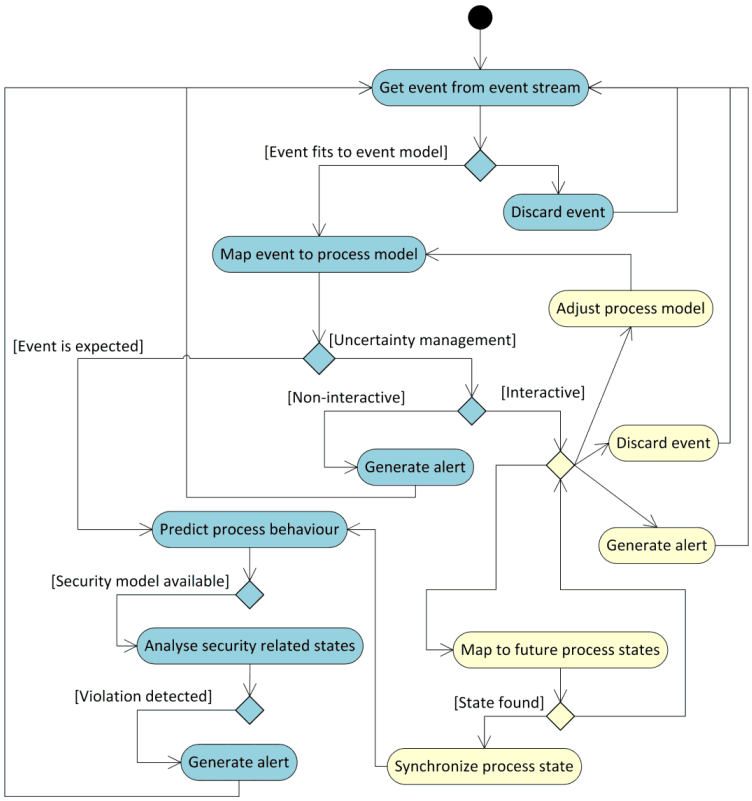

Figure 2.2: PSA Event Processing[27]

2.2.4 Attack Graphs

Security violations in a network can be caused by many factors such as security policy errors, system vulnerabilities, incorrect configurations and other security miscalculations. A malicious agent will use such vulnerabilities in the network as methods for penetration of the system. An assault follows a chain of attacks from one point of entry to the intended target machine. This includes different network resources and a myriad of different types of attack actions. The step-by-step compromise can realise different security threats and possibilities in network compromise.

An attack graph is a route calculation taken by an attacker through system vulnerabilities. The graph can be seen as a collection of scenarios showing the various methods in which a malicious agent can compromise a target system. The correctness of a graph refers to the mitigation of an attacker route. The execution is successful or correct if the attacker cannot reach the intended target of compromise. The level of correctness is defined as a security property[55]. An example of a security property in a computer network would be a statement like the intruder cannot get root access on the web server. similiar to the security requirements required in the process mining tool process discussed earlier.

In SIEM systems, the security administrator is tasked with checking network configurations and security procedures to mitigate the level of vulnerability. Through the use of these graphs

the administrator can perform security checks at various stages:

Exploitation Stage Current security event and alerts can be taken into account as well as changes in the configurations of computer networks. This includes the identification of new vulnerabilities, attack exploits and services to be added. The analysis is a continuous process of network monitoring, vulnerability assessment and security level evaluation.

System Design, First Stage Analysis The specifications of network configuration and security policies facilitate the main input for performing security analysis.

Operational Stage The main input is the actual parameters of the network configuration and security policy including the alarm and security event sequences.

Therefore, system administrators would use such a tool to determine the security measures that need to be deployed to patch the detected vulnerabilities[55].

2.3 The Power of GeoLocation

A person's location can be just as, or more valuable, than a persons identity for either justice or the opposite depending on the application. If applied to computer science this refers to the two main fields, Security and Privacy.

To support the motivations of this research, the following questions with regards to geographic data are raised:

1. What can geolocation (geographic location) provide for SIEMs? Is it significant?
2. How does geolocation feature in global privacy concerns?
3. Can geolocation augment tools already existing within the SIEM?

The investigation is lead through attempts to answer these queries, starting with an initial exploration of geographic data and its application in GIS applications.

2.3.1 Core Offerings

Geography, is an integrative discipline that can be applied in a wide range of areas dealing with spatial components. Geographic information systems(GIS) are versatile yet fundamental information management systems. They offer strong ability of data transferring, data management and data analysis [33].

The advantages of applying geolocation to provide various functionalities to a network or distributed system are summarised below:

- *Visualisation*
  Visualisation is a key aspect in the application of GIS, this generally infers that large sets of complicated data are represented graphically. By doing this, information displayed is actually absorbed into the human visual system directly rather than being read, interpreted, depicted, then processed, this encourages a much faster response and

analytic ability of the person. It also supplies a more accessible and realistic visual expression of static security assessment[33]. By exploiting this ability, information can be used more effectively and help a user make more sound decisions as well discover problems faster.

- *Dealing with large amounts of data*
  GIS has a powerful ability to display and manage large data. It offers an effective tool in dealing with large scales of data and engineering of data more directly[33]. Dealing with copious amounts of information is a long standing issue that drives the efficiency of a network. Information filtering needs to be balanced between what information is relevant at the time, what is contextually important and what is neither. GIS can be seen as an apt solution for such a situation.

- *Data transferring, data management and data analysis*
  A lot of the data is dull and redundant, with GIS one can engineer the data more efficiently as it proves to be an effective tool for data transferring, data management and data analysis [33]. The collective management of data and other aspects can all be facilitated within GIS.

- *Focus on the User*
  Having a friendly interface implies ease of use, data comprehension and can be seen in terms of a security officer having to analyse a situation in a timely manner. This is an important aspect for end user satisfaction as it determines how well or productive a system is. One can consider the creation of a user-specific GIS system. The versatility of such a system enables it to be used not only as an information management tool but also a decision support system. With more focus on the usability and the end-users needs more efficiency in response and decision making is advocated.

- *Selective display*
  Liu[33] states 'different information should be displayed in different ways'. The purpose of this is to allow administrative users to focus on related data in the area of concern enabling them to pay attention or point out the weak points of a system.
  The concept of *geographical perspective* is applied through the use of geolocation on networks or stand-alone systems. This perspective can also be used to examine the reliability and vulnerability of infrastructure[46].

GIS has also been applied to determine the efficiencies of a system. Using overlays of energy distribution within a physical system, information is gathered on where energy is being used and patterns of energy usage. Locational values assigned to each bus present on the system is geo-referenced with the energy map overlays resulting in correlated information that can help accomplish adequacy and overall security system benefit[23].

2.3.2 Where you are: Matter of Security and Privacy

The possible use of geolocation for supporting the security measures in a system and it's role in privacy is an important consideration best examined through the existing systems or

networks that apply this data in such a context.
The various implementations of geolocation through GIS that augment security to provide some of the benefits mentioned are discussed in this section.

GIS has been used to create a command and control infrastructure for the protection of critical infrastructures[64]. Using a geographical information system as the common semantic foundation in presenting and analysing its data. The concept of selective display is also used whereby information is visually shared with - where necessary - downgraded information depending on what needs to be imparted. Modeling and simulation of infrastructure elements as well as the interdependencies and risks to them is done on this system[64].

An example of GIS being integrated in a security system is ESRI with SAP[13]. It is used to visualise the state of the computer system instead of having physical detectors . The physical detection of systems using actual visualisation of the system can help provide better intrusion detection [13]. The concept of using GIS for the protection of critical infrastructures had already been introduced in 2003 when the Open GIS Consortium formed a pilot initiative for a limited regional area (north-eastern US and south-eastern Canada)[64].

The application of a visual mathematical model for intrusion detection using GIS[59] has been implemented. Though it should be noted here the GIS aspect is simply the final phase using the basis of location to relate all models in the network. This facilitates a lot of collaborated useful information as a result. The intrusion model itself uses an abstract geometric approach to intrusion detection. Geometric models enable the possibilities of combination and prediction. Information that is visually presented is much more naturally and easily processed by humans. The model is suggested to improve the performance of an intrusion detection system. It proposes a standard by which to encode and characterise complex systems as well as a visual model to present information in a context - where data is presented as relationships between components[59]. To limit the computational requirements current intrusion detection systems tend to focus on a limited set of system and user attributes, thus, the dynamic behaviour of a hostile user across a networked system is hard to encapsulate and characterise with such a limited set[59]. The model attempts to overcome this with a completely different approach to how systems security data are modelled. This visual model can be created for all computers on the network with a semantic basis of physical geographical location to view the activities of the network nodes. The collaborative information drawn from this can help trace dynamic intrusion detections on the network. Using these behavioural patterns determined from dynamic behaviour one can further help the early detection of intrusion and ensure network protection.

One of the primary uses of GIS-based systems is the ability to visually correlate information.It has a foundation for integrating varied types of information that needs to be selectively displayed for decision makers, enabling them to quickly judge a situation or potential risk before time[64]. An important aspect of GIS is the capability of providing situational awareness, providing invaluable context to a situation to enable timely reaction and enable more accurate decisions.

An example of the potential applicability of GIS was the previously mentioned power system static security assessment[33]. Using GIS as its base for security analysis, it highlights very

important advantages which current SIEM Frameworks can be evaluated against. The main aspect of GIS applied was 'Visualisation', displaying large complicated data information with visual figures[33]. It can abstract useful information from complicated data quickly. It needs to be determined what existing tools in SIEMs convert complicated data into understandable data and how the data is currently given and summarized to users. After which it needs to be evaluated whether GIS can contribute better results. Aspects such as selective display and friendly user interface are also very important aspects to take into account. Customising the data according to user needs can help form an invaluable contribution to better decision making. The creation of general views allows users to hide lower level data and pay attention to the weak points of a system in relation to its overall position, this can give key information to structural weaknesses. The contextualisation through geolocation can be further extended to represent the vulnerability of infrastructure components to a far wider range of potential threats. Taking into consideration the fact that vulnerability is dependent on the type of attack, it is also possible to change the level of vulnerability enabling dynamic calculations of effects on the system. Thus, GIS approaches can be utilized to promote a more 'joined up' approach to resilience planning for physical assets such as critical infrastructures[29].

Currently, within existing SIEM frameworks the data generated to provide security information has considerate potential for strong security enforcement depending on how this information is exploited. Additional information that can be gathered from the low-level devices is geographic location but it is not yet exploited to much extent in SIEMs. Without the analysis of data with respect to location a significant gap in the aspect of efficient security is present.

The security event management component in a SIEM consists of processing events from security and network devices. Event information gathered from various devices would become more useful if supplied with location data. This can then be mapped using GIS techniques to help relate and monitor the physical status of network and the consideration of compromise at any point.
SIEM data can be related with geolocation to formulate contextual input which encourages speedy analysis for large systems. It can also aid security assessment reporting in line with other assessments in an existing SIEM.
It can facilitate an acceleration of behaviour in the areas of detection, preparedness, prevention, response and recovery. This is due to its ability to provide 'situational awareness' which has become a fundamental aspect for better decision making of large systems.

SIEM Frameworks retrieve masses of feedback on all components and security events within a system, the ability to filter and evaluate this data needs the highest attention to enable timely reaction and prevention. Unarguably, the effectiveness of SIEM depends on this very aspect, which could possibly be assisted by the integrative discipline of GIS.

The assuaging proposition of geolocation in SIEM, through the advantages discussed, brings in the second consideration - the privacy concerns surrounding geographic data location. Geographic data in some country regulations is considered personally identifiable information as sensitive as ones own name and surname. The incorporation of geolocation in SIEM needs to be evaluated with an assurance that data rights are kept secure while augmenting security in one or more of the ways towards advancing SIEM capabilities.

2.4 Summary

SIEM technologies can be defined as holistic approaches to security analysis and detection. The framework centralises and performs mining procedures on the collected data to produce insights into the conditions of all devices and systems being monitored. The methods applied for security data management and exploitation for analytics are a primary concern. Common security techniques used within SIEM are discussed and considered in their contributive abilities to incident detection and response in the context of a SIEM platform. This concludes the contextual description of SIEM from high level to component level, highlighting the various technical methods in security analysis used by these SIEM technologies. Geographic location has assorted applications in the areas of security and analytics filtering. The main contributive attributes of geolocation can be condensed into the following:

- Enhanced visualisation
- Aiding user preparedness and rapid response
- Selective display, isolating errors or situations
- Predictive modeling
- Better network analysis and simulation
- Improved decision making
- Facilitating dynamic visual intrusion detection
- Risk assessment of assets with physical relevance

The advantages of geolocation reviewed from various applications in security systems indicate the possible augmentation of security within a SIEM. In the chapters to follow, geolocation is considered for specified security challenges within a managed enterprise environment and in the considerations of privacy for data within a SIEM environment.

Chapter 3

Security, SIEMs and Location

"Hardware is easy to protect: lock it in a room, chain it to a desk, or buy a spare. Information poses more of a problem. It can exist in more than one place; be transported halfway across the planet in seconds; and be stolen without your knowledge."
– Bruce Schneier, Security Expert

3.1 Where is Security Now?

The problems faced by organisations, SIEM frameworks, and all security devices alike identify concerns that need to be addressed in current security battlefield strategies. At an RSA Conference, McAfee's Chief Security Officer summarised the key trends and drivers of security[11] today as;
(1) data mobility through applications/software,
(2) the advent of the cloud,
(3) concerns in regulatory compliance
and lastly,
(4) emerging threats in network and data security.

These four corners are reviewed to gauge the current position of IT security and the measures available applying to SIEM and other security management devices. Major threats identified particularly within the managed enterprise environment are also discussed as part of current security concerns. The use of geolocation to augment the identified problems in these areas is considered, and is regarded as the salient point of this chapter.

3.1.1 Mobility

Mobile devices, the cloud, social media and various other non-standard applications are increasing ease of data use. The provision of personal data transfer, constant data access, anywhere, any place are now seen as standard features in technology. The trend induces security implications at a large scale with wide-reaching data control and access impended to the average user. Security analysts require a contextual view of user data patterns and collective authentication methods. This is embedded through awareness of multiple factors applied centrally to a users behaviour. Thus, multi-level authentication requirements are a

result of the progressions emerging through technology shifts today.

Influentially, due to data mobility and the cloud, user identity and authentication have become chief focal points in security systems. The necessity of *situational awareness* propagated by the removal of physical restrictions in data realms, is a vital requirement. Situational awareness for security analysts can be seen as the ability to identify and comprehend all information/factors concerning the surroundings of the environments and user activity under observation. User authentication is supported significantly through this practice, the provision of context to an authentication activity gives administration better ability in discerning suspicious activity from normal traffic.

3.1.2 The Cloud

The cloud is a large contributor to technology change, the platform extends boundaries decreasing physical limitations and facilitating powerful resources, large capacity storage and other services. In the advent of cloud services achieving status as mainstream solutions in ICT, the issues concerning privacy and security increase[3]. The provision of Security-as-a-Service (SaaS) to enable the level of computation and resources for traffic analysis gives security an edge on much needed performance in high load capacities. As attacks grow in complexity and become more sophisticated, a prime outcome of security applied in the cloud is a form of "Global Threat Intelligence". A cloud-based threat intelligence tool will make use of threat information from users all over the world. The gathering of attack intelligence from a broad customer base, using internet traffic patterns from around the globe supports a stronger prevention approach from collective knowledge.

Many organizations however, are reluctant to embrace cloud technologies, at least in part, because of concerns about the security approaches to their data and applications. Perhaps foremost among those concerns is the risk resulting from digital assets residing on cloud servers in locations they may consider undesirable, such as countries with differing security and privacy laws. An organization may not be willing, or ready, to have its data subject to multiple sets of laws and to allow the location of its data to change without warning as dictated by the internal decisions of the cloud provider.

The general solution to this issue is the contractual agreement with cloud providers not to allow workloads to reside in certain geographic areas. Cloud geolocation techniques can facilitate this requirement through identification of approximate geographic locations of servers. This approach however is primitive and does not encompass security considerations, workloads can quite easily be shifted into undesirable servers through manipulation of the data on geolocation. The National Institute of Standards and Technology (NIST), developed a method for trusted geolocation in the cloud. The technique involves cryptographic hash generation representing geolocation information for a cloud server. The hash is stored within the server Basic Input/Output System (BIOS), securing it from alterations whilst allowing access by management processes. The server BIOS is considered a better source of trust than reliance on software-based information[54].

3.1.3 Regulatory and Compliance

Regulatory needs concern all national and international laws that require adherence from businesses. In addition to this, there are industry specific best practices that either specify or insist compliance with, in some cases this involves internal policies and requirements. The encompassing arena of law regulations range from privacy through the European Protection Directive, the Personal Information Protection Law in Japan to Financial Services Agency (Japan) banking rules, HIPAA rules for healthcare and international rules like that of the PCI in the payment card industry or other simple policies of enterprise[3].
Considerations with regards to data privacy and rights control has become a central focus for countries and conglomerations responsible for regulatory enforcement. Legislations such as the Data Protection Directive and Regulation Proposal provided by the European Union (EU) are major contributers, particularly in regard to the advent of privacy protection with mobility and the cloud.

Efforts in regulatory compliance for SIEM are considered in chapter 4 as a principle facet of privacy considerations, legislations such as EU legislative documents are discussed and their implications to a SIEM environment.

3.1.4 Emerging Threats

Despite the billions spent every year on IT security, over 80% of organisations *expect* to be breached every year[7]. Threats faced by networks and systems over the web are getting more sophisticated. Time decreases in exploits, strategic attacks targeting valuable resources and environments and Advanced Persistent Threat (APT)'s shaping the biggest problems of security organisations.
An APT is a network attack whereby a user intends to keep access to the network for a long period of time. Typically, a hacker attempts to get in and out as fast as possible without alerting an Intrusion Detection System (IDS) or administration but in this case the hacker wants to be in as long as possible - undetected[52]. The purpose is to steal information from target organisations, such hackers are paid for their services and are skilled adversaries.

3.2 Threats in Managed Enterprise Environments

Cyber attacks have continued to advance in frequency and intensity since last year with the focus shifting back to large organisations. In a 2014 survey by PricewaterhouseCoopers[1] the fraction of large organisations successfully hacked has risen up to almost a quarter this year i.e one in four large organisations partaking the survey reported penetration of their networks[7]. Clients in managed enterprise environments are typically in the large to really-large category. The attacks carried out within their networks have strong consequences, the requirement for timely detection and response being critical factors in such high data capacities.
As the scenario/data providers for this research, TSSA identified the following attacks as persistent problems within their environment:

[1] http://www.pwc.com

- **MC-5.5.1 Brute-Force Password Attack**
  The Brute force misuse case is a common method for password hacking in enterprise due to the many users being the likely weak point for infiltration. A Brute-force attack essentially follows its name in that it floods an account authentication entry with combinations till it discovers the username/password from sheer brute force guesswork applications.
  Due to factors such as end-user predictability and automated scripts success is quite often achieved and are usually the entry points to getting into a network to compromise an account with higher privileges and accesses.

- **MC-5.5.2 Unauthorized Login to a Computer System, Network or Application**
  Unauthorised login is when a malicious user compromises an account and uses it to change permissions, gain access to resources or information, amongst other activites of ill intent. This is the route taken by hackers carrying out APTs and is a prevalent issue in large organisations often keeping confidential information and other data of high business value.

- **MC-5.5.3 SQL Injection**
  This attack is commmonly used to extract information from a vulnerable website e.g, system configuration details, stored credit card data, passwords etc. Code injection essentially applies to any applicable code such as SQL, HTML, XML and OS command injection. Arguably, the most prevalent injection is the SQL injection. For SQL injections to work, the attacker has to jump out of the original SQL statement and append his own query[43].
  This is done with a meta character such as the ' inserted in fields where user data is meant to be received. The flaw depends on SQLs inability to distinguish between the control and data planes.

- **MC-5.5.4 Cross Site Scripting (XSS)**
  XSS is actually what can be considered as a subset of HTML injection. This attempt is usually the compromise of a system through malware installation or cookie hijacking, used to impersonate a user. An attacker redirects a users web browser to a site hosting malicious code and tricks him to execute it, which gives the attacker the ability to put malicious code into websites they do not own.

  Cross Site Scripting attacks work by embedding script tags in URLs/HTTP requests and enticing unsuspecting users to click on them, ensuring that the malicious javascript gets executed on the victim's machine. These attacks leverage the trust between the user and the server and the fact that there is no input/output validation on the server to reject javascript or other active code[43].

  A waterhole attack in XSS is finding a website that is trusted by users and injecting code into it, so the target user a hacker attempts to compromise clicks on the "trusted" page and becomes the entry point. One of the main problems with these attacks is that they can be easy to carry out using tools like the Browser Exploitation Framework[2]

[2]http://beefproject.com

- **MC-5.5.5 Worm Propagation**
  A self-propogating virus that spreads through a computer network, malicious code that users do not want running on their system. It manages to spread by replicating itself over the network, an example effect, through muliplication it uses up a computers resources and causes shutdown[17].

3.3 A proposal: Geolocation for Identification

The use of geotechnology has previously been limited to content delivery networks (CDNs) and for the purposes of target advertising[37]. In these instances it was required to determine the location of customers as accurately as possible to determine the best route for application requests to the nearest available data center, optimising user performance. This also applied in the case of advertising, using location to deliver context-aware advertisements for more effective results. With the increasing accuracy of geolocation through technology advancements more cases for applying the data have emerged.
While advertising and performance related considerations are still applied, location-based networking has encompassed a broader scope. The use of location-based access restrictions and context-aware security is a significant case of application, becoming more crucial with the increase of user mobility[37].

The following matrix is constructed (Table 3.1) indicating geolocation centric security procedures that can strengthen efforts in the four driving areas of security:

GeoLocation Security Method	Mobility	Cloud	Threats	Regulatory
Location-based Authentication				
Location-based Access Restrictions				
Location-based Policies				
Location-based Event Visualisation				
Rule detection based on geolocation				

Table 3.1: Geolocation Security Augmentation Matrix

Key:

	Applies
	Possible influence
	Not Applicable

The suggested geolocation security techniques are discussed and analysed in terms of its proposed effect in the areas of mobility, cloud service, threats, and compliance. For each technique that applies in the area of threat detection, a misuse table is listed indicating which of the misuse cases for the managed enterprise (in section 3.2) can be addressed by the solution.

3.3.1 Location-based Security Authentication

Using contextual information such as user geolocation as a way of corroborating a users identity is a key method of increasing the assurance level from an existing authentication procedure. With regard to user authentication through contextual data sourcing, Gartner predicts by 2016 more than 30% of enterprises will use 'contextual' analytics for remote access, rising significantly from 2% as of 2014[2].

Location-based Authentication	
Mobility	Using geolocation as a second-level authentication criteria as *step-up* or *progressive* authentication can facilitate stronger security for mobile users using the mobile factor as a positive security application. Alternatively a more granular approach towards *transaction* level verification could be considered. For example, if a users phone is out-of-bounds from a bank card the transaction can either be disallowed or further security authentication questions requested of the user. Mastercard embraced the use of such a geolocation strategy just this year, as an opt-in service users can carry out transactions when their mobile phones are switched on within a specific geolocation abroad[42]. Using mobile phones as One Time Password (OTP) tokens as an effort towards tokenless solutions of authentication mechanisms is a proabable direction in banking security[2]. Adaptive access control using mobility features and geolocation can gear a trust authentication approach. Focusing on what a user has, in addition to what a user is and knows aids collective trust. Various levels of security can be set and adjusted depending on the accredited items within possession of a user that can verify his/her identity[31].The user authentication level depends on the number of tokens they can provide when authenticating - this can apply to mobile-location-as-a-token.

<table>
<tr>
<td>Threats</td>
<td>
Using process mining tools described in subsection 2.2.3 business process rules can be identified taking user location into account as part of authentication process monitoring. An example of security requirements within a logon/logout authentication process:

A user logout must be preceded by a user login. (login count == 0)
A user should not login if they have not previously logged out. (login count <= 1)
A user logging in from IP address x with location z should not logout from location !(area range of z) within time T (where T is maximum time of physical possibility e.g 30 mins).
A user should not login if source of location z is not in accepted location for that user.
There cannot be two LogonIDs for a session.
Special privileges assigned to new logon cannot have preceding logon failure attempt event.

A second application considered in threat identification, is the use of location in procedures where physical access is a requirement. This refers to the example case of administrators working in a critical infrastructure environment such as a dam. The control systems are often automated to a certain degree and are known common hacker targets; this is also partly due to easy password setups by administrators, the nature arising from the need to log in easily when a system failure occurs in such critical environments. Since the administrator is expected to IN the control center of the dam building, the concept of in-region verification is feasible and can be significant in countering system compromise through password limitations.
An implementation to provide this method with geolocation can be facilitated through the use of a lightweight protocol called Echo. This protocol addresses in-region verification through sound and radio frequency signal propagation[53].

Applicable to:
<table>
<tr><th>MC-5.5.1</th><th>MC-5.5.2</th><th>MC-5.5.3</th><th>MC-5.5.4</th><th>MC-5.5.5</th></tr>
<tr><td>✓</td><td>✓</td><td></td><td></td><td></td></tr>
</table>
</td>
</tr>
</table>

Table 3.3: Location-based Security Authenticaton

3.3.2 Location-based Access Restrictions

Current location sources in some cases e.g IP addresses, are not a strong enough source for authentication. It is not suggested to authenticate a user solely based on their geolocation in such cases. Geolocation has a strong presence on the internet but it is not feasible to give it the power of aunthentication in all areas of application when location is not sourced from tamper-proof methods. However, *blocking* a location based on suspicious proxy use or activity can be seen as an acceptable method in attack prevention.

<table>
<tr><th colspan="2">Location-based Access Restrictions</th></tr>
<tr><td>Mobility</td><td>The alternative to contextual authentication is the denial or restriction of access based on location. This approach proves more feasible for fraud prevention purposes. The approach used by MasterCard can be seen as a combination of location-based authentication and access. If the user is outside (x,y) range he/she is restricted from performing transactions through the card.
The method of restriction applies to situations when avoiding known ‘bad’ locations is preferred to pinpointing the ‘right’ location. For example, restricting access from Russia, when the user is actively using their phone/devices from an entirely different country.</td></tr>
<tr><td>Threats</td><td>IP Reputation used in SIEMs like OSSIM to detect these attacks, can be supported using geolocation-based blocking. Traffic from certain places that exhibit obvious malicious behavior often such as SQL injection and XSS attacks, can be denied, e.g China locations with no client present within that region.
Often IP addresses are using proxies. The technical workaround for this is that proxies/CDN’s will add an X-Forwarded-For header that tells you the IP address of the actual user but this is not a sound solution for all contexts. Therefore, geolocation in these cases shouldnt be a criteria for giving access but that does not mean it isnt a valid criteria for blocking access.
Applicable to:
<table>
<tr><th>MC-5.5.1</th><th>MC-5.5.2</th><th>MC-5.5.3</th><th>MC-5.5.4</th><th>MC-5.5.5</th></tr>
<tr><td>✓</td><td></td><td>✓</td><td>✓</td><td></td></tr>
</table></td></tr>
<tr><td>Regulatory</td><td>Previously in crytographic concerns regarding SSL, technologies that enabled 128-bit key certificates were not allowed to be exported out of the United States. The restriction applied to all digital content including browsers. Restrictions such as these were obstreperous to enforce mainly due to the fact that geolocation technology was a primitive concept at the time and often inaccurate[37].
However, such restrictions can be applied to technologies specifically that in areas of security and cryptography aided through well-developed geolocation technology that can successfully enforce access restrictions</td></tr>
</table>

Table 3.4: Location-based Access Restrictions

3.3.3 Location-based Policies

There is a host of regulatory requirements needing consideration. Organisations in the past were required to prove their adherence to such rules, but this has changed to demonstrating continous compliance. The reduction of the complexity can be aided though the use of geolocation corroborated with security management frameworks, like SIEMs[3].

Location-based Policies	
Cloud	Cloud workload distribution with geolocation can help and ensure data is within the desired country to keep 'within' law and rights of specific country data rights. Another challenge that can addressed is that of cloud computing. Global load balancing configurations can be levered through the use of geolocation data considerations. This can aid public and private cloud implementations in essential purposes such as disaster recovery[37]. With cloud services reaching mainstream status, they are now targeted by organised crime or used themselves *to* carry out crime. In both cases public enforcement agencies will require access to data held on systems for forensics or to collect information on suspects. Such forensic data can often be located in foreign jurisdictions or unknown locations. Building cloud services with location-aware structures aids the enforcement agencies in tracking and jurisdiction consideration awareness for forensic evidence. In the same regard, law enforced access can cause security and privacy concerns for users, such as in the case of Microsoft Office 365 in 2011. Microsoft was unable to guarantee european customers that their data wouldn't be accessed by agencies under US jurisdictions. Similiarly the dutch government suggested 'US'-based cloud service suppliers be exlcuded from handling government or citizen data due to the risk of access by US authorities[60]. These risks can be mitigated providing assurance to users and government organisations through the enforcement of location-based policies and practices within the cloud and its security sectors.
Regulatory	When talking compliance with regards specifically to management of an organisation's requirements in log retention, SIEMs have a stronghold in best fulfilling that purpose. Geographic restrictions with data movement arising from such compliance requirements are salient matters of concern. Examination of the locations of log data to prevent compliance invalidation need to be performed. Creating security policies/rules to verify geolocations within security frameworks such as SIEMs supports the compliance for the entire organisational system, enforcing it in a collateral effort of organisation protection.

Table 3.5: Location-based Policies

3.3.4 Geographic Visualisation

A visualisation method puts the data information into the human visual system directly, through this users can abstract useful information from complicated data quickly[34].
The presentation of the collected data on infrastructure elements and environmental conditions as well as the integration into relevant information that is immediately required by decision makers can occur in a number of different views depending on the task at hand.
Geographical information systems can provide this contextualization as well as a foundation for integrating the varied types of information that must be aggregated and selectively displayed for decision makers.

<table>
<tr><th colspan="2">Geographic Visualisation</th></tr>
<tr><td>Cloud</td><td>Just as accurate geolocation data has valuable benefits in terms of security and performance of web applications and resources, it also provides greater business value and insight through enhanced visibility. Business value and insight come from discerning the clients location and from additional data provided by geolocation. For example, geolocation can be used in defined areas, such as those established by Designated Market Areas (DMAs) and Metropolitan Statistical Areas (MSAs), to derive deep demographic data that becomes part of the application request context and can be subsequently incorporated into analytical evaluation of visitor and customer web application interaction[37].
The provision of situational awareness for security analysts of networks and communications on the cloud is a large contributory factor to incident detection and system monitoring capabilities</td></tr>
<tr><td>Mobility & Threats</td><td>The provision of situational awareness for security analysts for mobile devices and other applications plays a role detection of suspicious user activity. Enhanced visualisation through geolocation allows analysts to evaluate user movement patterns against other user data that correlate with these patterns. Anomalies that arise from conflicting data can flag unauthorised users roaming the networks through compromised accounts.

Applicable to:
<table>
<tr><th>MC-5.5.1</th><th>MC-5.5.2</th><th>MC-5.5.3</th><th>MC-5.5.4</th><th>MC-5.5.5</th></tr>
<tr><td>✓</td><td>✓</td><td></td><td></td><td></td></tr>
</table>
</td></tr>
</table>

Table 3.6: Geographic Visualisation

3.3.5 Rule Detection based on Geolocation

The exploitation of user geographic location can be carried out through the definition of the conditions that characterise such data, in terms of limitations and behavioural patterns. Using the specifics of the data type to aid the identification of anomalous user activity in the geographic paradigm further supports the isolation of threat activity and narrowing down of false positives.

<table>
<tr><th colspan="2">Rule Detection based on Geolocation</th></tr>
<tr><td>Threats</td><td>The exploitation of geolocation as a verification mechanism is a viable route of investigation considering the data availability and significance in trends such as mobility and the cloud. Using client geolocation data, security rules can be written to trigger anomalous behaviours in geographic footprints. An example extract application summarised in pseudocode for rule detection to trigger fraud possibilities:

if (this.user.calculateLocation() != Given Location)
trigger alarm(Location Mismatch)
threatlevel ++
sendAuthenticationRequest(this.user, 3)

Depending on the type of organisation, geolocation can be used for many detection rules taking this data as a method of continous verification considering the convenience of location-enabled mobile phones on the rise.

Applicable to:
<table>
<tr><th>MC-5.5.1</th><th>MC-5.5.2</th><th>MC-5.5.3</th><th>MC-5.5.4</th><th>MC-5.5.5</th></tr>
<tr><td>✓</td><td>✓</td><td></td><td></td><td></td></tr>
</table>
</td></tr>
</table>

Table 3.7: Location-based Rules

3.4 Geographic Data Accuracy

When geolocation data is highly accurate, it can be employed across a broader set of functions that might depend on or be enhanced by having access to that information.
The key is to ensure the geolocation technology is, in fact, as accurate as possible. This often requires that a solution wishing to take advantage of geolocation capabilities must look to an outside source. The traditional methods of geolocation have depended upon public IP address registries, which are now highly suspect in regard to accuracy and thus cannot be depended upon to provide valid location information. Using a trusted third-party source for location determination enables solutions to apply location-based policies with a high degree of assurance that the data is accurate. This level of accuracy permits a broader set of uses for geolocation technology[37]. In security, the SIEM technologies would need to locate without any dependance on the client to provide this information, of course to rule out the chance of being provided with false information, and alternatively proven legitimately accurate through the system.

The most common and effective accuracy methods in determining a users location from IP address are:

1. **Constraint-Based Geolocation (CBG)**
   This involves measuring delays from active vantage points, places that we *know* the exact location of. The distance is calculated using the time delay difference from the place we know and the place we want to know, narrowing down into a radius of distance.

2. **Topology based geolocation**
   This involves CBG that considers network topological information in calculations.

3. **Octant (Cornell University)**
   This involves CBG, router locations, geographical and demographic information in estimations.

All of the above mentioned are effective accuracy determining methods that can be applied across the globe, however the problem here is regarding our specific application. The level of accuracy needs to be able to pinpoint a subject in some security rule applications within a distance radius of a few metres. In 2011, a method was discovered by Yong Wang[61], that builds on the above mentioned methods going further to street-level accuracy client-independant IP geolocation.

3.4.1 Wang's Accuracy Method

The method determined provides geographical data with a street-level accuracy utilising just the Internet Protocol (IP) address of the target computer. The procedure is as follows:

1. The initial step towards narrowing down the accuracy of a users location focuses on Point of Interest (POI)'s. Organisations typically host their websites on servers stored on the premises, the IP addresses of the servers are therefore tied to their physical location. Through the use of Google Maps, the team extracted web and physical addresses of such organisations accumulating data points of 76000 places as landmark points needed in the second step.

2. The target computer is pinged, the time taken to send a data packet to the target is used to calculate the distance. This is the CBG technique narrowing the possible location down to a radius of around 200 kilometres.

3. Within the radius determined from step 2, data packets are sent to landmark servers contained in that geographic range to find the routers through which they go through.

4. If a landmark machine and the target computer share the same router, comparisons can be made between the time taken for a data packet to reach each machine from the router, converting this to distance estimates narrows down the search area significantly further shrinking the area size.

5. Finally the landmark search is repeated at a granular level, determining the closest landmark server to the target.

On average the results fall within 690 metres of the target up to as close as 100 metres, sufficient enough for identification of the target computer at a physical location within a few blocks[61].

Limitations and Workarounds

Regardless of the geolocation accuracy method used, proper identification can be avoided by routing traffic through a proxy server which will provide a geographic location elsewhere. According to MaxMind[3], a open-source provider of geolocation information for IPs, the average statistics shown in Table 3.8 of data they have collected indicating the sources of bad reputation IP addresses are:

Bad IP Reputation	
7%	High Risk Countries
25%	Country Mismatch
39%	Proxies

Table 3.8: MaxMind bad IP Reputation source statistics

With proxies, at the highest of 39% we can see that location verification can be hindered most often through efforts such as suspicious proxies. Wang affirmed the inability to get around proxies through his method but can facilitate the detection of a proxy, preventing the return of false positives and highlighting addresses that are using proxies.

Therefore, identification of IP address locations can encompass checks for proxy, if there is indication of a proxy, the proxy address can be run against collected information of anonymous proxies through global intelligence data of bad IP reputation databases. Depending on the reports received from these sources, the IP can be blocked as the security measure. There may be something wrong with authenticating a person based on their geolocation but there isn't anything against *blocking* a suspected suspicious location(discussed in Table 3.4) or proxy.

3.5 Summary

The driving areas of security today are subcategorised into four main areas of concern - data mobility, the advent of cloud technology, compliance in legal and regulatory facets, and the emerging threats in the information age. Using these four corners as points of reference, the rising trends, concerns and methods surfacing to address the raised considerations are examined.

Mobility provides the challenge in security with respect to user identity and authentication, the flexibility of multiplatform use introduces various security implications and considerations. The requisition of situational awareness is encouraged to support security analyst decisions in minimal time spaces when evaluating the criticality of a situation.

Secondly, the cloud is a large contributing factor to technology change, extending physical and virtual boundaries. Many companies however are reluctant to embrace the concept due to the surrounding concerns of legal implications in data ownership. The suggested solutions are based around contractual agreements with cloud providers or trusted cloud based procedures

[3]http://www.maxmind.com, provider of geolocation estimates of IP addresses

using techniques such as cryptographic hashing.

Regulatory compliance is the third main area of concern, whereby many legalities are affected by the implications of the preceeding trends, mainly mobility and cloud technology. The chief issue rotating around the concerns of user privacy and data ownership.

The fourth area is the evaluation of the current attack trends affecting organisations, enterprises and networks today. Of these advanced persistent threats are highlighted as a critical issue in todays security. Threats within managed enterprises identified from the real enterprise in which this study is based, are examined. These misuse case identified main attacks that need to be addressed better in such environments, which are; brute-force attempts on passwords, unauthorised log in attempts, cross-site scripting techniques, SQL injection and worm propagation.

Geolocation for identification is proposed as a solution in an effort to mitigate these attacks. A matrix of security techniques are mapped to addressing the driving areas of concern in security discussed earlier. Each geographic-centric security approach is collated with applicable areas and misuse cases identified from the managed enterprise. The application of the solution to the applicable issues are given.

Lastly, an investigation into determining the accuracy of geolocation data, to aid useful application of this field to produce correct results is discussed. A summary of techniques are provided with the best solution determined through a combination of all methods, the limitations and workarounds of solution are also provided.

In conclusion, geolocation can be applied in many intuitive ways to enhance SIEM in security. A few of the main applications were considered for security enhancement and detection. The most important outcome for this chapter of research is location-based identification and authentication procedures for alarm triggering within a SIEM. Visualisation for data analysis was also discussed as a contributive factor with aiding security analysts in their filtering techniques.
While solutions for future cloud-based SIEMs using geolocation was briefly discussed here, it is introduced to support the research of geolocation in security, further investigation into the application in a cloud is considered future work and out of the current research scope.

Chapter 4

Privacy, SIEMs and Location

Sed quis custodiet custodes ipsos?
"But who will guard the guardians?"
– Juvenal (Satire VI, lines 347-8)

4.1 Our Data

The position of data privacy and its role in the information age is argued by renowned security expert Bruce Schneier as being an enforcer of a critical sense of self ownership. Users leave traces of their data in daily activities like ATMS, tolls, and surfing the internet, these traces are the *data shadow* of a user. An entity withholding this information can directly influence the life of the user, having removed the physical limitations such as the need to break into a home to find infringing information. One's data determines if a user can get a loan, get on a plane, get into a country. A stronger hold on user data and its basic rights is required in the overwhelming data capacity of the internet and its continuous expansion.

Discussions around privacy with arguments against the implied criticality of the responsibility is often given as a "I have nothing to hide" argument. The argument expresses a form of personal responsibility as the solution, the onus being on the user to be careful and have no incriminating activities. Schneier regards it as the most common response used against advocates for privacy. The general public most often classify the stressed priority given to data privacy as exaggerated concern; when it should be perceived as the second half of internet security, with the unified aim of protecting an individual from adversaries they can and cannot see. Users can be misused with, what privacy expert Paul Ohm coins our *database of ruin*[49], whereby sensitive information can be used to deframe or blackmail using the simple power of that information we dont want to be known. To get such information a physical break-in would have been necessary, but today we are addressing the simple urgency that those physical boundaries we depended on need to be modelled in the virtual world to every system that stores our data, and control who has the keys to it.

Unfortunately it is not a simple procedure, the process of defending data rights, in the virtual realm of the 'internet of things'. Techniques employed for the free flow of data on the internet, pre-processed with privacy-ensuring procedures, have been developed mainly to encourage the ongoing distribution of information while compensating for individual privacy considerations.

Among these procedures, the most common application is, or a variation of, data anonymisation. An analysis into the effectiveness of such procedures is needed to determine if they manage to fulfill the requirements of their implementational purpose.

4.1.1 Anonymisation

Anonymisation allows user data to retain some of it's information whilst screening the properties one can identify. It is used to provide statistics while making efforts to keeping user privacy intact. This concept has been well received by the various groups involved in pursuing the privacy of individuals and is applied in many fields at many levels of internet security. However, the success of this method when applied in industry requires further investigation.

One of the main issues that arise in industry, is determining what part of the user data falls under the definition of "personally identifiable" that needs to be protected through procedures such as anonymisation.
The concept of "personal" information is not a straightforward classification, almost all information can become "personal" when combined with other relevant data fields. The comparisons of collected seemingly generic information can make up the head and arms of a body of sensitive information. This issue was proven in the case of Latanya Sweeney[56], who took data released to the public and combined them to find the health records of the Governer of Massachusetts. The only information needed was the zipcode, birthdate and gender to identify the target person and marginalised the search scope through the identification of combination anomalies.
It is for this reason, the fields of data available need to be evaluated in relational terms that could aid re-identification from a culminated field effect.

With regards to the definition of personal information, geographic location is not considered unilaterally as a type of personal data. Systems that utilise geolocation in networks and devices often store this information in the clear. A 2013 study[14] by MIT and the Catholic University of Louvain demonstrated the pitfall of incomplete privacy measures, whereby people were identified from a collection of anonymised data through the use of *only* their geolocation data fields present in the data sets. The researchers used 15 months of user mobility geolocation data co-ordinates of test range of 1.5 million people on an area within a radius of 100km. The results revealed there is evidently a close correlation between peoples identity patterns and their movement patterns. The uniqueness and predictability of a user's geographic footprints makes it fairly straightforward to reverse the efforts of the anonymisation through this factor, resulting in a user's re-identification.

The application of a privacy-enforcing method on geolocation is necessary to avoid this re-identification that causes the failure of anonymisation entirely on the remaining fields. A further investigation into anonymisation is required to determine a method of anonymising the field of geolocation to prevent re-identification or otherwise determine if an alternate approach is required.

As in the case of the study by MIT, the effectiveness of anonymisation raised many questions in privacy debates and was arguably claimed as not sound. In 2009, Ohm evaluated the workings of anonymisation to determine the concerns on the solution in ensuring complete privacy.

Ohm confirmed, re-identification from anonymised data was possible due to the paradoxical situation arising of data utility over data privacy. It is made apparent that either utility or privacy can be provisioned for but not both[62], anonymisation can only be successful if it is completely curtaining data details, but this renders it unproductive.

As advocators of privacy, the debate is analysed in the context of utilising geolocation in SIEMs. First, SIEMs are essentially meta-systems, collecting data about data. The NSA court order discovered to freely accessing user metadata, was justified through the argument that only meta-data was collected and not the actual data[45], however, mathematician Susan Landau explained metadata is much more intrusive than the actual personal content. One does not need the actual data most often it is the patterns - the 'organised' version describing this data that provide all the necessary information.
If in the future, SIEMs are to be migrated into the cloud as a security-as-a-service, one of the biggest areas of concern for this transition are the frameworks adaptability to legal considerations in terms of privacy and various legal jurisdictions.

With regards to the SIEM in the existing scenario environment, the application of techniques such as pseudonymisation have been applied to the data, with further data enforcements such as data encryption and limited user access for different levels of personnel handling this data. Geolocation is often present in the masses of information collected through devices from the enterprise network and sent to a central management system for advanced analysis. The geolocation used in these SIEMs is not obfuscated in any manner and is in fact processed and viewed in the clear. SIEM frameworks need to introduce efforts in privacy of the raw data collected by the system, such as geolocation, regardless of whether the data is exploited for analysis or other procedures.

This discussion unfolds two main considerations of the research objectives in privacy; to determine a suitable method of ensuring privacy enforcement on geolocation, and an approach to mitigate the failures of anonymisation in SIEM frameworks as a whole that hold copious amounts of information from every section of a monitored organisation.
The concern towards the privacy of SIEM frameworks as a whole is addressed in section 4.2 followed by the approach of geolocation privacy in section 4.3.

4.2 A SIEM Privacy Model

As it stands, for SIEMs and other data systems alike, techniques like anonymisation provide proven limitations in their protection of user data. To enforce data privacy rights on these systems the need for regulatory compliance is required. The law can aid the global adherence to user protection as well as the method of approach in cross-border data transmission considerations. Where the boundaries of technology are reached in ensuring privacy, regulations need to play a role.

The managed enterprise scenario considered in this research is based on a managed IT outsource environment, where events from multiple sources are collected centrally. The collection of security events by SIEMs is becoming more widespread and can be identified under the

growing notion of a type of big data in the security field[1]. The challenge with shared SIEM services occurs where monitoring services may be provided by a third party organisation or in a different country. This requires levels of event information sharing, consolidation and aggregation.

There are many regulatory documents available to ensure various levels of privacy in systems, though none as such directed to SIEMs themselves as a whole but rather components. Therefore, in this section a privacy guideline is constructed based on existing approved benchmarks of privacy enforcing legislations. Addressing the concepts of data protection and privacy within this environment, relevant legislative documents are identified and summarised to rules that apply to SIEMs.

4.2.1 Legal Documents: The European Union

The European Commission plays a major role in the privacy battlefield, with countries such as Germany advocating the need for user protection in international data transfers and within the country itself through facilities such as an*opt-in* as opposed to the *opt-out* approach. The EU 2012 Privacy Regulation Proposal[9] COM 2012/0011, Personal Data transfer Sect 3.4.5 (V), Article 45, stipulates the conditions for information transfer to regulate the privacy rights in the movement of data.

The relevant documents introducing data protection principles to be applied to organistions, published by the Commission are;

- EU Data Privacy Directive 2009
- The proposed EU *Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data* Data Protection Directive 2012[15]
- Communication on "Safeguarding Privacy in a Connected World - A European Data Protection Framework for the 21st Century" [10]
- EU 2012 Privacy Regulation Proposal[9]

The privacy requirements outlined in these documents need to be applied in the context of a SIEM framework. The process of ensuring privacy can be enforced throughout the framework by following a legislated standardisation. The implementation of the EU Directives varies and can be interpreted in alternate ways, the existing Data Protection Directive is implemented by member states in very different manners[30]. However, the same cannot be said of the legislative EU Regulation documents. This differentiation occurring from the aim of the respective documents, the purpose of an EU Regulation is to encourage a harmonisation among member states, needing the application to thus be followed precisely[30].

The proposed EU Directive[15] is currently in the European Parliament and is yet to be agreed on for adoption by the Parliament and Council. The guidelines set forth, are based on this document as a commencement of SIEM technologies in applying legislative structures

based on the most recent suggested privacy documentations. The EU legislative documents are widely considered the strongest approaches of enforcing privacy than any other legal documents put forth in the areas of privacy preservation. A summarised overview is provided in Figure 4.1, indicating the general areas stipulated in the proposed EU Directive [15].

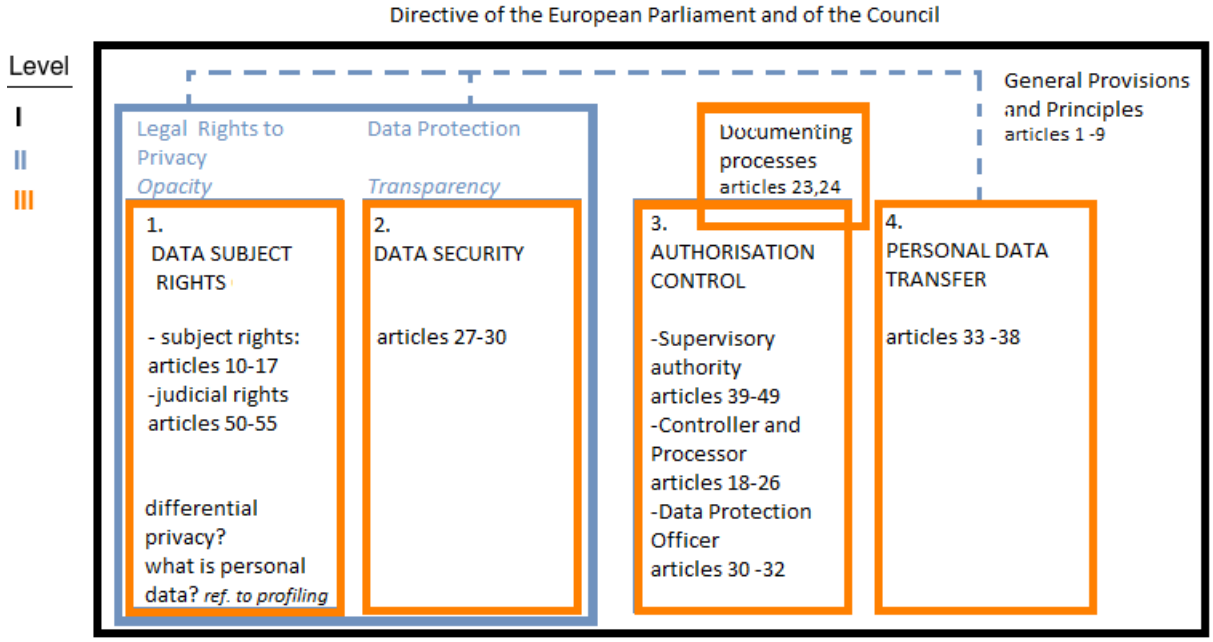

Figure 4.1: Categorical Overview of the EU Directive[30]

The sections highlighted in Level 3; Data Subject Rights, Data Security, Authorisation Control and finally, Personal Data Transfer are discussed in their implications to SIEM management approaches and their responsibilities as handlers of sensitive data.

Data Subject Rights

The foremost specification requirement of this section, is the explicit definition of what information flowing with the SIEM is classified as "personal" data. As discussed earlier, failure to correctly classify all data types in their sensitivity can lead to serious breaches of user privacy. The SIEM framework needs to consider the following[30]:

- The classification of personal data, applying to all data types and formats within the framework.
- The access and usage rights of the classified data. For example, IP addresses, RFID tags, cookies storing user information.
- The examination of data held within the framework and the levels of privacy assured to them depending on their classification.

- The need to store user identity information in cases where the data is simply mined for statistical evaluation (SIEM security reports) need to be evaluated.

The implementation of these requirements can be carried out through the construction of access rules which allow the use of information to be governed under various privacy levels. Differential privacy can be implemented in the DBMS layers, or through the use of cryptographic keys at the various management levels[30].

Data Security

The articles specified under this section require the adherence in terms of automated data processing. Specifications regarding the various areas of concern are listed in the following areas:

(a) The control of access to all equipment

(b) The control of data media handling

(c) The control of storage access and changes

(d) The control of users utlising the SIEM framework. In this case, these are the security analysts and administrators.

(e) The control of access to data

(f) The control of communications within the system

(g) The control of input feeds received by the system

(h) The control of transport of data, communication, and storage.

(i) A recovery control process.

(j) Measurable reliability and integrity

The SIEM framework is required to assess if all the above principles are enforced in a log system that records activity across all areas. This ensures the internal circulation of data maintains a privacy trail in all communications.

Authorisation Control

The articles categorised in the area of authorisation control consist of specifications regarding the person hierarchy to be implemented to oversee t he control systems specified within the *Data Security* specifications stated previously.
The control authorisation to be monitored should consist of the following members[30]:

- **Controller**, a Controller is defined in Article 3, with the related responsibilities stated in articles 18-24, with regards to rights and needs for process documentation.

- **Processor**, the definition of Processor is also defined in Article 3, and applicable requiremens in articles 18-24.

- **Supervisory Bodies**, A board of supervisors assigned for the SIEM framework need to be assigned; these members oversee the data Processors ensuring they don't abuse their permissions and rights on data processing[18]. They also ensure the rights of data subject are heard and acccounted for. There may be more than one supervisory body assigned to different areas of framework (for example, incident detection supervisory and system analyst supervisory), if so the mutual assistance must be facilitated across all bodies. There can be more than one supervisory bodies to ensure collaboration between boards and if so they must ensure mutual assistance amongst themselves.
- **Data Protection Officer (DPO)**, described with regulation requirements stated in articles 30 -32 needs to be assigned to ensure data protection and privacy with the SIEM framework[30].

The documenting of all processes and communications between these members need to be documented and recorded.

Personal Data Transfer

The definitions for personal data transfer concern the movement of data, within or external to the current environment. The following considerations are necessary with regards to the consistent application of privacy[30]:

- The Data Protection Officer assigned is particularly responsible to ensure the consistent application of data privacy, a global DPO can be assigned to check for the privacy is ensured through the transfer, receiving reports from all relevant officers in the transfer.
- The biggest concern is the harmonisation across different jurisdictions in terms of data privacy and handling. The approach of using a global DPO and where possible only providing the data that is necessary(required by the receiving end), can aid this consideration.

The most important concern stressed in the articles of this section is the harmonisation of data privacy internally and specifically with cross-border situations; a common occurence when overseeing very large enterprises, such as those typically monitored by a SIEM framework.

In summary, the explicit implications towards SIEM environments extracted from the data directive, are critical components in ensuring a privacy infrastructure support for an entire monitored enterprise.
The implementation of data protection and privacy is needed for Security Information and Event Management (SIEM) Frameworks, we need to make explicit the associated requirements for an SIEM framework. In particular, roles need to be clearly defined for an SIEM system (for example processor, controller etc.) and the SIEM itself needs to be treated as a processing system of an organisation. This means that the data residing in the system needs to be made explicit, with the retention and storing or processing purposes made known and documented. A data protection officer should be assigned, with responsibility for the SIEM system in the same manner as other systems.

Given the potential sensitivity of data collected by a SIEM this needs to be done very carefully. Techniques of full anonymisation or aggregation can also be enforced where applicable. Rules for processing jurisdiction could either be agreed contractually with SIEM providers, or there may be ways to embed such meta-information into the processing rules of cloud or other service providers so that Service Level Agreements can be implemented to guide and control how security processing is conducted.
Further requirements specified by the Regulation include the establishment of a European Data Board, enforcing the explicit consent of a data subject, possibly by existing methods such as 'opt-in', enforcing consistency mechanisms, data protection certification, codes of conduct for the SIEM workforce, time limits for the processing of data, further authorised member documenting, identifying types of personal data and treating them with different levels of privacy and finally stricter enforcements of data subject rights.

4.3 Defending the Defender

Internet monitoring often results in employers versus employees, where both parties are fighting to protect their personal interests. Employees desire the conservation of their privacy while employers are interested in protecting their company from misuse. SIEM log data collects a range of information, with many streams of traffic containing traceable user footprints. Therefore, server logs, email activity and other data requires a monitoring approach governed under defined enterprise policies to avoid indiscriminate employee monitoring. A second legal concern in SIEMs is the capturing of data but failure to exploit it. There may be political or legal ramifications to having data and failing to take action on it. Though it is known that not every piece of information cannot be thoroughly analysed there must be enough justification present to withstand the questions on inability to act. In some cases concerning legal consequence, to not have the data is a better option than to have it and be unable to act upon it[3].

Using legislation to shape standards at which one must aim towards for securing user data interests is a necessity, being fully aware that in large enterprises cross-border data transfer is also not uncommon. With this in mind, the application of adherence to privacy through legislations is introduced. In addition to this, the second privacy consideration addressed is that specifically of geolocation. This data type is often left in the 'clear' in source logs and not catered for through encryption or any other privacy technique employed in system processes. The failure to commit this data under privacy procedures abrogates any levels of anonymisation carried out on the related fields of data containing it. The possible methods that can be applied on this data are further discussed, for application in this research.

4.3.1 Geolocation Anonymisation

The consideration of privacy procedures to be applied to geolocation information differs from the application to standard data field types such as *username*, *date of birth*. Anonymisation for example, when applied in geolocation is not re-identifiable through collective fields, such as the case discussed with Sweeney and the governer[56]. This is due to the nature of the data type itself, geolocation is an identity 'at a point in time' and is most often given as a range due to the accuracy radius. Anonymisation can be correctly applied to geolocation in efforts to ensuring privacy, based on this dynamic nature of the data type. The approach used to facilitate this is defined as *spatial cloaking*. The general strategy of this approach

is to increase the range of geolocation data. Rather than an exact data point, it becomes a data point estimate within a radius or cloaking region. Identified methods applying varied approaches to spatial cloaking are summarised with their advantages and costs.

- Two-tier spatial approaches involve only the user and the LBS provider[19]. Privacy/-cloaking can be created through the use of redundant queries. A user u can generate r random queries to the LBS, in addition the original query, thus *hiding* through the creation of decoy locations. These transformation methods tend to incur low overhead but provide privacy through a limited set of assumptions. This results in provision of privacy only against attackers with little background knowledge.
- Cryptographic private information retrieval technology (PIR) allows a user to privately retrieve information (through the use of an encrypted request) from a database , without the database learning what information was requested by the user. It relies on the fact that it is not computationally possible for an attacker to discover the value of i given *q(i)* provides a high level of privacy that can even hold against skilled attackers, but this comes with a significant overhead cost especially for large data sets. In terms of SIEM applicability, the method can be good but it is not feasible if it affects performance.
- Hybrid approaches utilise a combination of geographic and cryptographic combinations, allowing the advantages of both approaches to be exploited. The strategy seems to be the best direction to follow, because it provides strong cryptographic guarantees as long as the user moves within a certain range of the dataspace. With a properly chosen range, it can be ensured that no information about an individuals whereabouts is released that can be used to infer sensitive information, and at the same time the overhead of expensive cryptographic operations can be held in check[19].
- As a third dimension of the required properties of query protection methods, hybrid approaches also bring the benefit of protecting excessive disclosure of data from the provider, which not only protects the privacy of the users, but also the commercial interest of the service providers[19].

A preferred Generalization

The anonymization process is based on the concept of k-anonymity. K-anonymity protection means that the information for each person in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release[56]. Figure 5.2 shows the cloaking region (CR) achieved when k = 4. The requirement is thus achieved by generalizing, and possibly suppressing, information upon release[56] through ambiguity.

Three-tier spatial transformations implement the spatial k-anonymity paradigm. A contains *k - 1* users in addition to the query source (a *k-CR*) is generated, and the location-based service(LBS) processes the query with respect to the CR. Since all the k locations enclosed by the CR correspond to actual users (as opposed to“fake locations in the previous category), the probability to identify the query source is at most 1/k, even if the attacker has knowledge about exact user locations.

A trusted centralized anonymizer acts as an intermediate tier between the users and the LBS. All users subscribe to the anonymizer and continuously report their location while they move. Each user sends his query to the anonymizer, which constructs the appropriate CR and contacts the LBS. The LBS computes the answer based on the CR, instead of the exact user location; thus, the response of the LBS is a superset of the answer. Finally, the anonymizer filters the result from the LBS and returns the exact answer to the user[19].

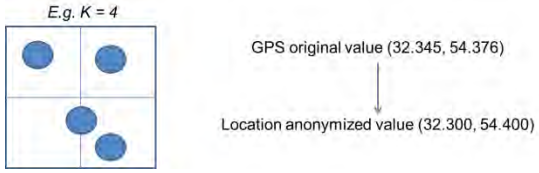

Figure 4.2: Anonymisation technique

The anonymity requirement is expressed by specifying a quasi-identifier and a minimum number k of duplicates of each released tuple with respect to the attributes of the quasi-identifier.

4.4 Summary

This chapter concerns the implementation of privacy within SIEMs. An introduction to the fundamental concerns in ensuring privacy and the position of this concept in the current information age is provided. The need with respect to data privacy is highlighted, techniques such as anonymisation employed to instill these attributes is discussed. The failure of anonymisation is explained, identifying the flaws of the technique when it is incorrectly applied.

A privacy model based on EU legislation providing a legal guideline by which SIEM technologies need to adhere, is given. The necessity of technologies like SIEM containing meta-data of masses of information to comply with privacy, can be facitated through this guideline.

The position of geolocaton in privacy is assessed, and the implications towards privacy through its inclusion. Anonymisation is discussed in its application of geolocation, proving to be a useful privacy technique for this data type, albeit not a good approach for other forms of personal data.

Finally, the various spatial cloaking approaches are discussed, including the method of generalisation to be used in the prototype implementation.
In conclusion, the SIEM solution we aim towards, is to provide anonymisation techniques which can ensure privacy as far as technology boundaries can facilitate. This can be encompassed in regulations that provide an optimal privacy solution using both law and technology. Anonymisation will be applied to geolocation, the characteristic of geolocation to be able to adjust in ranges, facilitates its ability to be sufficiently useful even after procedures of anonymisation. Thus, in fact advancing our SIEMs in privacy and utility through geolocation.

Chapter 5

Design and Architecture

This chapter concerns the method and approach used to validate the research discussions surrounding geolocation in SIEM technology and the concerns of data exploitation within boundaries that encapsulate privacy concerns. The opening section discusses the experimental objectives in terms of the aims that need to be demonstrated to support this study. These objectives are then applied to the available tools, addressing tool feasibility and the overall approach. The conceptual testbed accommodates the tools and framework justified with individually determined purposes combining them into a process flow.
Finally, the integration and application strategies are determined towards committing the determined justifications through the collective processes. The experimental design will lead into the next chapter focusing on the complete implementation of a simulated testbed.

5.0.1 Misuse Cases of Managed Enterprise

The misuse cases (discussed in section 3.2) in the managed enterprise environment concerning common issues faced in such a scenario have been examined as current security challenges of enterprise. Each identified case and its applicability to the research investigations are considered in Table 5.1 to Table 5.4.

MC-5.5.1: Brute-Force Attack	
Description	Multiple attempts to log into a user account with all possible combinations till the correct combination is guessed, from sheer brute-force password testing.
Geolocation Applicability	The use of geolocation in such an attack can narrow down large analysis sets to physical areas of concern. This applies to the case of physical buildings that allow administrative use only *within* the building, making an outside location an immediate trigger factor.
Possible Solutions	For this attack location-based identification or authentication confirmation is considered a viable solution in efforts to re-affirm a genuine user or quickly identify a malicious one.

Table 5.1: MC-5.5.1 Brute-force solution using location-based authentication

MC-5.5.2: Unauthorised Login	
Description	A malicious user gains access and privileges to a system through the compromise of an authentication procedure of an account belonging to a valid user.
Geolocation Applicability	A scenario pattern to consider, log in attempts from different locations, if a user has been verified to be logged in from a certain location and an attempted successful login is made from a different location whilst still logged in at location 'workstation', this identifies the anomaly that more than one person is using the account and alerts the admin of the double usage hinting to possible account compromise.
Possible Solutions	Using geolocation as the second factor in a 'two-factor' authentication approach, should a user successfully log in but their source location not an expected location, a flag is raised for additional security checks to be made of the user e.g answering security questions.

Table 5.2: MC-5.5.2 Unauthorised login solutions using geo-fencing

MC-5.5.3: SQL Injection & **MC-5.5.4:** Cross Site Scripting (XSS)	
Description	SQL injection is commmonly used to extract information from a vulnerable website by getting SQL queries to execute through it. XSS is the case of an attacker attempting to get a valid user to execute malicious code giving the attacker rights into a website she does not own
Geolocation Applicability	IP reputation is used to circumvent attempts of these kind of attacks where the intent is obvious and cannot be done in err, geolocation can enforce IP reputation by facilitating geolocation-based blocking in certain areas where attacks are habitually originating from, in a sense a form of *geo-reputation*.
Possible Solutions	Geo-reputation used as a consideration for user authentication procedures, as well as flagging suspect users that have high chances of performing malicious activities in the system network.

Table 5.3: MC-5.5.3/4: SQL injection and cross site scripting solutions using location restrictions

MC-5.5.5: Worm Propogation	
Description	A malicious user creates the spread of a self-propogating virus that infect computers in the network, such code can eat up resources and many other damages.
Geolocation Applicability	Worm propogation gives messy results. We can encourage their circumvention in a broader security technique that restricts the possibilities of receiving such traffic in our network. When connected to the internet, an enterprise is exposed to incoming connections from all over the world. Enforcing policies that allow incoming connections to employee networks or customer portals depending on geographic location is a start in that direction. This can greatly reduce the exposure of an enterprise to dangerous zones that are key in producing such attacks, graciously reducing overall unwanted traffic in the process.
Possible Solutions	Location-based policies can be implemented to flag traffic from areas that repute illicit behaviour e.g areas of China consistently demonstrating malicious intentions.

Table 5.4: MC-5.5.5: Worm propogation solution using location-based policies

To simulate the attack formation within a SIEM context, the source environment and data collected for the test experiments need to be examined. The events can emulate one or more of the misuse cases depending on the source it retrieves logs from, for example - Windows/Linux servers, intrusion detection systems or anti-virus solutions.

5.1 Technical Objectives

In chapter 3 regarding SIEM security, the current challenges faced in the field of IT security were discussed. Within enterprise security specifically, five main cyber attacks of a managed enterprise were highlighted and examined. These attacks (addressed as misuse cases) in an enterprise scenario, are fundamental issues that need to be addressed. Security measures have been put in place to counter these problems, but these attacks still remain a legitimate concern for many enterprises today. Possible methods of augmenting the existing security threat detection methods with geographic location data was considered, from aiding initial event analysis to stronger applications such as adding a second layer of user authentication.

In summary, there were many effective applications deliberated of geolocation, towards identification and authentication using smart security rules and threshold techniques. Additionally, enhanced visualisation for data analysis with geographic perspective was highlighted in helping security analysts for filtering techniques.

Privacy investigations of SIEMs and geolocation in chapter 4 determined the need for data manipulation procedures such as pseudonymisation and anonymisation, for the protection of user data. The SIEM solution must provide such techniques for location data to ensure user

privacy irrespective of whether the data is exploited or not. The use of anonymisation as a privacy-enforcing technique for geolocation is justified in subsection 4.3.1 not falling under the realities of 'anonymisation is a failure' concerns. This encourages the concept of 'defending the defender' which concerns the best practices SIEMs follow as meta-systems. These data managers hold copious amounts of user data that is often cross-border falling over different regulations of countries. Regulation compliance is a significant concern for international considerations. Not all sensitive data can guarantee user privacy through undergoing technical techniques like anonymisation, as discussed earlier, though we saw it suffices in the case of data like geolocation. Encompassing this functionality with our guideline privacy model for SIEMs will provide an optimal privacy solution - using law to step up and enforce where technology approaches are limited.

The deliberations of the preceeding chapters brings forward important security research considerations. We epitomize the perspectives into research validation goals. These are summarised as:

1. Augmenting existing security methods in SIEM technology *using geolocation data.*
2. Provision of anonymisation in SIEMs for geolocation data, as this data is considered private information. Anonymisation has been justified as the privacy approach for this kind of data.

The misuse case attacks identified for the managed enterprise are used as the platform for demonstrating the objectives stipulated above. The attack detection approaches used in enterprise are a constructive point of focus to develop the research validations around. We are to develop an attack scenario for location security measures to demonstrate abilities in augmentation as well as showing feasibility of its integration. This simulation must be carried out within the umbrella of privacy-enforced procedures to support the aims of providing cross-border friendly mining of user security data.

5.2 Source Data Environment

To replicate the misuse cases specific for this environment further investigations into the source data system architecture and data sources is required. The security system in the scenario environment is based on the commercial SIEM solution, IBM's Tivoli Security Operations Manager (TSOM) which T-Systems SA provides for clients. The client security management environment based on TSOM consists of three main components, the event aggregation module, the central management system and the database.

1. **Event Aggregation Module (EAM)**
   This module is the entry point for event sources. It gathers data from network devices using conduits, which it then normalizes, filters and batches into incoming data streams. It then transmits the data to Central Management Systems.
2. **Central Management System (CMS)**
   This system performs various event process activities. It receives data streams from multiple EAM servers, correlates the data and categorizes the events. It performs calculations on the threat the event poses to the destination. Finally, it applies rules to respond to specific attack signatures.

3. **Database**
   An Oracle Database to store persistent data information.

A profile carried out on the TSOM solution in the source environment identified the basic architectural setup of the source environment, illustrated in Figure 5.1.

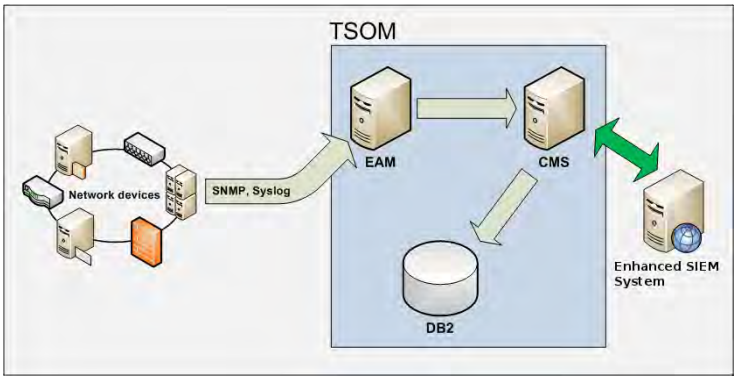

Figure 5.1: MESI event environment

Within the scenario the solution carries four EAMs, one CMS and database. Each EAM server uses specific conduits to gather data streams from various devices, with the following devices present in the profiled environment:

Sensors	Network Device	Conduit
3	McAfee ePolicy Orchestrator	SNMP v2
2	Tipping Point Security Management System	Syslog
120	AIX Servers	Syslog
47	Window Server 2003 & 2008	SNMPv2 & UCM
117	Solaris 9 Server	Syslog
108	Cisco devices	Syslog

Table 5.5: Data Sources within Managed Enterprise Environment

The effectiveness of the SIEM depends on the accuracy of the analysis on these events sourced from various inputs. As shown in Figure 5.1 source flow events can be collected from these devices through the use of the logging standard Syslog or the networking managing protocol SNMP.

SNMP is a transport protocol that manages networks and devices. It works on the application layer of the Transmission Control Protocol (TCP)/IP model and is used by TSOM itself to capture events from various sensors. It consists of two components, an SNMP agent/trap and an SNMP Manager.
The SNMP agents are essentially event notifiers, which are encapsulated in the User Datagram Protocol (UDP) header of an IP packet. It agent contains an Protocol Data Unit (PDU) which holds the contents of the SNMP trap in a element called a *variable binding*.

The variable-findings field is encapuslated in managed objects such as the $EVENT.INFO token which contains the raw data collected by the sensor and any other information attached to the event[4].

Syslog is a client/server protocol transmitting text messages to the syslog receiver commonly known as syslog daemon/server called *syslogd*.

Two devices/conduit sources channeled from the CMS are considered for the research testbed:

- **McAfee ePolicy Orchestrator**
  Data feeds from McAfee anti-virus and anti-malware are sent to McAfee ePolicy Orchestrator, which feeds into TSOM.
- **Window Server 2003**
  Responsible for monitoring the network for any suspicious activities, attacks and other events which may be of interest. Sensors used by TSOM can communicate using conduits such as SNMP, Syslog or XML.

	TSOM Data Source	**Sensors**	**Events**	**Total**
Windows Server	Windows Log Parser	7	8,612,030	16,745,426
	Windows Event Log	40	8,133,396	
McAfee Anti Virus	McAfee ePO 4.x	3	132,322	132,322

Table 5.6: Event Statistics

Events captured from the CMS sources covered a time period of 79 days, the total dataset contained 38,509,400 events from 322 individual sensors. For the experimental purposes of this research a subset of 16,877,748 events from 50 sensors was extracted. This consisted of events from Windows Server(s) and McAfee Anti Virus sources.

5.2.1 Event Schema

The events received are not in their source format but operationally modified for two reasons; they are the output of event data from a live enterprise environment and therefore have undergone obsfucation procedures performed on user names and other personally identifiable data. The environment contains an existing SIEM (TSOM) which has performed its own analysis procedures on the data.
The purpose is not to obstruct, repeat or undo the procedures performed on these data sets, but to extrapolate more security data from patterns in these sets. The following procedures were carried out on the collected datasets before use:

1. The current schema has undergone pseudonymisation of certain fields, as this test data is captured from a live environment and needs to privitise information.
2. In addition, to this it has already have undergone procedures of event normalisation by the environment SIEM.

Processed Format

The normalisation technique applied on the data needs to be reviewed to understand the context of the information that is to be sent to the test environment created using OSSIM and selected tools of MASSIF. The process undergone on the data events is summarised in the diagrammatic flow shown in Figure 5.2.

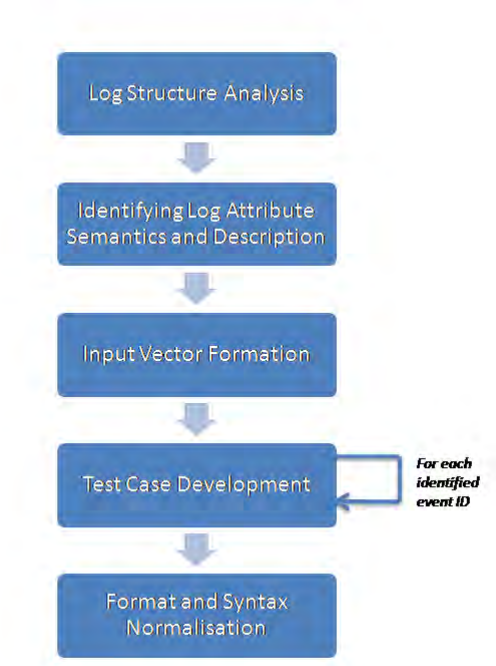

Figure 5.2: Anonymisation technique

The most significant event attributes were identified and combined in the definition of a well-structured table referred to as the input vector. This input vector describes how the content of the attributes should be mapped. For example, event attributes with similar content but varied names are catered for accordingly, thus event attributes such as "account name" and "user name" both contain user names and are therefore mapped to one specific input vector attribute. Test cases were developed for each event ID of the identified Windows security logs. A test case specifies how the source event should be parsed and transformed according to the input vector.

Each test case consists of the source event (SNMP trap) and the corresponding input vector (Comma Separated Value (CSV) file). The test case identified various Windows event formats which differed from Microsoft's the documented format. After identifying test cases, normalisation was performed on certain input vector attributes. Multiple Windows event attributes were mapped to one input vector attribute (e.g. 'Username' provided by TSOM and 'Username2' from a native event), the syntax and format had to be normalised. Different letter cases, IP address formats, domain name formats and user name formats were identified. As an example, a user name could appear as "**user@domain.com**", "**domain.com/user**",

"**user@domain com**". This was thus normalised to "user". The same principle applies to other Windows event attributes[4].

The input vector for Windows events (shown in Table 5.7) consists of two parts, TSOM-specific and Windows-specific. Most of the TSOM-specific attributes are provided by TSOM itself (e.g. SensorName, SensorType). Most of the Windows-specific attributes were exported from the native Windows event which is contained in $EVENT.INFO (Event information, often contains the native event as reported by the sensor) token.

Field	Field name	Pseudonymised	Comment
TSOM Specific			
1	<InternalSequenceNumber>	-	The internal sequence number used by ADEWaS
2	<TsomID>	-	TSOM-internal event ID
3	<TrapRequestID>	-	The SNMP request ID of the SNMP header
4	<InternalTimestamp>	-	The time the event was received by ADEWaS
5	<TsomTimestamp>	-	The time the event was received by the EAM. The time format is Unix time (the last three digits are milliseconds)
6	<SensorTimestamp>	-	The time the event was recorded by the sensor (e.g. a Windows server). The time format is Unix time (the last three digits are milliseconds)
7	<SensorName>	Yes	The name of the sensor which originated the event
8	<SensorNameDescription>	-	Describes if <SensorName>is a domain controller, member server or workstation
9	<SensorType>	-	The type of sensor which originated the event
10	<EventType>	-	Event type as defined, if possible, by the device generating the event. The type is otherwise derived from the reported event
11	<EventValidility>	-	A validity rating of the reported event by TSOM on a scale from 1 to 100
12	<EventClass>	-	Identifies the event class assigned to the event as defined by TSOM
13	<SecurityDomain>	Yes	The security domain to which the event was assigned
14	<UserName>	Yes	User name associated with the event
15	<Domain>	Yes	Operating system domain or internet realm if provided by the device reporting the event
16	<SourceIP>	Yes	The IP Address of the source host
17	<SourceIPPrivate>	-	Binary indicator which specifies if <SourceIP>is a private IP address, as described in RFC1918
18	<SourceIPGeoCC>	-	Two letter country code of <SourceIP>or string "1918" if it is a private IP address
19	<SourceIPGeoASN>	-	ASN of <SourceIP>or empty string if it is a private IP address.
20	<SourceIPGeoLat>	-	Latitude of <SourceIP>or empty string if it is a private IP address
21	<SourceIPGeoLong>	-	Longitude of <SourceIP>or empty string if it is a private IP address
22	<SourcePort>	-	The port of origin of the event from the source host
23	<DestinationIP>	Yes	The IP Address of the destination host

Field	Field name	Pseudonymised	Comment
24	<DestinationIPPrivate >	-	Binary indicator which specifies if <DestinationIP>is a private IP address, as described in RFC1918
25	<DestinationIPGeoCC>	-	Two letter country code of <DestinationIP>or string "1918" if it is a private IP address
26	<DestinationIPGeoASN>	-	ASN of <DestinationIP>or empty string if it is a private IP address
27	<DestinationIPGeoLat>	-	Latitude of <DestinationIP>or empty string if it is a private IP address
28	<DestinationIPGeoLong>	-	Longitude of <DestinationIP>or empty string if it is a private IP address
29	<DestinationPort>	-	Port on the host at which the event was directed
30	<SourceThreat>	-	The threat level assigned to the source host
	Event Specific		
31	<DestinationThreat>	-	The threat level assigned to the destination host
32	<InternalSubScenario>	-	Internal identification string of the sub scenario.
33	<InternalUseCase>	-	Internal identification string of the use case.
34	<EventID>	-	The unique ID of an event.
35	<TypeOfAction>	-	A short description of the event. It further reveals if it was a success or a failure.
36	<Workstation>	Yes	Specifies the NetBIOS name of the remote computer that originated the logon request.
37	<OSType>	-	The type of the OS (e.g. WIN-XP, WIN-Vista, WIN-7, WIN-2003 or WIN-2008).
38	<SourceIP2>	Yes	The IP Address of the source host.
39	<SourceIPPrivate2>	-	Binary indicator which specifies if <SourceIP2¿ is a private IP address, as described in RFC1918.
40	<SourceIPGeoCC2>	-	Two letter country code of <SourceIP2¿ or string "1918" if it is a private IP address.
41	<SourceIPGeoASN2>	-	ASN of<SourceIP2¿ or empty string if it is a private IP address.
42	<SourceIPGeoLat2>	-	Latitude of<SourceIP2¿ or empty string if it is a private IP address.
43	<SourceIPGeoLong2>	-	Longitude of<SourceIP2¿ or empty string if it is a private IP address.
44	<SourcePort2>	-	The port of origin of the event from the source host.
45	<UserName2>	Yes	The name of the account that authenticated.
46	<SecurityID>	Yes	The Security ID of the account that authenticated.
47	<Domain2>	Yes	The domain of the account for which logon is requested. The account is related to the field<UserName2¿.
48	<AccountType>	-	Identifies if the account type in <UserName2>is Privileged, SystemAccount or ComputerAccount.
49	<LogonType>	-	The logon type reveals how the user authenticated (e.g. 2=interactively, 3=network or 10=remotely).
50	<LogonID>	-	The unique logon ID of the logon session.
51	<LogonGUID>	-	For logons that use Kerberos, the unique logon GUID can be used to associate a logon event with a domain controller.
52	<LogonProcess>	-	The process performing the logon.
53	<AuthenticationMethod>	-	The security package called to authenticate the account.
54	<ServiceName>	Yes	Indicates the service for which a Kerberos Ticket Granting Ticket (TGT) or a Kerberos service ticket was issued.
55	<ServiceType>	-	Identifies the service in<ServiceName¿ (e.g. tickets for domain controllers should be distinguished from tickets for member servers or workstations).

Field	Field name	Pseudonymised	Comment
56	<UserPrivileges>	-	Privileges of the account that authenticated.
57	<CallerDomain>	Yes	The domain of the account on the local system which requested the logon. The account is related to the field <CallerUserName>.
58	<CallerUserName>	Yes	Account name on the local system which requested the logon.
59	<CallerLogonID>	-	ID of the logon session for the account mentioned in <CallerUserName>.
60	<FailureReason>	-	The reason why the authentication attempt failed.
61	<StatusCode>	-	The status code indicates why the authentication attempt failed.
62	<SubStatusCode>	-	Further details on why the authentication attempt failed.
63	<ResultCode>	-	Result code of the Kerberos ticket request. Indicates if it was a success or a failure.
64	<CodeDescription>	-	Description in words of the code in <ResultCode>.

Table 5.7: Windows Server Event Fields from Normalisaton

The input vector for McAfee events (shown in Table 5.8) contains fields which were directly exported from the epoEvents database, except for fields which provide geographical IP information (e.g. field no 23, 24, 25, 26 and 27).

Field	Field name	Pseudonymised	Comment
1	AutoID	-	Unique event id(primary key)
2	Auto GUID	-	-
3	ServerID	Yes	Computer name of the ePO server
4	ReceivedUTC	-	Timestamp the event was received by the ePO server
5	DetectedUTC	-	Timestamp the event was generated on the agent.
6	AgentGUID	-	Unique ID of the McAfee agent. An agent is a computer the McAfee antivirussoftware is installed on.
7	Analyzer	-	The analyzer which was used by the agent (e.g. VIRUSCAN8700)
8	Analyzer Name	-	Detailed name of the analyzer used by the agent (e.g. VirusScan Enterprise).
9	Analyzer Version	-	The version number of the analyzer.
10	AnalyzerHostName	Yes	Name of the computer the McAfee agent is installed on.
11	AnalyzerIPv4	Yes	IP address of the computer mentioned in <AnalyzerHostName>.
12	AnalyzerIPv4Private	-	Binary indicator which specifies if <AnalyzerIPV4>is a private IP address, as described in RFC1918.
13	AnalyzerIPV4GeoCC	-	Two letter country code of <AnalyzerIPV4>or string ”1918” if it is a private IP address.
14	AnalyzerIPV4GeoASN	-	ASN of <AnalyzerIPV4>or empty string if it is a private IP address.
15	AnalyzerIPV4GeoLat	-	Latitude of ¡AnalyzerIPV4¿ or empty string if it is a private IP address.
16	AnalyzerIPV4GeoLong	-	Longitude of ¡AnalyzerIPV4¿ or empty string if it is a private IP address.
17	AnalyzerMAC	-	MAC address of the computer mentioned in ¡AnalyzerHostName¿.
18	AnalyzerDATVersion	-	Version number of the analyzer’s signature files.
19	AnalyzerEngineVersion	-	Version number of the engine used by the analyzer.

Field	Field name	Pseudonymised	Comment
20	AnalyzerDetectionMethod	-	Detection method used by the analyzer. Most common detection methods are on-demand scan (ODS) and on-access scan (OAS).
21	SourceHostName	Yes	Computer name the threat originated from.
22	SourceHostIP	Yes	IP address extracted from <SourceHostName>(if it contains an IP address instead of a computer name).
23	SourceHostIPPrivate	Yes	Binary indicator which specifies if <SourceHostIP>is a private IP address, as described in RFC1918
24	SourceHostIPGeoCC	-	Two letter country code of <SourceHostIP>or string "1918" if it is a private IP address
25	SourceHostIPGeoASN	-	ASN of <SourceHostIP>or empty string if it is a private IP address
26	SourceHostIPGeoLat	-	Latitude of <SourceHostIP>or empty string if it is a private IP address
27	SourceHostIPGeoLong	-	Longitude of <SourceHostIP>or empty string if it is a private IP address
28	SourceIPV4	Yes	IP address of the computer mentioned in <SourceHostName>
29	SourceHostIPv4Private	Yes	Binary indicator which specifies if <SourceHostIPv4>is a private IP address, as described in RFC1918
30	SourceHostIPv4GeoCC	-	Two letter country code of <SourceHostIPv4>or string "1918" if it is a private IP address
31	SourceHostIPv4GeoASN	-	ASN of <SourceHostIPv4>or empty string if it is a private IP address
32	SourceHostIPv4GeoLat	-	Latitude of <SourceHostIPv4>or empty string if it is a private IP address
33	SourceHostIPv4GeoLong	-	Longitude of <SourceHostIP>or empty string if it is a private IP address
35	SourceUserName	Yes	User account on <SourceHostName>.
36	SourceProcessName	Yes	User account on <SourceHostName>.
37	SourceURL	-	-
38	TargetHostName	Yes	The IP Address of the source host.
39	TargetIPV4	Yes	IP address of the computer mentioned in <TargetHostName>
40	TargetHostIPv4Private	Yes	Binary indicator which specifies if <TargetHostIPv4>is a private IP address, as described in RFC1918
41	TargetHostIPv4GeoCC	-	Two letter country code of <TargetHostIPv4>or string "1918" if it is a private IP address
42	TargetHostIPv4GeoASN	-	ASN of <TargetHostIPv4>or empty string if it is a private IP address
43	TargetHostIPv4GeoLat	-	Latitude of <TargetHostIPv4>or empty string if it is a private IP address
44	TargetHostIPv4GeoLong	-	Longitude of <TargetHostIP>or empty string if it is a private IP address
45	TargetMac	-	MAC address of the computer mentioned in <TargetHostName>.
46	TargetUserName	Yes	User account on <TargetHostName>.
47	TargetUserIP	Yes	P address extracted from ¡TargetUserName¿ (if it contains an IP address instead ofa computer name).
48	TargetHostIPPrivate	Yes	Binary indicator which specifies if <TargetHostIP>is a private IP address, as described in RFC1918
49	TargetHostIPGeoCC	-	Two letter country code of <TargetHostIP>or string "1918" if it is a private IP address
50	TargetHostIPGeoASN	-	ASN of <TargetHostIP>or empty string if it is a private IP address
51	TargetHostIPGeoLat	-	Latitude of <TargetHostIP>or empty string if it is a private IP address

Field	Field name	Pseudonymised	Comment
52	TargetHostIPGeoLong	-	Longitude of <TargetHostIP>or empty string if it is a private IP address
53	TargetPort	-	The port targeted by the threat.
54	TargetProtocol	-	The protocol targeted by the threat.
55	TargetProcessName	-	The process name which was the target of the threat.
56	TargetFileName	Yes	The name of the file which was the target of the threat.
57	ThreatCategory	-	Common threat categories are anti-virus detection (av.detect) or firewall detection(fw.detect)
58	ThreatEventID	-	Unique event ID of the logged threat.
59	ThreatSeverity	-	Severity of the threat
60	ThreatName	-	Description of the logged threat (e.g. name of the virus).
61	ThreatType	-	Type of the threat (e.g. buffer overflow).
62	ThreatActionTaken	-	Description of the action taken (e.g. deleted).
63	ThreatHandled	-	Indicator if the threat was handled or not
64	The Timestamp	-	Sequence number the ePO expects from a connecting agent. If the sequence number does not match, communication is rejected.

Table 5.8: Input Vector of McAfee Events

The events were transformed according to the above specification and test cases, and produced a normalised set of logs. The resulting event log is an output from the input vectors in CSV format.
The highlighted fields indicated the geolocation information present within the logs. The geolocation data is identifiable for source and destination of events. The Windows events, contain two sources for geolocation as shown in the table. One source is of geolocation is from the TSOM sensor itself and the second is from the windows raw events, extracted from the $EVENT.INFO token.

In Figure 5.1 we highlighted the connection between the TSOM environment output and our experimental solution - the data received from the CMS was retrieved, normalisation performed according to the input vectors specified above and retrieved for use in this 'enhanced' SIEM testbed to carry out the determined technical objectives.

5.3 SIEM Tools and Frameworks

The primary design step concerns analysis of the SIEM architectures and their technologies. We take a look at the MASSIF tool framework and the OSSIM solution to determine the elements of these SIEM implementations applicable for the research objectives and demonstrative results.

5.3.1 MASSIF

MASSIF aims to develop new generation SIEM technology for service infrastructures. The technical design is 'requirements-driven' focusing on providing high interoperability, high scalability and high elasticity.
It focuses on four industrial domains to serve as a source for requirements and to validate the project results. These are the Olympic games, mobile phone based money transfer service,

critical infrastructure process control and managed enterprise service infrastructures. The data used is from a managed enterprise environment therefore, the considerations are focused on the tools applied within that scenario.
As a next-generation SIEM framework support for service infrastructures it encourages intelligent, scalable, multi-level/multi-domain security event processing and predictive security monitoring[40]. Figure 5.3 provides an architectural overview of the MASSIF framework at a high level.

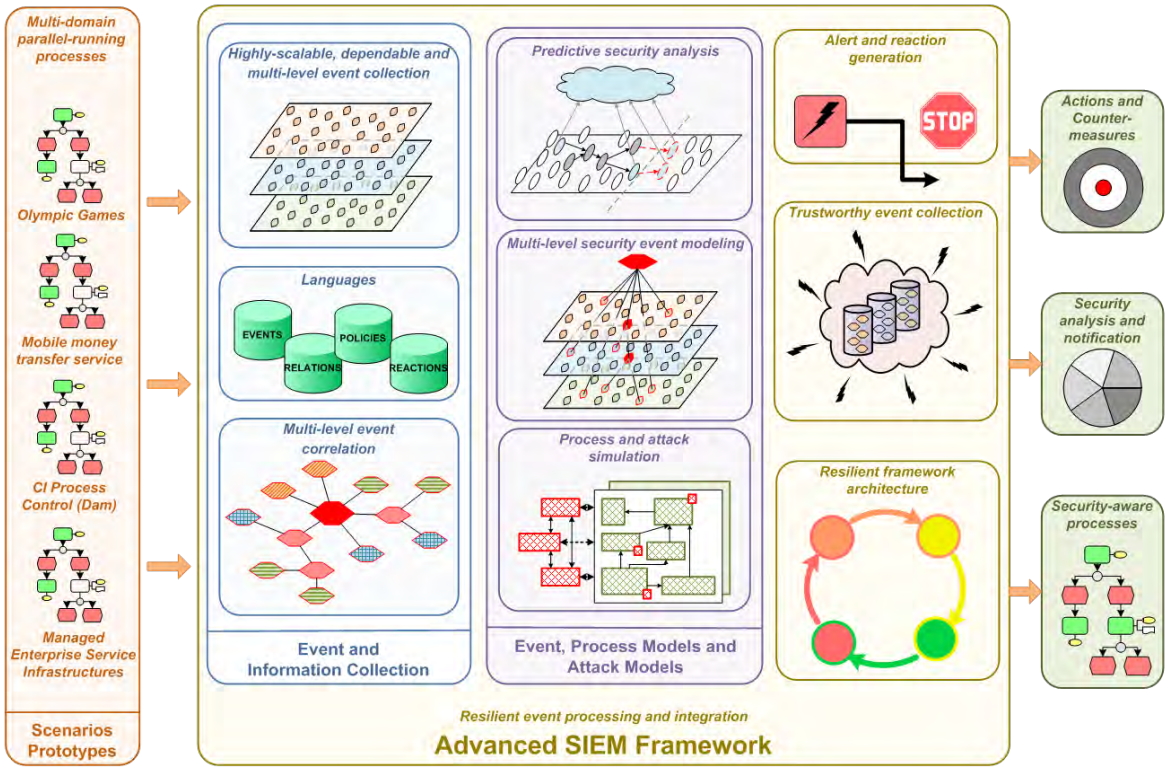

Figure 5.3: MASSIF Architecture Overview[40]

On the basis of proper multi-level event correlation MASSIF can provide innovative techniques to enable the detection of upcoming security threats and trigger remediation actions even before the occurrence of possible security incidences.
This service-level SIEM technology involves the modelling and formal validation of security, including trusted computing concepts, architecture for dependable and resilient collection of service events, supported by an extremely scalable and high performance event collection and processing framework, in the context of service-level attack models[40].

The tools and elements of MASSIF can be integrated with two SIEM solutions, these are Alienvault OSSIM and 6Cure. The defined elements within the MASSIF architecture are modular, allowing a tool or combinations of tools to be used with other systems. For this

research this ability is exploited in the consideration of integration with OSSIM.

Technical Summary

The modularity of the tool structure that defines the MASSIF framework covers a range of software requirements. However, the differences are not incompatible and easily deployable in many environments. In terms of installation and implementation requirements, most tools just require a Java friendly environment with the appropriate runtime environment version installed. A summarised technical breakdown of the tools and elements of the conceptual framework is shown in Table 5.9. The functionality area and it's technical requirements can assist an evaluation of compatibility for a testing environment and integration with other software.

Component/Functionality	**General Tool Specifications**
Event and Information Collection	
Multi-Level Event Collection	Java Runtime Environment (JRE) version 1.7
Languages	SQL, Java, Python, UML
Multi-level event correlation	JREv1.7, JDKv1.6, Derby DBMS v10.10.1.1, Apache Tomcat 6.0.20 or later
Event, Process and Attack Models	
Process & Attack Simulation	Syslog, JREv1.7 , Linux MOTIF Library[1], Ruby
Multi-level security event modelling	JREv1.7, Python, Apache Felix 4.0
Predictive security analysis	Python, Netbeans IDE, JRE v1.7
Other tools	
Alert and Reaction Generation	Python(requiring packages e.g python-lxml,python-mysqlb, python-setuptools, figlet), JRE1.7
Trustworthy Event Collection	JRE 1.7
Resilient Framework Architecture	JRE. 1.7

Table 5.9: MASSIF tools software specifications

5.3.2 OSSIM Alienvault Solution

OSSIM is a fully featured SIEM solution offering all necessary functionality from event collection and incident detection to alarm response and feedback visualisation. The foundation for Alienvault's security management solution is Open Source SIEM(OSSIM) which provides SIEM vulnerability assessment, network and host intrusion detection, and file integrity monitoring. Alienvault markets and supports commercial software or application offerings that extends OSSIM with enhancements in scaling, log management, reporting, administration and multitenanting for managed security service providers (MSSPs). Therefore, OSSIM works as an open-source framework as well as commercial depending on the level of SIEM tool provision. Enabling the facility of open-source allows users to experiment with the current solution

and help suggest even better adjustments to this framework.

In a summarised overview the solution provides the following features;

- Low level, real-time detection of security events
- Compliance automation
- Data log management and storage
- Event log and alarm visualisation
- Security intelligence that enhances the accuracy of threat detection
- Network behaviour analysis and situational behaviour awareness

Figure 5.4 provides a high level overview of OSSIM solution, it's capabilities and security event flow.

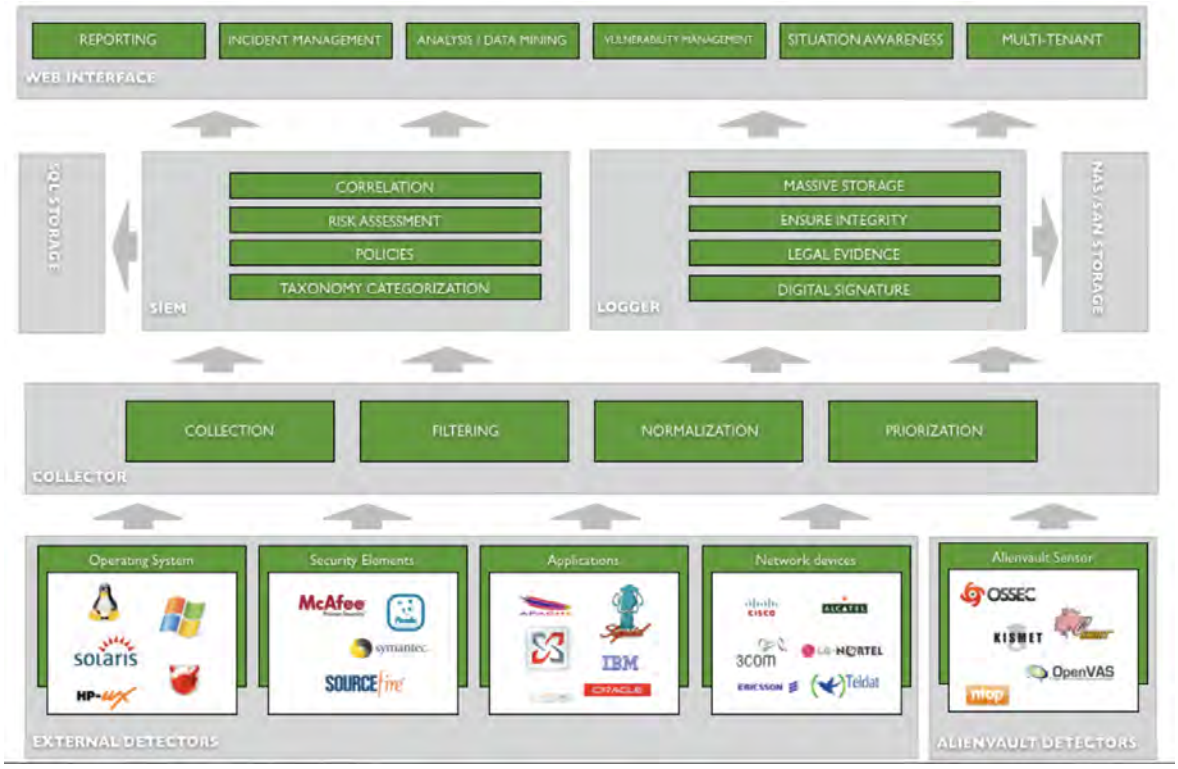

Figure 5.4: High level view of the OSSIM Architecture[36]

Technical Summary

The OSSIM solution based on the Linux Debian operating system, is available as a complete solution using virtualisation software, such as VMware[2]. The entire solution can then be

[2]www.vmware.com

installed in the virtual environment and accessed through its assigned static IP specified for that machine instance. The system requirements for testing purposes use a minimum of 8GB RAM, 2 CPU cores, with at least 250GB storage capacity.

This architecture is based and centered on the efforts of four main components:

- OSSIM sensor: The sensors collect and normalise the events generated by the different security equipment. One can deploy as many sensors as needed
- OSSIM server: The server receives events from sensors, and does Risk Assessment and Correlation tasks
- OSSIM database: For storage purposes, there is a provision of a MySQL database that stores the events, configurations, and useful information.
- OSSIM framework: The framework encapsulating the solution is the PHP/Python code that "serves" the information to the webfront-end.

Events are received by the sensors, once they are normalised on the sensor they are sent forward to the remote or local OSSIM server[21]. The server collects the normalised events and applies user defined policies on them. Depending on the policy definition the event is sent for correlation in the correlation engine or not. The correlation engine is a powerful feature of the solution, providing users the ability to write correlation rules using XML to parse the events and trigger certain responses from the SIEM solution. After policy enforcement, the next phase is risk assessment, then finally events are sent to the correlation engine. Correlation of events is performed, which enables the generation of alarms that are visualised in the interface. Any alarms generated are stored in the OSSIM database for a certain time period.
The solution allows tight control over widely distributed enterprise networks from a single location. In summary, the system considers the context of each threat and the importance of the assets involved, evaluates situational risk, discovers and distinguishes actual threats from the thousand of false positives that are produced each day in each network.

5.4 Tool Discussion

In section 5.1 we identified the main objectives of implementation needed to justify the research implications. Using the collected data sets, the strategy of committing the results to an implementation required examination of feature/tool(s) applicability in OSSIM and MASSIF.

Objective 1: *Augmenting existing security methods using geolocation data*
The first requirement would be for the SIEM to receive geolocation data in a usable format for event analysis. These analysis tools need to expend it in authentication/alarm triggering procedures and (possibly) visualisation. This requires an implementation of security procedures, if considering visualisation, an interface for the SOC to view attacks and see alarms for locations. These requirements are applicable to the interface of OSSIM which has excellent alarm triggering visualisation and facilitates the analysis of any data regardless of the format through the use of plugins. If a plugin for a specific data log format is not available, OSSIM

facilitates the creation of a custom plugin by users themselves.

MASSIF in this regard is the lesser alternative to take, mainly due to the fact that OSSIM gives easier visualisation at SOC level providing a unified visually appealing interface and alert facilities as expected from a industry product. MASSIF can provide for this but as it is currently functionality oriented and research driven its power lies in the tools more than the visual asthetics.

Objective 2: *Provision of anonymisation in SIEMs for geolocation data*
This is an advanced expectation - the offering of anonymisation procedures in event processing tools. This requires a SIEM with potential provisions and tools for untried manipulation of data. In this case, MASSIF will be the preferred framework of operation to facilitate this. MASSIF aims to realise the expectations of next-generation SIEMs, requirements such as novel processing applications on data can be facilitated through the tools of MASSIF. These tools are developed to experiment with data techniques and processes from different scenarios.

OSSIM in this case, is not the preferred SIEM, with MASSIF having advanced tools ready to implement this without the need for writing an entire anonymisation procedure by the user which would be necessary and more work in the case of OSSIM.
Both OSSIM and MASSIF have their strong points and provisions in feature delivery. MASSIF is a modular collection of tools that combine together to form a conceptual framework. The tools can be used individually, with the necessary elements used as additions to existing SIEMs. This is a feasible option for implementation purposes, utilising the required elements of MASSIF to 'enhance' the OSSIM solution used for the experiment visual front-end. We use the flexible properties of OSSIM and the advanced tools of MASSIF to create a testbed for the initial hypothesis, using geolocation to enhance SIEM capabilities.
Thus, the implementation will be dependant on the integration efforts of an existing SIEM, in this case OSSIM, with chosen elements of MASSIF.

5.4.1 MASSIF: Applicable Elements

The MASSIF conceptual framework has developed many tools that combine together in efforts to creating novel or enhancing security procedures. These tools range in application from pre-analysis, real time, detection and reaction feedback. A selection of tools were utilised from the MASSIF framework in order to address the misuse cases and study, their functionalities and provision towards demonstration validation are clarified.

Generic Event Translator (GET)

The GET is used for event aggregation, collection, normalisation and techniques such as encryption, anonymisation or where applicable pseudonymisation. The purpose of this tool is to act as the gateway for the entire framework to using data from any source or device, with the capability to provide the handling of wide ranges of event formats produced by these sources. It is able to convert these events to a standard format, such as the MASSIF event format.
This data format conversion ability open doors for other uses besides standardisation, other than pulling data from log lines into uniform schemas, it also facilitates the enforcing of

security integrity on this data. The data is additionally encrypted and is sent along with the standardised log to the relevant server.

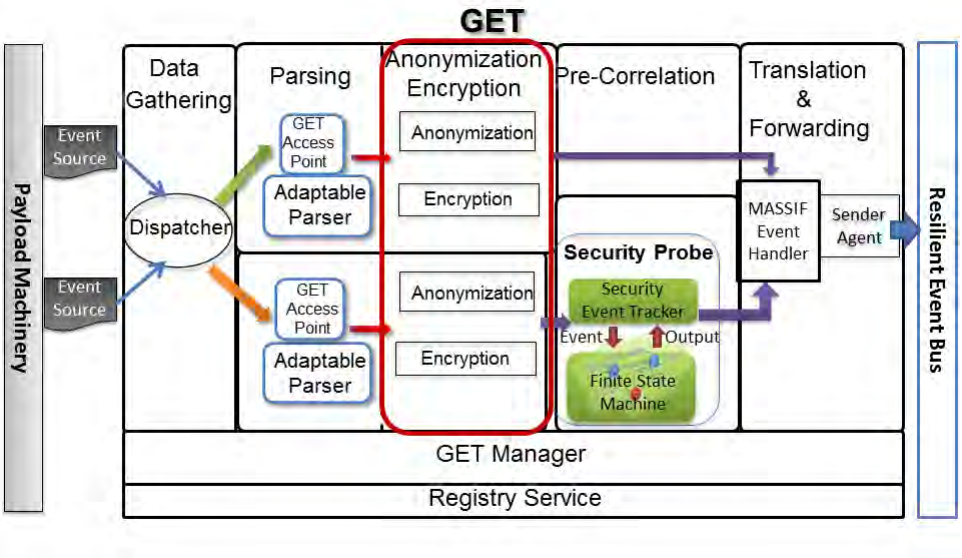

Figure 5.5: MASSIF: The GET tool[27]

The tool primarily works as a normalisation tool, but in this scenario in particular it can be seen it will not be necessary as the data has been normalised. Thus, the focus can be made specifically on carrying out SIEM procedures that are novel and have not been applied, without the need for simulating the common SIEM procedures again.
The conversion abilities from one format to another, is a similiar process to modification of data fields, if the conversion formula contains an anonmyisation algorithm, this enables the anonymisation to be performed on the necessary fields smoothly.

The GET requires an operating system with Java Platform Standard Edition (JSE) Runtime Environment version 1.7.0_03 (build 1.7.0_03-b05) or later. The tool is a Java-based platform for ease of use on any platform and portability.

Resilient Event Storage (RES)

The RES is an intrusion and fault tolerant data storage facility designed to ensure integrity and unforgeability of events to be stored even if some components of architecture are compromised. Alarms and Events that are stored in the RES cannot be modified, also the data is stored in encrypted form using a Password Based Algorithm technique. Should admin/ forensic authorities require the event data in original format, it can be retrieved through the input of the correct password.
The purpose of the RES is to ensure integrity of the original data is maintained, as parts of the data will be modified by undergoing anonymisation procedures. The secure storage of alarms and events is also a needed facility to ensure the analysis results are not modified or tampered with, and are in fact a result of the simulated attack.

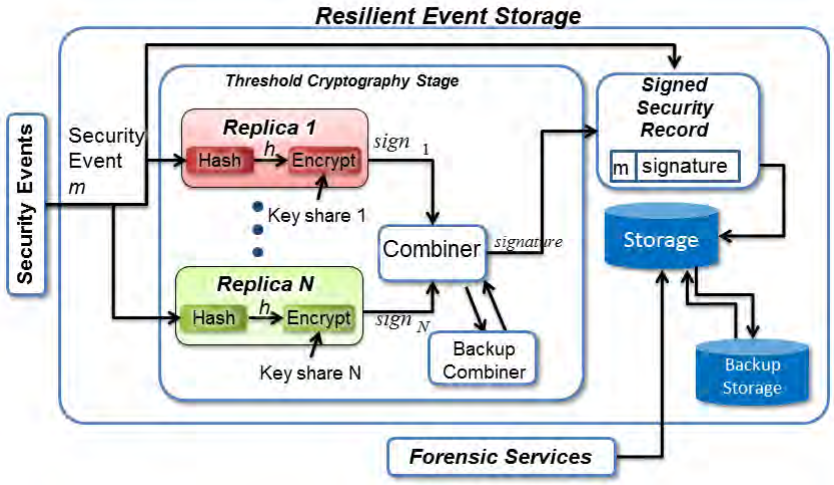

Figure 5.6: MASSIF: The Resilient Event Storage[22]

The RES facilities are composed of a number of nodes that implement resilient storage mechanisms based on threshold cryptography algorithms. The RES nodes might be corrupted, but the RES should provide the ability to detect the corruption and retrieve the correct event[16]. The RES requires an operating system with Java Platform Standard Edition (JSE) Runtime Environment version 1.7.0_03 (build 1.7.0_03-b05) or later. The tool is a java-based platform for ease of use on any platform and portability.

Resilient Event Bus (REB)

The Resilient Event Bus provides resilient communication among a set of nodes (see Figure 5.7). The use of nodes as an overlay network with communications performed on top of the standard protocols like UDP/IP, allow varied network settings to be supported. The REB uses an application-level one-hop source routing policy that forwards messages towards a destination, instead of following only the routes set by the network level routing[48]. It also takes advantage of coding techniques and the available redundancies of a network, for example, when a node has multiple network connections (as in multihoming, for instance), to ensure a very high probability of secure delivery of messages to the required destination[48]. The purpose of the REB is to securely transmit events between components, a managed enterprise environment is a heavy duty network, event flows are high and fast, a channel that can ensure the nature of this scenario in a typical live environment is an important consideration.

The REB requires an operating system with JSE Runtime Environment version 1.7.0_03 (build 1.7.0_03-b05) or later. The tool is a java-based platform for ease of use on any platform and portability.

5.4.2 OSSIM: Applicable Features

The simulation of use-case can be facilitated through the use of writing a correlation rules matching the specfication of the required attack. Events received or collected from agents

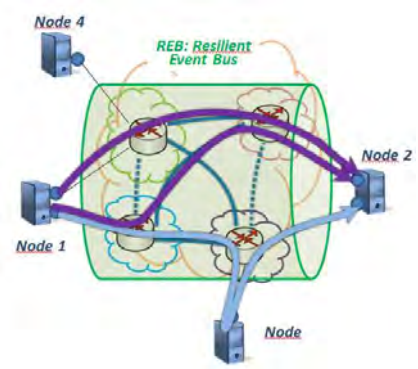

Figure 5.7: MASSIF: The Resilient Event Bus[48]

go through various security processes to determine the generation of alarms from policy comparisons and correlation checks, if an alert is triggered the incidents are stored in the OSSIM DB. The data security flow that concerns us, is shown below[21]:

Correlation Directives

A SIEM correlation rule can be used to automate parts of the system login and authentication monitoring process[8]. The rule encapsulates patterns of events that indicate suspicious or malicious behaviour. The nature of the activity marked from the sequence of events from a specific source towards another node on the network can highlight the intention of that user and the predicted outcome.

The following examples of threat situations can be defined through the use of correlation mechanisms to trigger an alarm to administration of suspected illicit activity:

- A single system attack when an attacker tries all credentials on one system
- A string of several login failures immediately followed by a success
- Authentication sweep attack (a user trying the same credential on all systems)
- Successful login at unusual times (for the user or for the system)
- Successful login from unusual locations (for the user or for the system)

A correlation directive, that can be defined in OSSIM, is a sequence of rules for the OSSIM server to follow in order to correlate events. In each rule, one can specify at the very least[25]:

- the event's signature to detect;
- the reliability of the rule;
- the occurrence of that event along with timeout;
- the source and destination ports/IP addresses;

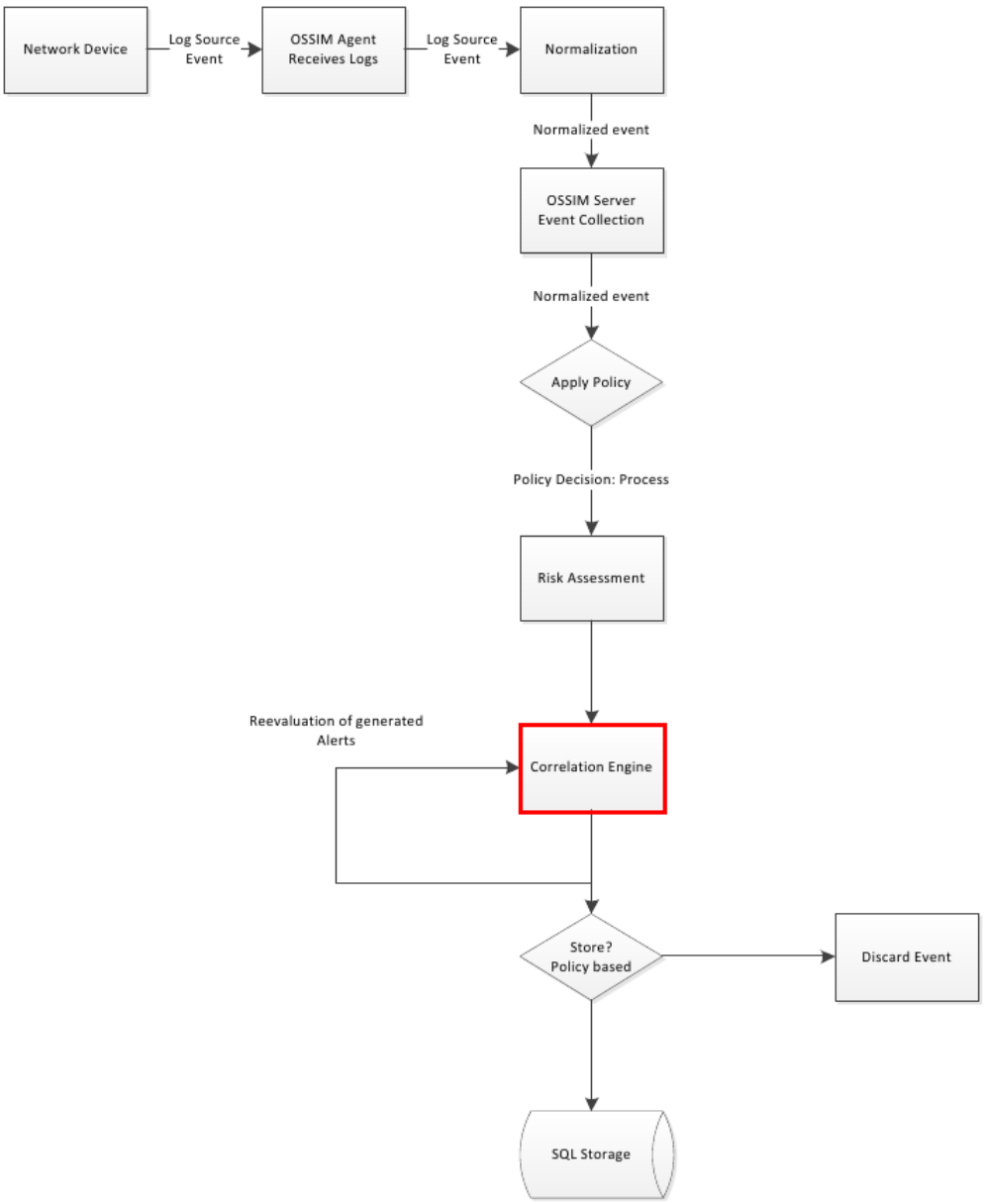

Figure 5.8: Security Information Flow within OSSIM[21]

For a correlation directive to raise an alarm, the rule detects event flows which satisfy conditions leading to an increased risk factor. A risk factor of one or higher results in the generation of an alarm.
Risk assessment depends on three variables, asset value, event priority, and event reliability. Assets are key structures for SIM systems. An asset is an equipment that one wants to secure or protect. Examples of assets are host, firewalls, databases and other company devices holding value to the company in some way. For each asset that is defined in OSSIM, an asset-value between 0 and 5 that has to be specified.
This value estimates the asset's importance to the client/company itself. Priority and relia-

bility are both values relative to the actual event, the values can be assigned within the range of 0 to 10. Priority measures the relative importance of the event seen as an attack, while reliability measures the probability of the event itself being reliable and true. For instance, intrustion detection systems are known to generate many false positives in some instances[25]. In such cases, the reliability value needs to be set as low.
Each of those three variables are then calculated together to determine a final value defined as the 'risk'. This risk factor is calculated with the following formula:

$$Risk = \left(\frac{priority \times reliability \times asset}{25} \right) \tag{5.1}$$

If the resulting value bypasses the value of 1, the server will generate an alarm to notify the respective security manager. To illustrate the method of risk calculation let us suppose the asset of IP address 11.22.33.44 is an asset belonging to the company. The asset value assigned is given a value of 5. The agent/protocol generating the security event that is received in OSSIM has been given a reliability value of 2 and due it's nature, a priority of 5. The calculated risk for this IP in question is $R = (5 \times 5 \times 2)/25 = 2$. Therefore, an alarm will be generated. If the device is not specified to be a company asset however, it will be assigned a default value of 2, in which case $R<1$ and an alarm will not be raised.

Alarm Generation

OSSIM triggers alarms visually in the user interface to warn the administration. For example, if a pattern is matched within a correlation directive of a specific this invokes the creation of an OSSIM incident, and an alarm at the alarm database. Both the generated security incident and alarm are stored in a secure backup and storage resource to ensure responses cannot be manipulated within the SIEM.

5.5 Integrated Concept

The conceptual integration aims to encompass the following processes to be applied on the required managed enterprise data, towards satisfying all objectives of this study;

- The demonstration of confidentiality preservation capabilities throughout the process flow of the concerned events (carrying geolocation information).
  - At the edge side of event processing, the following needs to be performed:
    Anonymisation of event data
    Encryption of the entire event payload
  - At the core side of event processing, the following is required to finalise the validity of output:
    Secure storage of all relevant security events, i.e. alarms generated and incidents created.
    Restricted access to the original event content through the use of authentication, to ensure the results have not been tampered with - particularly for forensics.

- The use of secure resilient mechanisms responsible for the transfer of events. This is to validate the process flow ability to deal with the high volume nature of SIEM environment information flows.
  Tool communications based on the Resilient Event Bus mechanisms

- Alarm storage based on Resilient Event Storage facilities, post-analysis.

- Usability vs. Privacy
  The use of MASSIF anonymisation techniques allows location data to be used in security analysis procedures whilst preserving user privacy.

- Feasible integration between MASSIF data layer components and the OSSIM SIEM solution

- The demonstration of the detection of a MESI scenario misuse case, from the entry of the attack sequence to the output of alarms identifying the malicious attempt through alarms and incident generation, in a complete entry-to-discovery cycle.

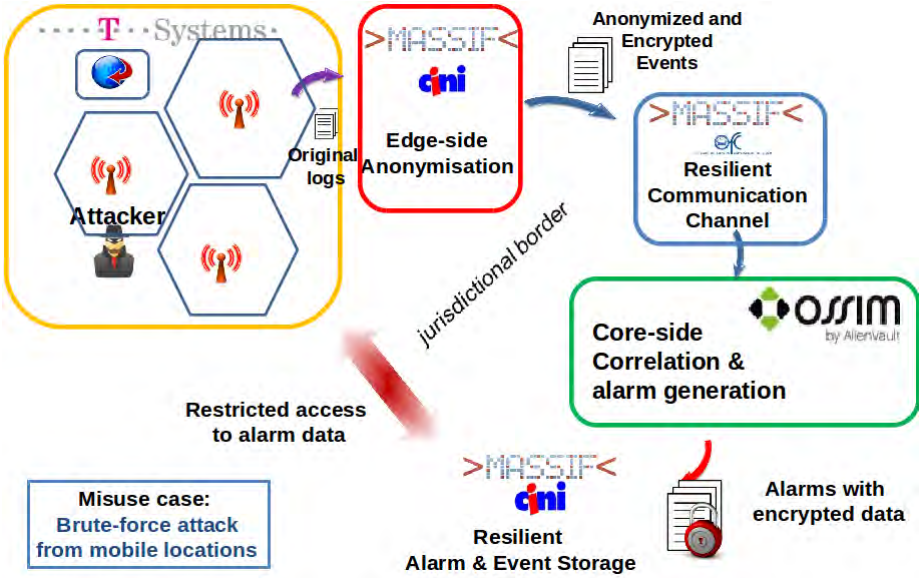

Figure 5.9: MASSIF Conceptual Diagram

5.6 Experimental Design

In Figure 5.10, an experimental solution of the testbed depicting the relevant tools and the flow to security events through the system is shown. The flow can be analysed to oversee the conversion process of the location data used with the selected MASSIF tools and the OSSIM solution.

1. Data layer: The GET constitutes the point of access to the MASSIF system. It translates the MESI events from the source with anonymisation rules for data fields requiring privacy. The events are then translated into OSSIM format.
2. Event layer: the REB transports these events from GET, adding the security provided by the resilient framework, ensuring communication resilience.
3. Application layer: OSSIM processes the events and generates alarms for the end users to check for irregularities that indicate attack attempts in login cycles using the correlation directive. The directive is matched against the incoming events to be triggered if there is a matching sequence of suspicious activity trigged from anaysis of the location data. RES stores the alarm data for integrity and forensic purposes to support the process of privacy ensuring techniques; committed from the beginning of the locational information intake process by the GET. This ensures the privacy conditions are maintained throughout the entire process of location data use.

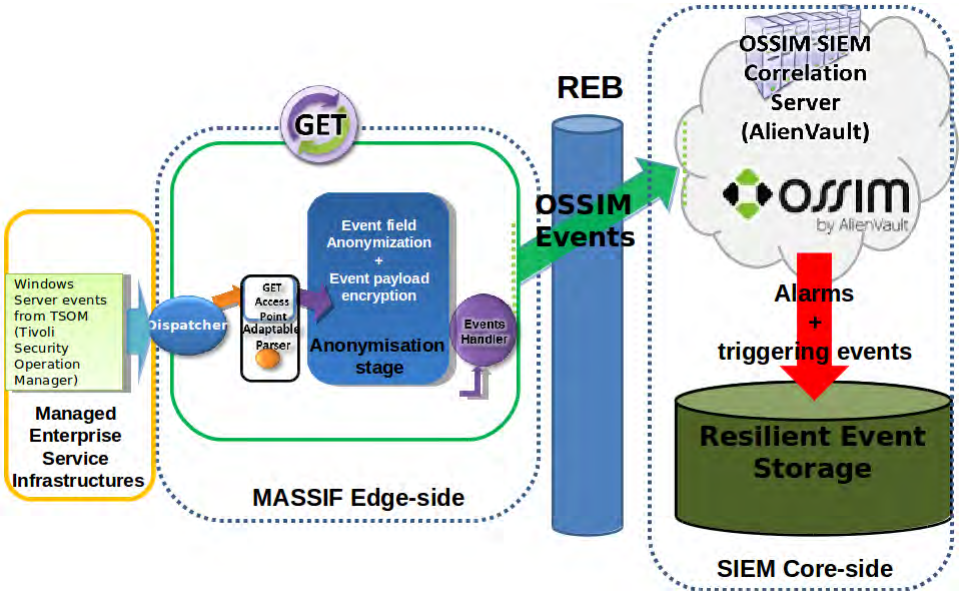

Figure 5.10: MASSIF Experimental Diagram

5.6.1 Technical assumptions

The assumptions made through the test implementation, of the attacker and the operative environment of the MESI systems, are as follows:

- The network has accurate location information on user access points in network
- Other user-sensitive log fields are pseudonymised to ensure privacy.
- The attacker attempts a brute force attack with consecutive login attempts from location not identified as valid for user X.
- The target is a machine with admin rights.

5.7 Summary

To validate the objectives of this research in terms of SIEM application of geolocation in security, privacy enforcement of geolocation information and the feasibility of these suggestions in a such a large framework with a high level of technicality. An experimental set up is proposed, as a proof-of-concept implementation that can achieve the propositions of the study. In summary, the experimental simulation will demonstrate the following features:

- **Possible Attack.**
  An attack, for example, the brute force attack isolated from a specific area. Location identification can be a good way to narrow a field of security analysis down.
- **Data Privacy**
  Data crosses international boundaries, requires privacy considerations
- **Geolocation for authentication**
  Location data can provide a second layer of verification of users permissible login but for this anonymised location information must still be usable.
- **Log Generation**
  source events are provided with location data for user logins on a windows server 03. Login attempts with location either derived from sensors, WiFi, or IP address.
- **Reality Vs Simulation**
  The generated log similarity to real-life and response. Login attempt pattern follows definition of Brute-force attack. Possible system response, admin block user access from location A.
- **Limitations**
  Events are batch collections and not real time, simulation of real-time events will be required.
- **SIEM Integration**
  Attack Detection through the correlation abilites of OSSIM. Additional clause, anonymised location information is still to generate alarms for geolocation rules.

The next chapter discusses the approaches taken to implementing this conceptual solution and the resulting configurations and modifications required for the successful implementation of the above listed objectives.

Chapter 6

Implementation

Following the discussed architecture, this chapter presents the implementation logistics, with integration of all discussed components. The complete integration is run in a demonstration setup with the use of virtual machines to facilitate the various tools and their capacities. A SIEM environment is not a small deployment, it is a collection of tools connected with various configurations and instructions, and connected resources. Therefore, the integration of the required components to produce the needed process flow requires an examination from high to low level, surrounding the test data itself.

6.1 Event Analysis

The events extracted from the managed enterprise environment are collected from two main sources. For the purposes of this research, these two sources provide the test area for the misuse cases identified in the corresponding managed enterprise scenario as the prevalent attacks of that environment. These sources are; data events collected from the McAfee ePolicy Orchestrator and from Windows Servers 2003/2008.

6.1.1 Event Set A: Windows Server 2003/2008

This source event data focuses on all account logon activities, its attribution is focused towards simulating the following attack misuse cases:

- MC 5.5.1 An Unauthorised Login to a computer system, network or application
- MC 5.5.2 Brute force attack for attempted login

Listing 6.1-1 extracted from the month of May is a sample Windows event. The data pseudonymisation on user names and identities must be noted and the CSV format it was output in after procedures of normalisation (for field descriptions see subsection 5.2.1).

Listing 6.1-1: An example MESI Windows event

```
1 TBS;;1124490656;1325200062;1306496120083;1306496120000;500f7b56f8805d3fd2e4
2 8e36;;Windows EventLog;Security.Failure Audit;66;failed , failed , login ,
3 user;91734d39b55cb189ffc47c27;e0559ee6821fb98336ece5a5;;70.58.17.106;;;;39.
4 91470000000001;-105.0809;0;70.38.124.46;;;;45.5;-73.5833;0;66;66;S-6.1;U-6.
5 1.17;680;Logon attempt by;;WIN-XP/WIN-2003;;;;;;;;;;;;;;;;MICROSOFT_AUTHENT
6 ICATION_PACKAGE_V1_0;;;;;;;;0xC0000064;;;;
```

The first consideration on the analysis of the event information is to determine which elements of the event are relevant for processing and collection by the SIEM. A SIEM processes huge volumes of information with a typical collection rate reaching around 20 000 events per second. This mandates the need to ensure only required information is collected to aid the processing utility of the SIEM. The relevant fields deemed necessary for further analysis through MASSIF and OSSIM from the windows events are listed in Table 6.1.

Field	***Type***	***Value***
InternalTimeStamp	TimeStamp	1325184445
The time the event was received in the normalisation		
SensorName	String	fb2f0416b09c5e0611ee5319
The name of the sensor that orginated the event(example shown is in pseudonymised format)		
EventType	String	SE_AUDITID_UNKNOWN_USER_OR_PWD
Event type as defined, if possible, by the device generating the event. The type is otherwise derived from the reported event.		
UserName	String	675aa07dbab86489548c99c8
The username associated with the event(example shown is pseudonymised format)		
SourceIP	String	x.x.x.x
The IP address of the source host)		
SourcePort	Integer	4958
The port of the origin of the event from the source host		
DestinationIP	String	y.y.y.y
The IP address of the destination host		
DestinationPort	Integer	0
Port on the host at which the event was directed		
EventID	Integer	529
The unique ID of an event		
SourceIPGeoLat	Integer	529
Latitude of <SourceIP>or empty string if it is a private IP address.		
SourceIPGeoLong	Integer	529
Longitude of <SourceIP>or empty string if it is a private IP address.		
DestinationIPGeoLat	Integer	529
Latitude of <DestinationIP>or empty string if it is a private IP address.		
DestinationIPGeoLong	Integer	529
Longitude of <DestinationIP>or empty string if it is a private IP address.		

Table 6.1: Windows Event required fields

In the Windows domain architecture, the procedure of logon and the process of authentication are treated as separate concepts. A workstation logon through a domain account requires the

workstation to then authenticate with the Authentication Server (AS) on the domain controller. Two categories of security events allow the tracking of these events for both activities; the *Logon/Logoff* category records logon activity, and *Account Logon* track all authentication events. These are the nature of the events that will be analysed in the SIEM environment.

Taking a look at the chosen fields, it is necessary to point out the *EventID* field is the only event field, along with *EventType*, extracted of all the provided threat information fields. This is because SIEM managers like OSSIM permit storage of event rule ID and their information within databases, the log data sent through requires just the ID, while the rest of the data, such as event descriptions can be pulled through from an request to the database. This allows and lessening of stress in the volumes retrieved within the collection process of the analysis cycle.

6.1.2 Event Set B: McAfee ePolicy Orchestrator

The McAfee event data contains threat information identified by McAfee such as threat type, severity, name, action taken. This information applies mostly to assist detecting the following attack misuse case;

- MC 5.5.5 Worm propagation

Listing 6.1-2 is an example event from the data set, note the pseudonymisation effects and the McAfee event information regarding the action to be taken and identified threat.

Listing 6.1-2: An example MESI McAfee event

```
1 6370207;{BCD22761−E4B7−42AF−8063−DF6CB341A55A};460bfe9ebd6978c9bb94bd44;2011−11−22
    00:00:08.887000000;2011−11−21 23:55:46;{4DA35B52−AEB5−47DD−9454−3F821F75F3B8};
    VIRUSCAN8700;VirusScanEnterprise;8.7;c1c731a7a6cd59e353936196;x.x.x.x;1;1918;;;;;;;OAS;;;;;;;;x.x.x.x
    ;0;ZA;3741;−33.9167;18.4167;;;ba6071f5f175251b2877fdb6;;c1c731a7a6cd59e353936196;y.y.y.y;1;1918;;;;;
    e0559ee6821fb98336ece5a5;;;;;;;25;;;;fw.detect;1096;5;Anti−virus Standard Protection:Prevent mass
    mailing worms from sending mail;access protection;would block;True;00000000028749EF;
```

Using threat information from Mcafee propagations can be traced in efforts to tracing the source of the malicious spread within the network. We select the relevant fields from the data that apply to the required information needs:

Field	***Type***	***Value***
AutoID	Long	6370207
The unique id of an event(primary key)		
ThreatEventID	String	460bfe9ebd6978c9bb94bd44
Unique event ID of the logged threat.		
SourceIPv4	String	x.x.x.x
The ip address of the source computer)		
TargetIPv4	String	y.y.y.y
The ip address of the targeted computer		
TargetUserName	String	e0559ee6821fb98336ece5a5
Username on the computer the threat targeted (example shown in pseudonymised format)		

TargetPort	Integer	25
The port targeted by the threat		
EventID	Integer	529
The unique ID of an event		
SourceIPV4GeoLat	Integer	529
Latitude of <SourceIPV4>or empty string if it is a private IP address.		
SourceIPV4GeoLong	Integer	529
Longitude of <SourceIPV4>or empty string if it is a private IP address.		
TargetIPV4GeoLat	Integer	529
Latitude of <TargetIPV4>or empty string if it is a private IP address.		
TargetIPV4GeoLon	Integer	529
Latitude of <TargetIPV4>or empty string if it is a private IP address.		

The same applies to the *ThreatEventID* field in the McAfee events, the event identifier is linked to its information which is stored in a lookup table within OSSIM. This provision enables the less parsing and handling of event information, minimising the parsing and collection time effort, particularly beneficial optimisation for a framework handling data collection at such great scales. Both event sets pull through IP sources, targets, ports and all geographical information from event sources as the critical identification information.

6.2 Event Pre-processing

Prior to concept simulation, the data is evaluated on its content fields, the attestation of results can only be validated depending on the provision of usefulness coming from the extracted fields. The main fields identified were, IP addresses, ports and geographical data of source and targets.

Regarding geographical data, TSOM records this information recorded by sensors, however the SNMP send trap doesn't offer tokens yet for this purpose, the datasets therefore contain partial geographical information and is not consistently present throughout the collected events, for the required level of exploitation.
A workaround for this is to populate events that did not manage to obtain geolocation, with a calculated estimation through the use of the IP address.
Of course, an assumption of this research is that location data can be provided directly into the SIEM through conduits expressing this data. But in the event it does not, and in this case where the data collection technique unrelated to SIEM collection itself commits partial collection, this method is preferable. The great advantage as discussed in Chapter 3, is this very ability to calculate additional geographical information from a such a fundamental field in security, gaining more data for exploitation from existing ones.

The second concern is more specific to the SIEM applied in the investigation. OSSIM will be used as the detection and response end of the resulting coagulation, the way it works in terms of data management of sources is a very centric vision on the real-time aspect of events

coming in from various sources. The events are pulled in and treated from the timestamps, events older than three days are discarded from the event database making it possible for the large flow of events to consistently stream in while avoiding an overflow in storage resources. The datasets collected from the enterprise environment do not fall within this time limit, as they have been collected in a much earlier time frame, on analysis of the timestamps OSSIM will by default discard these events. There is no alternative available within OSSIM to change this feature, as it's purpose is focused on real-time analysis, not geared towards the processing of relatively 'old' data.

These two issues need to be addressed to enable the test implementation to detect and utilise events in the manner required. To facilitate the current limitations of trap collection and OSSIM with respect to the data set of the managed enterprise, the logs undergo processing to achieve the following:

- Real-time effect within a SIEM
- Addition of any missing geographical data through the events IP addresses

A script is written to facilitate this, and is run on the data sets before activation of the plugn through OSSIM for detection of new events. The script procedures, written in python, applied to the data sets are shown below;

6.2.1 Real-time Effect

Events falling in the same day need to be adjusted to the same day in the present time frame. Therefore, each group of events is adjusted per date to the day of testing, the time in the *datetimestamp* of the original events is not to be tampered with, only the date is extracted and converted.
The function shown in Listing 6.2-3 is used if when using OSSIMs plugin method. The approach exercises regular expressions (discussed later) to map a log event type to its own schema. It would match the timestamp and map this directly to the timestamp that is analysed if within real-time spans.

Listing 6.2-3: Real time datestamp conversion

```
1  #!/usr/bin/en/python
2
3    global datematchto
4    global datechangeto
5    global timestamp
6
7    if(matchfrom != datematchto):
8      datematchto = matchfrom
9
10 #If the event date differs from preceeding event, prompt user for a new date, else change date to same date
       set for preceeding event.
11     if(timestamp == True):
12       datechangeto = input('Change the current field from %s to? ' % time.strftime('%Y-%m-%d %H:%M
             :%S', time.localtime(int(matchfrom))))
13       return datetime.datetime.strptime(datechangeto, '%Y-%m-%d %H:%M:%S').strftime('%s')
14     else:
15       datechangeto = input('Change the current field from %s to?' % matchfrom)
```

```
16          return datechangeto
17      else:
18        if(timestamp == True):
19          return datetime.datetime.strptime(datechangeto,'%Y-%m-%d %H:%M:%S').strftime('%s')
20        else:
21          return datematchto
```

6.2.2 GeoIP Function

The function written in Listing 6.2-4 searches for the geolocation fields within the data. If the field is found empty, it performs a geographic-IP lookup to fill these fields. The tool used is MaxMind's downloadable GeoIP database. MaxMind[1] is an open source tool collecting geographic information on a global scale of IP addresses, stored in a database that can be queried by this tool. The use of this information is free to all users and can be either queried directly or downloaded with periodic updates. This procedure can easily be mimeographed with a more advanced and accurate IP-to-Location tool such as Yong Wang's street level technique[61] discussed in Chapter 3, available in the form of paid services. For this research however, an open source tool is considered sufficient for demonstration.

Listing 6.2-4: IP address to Geographic Location Conversion

```
1
2   #!/usr/bin/env python
3
4   gi = pygeoip.GeoIP('c:/Users/Herah/Documents/Mine/GeoIP/GeoLiteCity.dat')
5
6   for line in args.infile:
7         if (args.logsource == 'tsom'):
8           timestamp = True
9           old_date_array = re.split(';', line)
10          arraylen = len(old_date_array)
11          if(arraylen > 3):
12            print(old_date_array)
13            if (len(old_date_array[15]) > 1) :
14              geosrcinfo = gi.record_by_addr(old_date_array[15]) # perform lookup from IP address
15              print(geosrcinfo['latitude'])
16              old_date_array[19] = str(geosrcinfo['latitude'])
17              old_date_array[20] = str(geosrcinfo['longitude'])
```

6.3 Core Processing

OSSIM is needed to aid the scenario objective in augmenting attack detection with geolocation. The adaptation necessary for scenario events to channel into the body of OSSIM can be done in two ways. The first, is to use the scripting plugin of OSSIM and pull in the data from a specified directory with regular expression mapping techniques. The second, is to use MASSIF to relay the data via a method such as syslog to OSSIM listening in on the specific port. Both methods are discussed, with the latter eventuating to be the more apposite of the two.

[1] *See* http://www.maxmind.com/en/geolite_city_accuracy for accuracy details per country.

6.3.1 Pilot Approach: Custom OSSIM Plugin

OSSIM encourages ease of integration with devices by providing users the means of creating custom plugins. The rudimentary requirements for writing a plugin is the instantiation of a plugin name and an unused plugin id. The pattern matching of the log event format into predefined OSSIM variables (event mapping) is carried out and the definition of a source, either process or directory from where the plugin is to periodically comb for new events.

The procedure required to initiate a plugin to connect the test data events to OSSIM for analysis and correlation is described:

1. A plugin entry is made into the configuration file for the OSSIM agent (located in /etc/ossim/agent). The entry is used to inform the OSSIM server to enable the plugin, with its details defined in step 2, when the server is restarted.

2. The scripting of a new plugin file under the name defined in the configuration entry made. Listing 6.3-5 is a plugin script written for parsing a sample data file extracted from the Windows test data:

Listing 6.3-5: OSSIM plugin script

```
  [DEFAULT]
plugin_id=9010

[config]
type=detector
enable=yes
...
source=log
location=/var/log/herahlogs/brutenewwithgis
create_file=false

[ossim-herahk-format]
event_type=event

regexp= ...[insert regular expression to pattern match log into placeholders e.g (?P<eventid>[^;]*)]...

#ossim variables that map from the regex
plugin_id=9010
plugin_sid={$eventid}
...etc
```

3. The creation of a regular expression specific to event formats, to map the meta-tags to the incoming content.

   The core of the plugin, is based on the regular expression mapping, if this is not correctly created, the plugin will not retrieve any data. OSSIM stores plugins with the correctly configured regular expressions specific to the event source of typical sources, such as mcafee and windows server, however, in this case the events have been pre-processed normalisation procedures as discussed earlier, modifying the event format.

Regular expressions

A regular expression is a text pattern that is made of ordinary characters(for example, alphabetic letters a to z) and special characters sometimes referred to as *metacharacters*. The pattern this string makes defines what strings can match the given expression[44]. The resulting regular expression shown in Listing 6.3-6 is derived to map the windows event logs.

Listing 6.3-6: Regular expression matching Windows events

```
# event log format: TBS;123423434;;3345454534;223432;...etc basically a semi−colon csvd version of
    logs
regexp="^([^;]*);([^;]*);([^;]*);([^;]*);(?P<datee>\d+);([^;]*);([^;]*);([^;]*);
([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);(?P<user>[^;]*);([^;]*);(?P<src>[^;]*);
([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);(?P<dst>[^;]*);([^;]*);([^;]*);
([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);(?P<eventid>[^;]*)
;(.*)$"
```

The validation of a sequence can prove manually difficult, therefore must be run against an expression checking tool to dissect the problems, available through python scripting checks using its existing regex libraries(regexp). Table 6.2 is a reference table highlighting sample expressions and meanings of regex(regular expression) syntax.

Regex	Meaning
^	Indicated the beginning of a string
$	Has to occur at the end of a string
.	Matches any line
*	Preceeding item must match zero or more times.
[ ]	Bracket expression. Matches one of any characters enclosed.
[^]	Negates a bracket expression. Matches one of any characters EXCEPT those enclosed.
()	Parentheses. Creates a substring or item that metacharacters can be applied to
^\s*$	A full regular expression that matches a blank line

Table 6.2: Regular expression common syntax

Mapping this is fairly straightforward, e.g *plugin_sid*={*$eventid*} and so forth. This is to be stored within the assigned directory for plugins of OSSIM, and activated in the OSSIM framework script.

4. The creation of a plugin_sid SQL file to map the OSSIM database to the data source is carried out, shown in Listing 6.3-7. This step is responsible for mapping the event

IDs to all the necessary event information needed by the SIEM and analyst in order to evaluate the context and event activity.

Listing 6.3-7: SQL sequence for database population

```
1  -- Herah Test TSOM logs
2  -- Plugin id: 9010
3
4  DELETE FROM plugin WHERE id = "9010";
5  DELETE FROM plugin_sid where plugin_id = "9010";
6
7  INSERT INTO plugin (id, type, name, description) VALUES (9010, 1, 'herah-tsom', 'Herahs TSOM
       log testing');
8  INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES
       (9010, 528, NULL, NULL, 'Successful logon', 1, 3);
9    ...
```

5. Population of the OSSIM database on plugin data

   On the OSSIM server, the SQL script created is run on the OSSIM database. Once the server is restarted the plugin should be working and present with the OSSIM interface as a recognised source of security input. The data to be collected through the plugin will be identified through the OSSIM database, and enriched with event descriptions on the OSSIM server, after the collection process.

This approach allowed OSSIM to retrieve geolocation data from the collected data sets for exploitation in the event analysis tools in the correlation phase. This satisfies the aim of enabling geolocation for security augmentation techniques, as required in the design objectives. The limitation realised through this method however is the lack of privacy-enforcing procedures performed on the geolocation before passing it through to the OSSIM correlation devices and interface. In this manner, the geolocation data is managed within the SIEM in the clear, a significant issue in regulation and compliance of user data from various jurisdictions.

6.3.2 Final Approach: Integration with MASSIF

The OSSIM plugin approach is feasible but is not a complete solution to the objectives of the study. To satisfy the privacy objective for geolocation the data needs modification procedures that ensures this condition and keeps the original event version in a secure manner for integrity purposes.

The conceptual design shown in section 5.5 is the final approach chosen where the GET is used as the anonymiser of geolocation data *before* reaching OSSIM. The plugin approach is modified to use the GET. The basic adjustments are:

1. The plugin source: Modified from *log file* to *syslog*, with events being sent from the GET. Sending through the Syslog server of the GET enables event replay and the environment to simulate real-time.

2. The format: From windows normalised format, the events are modified through the GET and conveniently converted to send directly in OSSIM format,decreasing conversion time when received on the OSSIM server.

This approach[20] attempts integration of selected MASSIF components in a non-invasive way using the existing methods and interfaces of the OSSIM platform.
Since the GET is to be used in this experiment for applicable privacy-enforcing procedures it is convenient to utilise its sender agent and replay tool, events can be sent to OSSIM through a secure channel like the REB from the GET. Therefore, the task of channeling the events to OSSIM is shifted to the shoulders of MASSIF with OSSIM to function on the receiving end. The replay tool makes use of the Syslog protocol(briefly discussed in section 5.2) more specifically, the RFC 3164 BSD Syslog Protocol, for event replay.

The protocol architecture can be seen as consisting of the following definitions; device, relay and collector. The **device** is typically known as the machine that generates messages. The **relay** is a machine that can receive and forward messages to another machine. The **collector** receives messages but does not forward it to any machine, it is also referred to as the syslog server. Some machines can hold back a section of messages, thus acting as both collector and relay[35].

Replaying the event data through syslog removes the real-time issue incurred earlier, as the OSSIM agents plant their own timestamps on the data, in which case it appears real-time. But, this brings along another issue which is that of the actual time differences in events not being taken into account. This can be solved for the simulation, by getting the timing to mimeograph the attack timing patterns, through the use of delays. This is preferred over the first attempt as a smooth flow of event transmission and approach to testing.

6.3.3 Event Correlation

Various patterns can be detected through the use of correlation of data. The link between activities in a time sequence can provide insight into the intention of a user. The following two implementations of correlations listed, were constructed to identify an attack while applying geolocation. The first is applied within a brute-force attack detection, the second applies to an unauthorised login detection.

The rule of suspicious activity from a region that consistently records a high level of malicious activity (poor geo-reputation), can be applied in a correlation rule. The entire correlation rule sequence for brute-force attacks triggered from areas with malicious geo-reputation is demonstrated in Listing 6.3-8.

Listing 6.3-8: Brute-force Geo-centered Correlation Rule

```
1 <directive id="500000" name="Brute Force Attack from US Denver City location Against DST_IP"
      priority="4">
2         <rule type="detector" name=" Authentication failure"
3               reliability="0" occurrence="1" from="ANY" to="ANY"
4               port_from="ANY" port_to="ANY"
5               plugin_id="9010"
6               userdata1="FIND:39.9"
```

```
7                   userdata2="FIND:-105.1"
8                   plugin_sid="529,4625,530,531,532,533,534,537,539">
9             <rules>
10                  <rule type="detector" name="Successful Auth (After 1 failed)"
11                        reliability="1" occurrence="1" from="1:SRC_IP"
12                        to="1:DST_IP"
13                        port_from="ANY" time_out="15" port_to="ANY"
14                        plugin_id="9010" plugin_sid="528,4624,540"/>
15                  <rule type="detector" name="SSH Auth failure (10 times)"
16                        reliability="2" occurrence="10"
17                        from="1:SRC_IP" to="1:DST_IP"
18                        port_from="ANY" time_out="40" port_to="ANY"
19                        plugin_id="9010"
20                        plugin_sid="529,4625,530,531,532,533,534,537,539"
21                        sticky="true">
22                        <rules>
23                              <rule type="detector" name="Successful Authentication (After 1 failed)"
24                                    reliability="4" occurrence="1"
25                                    from="1:SRC_IP" to="1:DST_IP"
26                                    port_from="ANY" time_out="100"
27                                    port_to="ANY"
28                                    plugin_id="9010" plugin_sid="528,4624,540"/>
29                              <rule type="detector" name=" Account Access Denial (5 times)"
30                                    reliability="4" occurrence="5"
31                                    from="1:SRC_IP" to="1:DST_IP"
32                                    port_from="ANY" time_out="400"
33                                    port_to="ANY"
34                                    plugin_id="9010"
35                                    plugin_sid="531,532,533,534,539"
36                                    sticky="true">
37                                    <rules>
38                                          <rule type="detector" name=" Successful Authentication (After 1
                                                failed)"
39                                                reliability="6" occurrence="1" from="1:SRC_IP" to="1:
                                                      DST_IP"
40                                                port_from="ANY" time_out="150" port_to="ANY"
41                                                plugin_id="9010"
42                                                plugin_sid="528,4624,540"/>
43                                          <rule type="detector" name="Authentication failure"
44                                                reliability="7" occurrence="1" from="1:SRC_IP" to="1:
                                                      DST_IP"
45                                                port_from="ANY" time_out="4000" port_to="ANY"
46                                                plugin_id="9010" plugin_sid
                                                      ="529,4625,530,531,532,533,534,537,539"
47                                                sticky="true"/>
48                                    </rules>
49                              </rule>
50                        </rules>
51                  </rule>
52                    </rules>
53
54            </rule>
55  </directive>
```

The detection of a 'log in' from different physical locations with an impossible time frame for a user to travel between the places, for example, Zimbabwe and Russia, can be applied to a correlation rule, this is demonstrated in Listing 6.3-9.

Listing 6.3-9: Suspicious user location pattern triggering

```
1
2  <directive id="500001" name="Impossible login from location with DST_IP" priority="4">
3              <rule type="detector" name="First Successful Auth"
4                     reliability="0" occurrence="1" from="ANY" to="ANY"
5                     port_from="ANY" port_to="ANY"
6                     plugin_id="9010"
7                     plugin_sid="529,4625,530,531,532,533,534,537,539">
8              <rules>
9                     <rule type="detector" name="Duplicated Successful Auth"
10                           reliability="1" occurrence="1" from="1:SRC_IP"
11                           to="1:DST_IP"
12                           port_from="ANY" time_out="15" port_to="ANY"
13                           plugin_id="9010" plugin_sid="528,4624,540" userdata1="!1:userdata1"/>
14             </rules>
15                 </rule>
16 </directive>
```

6.4 Edge Processing

The GET Framework is mainly used for normalisation, event collection and aggregation. For this scenario however, the events received by the GET are already parsed and normalised. The purpose of adapting the GET is to determine whether further functionality can be provided than the current tools used by the existing SIEM. In this particular case we look at GET in terms of data anonymisation and pre-correlation abilities. Events from the MESI scenario will be retrieved and sent to OSSIM through the GET which will present additional possibilities of anonymisation and correlation not available through the existing OSSIM correlation engine mechanisms.

In order to address privacy concerns, the MASSIF GET applies the mechanism of location anonymisation in conjunction with pre-processing of data at the edge of the SIEM, such as data suppression, aggregation and filtering. This solution limits the disclosure of personal data that cross the event collection boundaries of companies or organizations. Also, in order to maintain the full evidence contained in the source logs, the original data is transmitted to the SIEM correlation engine after proper encryption.

6.4.1 Event Location Anonymisation

The events from the managed enterprise are anonymised and the original event payload is encrypted at the edge MASSIF architecture. This mechanism allows to avoid disclosure of sensitive data out of the security data collection systems of the MASSIF tools.

Approaches to provide K-anonymity in GPS based systems are given for instance in [24]. The main idea is to round off the location coordinates in order to cloaking the detailed position of users in a wider region including at least other k-1 users.

6.4.2 Event Parsing

The GET tool is a product of project partner CINI in MASSIF. They provide its modifications for a scenario such as the GET adaptable parsers (APs) that read in the event data for normalisation(in this case, anonymisation) procedures. For the managed enterprise scenario, the following parsers were created after providing them the relevant data schemas for simulation:

- Mcafee_v1.jar, a jar file applying the mapping from the provided schemas to OSSIM normalised format for McAfee events.
- WinSer_v1.jar, a jar file applying the mapping from the provided schemas to OSSIM normalised format for Windows Server 2003 events.

The mappings provided for the jar creation are discussed in the next section. In order to pass these events through to the OSSIM sensor that can receive the events, these applied parsers will output data in OSSIM normalized event format, more specifically the required field mapping to ensure OSSIM reads the data appropriately. In addition to this rules need to be defined in the OSSIM Database for the specific event types. The database will store the event type ID's and their information. These rules, seen in Table 6.3, are used to identify the type of events that stream through from these event ID's for the Windows Server logs.

Event ID	**Rule Description**
	Logon/Logoff
528/4624	Successful Logon
540/4624	Successful Network Logon
529/4625	Logon Failure: Unknown username or bad password
530/4625	Logon Failure: Account logon time restriction violation
531/4625	Logon Failure: Account currently disable
532/4625	Logon Failure: The specified user account has expired
533/4625	Logon Failure: User not allowed to logon at this computer
534/4625	Logon Failure: The user has not been granted the requested logon
537/4625	Logon Failure: An unexpected error occured during logon
539/4625	Logon Failure: Account locked out
576/4672	Special privileges assigned to new logon
538/4634	User logoff
551/4647	User initiated logoff
	Account logon
672/4768	A Kerberos authentication ticket (TGT) was requested
673/4769	A Kerberos service ticket was requested
675/4771	Kerberos pre-authentication failed
680/4776	A domain controller attempted to validate account credentials

Table 6.3: Windows Server Rules fed into OSSIM database for relevant event ID's

In subsection 6.3.1, steps 4 and 5 complete the rule definition for the event format. The SQL file constructed for the OSSIM plugin matches the specific event format IDs and type definitions for recognition within the OSSIM SIEM regardless of the source of the plugin.

6.4.3 Event Schema Mapping

To send the events through to OSSIM with the afore-mentioned parsers the event fields need to be mapped to the OSSIM normalized event format, to ensure OSSIM will read the data appropriately. Listing 6.4-10 depicts the typical format of an event converted from it's source into OSSIM normalized format.

Listing 6.4-10: Event in OSSIM normalized format

```
2010−05−30 13:15:49,441 Output [INFO]: event type="detector" date="1275239752"
sensor="192.168.178.201" interface="eth0" plugin_id="4003" plugin_sid="7"
src_ip="192.168.178.20" src_port="4445" dst_ip="192.168.178.200" dst_port="22"
username="root" log="May 30 13:15:52 dmz01 sshd[12980]: Accepted password for
root from 192.168.178.20 port 4445 ssh2" fdate="2010−05−30 13:15:52" tzone="0"
```

The necessary fields listed with their purpose are described in Table 6.4.

Attributes	***Description***
TYPE	Type of event: detector/monitor, reserved for OSSIM internal purposes
DATE	Date on which the event was generated provided by the event source
SENSOR	IP address of the generating sensor or the source
INTERFACE	Name of the network interface associated with the event
PLUGIN_ID	Identifier of the data source (plugin) that generated the event
PLUGIN_SID	Type of event specific to the data source that generated the event
PRIORITY	Event priority, used in risk calculation
PROTOCOL	The communication protocol used (TCP, UDP, ICMP, etc.)
SRC_IP	Source IP address of the generated event
SRC_PORT	Source port of the generated event
DST_IP	Destination IP address of the generated event
DST_PORT	Destination port of the generated event
LOG	the original log entry
DATA	Stores the event payload, can be used to store plugin specific data
USERNAME	User generating the event, mainly used in HIDS (Host-based intrusion detection system) events
PASSWORD	Password used in an event
FILENAME	File used in an event, mainly used in HIDS, can be also used in events where the normalizing plugin can recognize a generic filename
userdata1...userdata9	These fields can be used to store arbitrary data being relevant to the plugin

Table 6.4: OSSIM Event Format Description

6.4.4 Attribute Mapping

Table 6.5 and Table 6.6 show the field mapping required for the two main event sources that will be using OSSIM. Rows highlighted in red are OSSIM fields not used or are assigned by OSSIM rather than the event source.

Attributes	*Column*	*Fieldname*
TYPE		
DATE	4	internaltimestamp
SENSOR		
INTERFACE		
PLUGIN_ID		
PLUGIN_SID	34	EventID
PRIORITY		
PROTOCOL		
SRC_IP	16	sourceIP
SRC_PORT	22	sourcePort
DST_IP	23	DestinationIP
DST_PORT	29	DestinationPort
LOG		
DATA		
USERNAME	14	username
PASSWORD		
FILENAME		
userdata1	20	SourceIPGeoLat
userdata2	21	SourceIPGeoLong
userdata3	27	DestinationIPGeoLat
userdata4	28	DestinationIPGeoLong
userdata5	33	InternalUsecase
userdata6	50	LogonID
userdata7	51	LogonGUID
userdata8	30	SourceThreat
userdata9	31	DestinationThreat

Table 6.5: MESI Windows Server OSSIM Event Mapping

It is important to note EVENT_TYPE is not required to be mapped into OSSIM as the information can be retrieved using the EVENT_ID which maps from the event rule table already created in OSSIM for this event type.

Attributes	*Column*	*Fieldname*
TYPE		
DATE	5	internaltimestamp
SENSOR		
INTERFACE		
PLUGIN_ID		
PLUGIN_SID	58	EventID
PRIORITY		
PROTOCOL		
SRC_IP	28	sourceIP
SRC_PORT		sourcePort
DST_IP	39	DestinationIP
DST_PORT	53	DestinationPort
LOG		
DATA		
USERNAME	60	username
PASSWORD		
FILENAME		
userdata1	32	SourceIPLat
userdata2	33	SourceIPLon
userdata3	43	TargetIPLat
userdata4	44	TargetIPLon
userdata5	6	AgentGUID
userdata6	55	ThreatName
userdata7	57	ThreatCategory
userdata8	59	ThreatSeverity
userdata9	61	ThreatType

Table 6.6: MESI Mcafee OSSIM Event Mapping

Here again the EVENT_TYPE is not required to be mapped into OSSIM as the information can be retrieved using the ThreatEvent_ID which maps from the event rule table already created in OSSIM for this event type. In both event source mapping requirements, geolocation data is fed in through the OSSIM void type userdata fields

In conclusion, the event data manipulation performs the following cycle, summarised in Figure 6.1.

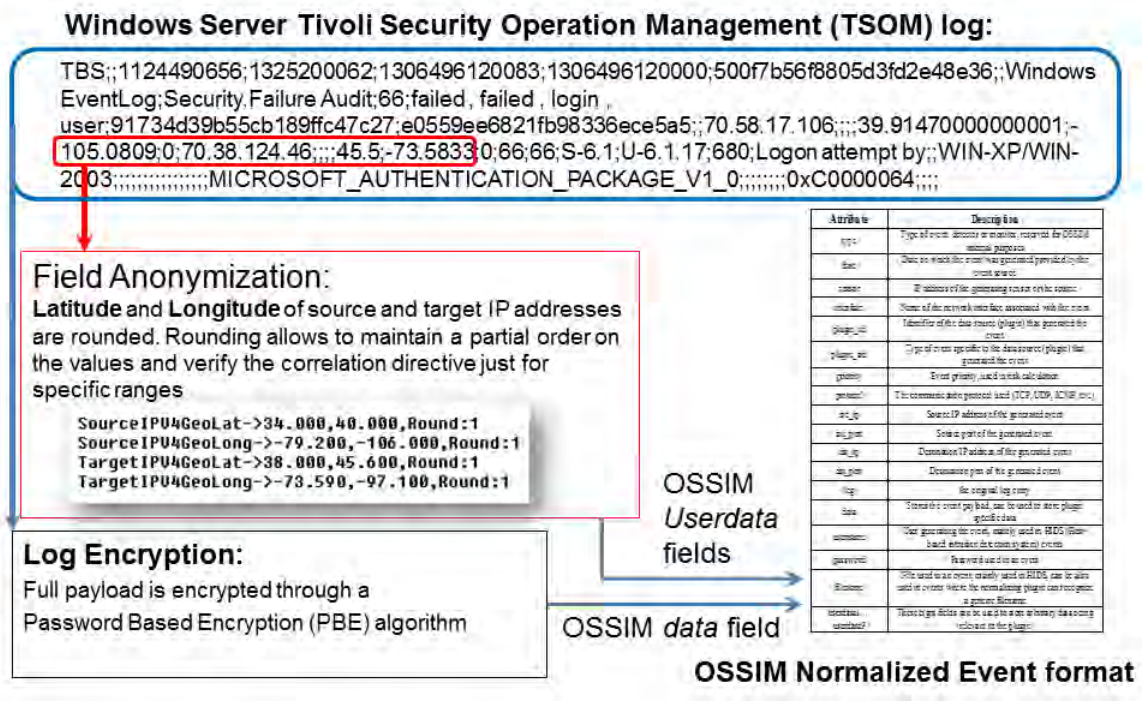

Figure 6.1: Anonymisation through Generalisation

This satisfies the design objectives determined for the events and format requirements through the MASSIF elements and within the OSSIM SIEM solution.

6.5 The Final Integration

The Misuse cases identified for the Managed Enterprise Service Infrastructure scenario can be reproduced though the use of existing correlation directives modelled for a SIEM system. The first step to reproducing these cases is to identify the sequence per use case. The Managed Enterprise Service Infrastructure scenario is based on a managed IT outsource environment, where events from multiple sources are collected centrally. The events typically include operating system, router, firewall and desktop security related activities. The brute-force attack was chosen, for the proof-of-concept demonstration of a test misuse case as it encompasses the study objectives and is applicable in several contexts.

6.5.1 MC 5.5.1 Brute-force Simulation

In order to show the MASSIF feature to protect locational data, we considered the misuse case defined as Brute force attack (Misuse case 5.5.1, MASSIF Deliverable D2.1.1 [41]). The Brute force is a common method for password hacking and is particularly relevant in large enterprises due to the many users being the likely weak point for infiltration.

The full details of the misuse case[41] simulated in the demonstration experiment is given in Table 6.7.

MISUSE CASE 2	Brute-Force Password Attack	
	1	System prompts user for login details
	2	Unauthorised user performs multiple attempts to log in using many username/password combinations
	3	System authenticates the authorised user's successful combination and creates a new session.
EXTENSIONS	**Step**	Branching Action
	-	
VARIATIONS	**Step**	Branching Action
	3a1	Detection Mechanisms discover a brute-force attack attempt and lock the computer to prevent any further combinations from being attempted.
Exceptions	-	-
Other Information	Most systems already have in place measures which identify failed authentication attempts within close succession. The result often locks the user from accessing the machine. Through early detection of port scans or automated login attempts, the damage from such an attack being successful can be reduced or eliminated	
OPEN ISSUES	N/A	

Table 6.7: Brute-force Misuse Case MC5.5.1

The directive (Listing 6.3-8), was created for the testbed simulation to verify the retrieval of geolocation through SIEMS, and their use in detection and reponse. SIEMs stucturally, are heavy frameworks and internally complicated. It's significant to consider the ease of integration of geolocation data into these tools and the OSSIM solution, to advocate the fundamental reasons in investing the efforts to exploiting the data.

6.5.2 The Testbed

The test solution makes use of three virtual machines named VM1, VM2, and VM3 respectively. These machines and their specifications are discussed in Table 6.8.

Virtual Machine	Operating System	Description
VM1: MASSIF-RES	Windows 7	Hosts Syslog tool for event replay and the RES visualization interface
VM2: OSSIM	OSSIM 4.0.2 Debian	The OSSIM solution with REB node
VM3: MASSIF-GET	Ubuntu 10.04	Hosts the GET Framework and RES tool

Table 6.8: Testbed virtual machines and their specifications

VM1 hosts the Syslog generator tool that replays logs generated by the MESI infrastructure. The logs contain evidence of brute force attacks and the GPS locations of users. VM3 hosts the GET framework configured with a proper parser for Windows Server events processed by the Tivoli Security Operation Manager (TSOM). The GET communicates with the OSSIM Server through a REB node. VM2 hosts a second REB node that collects the events sent by the GET and passes them to the OSSIM Server. Once alerts are triggered, the events are stored on the RES, which is installed on the VM3. Finally, the RES visualization interface is installed on the VM1 machine. Figure 6.2 demonstrates the experimental setup for scenario provider validation with MASSIF.

6.5.3 Execution Process

1. The Syslog client tool replays raw events, as shown in Figure 6.3, using the Syslog tool of the GET. The event sets contains patterns of a brute-force attack targeting a Windows Server machine in the environment.

2. Figure 6.4 demonstrates initiation of the REB server for event transfer to OSSIM. The GET, seen in Figure 6.5, is used to collect, parse and translate these events into the OSSIM Normalized format. Figure 6.6 demonstrates the replay of event logs using rsyslog to be fed into the GET tool. The cycle of events from GET to OSSIM with the REB tool is shown in Figure 6.7.

3. The GET encrypts each original log generated by the monitored infrastructure and converts it into the OSSIM Normalized event, the formatted logs have an anonymised version of the fields required for correlation on the OSSIM server. As seen in Figure 6.8, the REB then transfers these encrypted events in OSSIM format to the OSSIM server.

4. OSSIM receives these events through the REB and applies the correlation directive. The alarms triggered from the directive are stored both on the OSSIM alarm database and on the RES. The RES contains the alarms and the triggering events with the encrypted payload. These alarms stored on the RES, seen in Figure 6.9, can be decrypted only by authorized parties(password-based encryption).

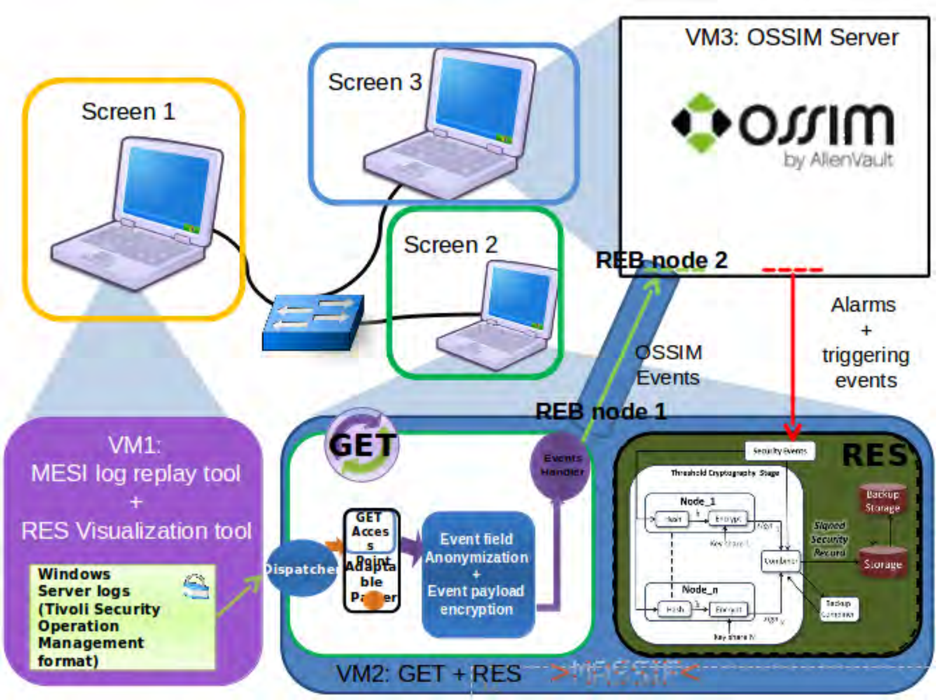

Figure 6.2: Experimental setup with MASSIF and Alienvault

Figure 6.3: Original raw event data showing accurate geolocation

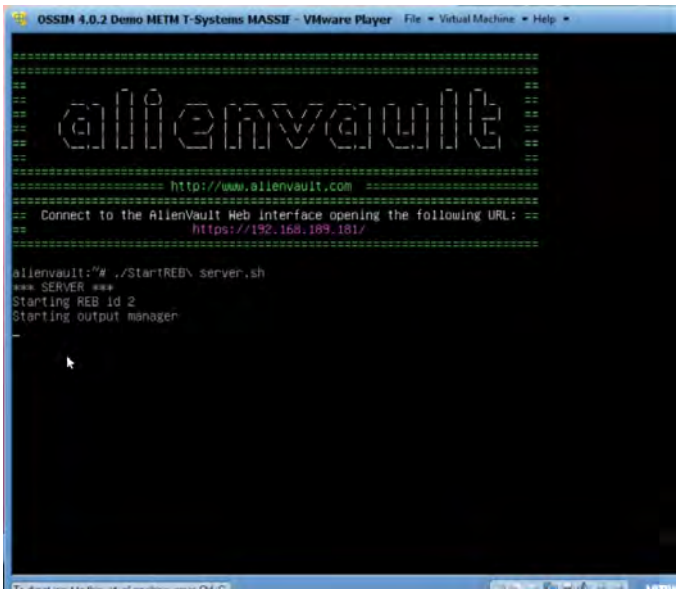

Figure 6.4: Starting the REB server, ready to channel parsed events from the GET to OSSIM

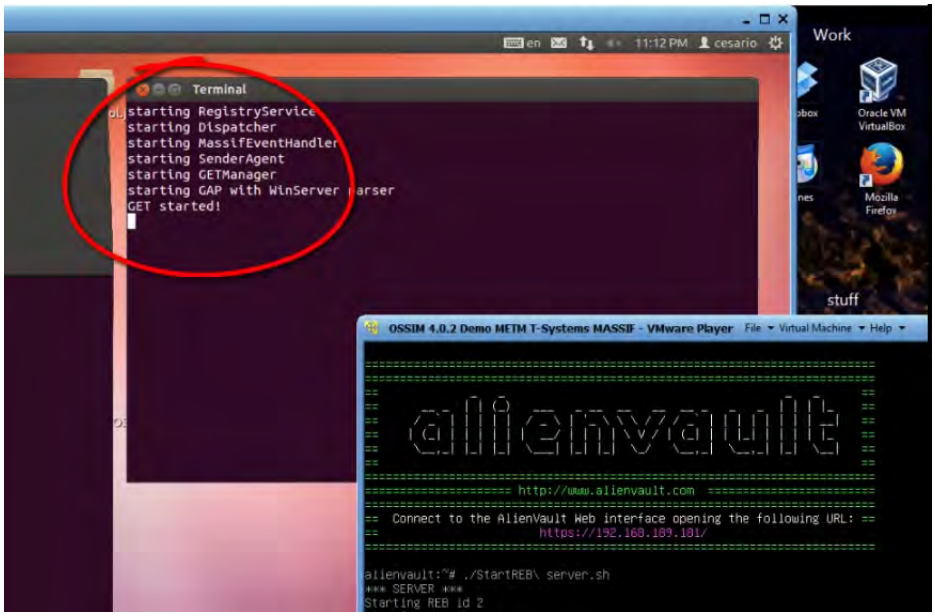

Figure 6.5: Starting the GET tool

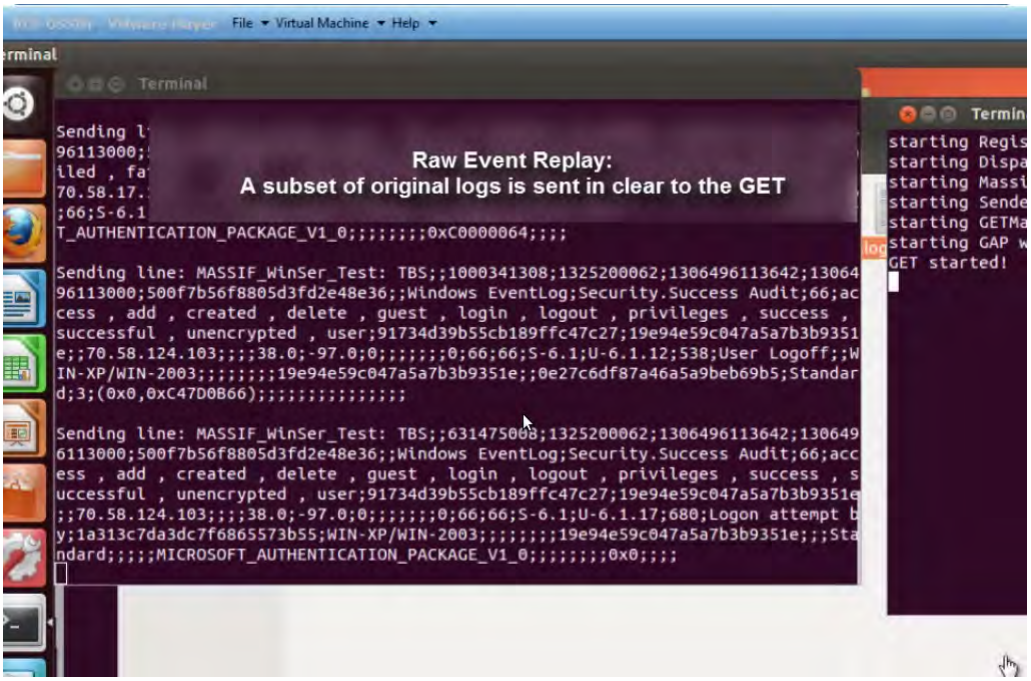

Figure 6.6: Event replay of these raw logs to be sent to the GET

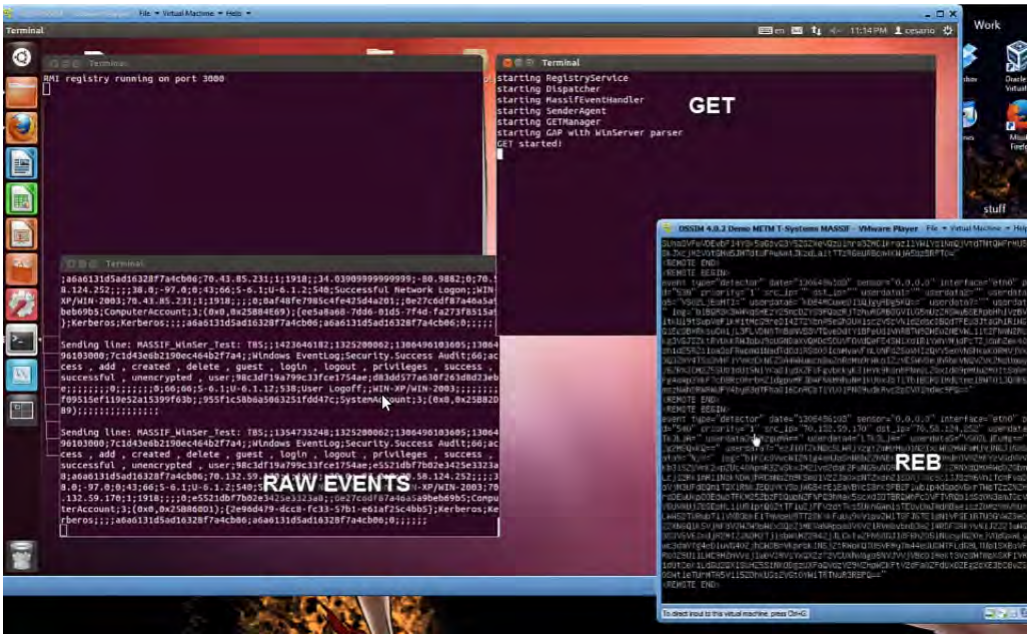

Figure 6.7: OSSIM, REB and GET

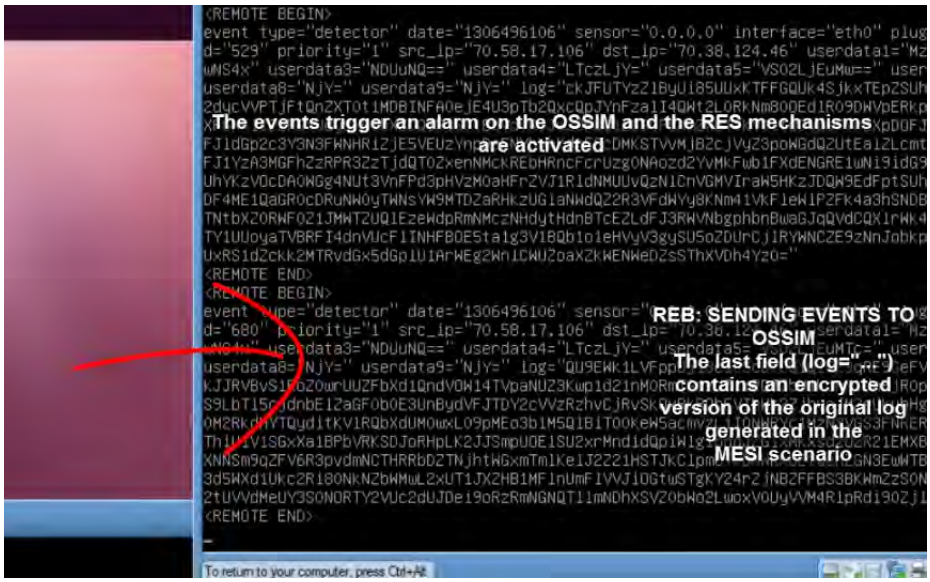

Figure 6.8: REB sending encrypted events in OSSIM format from GET to OSSIM server

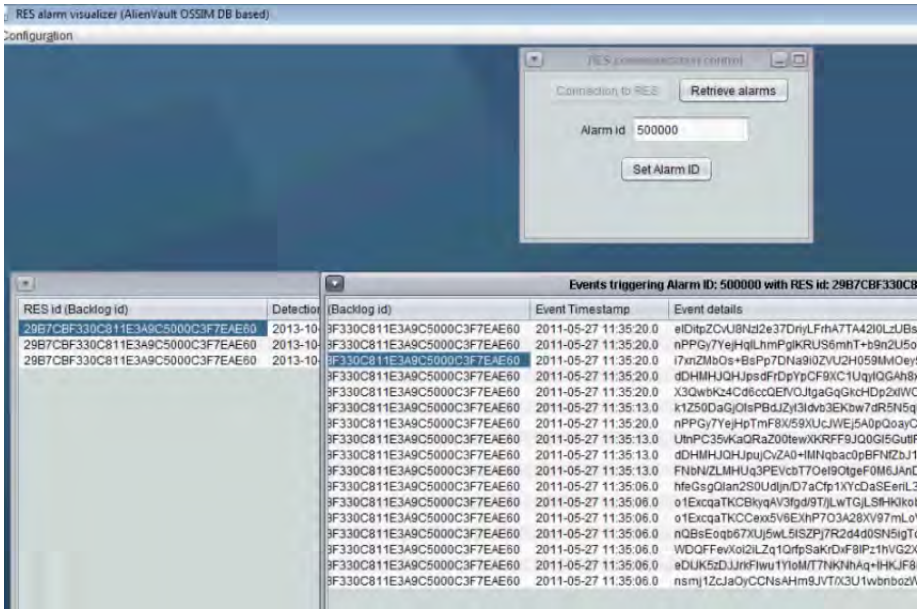

Figure 6.9: OSSIM generated alarms (left) and log events triggering the alarm (right) stored in RES and can be retrieved by plugin id(top)

6.6 Summary

With the Managed Enterprise simulation we aimed to determine novel security methods for SIEMs like anonymisation and whether these can be integrated in commercial SIEMs. The commercial SIEM considered in this demonstration was OSSIM by AlienVault.

The demonstration showed capabilities provided through the MASSIF GET tool to preserve privacy and confidentiality. Sensitive information was anonymised and the original events were encrypted at the edge of the SIEM by the GET. Data fields considered as 'private or potentially sensitive were specified to be geographic data, other data fields provided in the logs where encrypted as not to be utilised within the SIEM.

In particular, the need was identified to ensure privacy of users in this managed environment where services are outsourced and data handling is managed in areas of different jurisdictions.

Also, the demonstration used the operations of the RES system. The RES guarantees unforgeability of security events and implements the least persistence principle, i.e. just relevant security events are stored and maintained for forensic purposes. Moreover, data on the RES is demonstrated to be visible only to authorized parties.
The demonstration carried dissemination of events through the REB, which establishes a reliable communication between the GET component and the OSSIM Server.

Finally, the demonstration shows the OSSIM SIEM product seamlessly integrated with the MASSIF components mentioned above. The objectives of the demonstration that were achieved can be summarized as follows:

- Show confidentiality preservation capabilities:
  At the edge, encryption of the event payload is performed. At the core, security alarms are stored and restricted access is guaranteed.
- Show novel privacy implementation abilities:
  Event field anonymisation is applied to location data to ensure the privacy of users while still enabling the data to be used for security analysis purposes.
- Use of resilient mechanisms in the data layer of MASSIF:
  Communications between the edge and core sides relies on the Resilient Event Bus. The Alarm storage is based on Resilient Event Storage facilities(provided by CINI).
- Show ease of integrative abilities with an existing SIEM
- Integration between MASSIF data layer components and the OSSIM SIEM.

Chapter 7

Testing

The implementation discussed in the previous chapter considered a brute-force attack from a windows server, in a standard managed enterprise environment monitored by a SIEM. OSSIM is used to emulate a typical live SIEM environment and is enhanced with anonymisation capabilities and the accomodation to read geographical location data. The objectives of the demonstration centrally focused on the validation of the hypothesis supporting the purpose of this study. Security detection within the correlation engine - at the data and information layer, encompassed the application of geographical data in an attack detection strategy. Novel privacy implementation abilities were incorporated in the earlier stages of processing - at the information gathering and normalisation level, for this geographical data.

This chapter proceeds to assess the data sets used in the experimental simulation(section 7.1), the performance of the integration of various tools needed for the experiment, particularly the GET tool (section 7.2), the anonymisation technique implemented for the location data (section 7.3), and finally the application of geolocation in the the correlation phase of the SIEM(section 7.4).

7.1 Event Preparation

The data used in the experimental run were collected from two data sources in particular, Windows Log Parser and Windows Event Log. The collection amounted to 8,612,030 and 8,133,396 respectively. This totals the event sets retrieved for testing to 16,745,426 from Windows Servers, collected by 47 sensors.
The data was analysed in two main considerations, the misuse case brute-force attack and credible geographical content.

7.1.1 Misuse Case Data

The graph in Figure 7.1 shows the events statistics for a sample data set modelled to emulate the brute-force attack from a windows infiltration attempt. The data set is transported through the client-server protocol Syslog to the GET, which acts as the collector, peforms processing of events and sends it through to OSSIM, the acting SIEM for result and detection evaluation.
The number of events identified by OSSIM when received by the GET totalled 98%, OSSIM

did not recognise 2% of the received events. The loss was identified in events with unidentifiable event type IDs, that were not present in the the previously loaded event rules (see Table 6.3, section 6.4) in the event parsing phase of implementation into the OSSIM database.

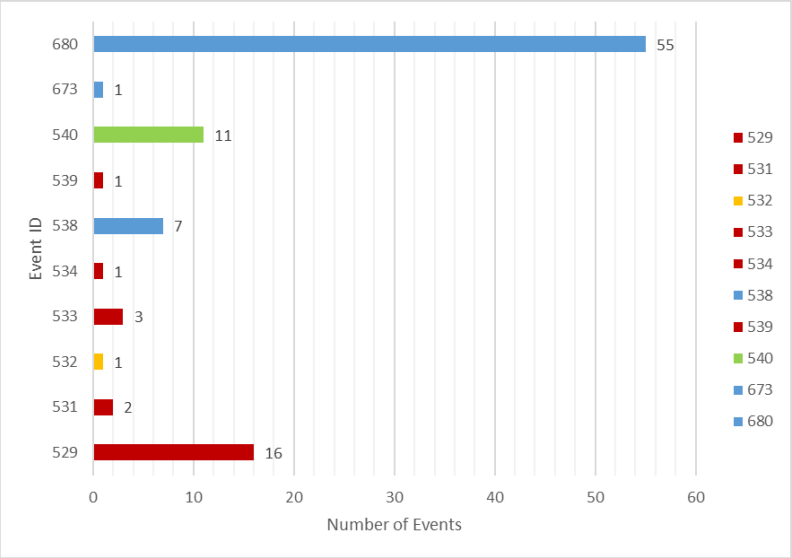

Figure 7.1: MC-5.5.1 Event Frequency Distribution

The foreign/nonsensical IDs ascertain the default behaviour when an unknown Windows log/ID is encountered by the SIEM. As observed through the test runs, the event is not matched to a default ID or included at all in the SIEM correlation rather, discarded, and can only be found through a manual analysis of the raw event logs. This is useful for noise reduction and clarity but is an important consideration for high volume data, ensuring all relevant event IDs for the specified data source are accounted for, particularly when changes or additions are made of the source format. As an example, the *Specified user account has expired* identifier for Windows servers is either 532 or 4625, both valid IDs for the rule.

The events colored in red indicate IDs in the *login failure* categories (539, 534, 533, 529), 16% of the data set indicated malfeasance or suspicious activity, this was the data needed for the identification of a brute force according to the definitions specified in the OSSIM correlation directive. The basic condition for the trigger required 15 consecutive login failures from the same IP address within a limited time frame. 62% of the data (680, 538, 673, 540) contained typical login data, authorised logins, logouts and authentication requests. The events indicated in green (540) identified successful authentications, and finally the suspicious events(540) such as requesting login through an expired account indicated in yellow.

7.1.2 Geographic Data

Geographic location data is comprised of a pair of co-ordinates. In the case of the event sets used, the data is retrieved in geographic lattitude and longitude format. The number of lattitude degrees changes in the direction of north-south and the longitude degrees in an east-west direction. The degrees of lattitude and longitude are the units for determining an

exact location on earth, and carries two attributes of consideration, precision and accuracy. The precision of the degree does not verify the accuracy of the location, and vice versa. Both need to be evaluated and considered seperately in their sufficiency for the purposes of experimentation.

Locational Precision

First, the level of precision necessary for geographic location data is discussed. In an effort to determine this, the physical distance between a 'degree' of longitude/ lattitude needs to be calculated. Consider the equator, the center of the two North and South poles, it is divided into 360 degrees of longitude. To ascertain the estimate metre per degree, the circumference of the equator must be divided by the 360 degrees. Therefore, where dm is estimate metres per degree, the calculation is :

$$dm = \left(\frac{2\pi \cdot radius}{360}\right) \tag{7.1}$$

The radius of the semi-major axis of the Earth at the equator is recorded at 6,378,137.0 metres[5]. Therefore, the calculated equatorial circumference is 40,091 146.8 metres, dividing by 360 gives the estimate metres for a degree of longitude, approximately 111.36km.

This calculation carries a certain margin of error, because the earth is not uniform and its attributes constant as the assumption this calculation makes. Considering the earth as a spherical shape, moving from the equator towards a pole a degree of longitude is multiplied by the cosine of the latitude decreasing the distance to approach zero at the pole. A degree of lattitude is therefore, either less or the same distance. When reaching the poles, the lattitudes shrink towards the poles decreasing the distance measure of a degree, but this is only significant over 80 degrees lattitude. Another factor is that the earth is in fact not purely spherical and is rather defined as an an oblate speheroid, this gives a discrepancy at an estimate error of 0.3%[58]. If one considers the margin of error in the context of determining the precision level of location, the offset can be ignored with the current estimation sufficient for judging the weight of geographic degree in physical distance.

Table 7.1 shows[63] the physical relation to the number of decimal places required for a particular precision.
In conclusion, the level of precision at four decimal places can be seen as a sufficient threshold requirement for determining a user location, with the precision working in a range of around 11 metres. For the purposes of this study data this is the maximum level considered in the experimental scenario. The geographic location of users with a precision of this level is enough indication for user identification should that be necessary in a jurisdication that explicitly permits the use of geolocation data for authentication procedures under user consent.

Digit	Approx. Distance	Size estimation
Units	111 km	Determines roughly the large state or country the point is in.
First Decimal (0.0)	11.1km	Can distinguish between cities.
Second Decimal (0.00)	1.1km	Narrows down to differentiation between villages.
Third Decimal (0.000)	110m	Can identify large facilties, such as a university campus or institution.
Fourth Decimal	11m	Identify a patch of land, the level of granularity is typical to GPS unit accuracy.
Fifth	1.1m	Can distinguish between trees, commercial GPS units can achieve this only with differential correction[63].
Sixth	0.11m	Can track the smallest details, especially in slow-travelling objects such as glaciers.
Seventh	11mm	Good for surveying needs, reaches the limit for a GPS-based instrument.
Eighth	1.1mm	Very sensitive data measurements, tectonic plate movement for example
Ninth	110microns	Level of microscopy, when determining a physical location, this is pointless precision.

Table 7.1: Reference table of degree precision to distance

Locational Accuracy

The managed enterprise data sets from the windows servers containing geographic co-ordinates fields, were obtained from sources typically connected to high-level accuracy sources, such as GPS and mobile devices or WiFi router data. The level of precision from these devices ranged between 0.4 to 0.6 decimal degrees.
However, some geographic data fields were not complete, in cases where some data sources did not publish the information. To handle this, events from the data sets that revealed empty location fields due to a missing sensor contribution or location source feed were populated with information from the online MaxMind database, using the conversion of IP address to geographical location. The accuracy of the data from this database varies depending on the country.

Table 7.2 shows the percentages of data accuracy, retrieved from the developers of Maxmind concerning the geo-database *GeoCity* used for the experimental evaluation.
If one compares the figures given for each country, the lowest accuracy is at 49% in Finland and the highest 98% for Singpapore.

Examining the data accuracy levels in Figure 7.2, the average accuracy is seen dependant stringently on the specific area of the globe the data comes from. Only one country is under the 50% accuracy threshold with 19 over 60% accurate. The level of incorrectly resolved information has a maximum of 34% with the average under 20%.

Country	Correctly Resolved	Incorrectly Resolved	Unresolved
Australia	66%	28%	6%
Austria	70%	15%	15%
Belgium	74%	4%	21%
Brazil	73%	19%	8%
Canada	84%	14%	3%
Denmark	83%	7%	10%
Finland	49%	14%	37%
France	63%	29%	8%
Germany	74%	19%	7%
India	50%	34%	16%
Italy	60%	27%	12%
Malaysia	71%	21%	7%
Netherlands	74%	6%	20%
New Zealand	66%	25%	9%
Norway	79%	10%	11%
Poland	56%	33%	11%
Singapore	98%	0%	2%
South Africa	71%	21%	7%
Spain	76%	15%	9%
Sweden	64%	14%	23%
Switzerland	59%	10%	32%
Turkey	77%	16%	8%
United Kingdom	71%	19%	10%
United States	81%	15%	4%

Table 7.2: Percentage data accuracy by Country, sourced from MaxMind

To summarise, the percentage level of accuracy estimated for IP address to geographic location is above 60 for 83% of the countries listed.
Considering all freely available IP-to-GeoLocation services, Maxmind has positive results considered sufficient for the purposes of this demonstration in the cases where geographic data must be estimated with just an IP address available.

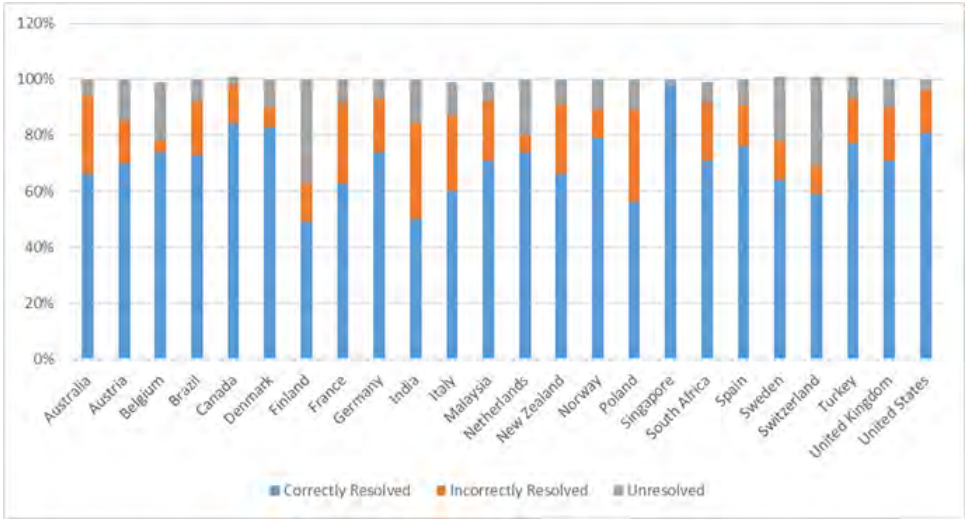

Figure 7.2: Geolocation data accuracy trend for selected countries

7.2 SIEM Performance in a Managed Enterprise

The experimental simulation was performed on a Windows i7 dual core, Intel HP laptop, with 8GB ram capacity, 500GB storage space. The tools of MASSIF and the OSSIM solutions were installed in Virtual Machines, running Ubuntu and Debian respectively, using VMWare virtual box software. The measured rates should be considered with these capacities to properly evaluate the performance of the tools and their integration.

7.2.1 Global Event Collection Rate

The first test concerned the rate of event collection, we assessed the number of events collected per time unit(for all collectors and log sources). This is to ensure the events replayed through the tools used (the GET tool), meets the performance requirement of a managed enterprise environment. The rate of collecting must meet the needs of a managed enterprise expectation to affirm the case of running this in a live environment. If bottlenecks, or event loss, occurs then this will be a negative factor while the ability to handle all submitted events will be a positive evaluation.

The results for the varied rate of submission are shown in Table 7.3. The prepared data sets were replayed to the GET at different submission rates to assess the collection ability of the tool and its effect in acting as the first point of entry and data collector for the SIEM.

In a typical managed enterprise environment, large reference clients are seen in the range of 10 000 - 20 000 EPS (events per second), or higher if network traffic is also included.
An estimate given by the source providers of the test data used in this study state around 30 000 EPS is seen sufficient to handle around 11 500 Desktops, 300 Servers and an estimate total of 225 network devices. A large industry reference SOC In the USA has reported peak processing in the order of 70 000 EPS. Commercial SIEM appliances typically handle 30 000 EPS before an additional appliance needs to be introduced.

No of events submission delays (s)	Collection time First event (HH:MM:SS)	Collection time Last event (HH:MM:SS)	Resulting time in (events per second)
100 Delay-0	18:28:34	18:28:35	100
506 Delay-0	23:48:52	23:48:53	506
1026 Delay-0	13:54:55	13:54:56	1026
1026 Delay-2	14:10:23	14:10:26	342
30000 Delay-0	13:33:28	13:34:01	909
30000 Delay-2	14:23:22	14:24:51	337
70954 Delay-0	14:17:33	14:18:28	1290

Table 7.3: Global Event Collection Rate Statistics

The observed contribution volumes in large enterprise environments are:

- 5 EPS multiplied by the number of servers in use with peak reaching 20 EPS at high volumes.
- 150 EPS multiplied by the number of Authorisation Service Control (ASC) Centers with peak of 1000 EPS at high loads.

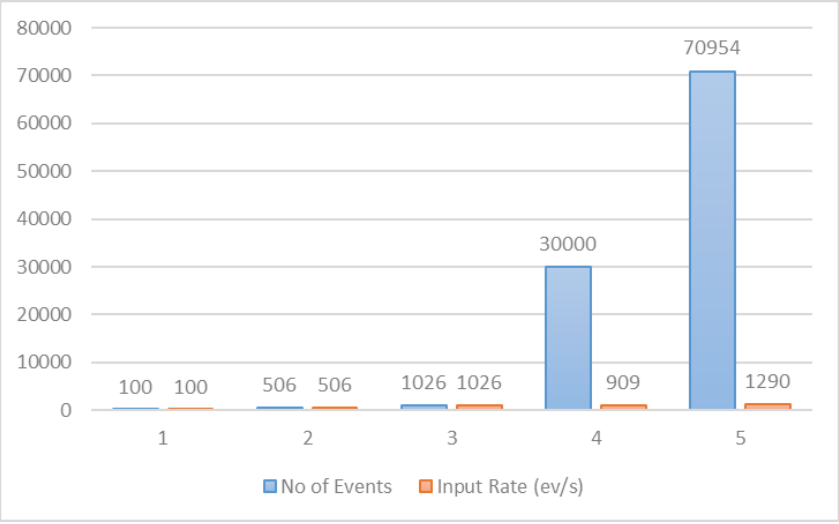

Figure 7.3: Global event collecting input rate

Figure 7.3 compares the rate at which data is sent into the system and the rate at which all events are collected in the system. It can be seen from the graph, the highest rate of events

coming into the collector is around 1000 EPS, the maximum rate at which the system can be seen to handle input rates.
Regarding the global event collecting input rate seen, for the current scenario the events are stored in batch format, the processing time in a batch was evaluated and seen to be sufficient for this rates of event delivery in the actual environment.

We observed a rate of around 1000 EPS per GET instance which satisfies the performance requirement of this scenario with further observing that multiple instances of the GET can be installed on the same machine. Thus the with two instances of the GET, and can further increase its scale to accommodate copious event rates through this use of multiple GET instances.

7.2.2 Output Rate for Output Processing

In the replay of sample log files we need to determine whether the tools used can process events at the offered submission rates. If bottlenecks occur, or processors exhaust memory, processing power before median loads are reached then this would indicate that the collector(in this simulation it is the GET) does not have the processing capacity to handle event submission for a managed enterprise service environment. Scalability techniques should be applicable to data from this scenario, and be able to process at the measured event rates.

Submission delays (s)	**Input time First event (HH:MM:SS)**	**Output time Last event (HH:MM:SS)**	**Resulting time in EPS(events per second)**
0	14:43:10	14:43:30	0.20
50	15:46:37	15:46:57	0.20
500	15:54:32	15:55:21	0.49
1000	16:16:13	16:17:36	0.83

Table 7.4: Output Rate for output processing statistics

A summary of results with test data sets for the MESI brute-force misuse case is shown in Table 7.4. Delays have been tested from the minimum (ideally 0, around 0.1ms) delay possible on a real machine to a maximum range which includes the possibility of slow attacks in addition to fast brute attack attempts.

Figure 7.4 compares the rate at which data is sent into the system and the resulting time taken for all events to be correlated once in the system. It can be seen from the graph the processing time is not dependant on delivery times for incident detection, the lowest EPS rate from 0.20 to 0.83 for the largest delay time.

Confirmation of the processing ability of the GET at varied rates can be deduced from these results based on the data set collected for the Managed Enterprise Service scenario.

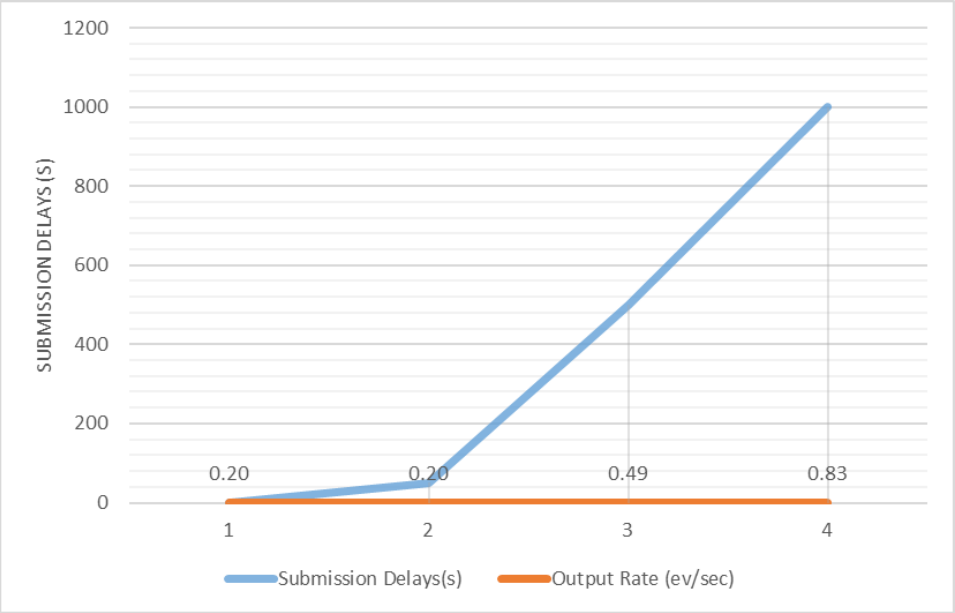

Figure 7.4: Output rate of processing comparison rates

7.2.3 Processing Time

The third performance test carried out was to measure the ability of the system to process input data and to provide meaningful results in real-time.
The duration of processing time is evaluated between the timestamp of delivery and detection in the event pattern sequence. This time represents the processing time of the GET tool to analyse and OSSIM to detect the event set anomalies.
Regarding the simulation, the time taken for test input data to be processed by the GET was recorded at an average of 20 seconds for the Brute-force misuse case. This statistic includes the identification of brute force patterns from correlation procedures by OSSIM.

7.3 Geolocation Anonymisation

The method of geolocation anonymisation implemented in this study, adopts a method of generalisation. Where k is the precision value and $k \in \{0, 1, 2, 3, 4\}$, defining the decimal degree of precision for a geographic co-ordinate. Therefore, an anonymisation level of 2, reduces the precision to a 0.2 figure. To determine the cloaking effect achieved for a user in the different level of granularities, the distance covering a range of precision needs to be calculated. The co-ordinate range for a location (x, y) where x is lattitude, y is longitude and $gen(p, k) =$ value of p at precision k, can be defined as;

$$(gen(x, k), (gen(x, 0) + (1 - gen(1, k))), gen(y, k) - (gen(y, 0), (1 - gen(y, k))))$$

For example, if k is 2 and (x, y) is (45.2314 , 87.3333), the range is determined from the above as (45.23, 45.99) , (87.33,87.99). The longest distance possible between any two points in this area would be the *diagonal* line calculated of the rectangular area covered,

this can be computed by finding the distance from the maximum bounds $(gen(x,0) + (1 - gen(1,k)), (gen(y,0) + (1 - gen(1,k))$ to the minumum bounds $gen(x,k), gen(y,k)$ of the co-ordinate range.
The area of cloaking for each level of precision (referred to as anonymisation level) can be evaluated using the distance in metres of this diagonal as a comparative measure. Assuming the basis of a spherical earth, the physical distance covered by th diagonal can be calculated through one or more applicable formulas discussed below.

7.3.1 Applicable Formulae

To determine the distance between two points on earth, depending on accuracy and complexity the following methods can be used:

- Equirectangular approximation
- Spherical Law of Cosines
- Haversine

The basic formula approach is the Equirectangular approximation, this uses pythagoras theorem on a geographic projection. This is most applicable for very small distances with more weight given to performance over accuracy. The accuracy is sufficient along the meridian but varies on other positions on the globe depending on bearing ,distance and lattitude[58].

$$x =\triangle\ \lambda \cdot \cos\varphi \tag{7.2}$$

$$y =\triangle\ \varphi$$
$$d = R \cdot \sqrt{x^2 + y^2}$$

Where φ is latitude, λ is longitude, and R is the earths radius.

The spherical law of cosines can calculate an estimated precision to a few metres on the earth[58], with more focus on calculation accuracy. The formula is simply;

$$d = acos(sin\varphi 1 sin\varphi 2 + cos\varphi 1 \cdot \cos\varphi 2 \cdot \cos\ \triangle\ \lambda) \cdot R \tag{7.3}$$

Where φ is latitude, λ is longitude, and R is the earths radius.

Both formulae discussed are a better preference over the Haversine which is far more complex numerical computation. The formula calculates the great-circle distance, 'as the crow flies', between two points[58].
However, the Haversine is the best option for accuracy even at small distances unlike the spherical law of cosines. As accuracy is a higher concern over performance in this case, to calculate the diagonal distance the Haversine formula is adopted to obtain accurate comparisons.

7.3.2 Haversine

The formula for calculating the great-circle distance is as follows:

$$a = \sin^2(\triangle\ \varphi/2) + \cos\varphi 1 \cos\varphi 2 \sin^2(\triangle\ \lambda/2) \tag{7.4}$$

$$c = 2atan2(\sqrt{a}, \sqrt{(1a)}) \tag{7.5}$$

$$d = \rho \cdot c \tag{7.6}$$

Where φ is latitude, λ is longitude, and ρ is the earths radius.

Assuming the radius of the earth is 6,378,137 metres, $\rho = 6{,}378{,}137$ m. The first step requires conversion of the lattitude and longitude to spherical co-ordinates. φ and λ can be converted to radians by multiplying by $2\pi/360°$, c would be the angular distance in radians.

Given two points in spherical coordinates $(\rho, \lambda 1, \psi 1)$ and $(\rho, \lambda 2, \psi 2)$, the arc made from connecting the points[12] is:

$$c = arccos(sin\varphi 1 sin 2 cos(\lambda 1 - \lambda 2) + cos\varphi 1 cos\varphi 2). \tag{7.7}$$

The great circle distance between the two locations is $\rho \cdot c$.

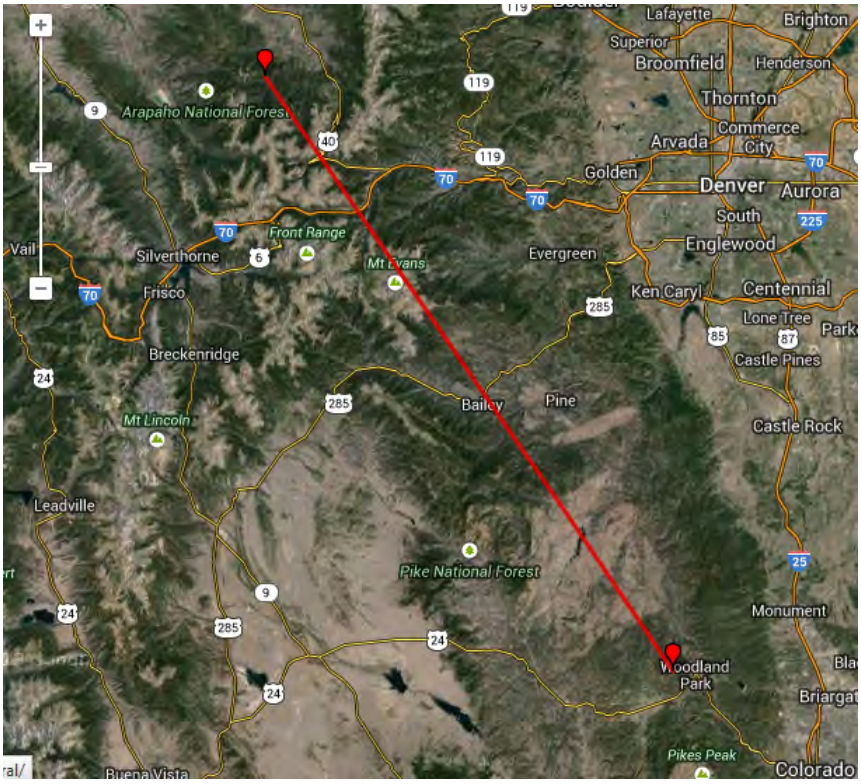

Figure 7.5: Distance between upper, lower bounds of cloaking range at Level 1

A visualisation of the a geographic data point present in the test data for Denver city in the United States of the great-circle diagonal is shown in Figure 7.5. The distance shown is

calculated using the Haversine formula to be 14 026.48m.
One can compare the calculated diagonal at each anonymisation level to ascertain the scope of spatial cloaking by the generalising anonymisation approach. In addition to this, the consideration of the actual position of the geographic data point weighs into the distance calculations. As discussed earlier in section 7.1, the distance between lattitude changes when reaching towards either poles as the lines shrink towards zero. A degree of lattitude is therefore not a consistent measurement when converted to metres.

To ensure the comparative assumptions made from our data sets are not subjective to these changes rather than the levels of precision, we need to consider the global effect in geographic terms of distance calculations. Therefore, in order to take all affecting factors into consideration, randomly selected data samples covering all known countries of the world is incorporated into the test range. The great-circle distance is then calculated at all levels of anonymisation for each of these locations covering the globe.

In the following tables we show the results found through the use of the Haversine formula to calculate the great-circle distance. The test geographic data points covered every populated country. The list of valid populated countries from which testing is necessary was obtained from the Google Maps API library, a source open to the public.

Country	**Code**	**Level 0**	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Afghanistan	AF	144942.7628	14463.3020	1446.3041	144.4971	14.3196
Albania	AL	145622.7637	14572.4153	1457.3453	145.6010	14.4289
Algeria	DZ	148196.0794	14847.0180	1484.6590	148.3331	14.6997
American Samoa	AS	154941.2127	15499.1515	1549.9263	154.8526	15.3458
Andorra	AD	138301.6409	13826.3875	1382.8651	138.1609	13.6916
Angola	AO	134340.0711	13430.1471	1343.1748	134.1954	13.2987
Anguilla	AI	153414.2997	15351.9631	1535.1629	153.3800	15.1998
Antarctica	AQ	138301.6409	13857.5032	1385.8199	138.4569	13.7209
Antigua and Barbuda	AG	134340.0711	13446.1826	1344.2170	134.2987	13.3089
Argentina	AR	143543.1308	14372.1595	1437.4058	143.6137	14.2320
Armenia	AM	157281.6147	15728.5360	1572.7409	157.1324	15.5717
Aruba	AW	155573.5623	15555.8585	1555.5196	155.4130	15.4013
Australia	AU	149955.9872	15009.5763	1500.6837	149.9336	14.8583
Austria	AT	134340.0711	13430.1471	1343.1748	134.1954	13.2987
Azerbaijan	AZ	144249.3331	14456.3680	1445.5416	144.4216	14.3120
Bahamas	BS	135938.0671	13582.0089	1358.0391	135.6781	13.4456
Bahrain	BH	152979.6997	15277.7416	1527.7502	152.6377	15.1263

Country	Code	Level 0	Level 1	Level 2	Level 3	Level 4
Bangladesh	BD	151040.4352	15096.1818	1509.2981	150.7942	14.9435
Barbados	BB	130321.0619	13060.3542	1306.0385	130.4891	12.9314
Belarus	BY	133537.1065	13333.7686	1332.8953	133.1732	13.19735
Belgium	BE	131927.9674	13188.9420	1319.1378	131.7963	13.0609
Belize	BZ	135140.6998	13494.2264	1349.4214	134.8211	13.3607
Benin	BJ	156352.4911	15638.56761	1563.8156	156.2403	15.4833
Bermuda	BM	145622.7637	14572.4153	1457.4123	145.6084	14.4296
Bhutan	BT	148799.8178	14877.0360	1487.7784	148.6442	14.7305
Bolivia	BO	136731.2669	13685.1135	1368.2699	136.7003	13.5469
Burundi	BI	157281.6147	15729.3500	1572.7734	157.1359	15.5720
Bonaire	BQ	155573.5623	15567.4736	1556.5358	155.5125	15.4112
Bosnia and Herzegovina	BA	137519.4134	13716.6706	1371.8196	137.0581	13.5823
Botswana	BW	133537.1065	13341.8087	1333.9405	133.2768	13.2076
BouvetIsland	BV	135938.0671	13574.0470	1357.0837	135.5834	13.4362
Brazil	BR	154941.2127	15502.4910	1550.1596	154.8763	15.3481
British Indian Ocean Territory	IO	156756.3532	15678.2504	1567.7103	156.6300	15.5219
Brunei Darussalam	BN	134340.0711	13454.1966	1345.3389	134.4100	13.3199
Bulgaria	BG	135938.0671	13566.0812	1356.4465	135.5206	13.4299
BurkinaFaso	BF	155573.5623	15564.6037	1556.3636	155.4953	15.4094
Cambodia	KH	155573.5623	15555.8585	1555.4017	155.4000	15.3400
Cameroon	CM	140604.4612	14079.3918	1407.6245	140.6324	13.9365
Canada	CA	142824.8336	14264.4291	1426.1330	142.4824	14.1199
Cape Verde	CV	135938.0671	13558.1115	1355.7292	135.4529	13.4232
Cayman Islands	KY	142095.1397	14183.8093	1418.3643	141.7054	14.0429
Central African Republic	CF	156923.1620	15689.9262	1568.9069	156.7500	15.5338
Chad	TD	154591.8383	15460.9652	1545.9395	154.4550	15.3067
Chile	CL	143543.1301	14343.6915	1434.0617	143.2768	14.1986
China	CN	136731.2669	13669.3035	1366.4516	136.5202	13.5291
Christmas Island	CX	156115.7489	15612.7823	1561.1502	155.9739	15.4569
Cocos(Keeling) Islands	CC	149386.6241	14935.7975	1493.2428	149.1896	14.7845
Colombia	CO	140604.4612	14056.7707	1405.5130	140.4282	13.9163
Comoros	KM	155856.0386	15581.4831	1558.0219	155.6618	15.4259
Congo	CG	157425.3797	15742.7200	1574.1318	157.2715	15.5855
The Democratic Republic Of The Congo	CD	157185.8933	15723.1507	1572.1882	157.0770	15.5662
Cook Islands	CK	152049.4647	15217.0158	1521.6368	152.0260	15.0656
Costa Rica	CR	156352.4911	15629.5144	1562.8224	156.1408	15.4734
Croatia	HR	135938.0671	13621.7559	1362.4104	136.1220	13.4896
Cuba	CU	152049.4652	15202.5238	1520.2371	151.8886	15.0520
Curaçao	CW	142824.8336	14307.8375	1430.3311	142.9018	14.1615
Cyprus	CY	143543.1308	14379.2459	1437.9723	143.6667	14.2372
Czech Republic	CZ	132732.7508	13245.2718	1324.6899	132.3477	13.1155
CôteD'Ivoire	CI	156756.3533	15674.7030	1567.3563	156.5937	15.5183

Country	Code	Level 0	Level 1	Level 2	Level 3	Level 4
Denmark	DK	127146.4387	12734.4411	1273.2122	127.2078	12.6061
Djibouti	DJ	136731.2669	13700.9028	1369.8489	136.8636	13.5631
Dominica	DM	154591.8383	15460.9651	1546.0831	154.4693	15.3078
Dominican Republic	DO	153414.2997	15330.7528	1533.0018	153.1620	15.1782
Ecuador	EC	145622.763	14579.1172	1457.4793	145.6157	14.4307
Egypt	EG	149386.6246	14918.3497	1491.8469	149.0530	14.7710
El Salvador	SV	135140.6998	13486.2261	1348.7815	134.7540	13.3540
Equatorial Guinea	GQ	157401.4037	15739.5315	1573.8096	157.2395	15.5823
Eritrea	ER	145622.7630	14558.9690	1456.0018	145.4674	14.4157
Estonia	EE	129520.9548	12964.2623	1296.4297	129.5299	12.8363
Ethiopia	ET	156352.4911	15642.9552	1564.1657	156.2757	15.4868
Falkland Islands(Malvinas)	FK	131123.7373	13092.4642	1308.7675	130.7577	12.9580
Faroe Islands	FO	149955.9872	15004.0080	1500.0148	149.8667	14.8517
Fiji	FJ	153828.0754	15372.6505	1537.2697	153.5892	15.2206
Finland	FI	137519.4134	13732.4164	1373.0003	137.1745	13.5938
France	FR	135140.6998	13534.1803	1353.4960	135.2258	13.4007
French Guiana	GF	157281.6148	15724.1079	1572.2835	157.0870	15.5672
French Polynesia	PF	153828.0754	15376.7249	1537.4326	153.6030	15.2219
French Southern Territories	TF	128724.4477	12884.5569	1288.6978	128.7510	12.7591
Gabon	GA	145622.7637	14585.8049	1458.3492	145.7033	14.4391
Gambia	GM	155268.5405	15528.4088	1552.7170	155.1324	15.3735
Georgia	GE	130321.0619	13068.3786	1306.7607	130.5580	12.9382
Germany	DE	141354.7687	14101.9287	1410.3285	140.9093	13.9640
Ghana	GH	156756.3533	15667.3267	1566.6012	156.5188	15.5109
Gibraltar	GI	142824.8336	14307.8375	1430.6912	142.9436	14.1656
Greece	GR	140604.4612	14094.4259	1409.1283	140.7864	13.9518
Greenland	GL	130321.0619	13068.3787	1306.7607	130.5580	12.9382
Grenada	GD	155573.5623	15567.4736	1556.7073	155.5301	15.4129
Guadeloupe	GP	149386.6241	14953.0887	1495.0310	149.3711	14.8025
Guam	GU	155268.5405	15528.4089	1552.7170	155.1321	15.3735
Guatemala	GT	149386.6241	14953.0887	1495.0310	149.3711	14.8025
Guernsey	GG	132732.7508	13277.4605	1327.5061	132.6307	13.1436
Guinea	GN	156352.4911	15624.8490	1562.3562	156.0949	15.4689
Guinea-Bissau	GW	155856.0389	15575.9473	1557.5800	155.6182	15.4216
Guyana	GY	157185.8934	15714.6363	1571.3049	156.9896	15.5575
Haiti	HT	153414.2997	15322.1234	1531.9661	153.0602	15.1681
Heard Island and McDonald Islands	HM	129520.9540	12988.2457	1298.4280	129.7287	12.8560
Holy See(VaticanCityState)	VA	142095.1397	14213.2766	1421.3101	142.0056	14.0726
Honduras	HN	154591.8383	15471.6426	1546.8662	154.5458	15.3153
HongKong	HK	151554.5959	15163.0153	1515.9393	151.4567	15.0092
Hungary	HU	134340.0711	13462.2081	1345.9798	134.4788	13.3267

Country	Code	Level 0	Level 1	Level 2	Level 3	Level 4
Iceland	IS	134340.0711	13454.1966	1345.1787	134.3948	13.3184
India	IN	139844.9788	14003.6718	1400.1271	139.8848	13.8625
Indonesia	ID	136731.2669	13708.7894	1370.6376	136.9376	13.5704
Iran	IR	145622.7637	14565.6993	1456.6070	145.5279	14.4217
Iraq	IQ	144942.7628	14511.4685	1451.1872	144.9891	14.3683
Ireland	IE	129520.9548	12956.2745	1295.7907	129.4644	12.8298
Isleof Man	IM	128724.4485	12892.5091	1289.2545	128.8082	12.7648
Israel	IL	139844.9788	13973.1388	1397.3031	139.6057	13.8348
Italy	IT	139077.1042	13880.7662	1387.7583	138.6536	13.7404
Jamaica	JM	153414.2997	15356.1425	1535.6636	153.4263	15.2044
Japan	JP	142824.8336	14300.6315	1430.2590	142.8975	14.1610
Jersey	JE	132732.7508	13293.5524	1329.5174	132.8325	13.1636
Jordan	JO	146939.9575	14690.8241	1468.7246	146.7400	14.5418
Kazakhstan	KZ	133537.1065	13390.0201	1339.1628	133.7921	13.2587
Kenya	KE	157425.3797	15742.7919	1574.1378	157.2721	15.5855
Kiribati	KI	157401.4037	15738.6927	1573.7166	157.2302	15.5814
Korea	KP	144249.3331	14393.3818	1439.1736	143.7896	14.2494
Korea	KR	143543.1308	14322.2146	1432.4156	143.1108	14.1821
Kuwait	KW	147575.9417	14767.0411	1476.7903	147.5483	14.6219
Kyrgyzstan	KG	139077.1042	13927.0928	1392.9305	139.16806	13.7914
Lao People's Democratic Republic	LA	146939.9575	14664.9257	1466.1330	146.4817	14.5162
Latvia	LV	127146.4387	12687.4896	1268.4400	126.7263	12.5585
Lebanon	LB	131123.7381	13116.5662	1311.5790	131.0419	12.9861
Lesotho	LS	147575.9410	14748.1980	1474.9708	147.3615	14.6034
Liberia	LR	156923.1620	15693.0501	1569.2023	156.7784	15.5366
Libya	LY	149386.6246	14947.3425	1494.6862	149.3338	14.7989
Liechtenstein	LI	131123.7381	13116.5662	1311.5790	131.0419	12.9861
Lithuania	LT	139077.1042	13911.6809	1391.3901	139.0111	13.7759
Luxembourg	LU	144249.3331	14456.3680	1445.4722	144.4188	14.3118
Macao	MO	142095.1397	14176.4162	1417.4033	141.6131	14.0338
Macedonia	MK	139077.1042	13896.2385	1389.8467	138.8569	13.7606
Madagascar	MG	153414.2993	15339.2993	1533.9407	153.2571	15.1877
Malawi	MW	155268.5402	15534.6656	1553.3113	155.1915	15.3793
Malaysia	MY	157185.8934	15721.1649	1572.0105	157.0600	15.5645
Maldives	MV	129520.9548	12940.3098	1294.1941	129.3049	12.8140
Mali	ML	153828.0758	15380.7783	1537.8384	153.6472	15.2263
Malta	MT	135140.6998	13486.2261	1348.3015	134.7052	13.3492
Marshall Islands	MH	136731.2669	13669.3035	1366.7680	136.5565	13.5326

Country	Code	Level 0	Level 1	Level 2	Level 3	Level 4
Martinique	MQ	154941.2130	15489.0006	1548.7779	154.7396	15.3346
Mauritania	MR	152049.4652	15226.5783	1522.7350	152.1353	15.0764
Mauritius	MU	152524.6304	15259.4184	1525.8277	152.4438	15.1070
Mayotte	YT	121193.4873	12136.98091	1213.345277	121.2271	12.0135
Mexico	MX	151040.4352	15096.1818	1509.5614	150.8205	14.9462
Micronesia	FM	149955.9872	15004.0080	1500.0148	149.8667	14.8517
Moldova	MD	140604.4612	14094.4260	1409.3535	140.8081	13.9539
Monaco	MC	137519.4134	13732.4164	1373.2363	137.1973	13.5961
Mongolia	MN	135140.6998	13494.2264	1349.5014	134.8251	13.3610
Montenegro	ME	135938.0671	13613.8151	1361.2195	136.0006	13.4775
Montserrat	MS	145622.7637	14572.4153	1457.0770	145.57689	14.4265
Morocco	MA	138301.6409	13834.1767	1383.6437	138.2379	13.6993
Mozambique	MZ	153414.2993	15335.0365	1533.3016	153.1919	15.1812
Myanmar	MM	152049.4652	15182.9257	1518.3288	151.6970	15.0330
Namibia	NA	139844.9788	13950.1519	1395.2351	139.4006	13.8144
Nauru	NR	157425.3797	15742.4323	1574.1049	157.2688	15.5852
Nepal	NP	148196.0794	14828.8071	1482.4730	148.1142	14.6780
Netherlands	NL	130321.0619	13060.3542	1306.0385	130.4883	12.9313
New Caledonia	NC	143543.1308	14336.5445	1433.4899	143.2167	14.1926
New Zealand	NZ	139844.978	13950.1519	1395.2351	139.4014	13.8145
Nicaragua	NI	155573.5623	15546.9106	1554.5060	155.3105	15.3911
Niger	NE	139077.1042	13911.6809	1390.6960	138.9463	13.7695
Nigeria	NG	156352.4911	15645.1142	1564.2954	156.2895	15.4881
Niue	NU	152979.6997	15317.7776	1531.6182	153.0242	15.1646
Norfolk Island	NF	147575.941	14785.7406	1478.4722	147.7169	14.6386
Northern Mariana Islands	MP	149955.9872	14992.8174	1499.0628	149.7711	14.8422
Norway	NO	124080.5404	12412.0176	1240.9035	123.9805	12.2864
Oman	OM	152049.4652	15202.5238	1520.2857	151.8929	15.0525
Pakistan	PK	146939.9575	14703.6821	1470.0757	146.8749	14.5552
Palau	PW	156756.3533	15674.7030	1567.3921	156.5982	15.5187
Palestine	PS	146288.7002	14599.1374	1459.7493	145.8452	14.4532
Panama	PA	139077.1042	13927.0928	1392.9305	139.1688	13.7915
Papua New Guinea	PG	138301.6409	13795.1642	1379.5098	137.8241	13.6583
Paraguay	PY	151040.4347	15106.6729	1510.5576	150.9211	14.9561
Peru	PE	135140.6998	13510.2177	1351.1801	135.0000	13.3786
Philippines	PH	155573.5623	15546.9106	1554.4758	155.3063	15.3907
Pitcairn	PN	135140.6998	13550.1381	1355.0117	135.3772	13.4157
Poland	PL	131123.7381	13076.4053	1307.8041	130.6591	12.9482

Country	Code	Level 0	Level 1	Level 2	Level 3	Level 4
Portugal	PT	140604.4612	14071.8607	1406.7205	140.5420	13.9276
Puerto Rico	PR	153414.2997	15351.9631	1535.1629	153.3800	15.1998
Qatar	QA	149955.9877	15004.0080	1500.2380	149.8890	14.8539
Romania	RO	135938.0671	13558.1115	1355.7292	135.4521	13.4232
Russian Federation	RU	123339.2316	12330.4807	1233.1211	123.2015	12.2092
Rwanda	RW	157401.4036	15738.2373	1573.6845	157.2270	15.5811
Reunion Island	RE	136731.2669	13669.3035	1366.5307	136.5297	13.5300
Saint Barthelemy	BL	134340.0711	13438.1660	1343.7360	134.2555	13.3046
Saint Helena	SH	128724.4485	12892.5091	1289.4136	128.8217	12.7661
Saint Kitts And Nevis	KN	153828.0758	15388.8214	1538.7238	153.7329	15.2348
Saint Lucia	LC	135938.0671	13558.1115	1355.8090	135.4561	13.4236
Saint Martin(FrenchPart)	MF	153414.2997	15360.3010	1535.7467	153.4375	15.2055
Saint Pierre And Miquelon	PM	135140.6998	13486.2261	1348.2214	134.7004	13.3487
Saint Vincent And The Grenadines	VC	144942.7628	14511.4685	1451.2556	144.9925	14.3686
Samoa	WS	155268.5402	15518.8563	1551.7300	155.0319	15.3635
San Marino	SM	134340.0711	13414.1026	1341.1694	133.9950	13.2788
Sao Tome and Principe	ST	157425.3797	15742.7680	1574.1338	157.2717	15.5855
Saudi Arabia	SA	146939.9568	14722.8536	1472.4394	147.1143	14.5789
Senegal	SN	131123.7381	13140.6816	1314.3120	131.3133	13.0130
Serbia	RS	146939.9568	14722.8536	1472.2485	147.0927	14.5767
Seychelles	SC	144942.7628	14470.2229	1446.9959	144.5683	14.3266
Sierra Leone	SL	132732.7508	13285.5067	1328.0693	132.6862	13.1491
Singapore	SG	157401.4037	15740.6102	1573.9178	157.2502	15.5834
Sint Maarten(Dutchpart)	SX	153414.2997	15360.3010	1535.9956	153.4603	15.2078
Slovakia	SK	133537.1065	13341.8087	1333.9405	133.2707	13.2070
Slovenia	SI	135140.6998	13542.1610	1354.0546	135.2864	13.4068
Solomon Islands	SB	135938.0671	13558.1115	1355.8887	135.4656	13.4245
Somalia	SO	134340.0711	13446.1826	1344.6177	134.3427	13.3133
SouthAfrica	ZA	146939.9568	14690.8240	1468.9181	146.7567	14.5435
South Georgia and the South Sandwich Islands	GS	128724.4477	12876.6093	1287.7440	128.6549	12.7496
South Sudan	SS	156756.3533	15667.3267	1566.5634	156.5156	15.5106
Spain	ES	139844.9788	13988.4222	1398.6021	139.7355	13.8476
SriLanka	LK	156756.3533	15669.2059	1566.7329	156.5326	15.5122
Sudan	SD	137519.4134	13716.6706	1371.4257	137.01559	13.5781
Suriname	SR	157281.6148	15724.1079	1572.3023	157.0883	15.5673
Svalbard And Jan Mayen	SJ	146939.9568	14671.4229	1466.8482	146.5544	14.5234
Swaziland	SZ	134340.0711	13470.2169	1346.7005	134.5508	13.3339
Sweden	SE	124080.5404	12434.4945	1243.5256	124.2380	12.3119

Country	Code	Level 0	Level 1	Level 2	Level 3	Level 4
Switzerland	CH	137519.4126	13740.2806	1373.6295	137.2397	13.6003
Syrian Arab Republic	SY	144249.3331	14400.4310	1440.2305	143.8952	14.2599
Taiwan	TW	151040.4352	15096.1818	1509.2454	150.7874	14.9429
Tajikistan	TJ	141354.7687	14109.4219	1410.7031	140.9460	13.9677
Tanzania	TZ	156923.1618	15694.5767	1569.2938	156.7874	15.5375
Thailand	TH	131123.7381	13148.7226	1315.0356	131.3888	13.0205
Timor-Leste	TL	135938.0671	13629.6922	1362.8865	136.1656	13.4939
Togo	TG	134340.0711	13422.1259	1342.4530	134.1257	13.2917
Tokelau	TK	156352.4908	15640.7730	1564.0351	156.2637	15.4856
Tonga	TO	135938.0671	13629.6922	1362.7278	136.1513	13.4925
Trinidad and Tobago	TT	156115.7492	15607.7946	1560.5256	155.9130	15.4508
Tunisia	TN	143543.1308	14350.8267	1435.2031	143.3894	14.2098
Turkey	TR	141354.7687	14101.9287	1409.9537	140.8696	13.9600
Turkmenistan	TM	135938.0671	13566.0812	1356.7651	135.5556	13.4334
Turks and Caicos Islands	TC	152049.4652	15197.6537	1519.4090	151.8044	15.0437
Tuvalu	TV	156756.3532	15676.4884	1567.4634	156.6046	15.5194
Uganda	UG	157401.4037	15740.6102	1573.9113	157.2495	15.5833
Ukraine	UA	133537.1065	13365.9213	1336.2712	133.5032	13.2301
United Arab Emirates	AE	151040.4352	15106.6730	1510.6621	150.9305	14.9571
United Kingdom	GB	127932.5896	12805.3006	1280.2177	127.9038	12.6752
United States	US	142095.1397	14242.5725	1423.8006	142.2552	14.0974
United States Minor Outlying Islands	UM	152979.7001	15309.0242	1530.6108	152.9244	15.1547
Uruguay	UY	145622.7630	14558.9690	1455.9344	145.4641	14.4153
Uzbekistan	UZ	139077.1042	13919.3907	1391.6214	139.0350	13.7782
Vanuatu	VU	154591.8380	15464.5461	1546.2262	154.4829	15.3091
Venezuela	VE	135938.0671	13597.9204	1359.3121	135.8101	13.4587
VietNam	VN	154941.2130	15509.1037	1550.7545	154.9347	15.3539
Virgin Islands	VG	153414.2997	15343.5415	1534.3219	153.2960	15.1915
Virgin Islands	VI	135938.0671	13582.0089	1357.7207	135.6471	13.4425
Wallis and Futuna	WF	154941.2127	15502.4910	1549.9597	154.8570	15.3462
Western Sahara	EH	150507.4152	15064.2585	1506.4785	150.5120	14.9156
Yemen	YE	154591.8383	15457.3624	1545.5792	154.4197	15.3028
Zambia	ZM	155268.5402	15537.7604	1553.6825	155.2289	15.3830
Zimbabwe	ZW	152979.6997	15317.7776	1531.7923	153.0411	15.1662
Åland Islands	AX	124080.5404	12434.4945	1243.1502	124.2005	12.3082

The tests show Mayotte Island with the smallest haversine distance in all levels of anonymisation, and Kenya with the longest haversine distance in all levels of anonymisation. A comparison of all levels for each country shows an approximate variance between levels by a multiplication factor of 10. This supports the initial calculation approximation in Table 7.1 for decimal degree precision area.

We assess the distance calculation difference in countries per level to ascertain the relevance of the effect on the tabulated comparison estimations. An abritrary point from each country was used to evaluate the pattern and effect of the physical distance between lattitudes on every inhabited place in the world. By doing this, it can be confirmed what distance can be cloaked at minimum and maximum at each anonymisation level.

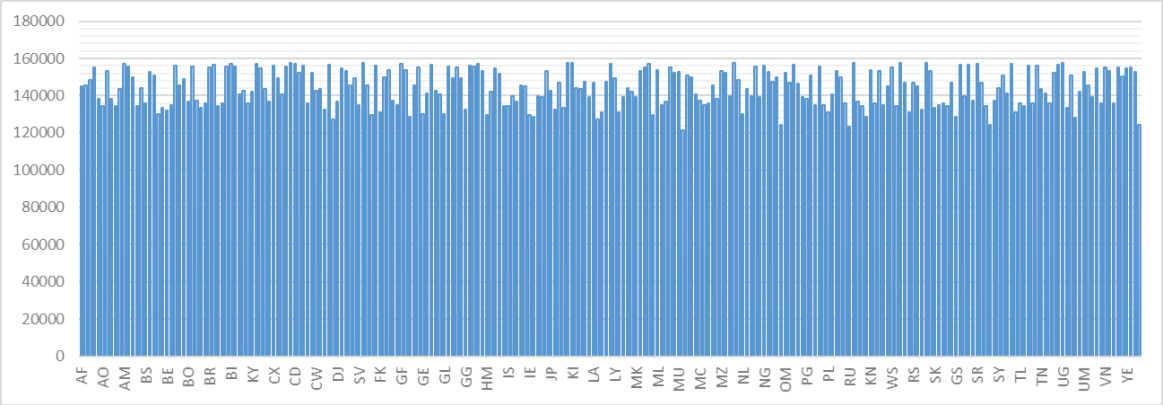

Figure 7.6: Global Geographic data, Anonymisation Level: 0

The dataset shown in Figure 7.6 recorded the distance ranges from the highest at 157425 metres to a lowest of the 121193 metres. The largest great path distance recorded to be 157425.3797 metres was located in Congo, Sao Tome and Principe, Nauru and Kenya. The average distance under the anonymisation cloak at level 0 for all countries was calculated at 144397.2822 metres. The place with the smallest distance margin of 121193.4873 metres was located at the French island Mayotte just off the coast of South-East Africa, between north-western Madagascar and north-eastern Mozambique.

The dataset shown in Figure 7.7 recorded the distance ranges from the highest at around 15743 metres to a lowest of around 12137 metres. The largest great path distance recorded to be 15742.7919 metres was located towards east africa in Kenya. The average distance under the anonymisation cloak at level 1 for all countries was calculated at 14441.3983 metres. The place with the smallest distance margin of 12136.9809 metres was located at the French island Mayotte.

The dataset shown in Figure 7.8 recorded the distance ranges from the highest at 1574.1378 metres to a lowest of the 1213.3453 metres. The largest great path distance recorded to be 1574.1378 metres was located towards Kenya. The average distance under the anonymisation cloak at level 2 for all countries was calculated at 1444.0303 metres. The country with the smallest distance margin of 1213.3453 metres was located at the French island Mayotte.

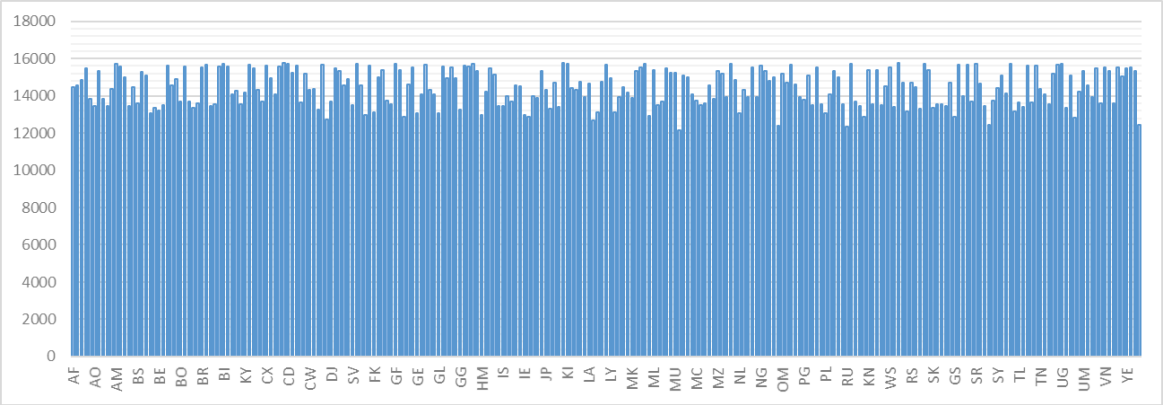

Figure 7.7: Global Geographic data, Anonymisation Level: 1

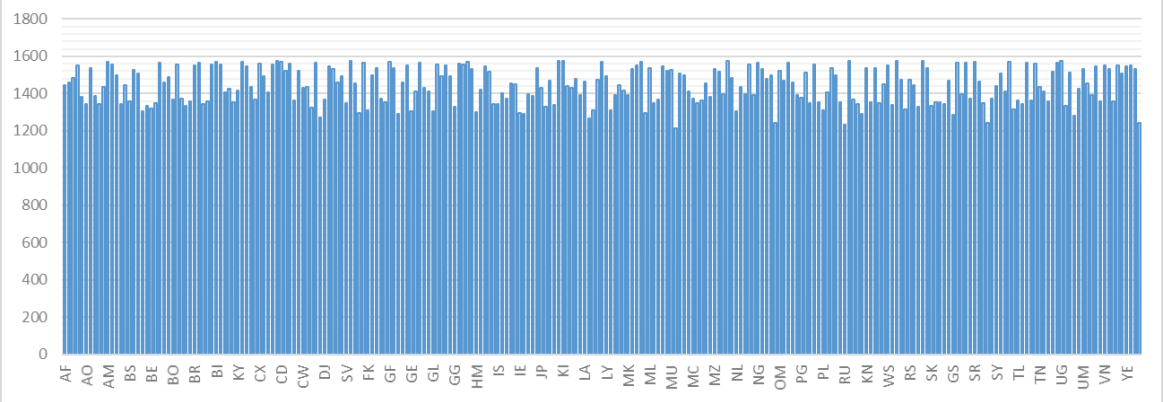

Figure 7.8: Global Geographic data, Anonymisation Level: 2

The dataset shown in Figure 7.9 recorded the distance ranges from the highest at 157.2721 metres to a lowest of the 121.2271 metres. The largest great path distance recorded to be 157.2721 metres was located once again, in Kenya. The average distance under the anonymisation cloak at level 3 for all countries was calculated at 144.2730 metres. The country with the smallest distance margin of 121.2271 metres was located at the French island Mayotte.

The dataset shown in Figure 7.10 recorded the distance ranges from the highest at 15.5855 metres to a lowest of the 12.0135 metres. The largest great path distance recorded to be 15.5855 metres was located in Kenya. The average distance under the anonymisation cloak at level 4 for all countries was calculated at 14.2973 metres. The country with the smallest distance margin of 14.2973 metres was located at the French island Mayotte.

At every level of anonymisation, the highest, average and lowest distances calculated matched the same country statistics. Kenya, identified as the highest distance in all levels, lies on the equator where the lattitude degree would be expected to be its maxmimum in metre distance. The relative variance in the distances depending on the country maintained a consistent pat-

tern at every level. This indicates the position on the globe does not interfere with the distance difference in the levels of precision of that geographic location. Figure 7.11 shows the relation between the anonymisation levels affected by geographic position when scaled in $1/(10^k)$ where k is the anonymisation level. The *y-axis* indicates the distance measured in metres, of the haversine diagonal distance for each location. The conclusions of the anonymisation effect on the geographic data is summarised in Table 7.5.

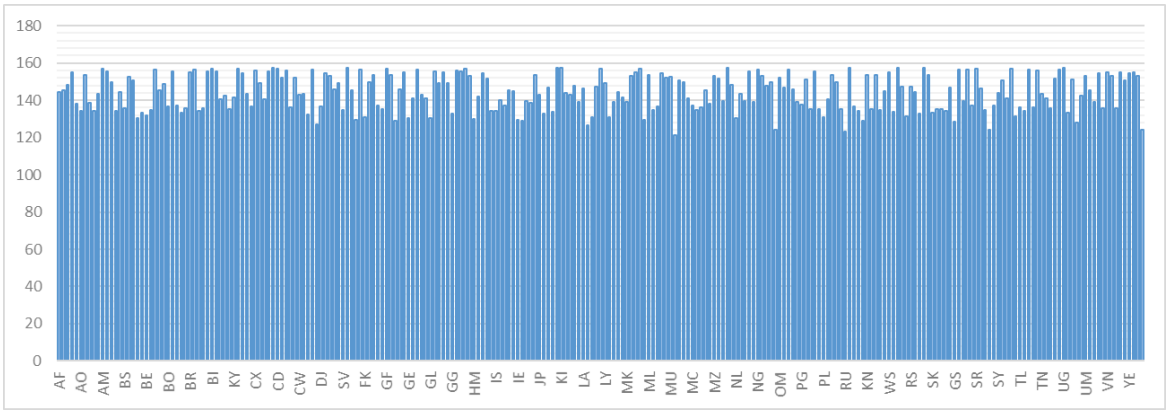

Figure 7.9: Global Geographic data, Anonymisation Level: 3

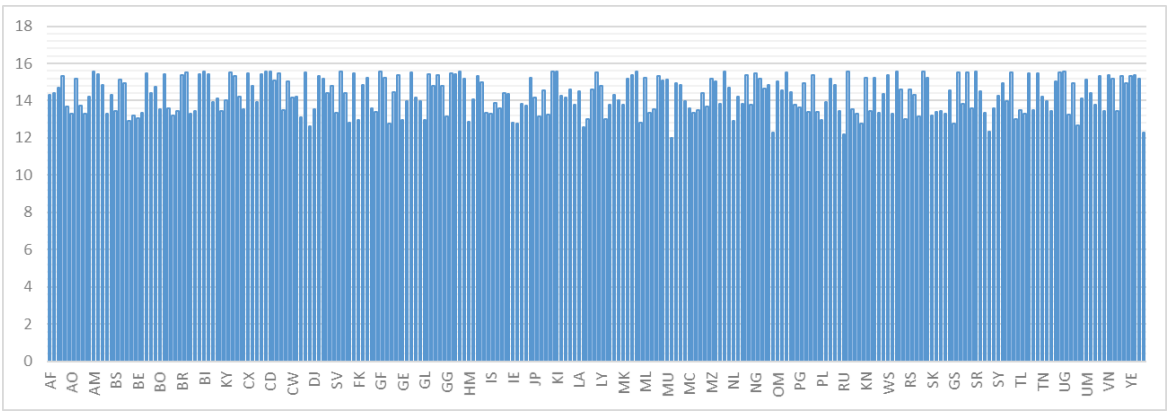

Figure 7.10: Global Geographic data, Anonymisation Level: 4

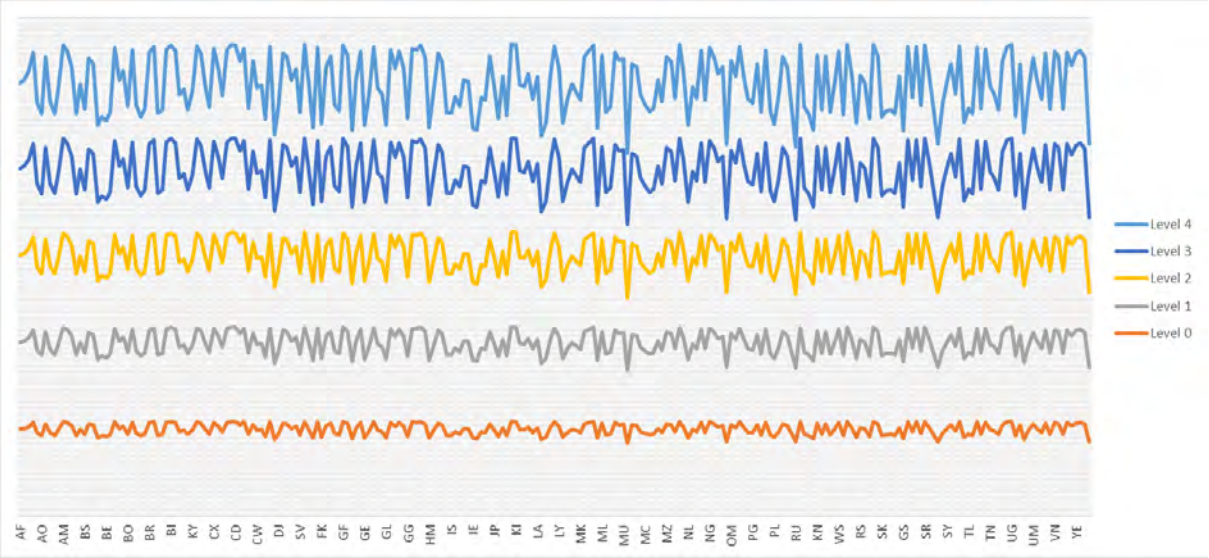

Figure 7.11: Haversine Distance Trend across countries for levels 0-4 of anonymisation

Level	Great-Circle Distance (Haversine)
0	The geographic location can be estimated within an area at a maximum distance of 121194m - 157425m from the actual location.
1	The geographic location can be estimated within an area at a maximum distance of 12137m - 15743m from the actual location.
2	The geographic location can be estimated within an area at a maximum distance of 1213m - 1574m from the actual location.
3	The geographic location can be estimated within an area at a maximum distance of 121m - 157m from the actual location.
4	The geographic location can be estimated within an area at a maximum distance of 12m - 16m from the actual location.

Table 7.5: Haversine effect on anonymisation depending on geographic location

7.4 Geolocation-based Security

The SIEM utilises various methods for detecting anomalies from log data. The method used in the implementation is the instruction of a correlation directive. The directive is essentially a collection of rules and conditions, if a rule is met it becomes the premise for one or more rules or invokes an alarm. The requirements set by the rule definitions aid security administrators in mining odd behaviour in a specific time-frame sequence of a user.
In the case of the brute-force attack implemented in this study, a consecutive number of login attempts within a very small time window indicates suspicious script attack behaviour.

Geographic location can be incorporated into a correlation rule definition as part of the condition for a trigger. The value of this inclusion can be specified in two areas. Alarms can be filtered through geographic context, promoting situational awareness and aiding prediction patterns using location as the basis of monitoring. Suspicious user location behaviour can indicate unauthorised access. An example rule definition would be, if user X logs out from location Z and logs in from location C within a time gap of half an hour where the physical possibility of this occurring is impossible (e.g Zimbabwe to Russia).

In order to verify the application of geolocation as a correlation rule, the additional condition for the brute-force attack added the requirement 'from a certain geographic location'. The location-based brute-force detection was written, using Denver, USA as the test location. The directive successfully triggered under the applied conditions, these confirmed results applied for all levels of anonymisation which was a fundamental concern of the test. The results of a directive alarm triggered from a level 1 test run is shown in Table 7.6.

Alarm	**No. of Events**	**Correlation Level**	**Reliability**	**Timestamp (HH:MM:SS)**	**Detection time (seconds)**	**DataSource**
Brute Force Attack from Denver City Location against 70.38.124.46	5	4	7	[2013-10-06 15:33:52]	1s	herah-tsom [9010:529]
Brute Force Attack from Denver City Location against 70.38.124.46	10	3	4	[2013-10-06 15:33:51]	30s	herah-tsom [9010:531]
Brute Force Attack from Denver City Location against 70.38.124.46	1	2	2	[2013-10-06 15:33:21]	37s	herah-tsom [9010:529]

Table 7.6: Details of triggered alarm - misuse case Brute Force Geolocation

The successful application of geolocation data in anonymised format for security rule detection indicates the feasibility of the anonymisation method, and validates the process of including location-based security procedures to capitilise on location for identification.

7.5 Summary

This chapter evaluated the outcomes of the implemented simulation, in terms of data accuracy and performance. The events sets were examined including geographic accuracy and precision of the location fields. The hypothesis of this study ascertains the effect of geographic data inclusion for SIEMs in a managed enterprise environmment. The performance focused on affirming the conditions and expectation of a running SIEM in such an environment. Finally, tests are carried out to evaluate the anonymisation effect on the geographic data, and the application of location-based security analyis.

Chapter 8

Analysis of Results

The previous chapter assessed the implementation approach supporting the utilisation of geolocation in security and privacy. We proceeded to validate the use of location in advancing an existing open source SIEM with the support of an EU initiated SIEM research framework which provided the necessary augmentation tools. The misuse case addressed in the test simulation was a location-based brute force attack in a managed enterprise environment. In addition to augmenting the SIEM security techniques with locational data, privacy procedures were incorporated for this data when used with the SIEM solution.

To state it briefly, the following solutions were proposed in this study towards advancing SIEMS in security and privacy with locations:

- A proposal for security augmentation with geolocation data, which is demonstrated with;
  Location-based incident detection in a managed enterprise environment.
- A SIEM privacy guideline which is demonstrated with;
  Location anonymisation within SIEM applications after normalisation.

The simulation was carried out incorporating these validation requirements as a proof-of-concept platform.
This chapter focuses on analysis of the conclusive results obtained from these research efforts towards supporting the aims of this study. The results achieved and the implications of these outcomes to the propositions made of applying geolocation data for security analysis with enforced privacy measures are discussed and evaluated.

8.1 Interpretation of Findings

The experimental setup was executed using the collected data sets from the Managed enterprise. The implementation proved successful and demonstrated feasibility in tool adaptation, modification, and integration of both OSSIM and the selected MASSIF tools with geolocation data. The results in terms of detection and experiment achievements are reported in this section.

8.1.1 SIEM Response

The OSSIM solution provides a web-based visualisation interface for viewing the events from various sensors, statistics generation, and most importantly incident detection and response through alarm generation. The test case(MC-5.5.1) executed a brute-force attack attempt, prompting the correlation directive stored on the OSSIM server. The additional condition of the directive was to examine the source location and raise the trigger only if it located to be from Denver city, USA.

The geographic-based filtering applied here functioned as a test of the SIEM solution ability to incorporate location in the correlation engine detection phases. Test data matching the required conditions triggered the following alarms seen in the OSSIM interface:

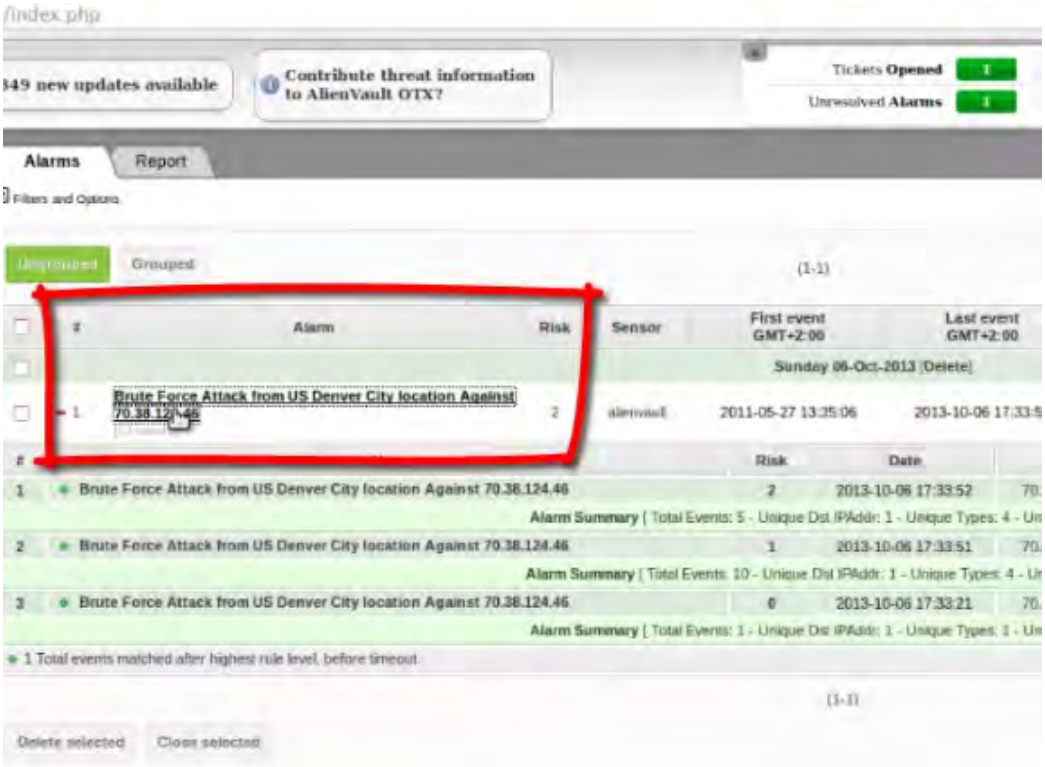

Figure 8.1: Alarms raised from brute force from set location

The default asset value applied in the duration of this execution was 2, the highest priority activated by the misuse data set was a level 4. Table 8.1 shows the resulting risk for the system reaching level 2, calibrated by the SIEM using $Risk = \left(\frac{priority \cdot reliability \cdot asset}{25}\right)$ as the method of calculation.

Calculated Risk				
Correlation Rule	**Asset Value**	**Reliability**	**Priority/Threat**	**Resulting Risk**
Level 1	2	2	4	0
Level 2	2	4	4	1
Level 3	2	7	4	2

Table 8.1: Calculated Risk by OSSIM for MC-5.5.1 Brute-Force

For every alarm generated on the OSSIM server when a security rule is triggered, the triggering events for the alarm can be examined, as shown in Figure 8.2, along with the level of correlation it satisified to aid identification of which rule within the relevant directive was satisfied by the event.

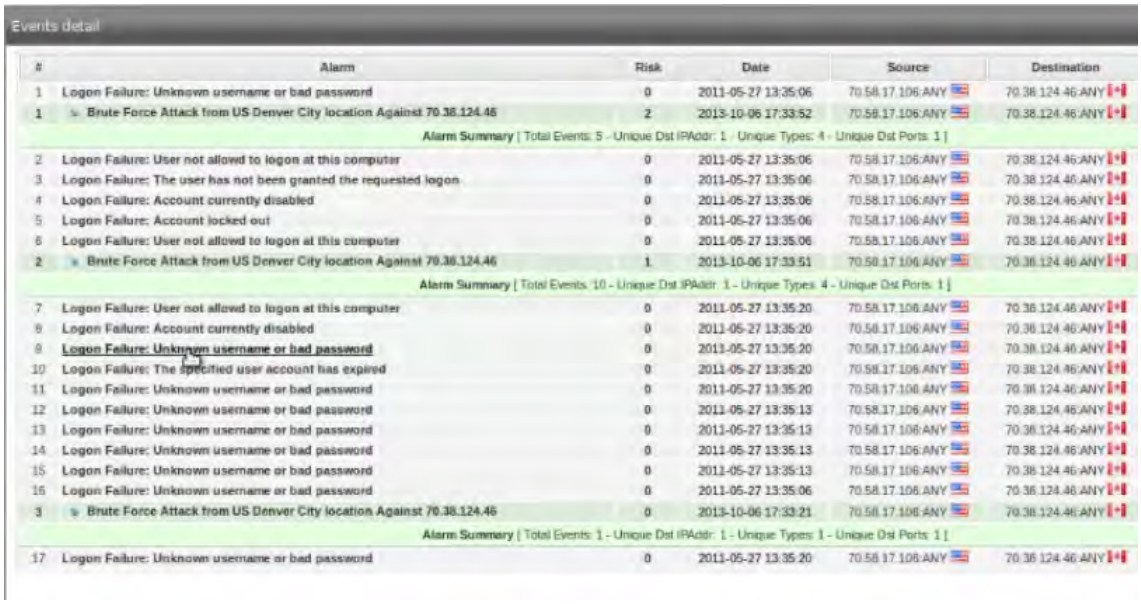

Figure 8.2: Triggering events for MC-5.5.1 from Denver, USA

The correlation rules specified for the brute-force attack has a collection of rules that need to be satisfied to trigger the alarm. The correlation level in the results indicated all rules were satisified within the directive, these are:

Initiation: 1 x *Authentication failure of user from location* (x, y) *where* $x \in \{39.9000...39.8999\}$ *and* $y \in \{-105, 1000... - 105.1999\}$

Level 1: WHILE T $< T^1$ **IF** 10 x *Authentication failure of user* [Reliability:2]

Level 2: WHILE T $< T^2$ **IF** Level 1 **AND** 5 x *Authentication failure of user* [Reliability:4]

Level 3: WHILE T $< T^3$ **IF** Level 2 **AND** 1 x *Authentication failure of user* [Reliability:7]

T^1, T^2 and T^3 refer to the time frame within the event must be received in before rule expiry, in this case these were set as 40, 400, 4000 seconds respectively. The actual events matching all rules discussed above are reported in Table 8.2, for the windows server test data with the correspondings level of correlation triggered and risk relation. The recorded detection time of the attack from the time the first event of the data set was received by OSSIM to the time of the alarm creation was recorded at 37 seconds.

The results are indicative of the success of applying location-based rules within a SIEM environment in processes surrounding user authentication. The application of location-based examinations for situational awareness and suspicious incident detection provide significant aid for security monitoring of global networks. In section 3.3 a matrix of geolocation based security rules was introduced and discussed towards improving security in the areas of cloud, mobility, advanced threats and regulatory compliance. Applying these rules in incident detection correlation procedures, as demonstrated in this study elevates the level of exploitation of existing data present in logs sent to the SIEM. Environments such as that of a managed enterprise where copious amounts of security data are received every day globally, are scenarios well applied for situational awareness and geographic compliance concerns.

Event Type and Details	**Risk**	**Correlation Level**
Logon Failure: Unknown username or bad password	0	4
ALARM: OSSIM generated Level 3	2	4
Logon Failure: User not allowed to log on at this computer	0	3
Logon Failure: The user has not been granted the requested login	0	3
Logon Failure: Account currently disabled	0	3
Logon Failure: Account locked out	0	3
Logon Failure: User not allowed to log on at this computer	0	3
ALARM: OSSIM generated Level 2	1	3
Logon Failure: User not allowed to log on at this computer	0	2
Logon Failure: Account currently disabled	0	2
Logon Failure: Unknown username or bad password	0	2
Logon Failure: The expected user account has expired	0	2
Logon Failure: Unknown username or bad password	0	2
Logon Failure: Unknown username or bad password	0	2
Logon Failure: Unknown username or bad password	0	2
Logon Failure: Unknown username or bad password	0	2
Logon Failure: Unknown username or bad password	0	2
Logon Failure: Unknown username or bad password	0	2
ALARM: OSSIM generated Level 1	0	2
Logon Failure: Unknown username or bad password	0	1

Table 8.2: Triggering event data for correlation level from the test MC-5.5.1

Field	Field name	Comment
TSOM Specific		
18	<SourceIPGeoCC>	Two letter country code of <SourceIP>or string "1918" if it is a private IP address
19	<SourceIPGeoASN>	ASN of <SourceIP>or empty string if it is a private IP address.
20	<SourceIPGeoLat>	Latitude of <SourceIP>or empty string if it is a private IP address
21	<SourceIPGeoLong>	Longitude of <SourceIP>or empty string if it is a private IP address
25	<DestinationIPGeoCC>	Two letter country code of <DestinationIP>or string "1918" if it is a private IP address
26	<DestinationIPGeoASN>	ASN of <DestinationIP>or empty string if it is a private IP address
27	<DestinationIPGeoLat>	Latitude of <DestinationIP>or empty string if it is a private IP address
28	<DestinationIPGeoLong>	Longitude of <DestinationIP>or empty string if it is a private IP address
Event Specific		
40	<SourceIPGeoCC2>	Two letter country code of <SourceIP2>or string "1918" if it is a private IP address.
41	<SourceIPGeoASN2>	ASN of<SourceIP2>or empty string if it is a private IP address.
42	<SourceIPGeoLat2>	Latitude of<SourceIP2>or empty string if it is a private IP address.
43	<SourceIPGeoLong2>	Longitude of<SourceIP2>or empty string if it is a private IP address.

Table 8.3: Event fields from sensor containing geographic information

8.1.2 Geographic Processing

The source data used in the test simulation considered event data from windows server 2003/2008 sources from a managed enterprise environment. The collected events retrieved through SNMP underwent a normalisation procedure mapping event attributes to an input vector. The input vector consists of two parts, TSOM-specific and Windows-specific. These relate to different sources containing geographic information for source/destination events.

Field Validation

The TSOM-specific attributes were pulled from Tivoli's central management system, containing information captured by the event aggregation module. The Windows-specific attributes are obtained from the raw windows event through the $EVENT.INFO token. Both sources capture geographic field data concerning the event, and at the very least, an flag indication if it is a private IP address. An extract of the fields containing location information in both sources is shown in Table 8.3. The fields highlighted in the table are the values extracted by the GET for anonymisation procedures from the event logs. The preferred source of geolocation was the data pulled from the TSOM rather than the raw events, on the basis of higher accountability of the TSOM sensors in providing valid information.

Process Validation

To examine the processing of these fields, consider the events from the test data with a geographic data point ρ from a location in Denver city, United States - the area chosen as the primal area of investigation for an analysis test case. The relevant fields were <SourceIPGeoLat>and <SourceIPGeoLong>. In the raw log, the fields were recorded as 39.91470000000001 and -105.0809, respectively.

Before any correlation process held in the SIEM solution, *all* events went through three transformation processes in the GET tool;

1. Encryption. The log was duplicated, encrypted using a password-based encryption algorithm.

2. Anonymisation. The highlighted fields in Table 8.3 were anonymised at a level k — where k could be any number from 0 to 4

3. OSSIM Schema mapping. The transformation vector for windows server events was applied, mapping selected fields into OSSIM event fields. In the case of the geographic data fields, they were placed in OSSIM **USERDATA** fields and the encrypted copy of the log in the OSSIM **LOG** field. The resulting event line sent to OSSIM is shown in Figure 8.3.

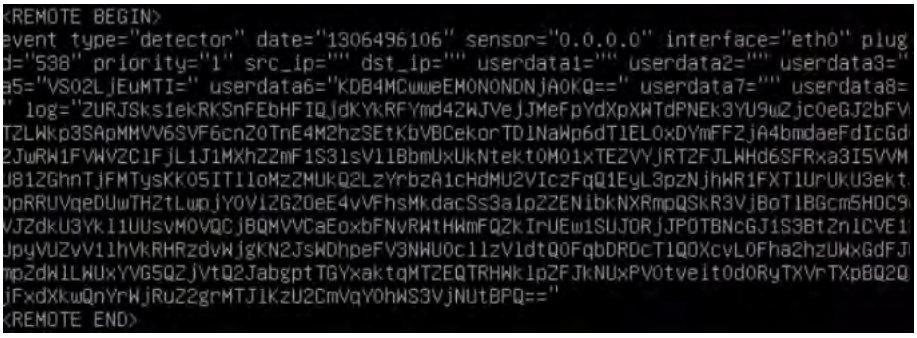

Figure 8.3: Format of test event sent to OSSIM

Depending on the level of anonymisation applied the geographic data fields received in OSSIM varied, Figure 8.4 shows the values for a Level 1 test run.
The rules applied from the correlation require the geographic range of lattitude of 39.9 and longitude of -105.1, this matches the range from the generalisation method applied by the GET in the case of level 1, which converts 39.91470000000001 ->39.9 and -105.0809 ->-105.1. However, the requirement for this process is to match the rule if the location is in Denver, regardless of the level of anonymisation applied. Therefore, the detection must work with all values within range bounds of anonymisation herefore for lattitude (39.9000, 39.9999...), longitude (-105.0000, -105.9999...) all values in this range will satisfy the trigger rule (note: decimal degrees past 0.9 can be considered pointless precision figures).

The successful triggering of the directive demonstrated at all levels in test runs validates the anonymisation implementation on the geolocation data. This satisfies the requirement of anonymisation of geolocation in the correlation process and its utilisation in the process of incident detection from a certain location.
Finally, Table 8.4 shows the point range provided through the anonymisation obfuscation method that can be obtained at each level. The maximum distance possible between two points provisioned by the cloaking area is an estimate 14 000 metres at Level 1 to just 14 metres at Level 4. The preferred range can be applied depending on regulatory compliance

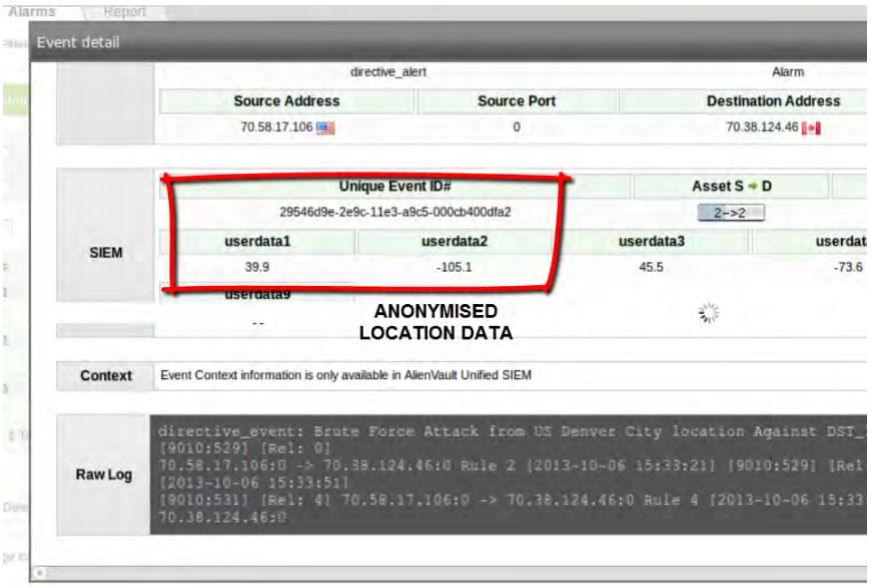

Figure 8.4: Geographic field data values received in OSSIM

rules applying to the source country.

A map visualisation of the ranges for ρ was implemented using *Google Maps JS API v3* to evaluate the context of physical range covered by the cloaking area, shown in Figure 8.5. The maxmimum obfuscation generalises to city level while the minimum is a granular level of couple of metres. This allows for flexibility in geolocation based security application, if the method is a geo-authentication procedure for a mobile user, the suitable level would be 4. If it is geographic load-balancing in the cloud, jurisdiction enforcement and/or a compliance procedure then a level 0 or 1 is sufficient.

Co-ordinates		Anonymisation Level	Maximum Distance(m)
39.xxxx	-105.xxxx	0	140 604.50
39.9xxx	-105.0xxx	1	14 026.48
39.91xx	-105.08xx	2	1 401.99
39.914x	-105.080x	3	140.15
39.9147	-105.0809	4	13.89

Table 8.4: Haversine distances for test Denver city data point

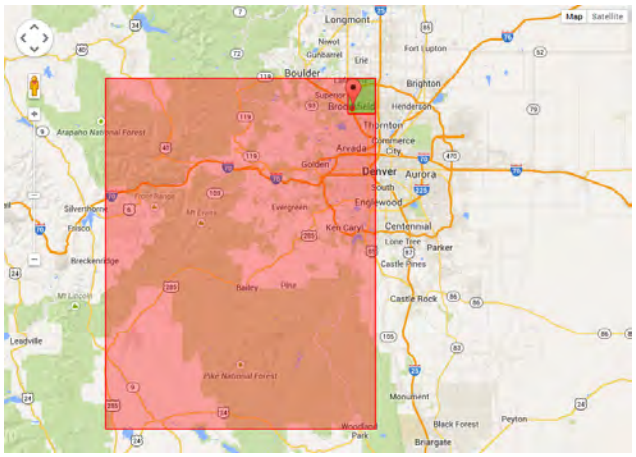

Figure 8.5.a: Red: Level 1, ∽ 14 000m

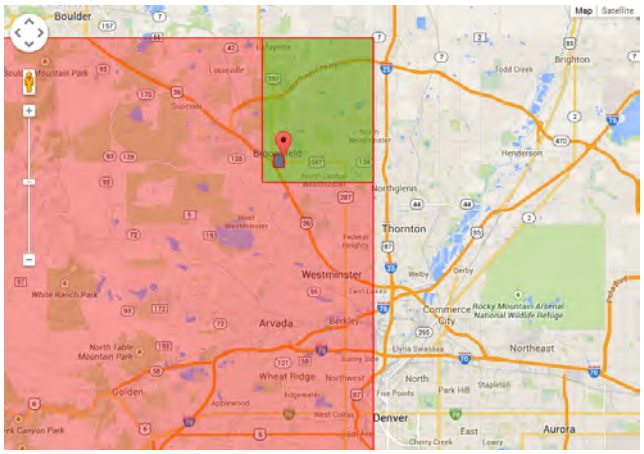

Figure 8.5.b: Green: Level 2, ∽ 1 400m

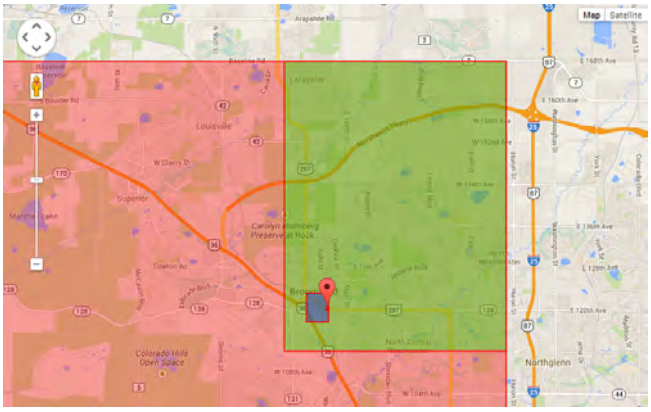

Figure 8.5.c: Blue: Level 3, ∽ 140m

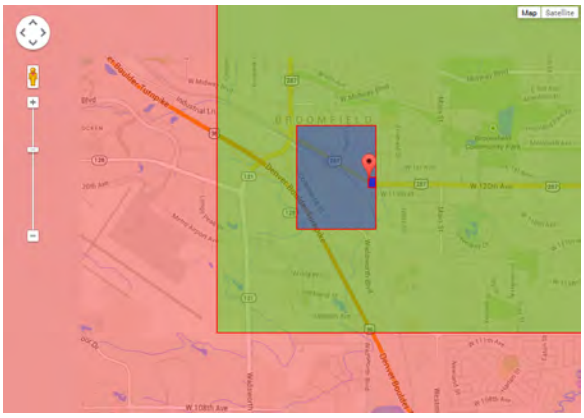

Figure 8.5.d: Dark Blue: Level 4, ∽ 14m

Figure 8.5: Spatial cloaking area for anonymisation of geolocation

8.2 Evaluation of Results

In the following section the effect of the integration of geolocation data into SIEM collection and analysis is evaluated loosely based on the Common Criteria [1] standard as the strategy of assessment. This criteria is a series of standards adapted to the evaluation of the security of information technology software and devices. It is applied in this context to evaluate a SIEM framework that applies geolocation security and privacy measures as a 'next-generation' SIEM solution.

For this evaluation, MASSIF as a framework and OSSIM as a solution are the two SIEMs assessed in their current abilities towards a 'next-generation' SIEM solution. The three categories of the evaluation criteria applying to this study are :

- Security
- Efficiency

[1]The Common Criteria(known as Common Criteria or CC) for Information Technology Security Evaluation is an international standard (ISO/IEC 15408) for computer security certification.

- Adaptability

These three areas are evaluated in the influence of geolocation on their quality of delivery as a SIEM. For each criterion C within these areas, there are two considerations; the support of the SIEMs of the criteria C and the effect of geolocation to criteria C. The assessment is carried out on the SIEM(s) that were applied for that criteria in the experiment implementation of this research.

A primary concern of the overall outcome of this proposed research is that the basis of the expectations of such integration do not compromise the existing SIEMs within these evaluatory facets, and that it encourages advanced functionality of the SIEMs.

A description of the criterion template used and its various applicable fields is given below:

ID:	e.g M.F.1.1.1	Name:	Name of the Criterion	Category:	e.g Security
Description:	Description of this criterion				
Rationale:	Brief description explaining the importance of this criterion in our evaluation.				
Metrics:	The metrics used to evaluate the product, this dependant on the defined criterion. The role of this metric is to aid the developer to define the score value. For example, to evaluate the processing rate the metrics can represent the number of processed events.				
Evaluator:	D	Rank	D	Score	100%
Evaluation assessment:	The assessment of the criterion with regards to the SIEMs				
Geolocation Applicability:	How the criterion is affected by the inclusion of geographic data in SIEM security techniques and/or it's privacy measures, and vice versa.				

Table 8.5: Evaluation Criteria Template

Additional details of some fields for the evaluation criteria can be listed as the following:

1. : ID: A unique identifier of that criterion.
2. Rank: Importance of the criteria: this field is used to specify the importance or ranking of each criterion. Each criterion may be assigned a rank of:
   - M: for Mandatory,
   - D: for Desirable, or
   - O: for Optional.
3. Score: Record the score assigned to the criterion after evaluation. For example, a scale value as 10/10 for fully satisfied, 5/10 for substantially satisfied, 1/10 for partially satisfied or 0 for not satisfied.
4. Evaluator: the person who will evaluate the product/results using this criterion and set up the score value. In this case, we are the evaluators (D), the partners of MASSIF confirmed the outcomes in any instance where the MASSIF tool(s) applied.

The remainder of this section henceforth, uses this template to describe and evaluate the SIEMs in security, efficiency and adaptability. The SIEMs, OSSIM and MASSIF framework are only evaluated in the criterion functionality in which they were *applied* in the test simulation.

8.2.1 Security

The following tables address criteria that concern the SIEM provision in the area of security. This functionality is referred to in terms of validity, protection and verification of data from damage or tampering. This encompasses the concerns of users with regards to data privacy rights and usage terms.

ID:	M.F.1.1.0	Name:	Data Authenticity	Category:	Security
Description:	System capability to provide unforgeability, non-repudiation and fault tolerance of stored data.				
Rationale:	This criterion aims at evaluating the robustness of the SIEM to guarantee authenticity, unforgeability and non-repudiation of stored data, when either faults or intrusions are affecting the system. The criterion provides a number representing the percentage of acceptable data loss as a consequence of stored data corruption. This percentage is the amount of lost data that can be recovered.				
Metrics:	Percentage				
Evaluator:	D	Rank	D	Score	100%
Evaluation assessment:	This can be assessed in the methods of data recovery put in place in the SIEM system. The Resilient Event Storage of MASSIF does not permit events to be edited that are in storage, to ensure data integrity and authenticity. Additionally, measures of encryption can be taken by the GET for every original event which can be stored as a copy along with the event sent for any processing after collection. Since the SIEM is implemented with an enterprise storage management solution beneath it, a virtualised replica of the store can be created (mirrored) on an ongoing basis. Given the integrity checks of MASSIF storage (from a security perspective) and underlying integrity of the Enterprise storage (from a resilience perspective) it is considered that no data corruption or loss should occur at the storage level.				
Geolocation Applicability:	The data recovery methods applied are of direct relevance to geographic location data stored in the event logs, particularly if this geographic data is used to structure visualisation interfaces within the SIEM to enhance situational awareness and for the enforcement of regulatory compliance through location filters or rules. The test simulation applied encryption to the entire event containing the location information. In addition to this, the utilisation of the Resilient Event Storage to store the data and alarms in the simulation affirmed the authenticity of the geographic data. This was facilitated through MASSIF tools and satisfies the data authenticity requirement.				

Table 8.6: M.F.1.1.0 - Data Authenticity

ID:	M.F.2.1.0	Name:	Privacy of forensic records	Category:	Functionality: Security
Description:	System reliability in guaranteeing privacy of forensic records				
Rationale:	This criterion aims at evaluating the reliability of the SIEM in providing forensic records only to authorized parties. Specifically, the criterion evaluates whether the SIEM allows the unauthorized parties to access the forensic records or not.				
Metrics:	Boolean Value: YES/NO				
Evaluator:	D	Rank	M	Score	TRUE
Evaluation assessment:	Access control external to a SIEM will prevent unauthorised parties from access and SIEM technologies such as those created in MASSIF can ensure privacy through encryption and anonymisation methods. Any data sent used for processing or security analysis is either encrypted or anonymised. Its only present in its original form upon entry through collectors such the GET tool of MASSIF. During any processing privacy of data is ensured and can cater for custom privacy requirements.				
Geolocation Applicability:	With regards to experimental simulation, the location fields were regarded as private information, the GET tool ensured the data reliability and privacy which can be used for forensic analysis, by encrypting the original field, and anonymising the content to be processed further. The issue of ensuring privacy of forensic records was supported by the geolocation implementation, as it included the condition of geo-anonymisation. If this information was exploited in the security analytic procedures without this consideration, the SIEM would be violating its advocation of privacy, considering the sensitive nature of geolocation data as explained in chapter 4. The simulation used anonymisation, data encryption and resilient storage on the geolocation information brought in, *regardless* of whether the data was to be used in security analysis or not. This is an important observation, bringing into cogitation that geolocation data most often present in logs from device networks, were not taken into consideration for privacy such as within the OSSIM solution. The experiment demonstrated privacy-enforcing procedures using the GET to enhance OSSIM in this regard. The techniques applied from the point of entry to it's storage and finally, deletion. This satisfies the SIEMs ability in privacy for it's forensic records, in particular geolocation, which was explicitly demonstrated and tested.				

Table 8.7: M.F.2.1.0 - Privacy of forensic records

ID:	E.F.3.1.0	Name:	Data Anonymisation	Category:	Functionality: Security
Description:	Sensitive corporate data within events needs to be anonymised.				
Rationale:	Anonymisation of event data is required, and a mapping from actual to anonymised (coded) data will be conducted before providing data sets. Suitable audit trails and evidence leading towards inferences will be required to ensure a legal basis for security findings from SIEM components.				
Metrics:	The event anonymization tool accepts CSV file input. Rows required to be anonymized must be identified.				
Evaluator:	D	Rank	M	Score	TRUE
Evaluation assessment:	Data with specification of which fields required anonymisation were given to the MASSIF collector. The result logs produced were compared against the output file results in terms of anonymisation and data integrity preservation of fields meant to be unaffected. The method of anonymisation used was custom to the type of field data requiring processing and ensured anonymisation of all sensitive data identified.				
Geolocation Applicability:	The advanced anonymisation shown with MASSIF is the customisable anonymisation ability of various fields whilst still enabling the data to be used for security analysis detection, thus ensuring privacy has not been breached while in use. This was applied to the geolocation information, with the customisation applying to data in the format of decimal degrees of a lattitude and longitude. The anonymisation applies to **all** geolocation source events. The level of anonymisation can be adjusted depending on the requirement from regulatory or user right specifications. The method applies a generalisation technique, decreasing the precision of a geographic data point. The generalisation can cloak with an offset of an estimate 11 metres up to an area covering a city. This specific SIEM requirement greatly satisfies the use of geolocation in SIEM through the advanced anonymisation provision demonstrated with the GET tool.				

Table 8.8: E.F.3.1.0 - Data Anonymisation

8.2.2 Efficiency

The following tables address the flexibility of the SIEM in aspects of data handling for the wide and growing range of network domains and sources. It also addresses the functionality of the SIEM with monitoring and defence mechanisms towards 'identified' adversaries.

ID:	M.E.17.1.0	Name:	Heterogeneous Data Source Support	Category:	Efficiency
Description:	Number of security event formats supported by SIEM				
Rationale:	The aim of this criterion is to evaluate the capability of a SIEM to cope with a large number of highly heterogeneous data sources originating from different operational domains				
Metrics:	Number of supported event formats				
Evaluator:	D,U	Rank	M	Score	2(out of 2)%
Evaluation assessment:	SIEMs such as MASSIF and OSSIM cater for this through the use of parsers, whilst OSSIM applies the parsing through the instruction of regular expression mapping, the GET tool creates a parser per event format and instructs modification through this method.				
Geolocation Applicability:	With regards to geolocation, two event formats were tested with using the GET MASSIF tool, both containing geolocation information, and proved to efficiently support them. OSSIM also demonstrated, in the case of geolocation, the ability to read and handle this information. The data was also applied in the correlation engine showing support from information gathering to the data-to-information layers (see Figure 2.1). The ability to handle different data sources supports the ability to accomodate sources of geolocation from sensors and other devices the are more equipped in it's provision. Therefore, SIEMs demonstrate strong capability to support geolocation inclusion, with complete support given for all scenarios and their unique data sources.				

Table 8.9: M.E.17.1.0 - Heterogeneous data source support

ID:	E.E.23.1.0	Name:	Track Logins	Category:	Efficiency
Description:	Track failed and successful logins.				
Rationale:	Login attempts which fail more than 3 times may be an attempt at a brute force attack. The source of the attempts must be recorded as well as the targeted user				
Metrics:	Login events must be identified as well as whether they are successful or not. Multiple failed attempts in rapid succession must be recorded. Users which thereafter succeed to login must be flagged.				
Evaluator:	D	Rank	M	Score	Possible
Evaluation assessment:	OSSIM provides visualisation, alarm generation and incident storage which facilitate all tracking of events. The correlation directive component of OSSIM can be applied to match attack patterns, a triggered directive initiates the creation of an incident event that is stored in the OSSIM database and an alarm is generated with levels of priority.				
Geolocation Applicability:	Geolocation factors can be applied using this functionality as seamlessly as it is applied for default patterns of known attacks like the brute force. Login attempts from an incorrect location for example, can flag a brute force at the beginning of the attack after three base failed attempts. The tracking of logins is particularly useful if the geographic data that comes with it is considered in security analytics. Particularly for detecting unauthorised access and physical areas of bad ip-reputation that can be blocked as pre-emptive measures. The applied SIEM greatly satisfies this requirement and is strongly supported by the use of geolocation for tracking suspicious behaviours.				

Table 8.10: E.E.23.1.0 - Track logins

8.2.3 Adaptability

The following table addresses the considerations of a SIEM with regards ease of adaptation and compatibility for various platforms. The scope available for adjustment is a critical component in assessing the feasibility of such a solution.

ID:	M.P.2.1.0	Name:	Parsing expressiveness and adaptibility	Category:	Portability: Adaptability
Description:	The capability of the SIEM to adapt to any new event format by creating a fully functional event parser.				
Rationale:	In order to improve the capabilities of the next generation SIEM, there needs to be the ability to seamlessly integrate any type of security tools/probes				
Metrics:	Capability of integrating any new type of data feeds				
Evaluator:	D,U	Rank	M	Score	2(out of 2)%
Evaluation assessment:	MASSIF is able to integrate any type of security event feeds with the understanding that the creator has a sound level of the associated grammar knowledge and of the environment. The integrative ability is facilitated through the GET, that works with the parsers. OSSIM is able to integrate any event format through plugins, provided the user/creator has a sound understanding of regular expressions and also, the understanding on the creation of plugins via the OSSIM specifications.				
Geolocation Applicability:	With regards to the test experiment, two fully functional event parsers were created and could be adapted to changes regarding the log information. This applies to geolocation due to the fact the adaptation ability is the critical requirement of the inclusive action of geolocation sources. The SIEMs satisfy this requirement and in doing so, support geolocation for exploitation possibilities.				

Table 8.11: M.P.2.1.0 - Parsing expressiveness and adaptibility

8.3 Summary

This chapter discussed the resulting outcomes from the simulation applied and the conclusions that can be drawn from the results. After the discussion of results, evaluation is carried out. The evaluation concerned the advancement of SIEM with geolocation security and privacy measures. Through the assessment of the SIEM provisions in the affecting criteria, the use of geolocation is evaluated and considered in its support and feasibility of application to future SIEM versions, towards a 'next-generation' SIEM. The effectiveness of the results of this study are therefore evaluated in the context of the applied SIEMs.

Chapter 9

Conclusion

This chapter provides a summary of the research findings and practically applied evaluations of geographic data in a SIEM context. The research highlights the need for expanding a SIEM security infrastructures intelligence through the exploitation of existing data, and has a strong focus on encompassing much-needed privacy compliance. The feasibility of geographic location is discussed and evaluated in these areas, its application within the framework, and the resulting outcomes. Finally, the results are weighed with the hypothesis put forth in this research; concluding the research achievements and providing recommended future work.

9.1 Research Summary

Security information and event management technology refers to tools and processes for the centralised real-time collection, integration, and analysis of log events occurring in a distributed system. The considerable advantage of centralised SIEM tools is the provision of unified interfaces to a variety of disparate data, while also allowing real-time correlation of the different events occurring in the collected parts, to effectively detect threats to the system[6].

This research involved an investigative study into the augmentation of SIEM technologies within an enterprise environment. A managed enterprise environment has many issues regarding managed security operating with clients of typically large enterprise level. An existing managed enterprise was considered as the source of identification of current security challenges still faced by these environments. The most common security issues discovered were presented as misuse cases;

- brute-force password attacks (MC-5.5.1),
- attempted unauthorised logins (MC-5.5.2),
- malicious SQL injection (MC-5.5.3),
- session hijacking through XSS (MC-5.5.4) and lastly,
- worm propogation (MC-5.5.5).

To determine the possibilities of advancing SIEM frameworks, by addressing these challenges and the main concerns affectings security; the research proposed the use of geographic data as

exploited in geographic information systems, to increase certainty of authenticity of certain individuals.

To evaluate this feasibility, an insight into SIEM architecture was required. SIEMs can be defined as holistic approaches to security analysis and detection. The framework centralises and performs mining procedures on the collected data to produce insights into the conditions of all devices and systems being monitored. The methods applied for security data management and exploitation for analytics are a primary area of consideration for geographic data exploitation. Such methods include the techniques of normalisation and pattern correlation detections. Attacks can be profiled and analysed using these correlative abilities to detect patterns of suspicious behaviour on the network. The method of correlation monitors incoming data provided in logs defining certain aspects of an activity, for example a login process would follow an authentication procedures. Deviations from the normal pattern of such a process raises flags for the attention of security administrators in discovering potential attacks from the mass of real-time collecting of data.

For inclusion within existing techniques such as correlation, geographic co-ordinate data and its application in various security-enforcing systems through GIS are investigated. The advantages of geolocation data identified from existing systems with applied geographic data can determine the augmentation potential of security within a SIEM.

To state it briefly the geolocation application investigated in the areas of security and analytics filtering, provided the following advantages;

- Enhanced visualisation[33], geographic perspective aided the creation of contextual viewing for administrators.
- Aiding user preparedness and rapid response, the use of a geographic context enabled a user to quickly identify an area of concern and seperate it from the unconcerning data.
- Selective display[33], isolating errors or situations through filters based on geographic location. This assisted in better analysis through isolation of specific areas for evaluation.
- Predictive modelling, data such as geographic location has an element of identification as a characteristic in certain situations. The behaviour of a target in terms of location with relation to time realises movement patterns, which can be used to create predictions for future events.
- Better network analysis and simulation, provided through the context of situational awareness provided through mapping from geolocation.
- Improved decision making, this is aided through the increased ability to evaluate a situation and a context, have a bird's eye view.
- Facilitating dynamic visual intrusion detection[13], a network system state visualised through physical locations, helps link a physical infiltration to the virtual context, allowing the security to be considered in context at all layers.

- Risk assessment of assets[47], locational context can provide the means to prevent transgression of failures and prioritising certain assets depending on their physical implications if compromised. For example, a dam is a high priority critical infrastructure, if an attacker infiltrated the automated control unit the repurcussions are extensive. The use of geolocation to collate the asset value to potential hazard is instrumental value.

The advantages of geolocation data identified were then discussed in the context of SIEM advancement in the two areas - security and privacy. The correct application of geolocation to augment existing security techniques present in SIEM tools with applied privacy considerations leads to the satisfaction of the hypothesis stated in this thesis. The hypothesis was stated as the following;

Location-based information enhances SIEM capability to perform advanced security detection. Privacy-enforcing procedures on geolocation in SIEMs and meta-systems alike are necessary and enforceable.

Towards SIEM security, a matrix of geolocation based security procedures was introduced that mapped their contributive abilities in the driving areas of security of today. The driving areas are mobility, the cloud, advanced persistent threats and regulatory compliance. User authentication through 'contextual' analytics such as location-based authentication has been predicted to rise in enterprises by more the 30% by 2016[2].
The introduced procedures can be applied within SIEM security through the integration of geolocation in correlation and analytic procedures. The effective results from application of these techniques are dependant on the accuracy of the user geographic data, in which case a set of geolocation accuracy techniques were evaluated. Wang's street-level client independant IP geolocation was resolved as the best solution, requiring only the IP address of the concerned user for a street-level estimation of users whereabouts. The method uses a combination of pinging and landmark range search using surrounding routers.

Towards SIEM privacy, a guideline based on the EU Data Protection Directive[15] was introduced. This considered the privacy implications from the Directive to a SIEM. SIEMs can be seen as meta-systems containing information of other systems, thus an extremely sensitive commodity of system or network. The enforceable requirements for SIEMs to adhere to in order to increase protection of user data rights and privacy are made explicit in this guideline. Regarding geolocation and privacy, techniques of anonymisation such as the use of generalisation approaches are discussed. The application of this technique to support the efforts towards enforcement of privacy is shown feasible for geolocation data.

Therefore, the inclusion of geolocation for security procedures through privacy-enforcing procedures supports both areas of argument presented in this research.
Once determined as a suitable solution, further study was carried out to explore the integration of geolocation into an existing SIEM. The implementation consisted the utilisation of an existing open-source SIEM, OSSIM, and selected tools of the MASSIF SIEM framework. Using a feasible integration, an integrated prototype was created and tested.

OSSIM was used to demonstrate the integration of geolocation in an existing SIEM for security analysis. The MASSIF tools were used to develop the privacy-enforcing technique

on geolocation information entering this SIEM. The MASSIF tools also ensured the forensic credibility of the data through protected storage. The privacy implications were addressed of geolocation data through complete application in the prototype experiment.

The results of applying geolocation in an incident detection procedure addressing the brute-force misuse case(MC-5.5.1) was fully accomplished. The data used in this test was anonymised by the MASSIF GET tool prior to it's exploitation within the SIEM. The anonymisation ranges were tested at various levels and evaluated in their feasibility.

The implementation was assessed in the context of the relevant SIEM solution and framework, how they augment functionalities of these SIEMs and to determine if the application is fully supported and does not mitigate SIEM standard of delivery.
In conclusion, both the security application of geolocation-based detection and the anonymisation procedures undergone on the geolocation data proved feasible and successful for addition within a SIEM.

9.2 Discussion

In the preliminary investigative stages of research, a collection of questions were raised, towards satsifying the aims of this research. These questions are reviewed here, to determine if they are satisfied supporting the achievement of the research hypothesis.

9.2.1 What can geolocation data provide for SIEMs? Is it significant?

As the accuracy and availability of geolocation data has increased, it's application can be seen in many avenues of security. The use of geolocation performs a provision of contextual analytics, a stringent need in large and distributed environments for better awareness within analysis of copious levels of security information.

Using geolocation as a second-level authentication or progressive authentication promotes stronger security for mobile users, using the mobility as the method of contextual identification. Process monitoring techniques can apply location in the authentication cycle, to support the user authentication with their location as part of their logon identity. This significantly curbs the possibilites of an unnauthorised user login. A compromised login is the gateway for larger security crimes such as advanced persistent threats. By addressing the first phase of attack through illegal user login the larger threats, harder to trace can be mitigated.

Another application within SIEMs is the use of location-based blocking, the unified approach of SIEMs enables analysts to identify the patterns in terms of bad reputation sources. Applying this through the concept of bad geo-reputation, analysts can identify physical locations on the globe which can be blocked for access based on users that repeatedly utilise suspicious proxies or malicious behaviour.

In terms of regulatory compliance by SIEMs, geolocation can successfully enforce access restrictions. Therefore, geographic restrictions with data movement by data managers, supporting compliance is a salient contributive factor. The creation of security policies to consider geolocations when handling data within security frameworks such as SIEMs supports the

compliance for the entire organisational system, enforcing it in a collateral effect of organisation protection.

The use of geographic data for contextualisation for better decision making within SIEMs is an important utility for security analysts. The provision of situational awareness plays a significant role in detection of suspicious user activity.
Finally, rule detection based on geolocation can provide insight into anomalous behaviour through the monitoring of geographic footprints.

9.2.2 How does geolocation data fare with privacy concerns?

Geolocation data is considered as personally identifiable information. It has been proven[14] that a user can be identified simply through an analysis of their geographic patterns. Therefore, this data needs to be ensured with privacy-enforcing techniques wherever the data is held or used.
Various cloaking techniques were discussed that can be applied to geolocation data in efforts to mainitaining privacy. The method of generalisation was the chosen anonymisation approach demonstrated in this research.

9.2.3 Can geolocation data augment SIEM event analysis tools that already exist?

The geolocation security techniques identified earlier, can be applied in the data-to-information layer of SIEMs. Correlation within the SIEM provides a base for creating rules such as those encompassing geolocation. The inclusion of the geolocation within these analyis engines was demonstrated to be possible. Additionally, it was demonstrated using the MASSIF SIEM tool, the GET, the performs normalisation and translation, was able to perform anonymisation procedures on the geolocation before entry into the centralised SIEM center.

9.3 Results Achieved

Returning to the hypothesis stated for this study; collated with the research efforts made in investigations and proof implementation, the achievements are reviewed;

- *Location-based information enhances SIEM capability to perform advanced security detection.*
  Location-based security techniques were brought forward from research investigations into existing applications of such methods in industry. The techniques were applied to current SIEM challenges towards solving them. A misuse case within an enterprise was successfully evaluated through the use of geolocation analysis filtering. The result highlighted the solution of increasing priority levels to events from high risk countries for faster response and remediation to alarms with low false negative probabilities.

- *Privacy-enforcing procedures on geolocation in SIEMs and meta-systems alike are necessary and enforceable.*
  The process of anonymisation was performed on geolocation data before sending the events to the OSSIM SIEM solution. The anonymisation allowed the geolocation data

to be used for security evaluation procedures while ensuring user privacy is not compromised simply for expediency.

In considerations where organisations inform users of geolocation tracking and retrieve their consent for use, multiple geolocation security enhancement methods have been discussed in the context of the four main driving areas of security. Augmentation of security in these areas using geographic data highlight significant security detection abilities in the case of unauthorised users and the provision of situational awareness to security management frameworks.

The enforcement of privacy-preserving approaches in SIEMs where users have concerns on personal geolocation privacy enforcement (for example on servers outside certain jurisdictions), is facilitated through the provision of anonymisation levels, while ensuring it can still be applied in security analytics even at the strictest level(Level 0) of anonymisation.

The study supports both dimensions of the research hypothesis through background investigations and fully integrated implementation analysed in the context of an existing SIEM, and in light of this is considered to be confirmed.

9.4 Future Work: Directions of the SIEM

Compliance requirements and growing concerns over more targeted and sophisticated attacks have boosted interest in security information and event management systems. Companies need to have greater ability to monitor their systems and generate compliance reports to meet regulatory requirements. SIEM systems are a good solution for this, but are typically expensive to deploy as well as complex to operate and manage[32].
However, organisations are looking to SIEMs in the cloud as a method of overcoming these challenges of cost and complexity.

Turning security and information management into a security cloud service will enable smaller companies lacking the high technical skills required to deploy a SIEM internally to reap these benefits. Its typical for a company to move from manual to management and finally to SIEM.

There are many applications of SIEMs to cloud technology, such as running SIEMS on clouds, SIEMs as cloud monitors for operators and users, cloud security services and cyber-physical systems. There are alot of advantages that come from this, if we look at SIEM as a managed security service in the cloud - it gains the additional cloud advantages of transportation cost cuts, unlimited storage and high computational resources. SIEM as a cloud management infrastructure component can provide analysis and reporting on services. And finally, SIEMs can be used as the core of a trust-enabler for cloud computing services.

For future work, it is suggested to use geolocation in aiding the transition of SIEM technology as a cloud-based service to harness the advantages of the cloud technology. There are many areas of concern with regards to cloud technology in terms of trust, data rights and usage.

Cloud providers have to do more to assuage the security concerns of potential customers, turning over internal security data to a cloud provider requires trust, and nearly half of all users of cloud services desire more clarity on providers security precautions, according to

Gartner[32]. The major roadblock to full adoption of cloud computing has been this concern regarding the security and privacy of information.

The National Institute of Standards and Technology (NIST) collaborated with representative from RSA and Intel Corporation to develop a method for trusted geolocation for cloud workloads. The proposition utilises the physical hardware as the root of trust that is monitored to ensure workloads are not placed in prohibited places, such as those falling out of a certain jursidiction. The solution takes a three-stage approach summarised below[54]:

- Stage 0: Trustworthy Platform Attestation. This prerequisite stage is making sure the platform a workload is placed on is trustworthy. By securing the cloud server and continous security configuration verification of the cloud server's BIOS and hypervisor while it's runnning.
- Stage 1: Secure Migration. After stage 0, the permission of workload migrations between homogeneous trusted servers – trusted servers in the same cloud with the same hardware and virtualization architectures.
- Stage 2: Geolocation-Based Secure Migration. This stage brings in the consideration of geolocation to Stage 1 and 0. Geolocation needs to be confirmed before placing a workload onto a server and continually throughout the run of the workload. This is done to ensure workloads are placed on servers with right jursidictions and should a geolocation policy change it can adjusted and triggered immediately from the geolocation checks on the running workloads. If a conflict arises, it can be reported immediately and owners of the workload can either cease the running of the workload or transfer it to another server.

The implementation of this proposition in the context of a cloud-based SIEM environment can be seen as the suggested future work, having incorporated geographic data within SIEMs and implemented the use with applied anonymisation techniques. The NIST technique can augment the SIEM security-as-a-service approach towards safer cloud security foundations. Ensuring trustworthiness of platforms and restricting workloads within geographic constraints to observe regulatory compliance is a salient consideration of future security.

References

[1] Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., et al. Challenges and opportunities with big data. *A community white paper developed by leading researches across the United States* (2012).

[2] Allan, A. 2013 Gartner Magic Quadrant for User Authentication. *Vasco* (March 2013). G00231072. https://www.vasco.com/images/12-09-2013_magic_quadrant_for_user_authentication_december_2013.pdf. Accessed online 5 January 2014.

[3] Alliance, C. S. SecaaS Category 7: Security Information and Event Management Implementation Guidance. https://cloudsecurityalliance.org/download/secaas-category-7-security-information-and-event-management-implementation-guidance/. Accessed online 15 August 2012.

[4] Assanger, S. Enhancing security information and event management capability using unsupervised anomaly detection, 2012.

[5] Australia, A. G. G. Geocentric Datum of Australia (GDA), January 2000. http://www.ga.gov.au/scientific-topics/positioning-navigation/geodesy/geodetic-datums/historical-datums-of-australia/australian-geodetic-datum-agd. Accessed online 8 August 2014.

[6] Bertino, E. Data Protection from Insider Threats. *Synthesis Lectures on Data Management 4*, 4 (2012), 1–91. http://www.morganclaypool.com/doi/abs/10.2200/S00431ED1V01Y201207DTM028.

[7] BIS, PwC, and InfoSecurity. Information Security Breaches Survey 2014 Technical Report. Department for Business and Innovation Skills. BIS/14/767. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/307296/bis-14-767-information-security-breaches-survey-2014-technical-report-revision1.pdf. Accessed online 15 August 2014.

[8] Chuvakin, A. Using SIEM Technology to Identify Unauthorised Access Attempts. *Gartner* (2014). TechTarget. http://searchsecurity.techtarget.com/tip/Use-SIEM-technology-to-identify-unauthorized-access-attempts. Accessed Online on 16 August 2014.

[9] Commision, E. Proposal for a regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on

the free movement of such data (general data protection regulation). *COM (2012) 11 final, 2012/0011 (COD), Brussels, 25 January 2012* (2012).

[10] Commission, E. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions safeguarding privacy in a connected world: A european data protection framework for the 21st century. *COM (2012)* (2012). Brussels.

[11] Conran, B. CyberSecurity: An Innovative Approach to Advanced Persistent Threats, February 2014. RSA Conference 2014. `www.rsaconference.com/writable/presentations/file_upload/ast1-r01-cybersecurity-an-innovative-approach-to-advanced-persistent-threats.pdf`. Accessed online 11 June 2014.

[12] Cook, J. Details for Computing Distance using Lat/Long Coordinates, 2014. `http://www.johndcook.com/lat_long_details.html`. Accessed online 18 August 2014.

[13] Coppolino, L., Antonio, S. D., Formicola, V., and Romano, L. Integration of a System for Critical Infrastructure Protection with the OSSIM SIEM Platform : A dam case study. 199–212.

[14] de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. Unique in the crowd: The privacy bounds of human mobility. *Nature srep. 3* (2013).

[15] Directive, E. R. E. Proposal for a directive of the european parliament and of the council on the promotion and use of energy from renewable sources. *COM (2008) 19* (2008).

[16] Gaber, C., Coppolino, L., Khan, H., Fonseca, R., Formicola, V., Vianello, V., Viinika, J., Goller, A., Chechulin, A., REpp, J., Arce, C., Torres, R., Gonzalez, S., and Diaz, R. Acquisition and Evaluation of Results, October 2013. Deliverable, D2.3.3, Project MASSIF EC FP7-257475.

[17] Gebhart, G. Worm Propogation and Countermeasures. *SANS Institute InfoSec Reading Room* (February 2014). `http://www.sans.org/reading-room/whitepapers/malicious/worm-propagation-countermeasures-1410`. Accessed online 1 August 2014.

[18] Gellert, R., and Gutwirth, S. Beyond Accountability, the Return to Privacy? *Managing Privacy Through Accountability* (2012), 261 – 283. Ed. Daniel Guagnin, Leon Hempel, Carla Ilten, Carla, Inga Kroener, Daniel Neyland, Hector Postigo.

[19] Ghinita, G. Privacy for location-based services. In *Synthesis Lectures on Information Security, Privacy, & Trust*, vol. 3188 of *Elisa Bertino and Ravi Shandu*. Morgan & Claypool, 2013.

[20] Goller, A., Ramos, M., and Charlton, D. OSSIM Integration, June 2013. Deliverable, D5.4.2, Project MASSIF EC FP7-257475.

[21] Goller, A., Viinikka, J., Neves, N. F., and Debar, H. Integration Specifications, May 2013. Deliverable, D5.4.1, Project MASSIF EC FP7-257475.

[22] Gonzalez, S., Montfaucon, A., Coppolino, L., Giot, R., Khan, H., and et al. Prototypes and Demonstrators Deployment, July 2013. Deliverable, D2.3.2, Project MASSIF EC FP7-257475.

[23] Grijalva, S., and a.M. Visnesky. The Effect of Generation on Network Security: Spatial Representation, Metrics, and Policy. *IEEE Transactions on Power Systems 21*, 3 (Aug. 2006), 1388–1395.

[24] Gruteser, M., and Grunwald, D. Anonymous Usage of Location Based Services through Spatial and Temporal Cloaking, July 2003. ACM/USENIX MobiSys.

[25] Hachicha, H. Hands-on OSSIM 2.3: Your quick and dirty guide to understanding and deploying OSSIM, 2014. `http://www.scribd.com/doc/112726589/OSSIM-Hands-On-pdf`. Accessed online 7 March 2013.

[26] Hutchison, A. Unlocking the Opportunity of the SIEM. *Information Security* (March 2012). Volume 2, pg 13 - 19.

[27] Hutchison, A., Khan, H., and et al. Tool Adaptation, May 2013. Deliverable, D2.2.1, Project MASSIF EC FP7-257475.

[28] internetlivestats.com. Internet Live Stats, 2014. `http://www.internetlivestats.com/internet-users/`. Accessed online 5 May 2014.

[29] Johnson, C. W., and Mclean, K. Tools for Local Critical Infrastructure Protection : Computational Support for Identifying Safety and Security Interdependencies between Local Critical Infrastructures retweepnd thategoryI incluencinras provwids. 1–6.

[30] Khan, H., and Hutchison, A. Data privacy implications for security information and event management systems and other meta-systems. In *Cyber Security and Privacy*. Springer, 2013, pp. 79–90.

[31] Lanzini, G., Bargh, M. S., and Hulsebosch, B. *Electronic Notes in Theoretical Computer Science 197*.

[32] Lemos, R. More Companies eyeing SIEM in the Cloud. `http://searchcloudsecurity.techtarget.com/news/2240147704/More-companies-eyeing-SIEM-in-the-cloud`. Accessed online 15 April 2012.

[33] Liu, Y. Visualization of power system static security assessment based on GIS. *POWERCON '98. 1998 International Conference on Power System Technology. Proceedings (Cat. No.98EX151) 2* (1998), 1266–1270.

[34] Liu, Y., and Qiu, J. Visualization of Power System Static Security Assessment Based on GIS. *IEEE Vol 4* (1998).

[35] Lonvick. RFC 3164: The BSD Syslog Protocol. IETF. `http://www.ietf.org/rfc/rfc3164.txt`. Accessed online 21 June 2014.

[36] Lorenzo, J. M. AlienVault Users Manual 1.0, 2011. `https://scadahacker.com/library/Documents/Manuals/AlienVault_Users_Manual_1.0.pdf`. Accessed online 11 May 2012.

[37] Macvittie, L. Geolocation and Application Delivery. *F5 White Paper* (2010).

[38] Maguire, D. J., Goodcgid, M. F., and Rhind, D. W. An overview and definition of GIS. *Geographical information systems: Principles and applications 1* (October 1991), 9–20.

[39] Marketwire. Frost & Sullivan: Greater Sophistication of Cyber Crimes Encourages Adoption of Security Information and Event Management, March 2011. http://www.marketwired.com/press-release/Frost-Sullivan-Greater-Sophistication-Cyber-Crimes-Encourages-Adoption-Security-Infor htm. Accessed online 14 June 2012.

[40] MASSIF, P. Description. Project MASSIF EC FP7-257475. http://www.massif-project.eu/description/. Accessed 26 May 2014.

[41] MASSIF Project, C. Scenario Requirements, March 2011. Deliverable, D2.1.1, Project MASSIF EC FP7-257475.

[42] MasterCard. Press Releases. MasterCard and Syniverse Deliver Peace of Mind for Mobile Users, 2014. http://newsroom.mastercard.com/press-releases/mastercard-and-syniverse-deliver-peace-of-mind-for-mobile-users/. Accessed online 11 August 2014.

[43] Meyer, R. Detecting Attacks on Web Applications from Log Files, 2008. http://www.sans.org/reading-room/whitepapers/logging/detecting-attacks-web-applications-log-files-2074. Accessed online 1 February 2014.

[44] Microsoft. Regular Expression Syntax. Microsoft Developer Network (MSDN). https://msdn.microsoft.com/en-us/library/ae5bf541(v=vs.90).aspx. Accessed online 10 January 2014.

[45] Miller, R. Metadata is Powerful Content, June 2013. http://www.fiercecontentmanagement.com/story/power-metadata/2013-06-103. Accessed online 5 May 2014.

[46] Murray, A. T., and Grubesic, T. H. 1 Overview of Reliability and Vulnerability in Critical Infrastructure. 1–8.

[47] Murray, A. T., Matisziw, T. C., and Grubesic, T. H. Critical network infrastructure analysis: interdiction and system flow. *Journal of Geographical Systems 9*, 2 (January 2007), 103–117.

[48] Neves, N. F., Kuntze, N., Sarno, C. D., and Vianello, V. Resilient SIEM Framework Architecture, Services and Protocols, September 2013. Deliverable, D5.1.4, Project MASSIF EC FP7-257475.

[49] Ohm, P. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review 57* (2010), 1701.

[50] ProM. Process Mining Workbench, 2014. http://promtools.org/prom6. Accessed 10 July 2014.

[51] Rogan, J., and Chen, D. Remote sensing technology for mapping and monitoring land-cover and land-use change. *Progress in planning 61*, 4 (2004), 301–325.

[52] Rouse, M. Advanced Persistent Threat(APT), 2014. `http://searchsecurity.techtarget.com/definition/advanced-persistent-threat-APT`. TechTarget. Accessed Online on 16 August 2014.

[53] Sastry, N., Shankar, U., and Wagner, D. Secure verification of location claims. In *Proceedings of the 2nd ACM workshop on Wireless security* (2003), ACM, pp. 1–10.

[54] Scarfone, K. NIST Cloud Security Spec Addresses Cloud Geolocation, Data Security. `http://searchcloudsecurity.techtarget.com/news/2240147704/More-companies-eyeing-SIEM-in-the-cloud`. Accessed online 17 May 2014.

[55] Sheyner, O., and Wing, J. Tools for generating and analyzing attack graphs. In *Formal Methods for Components and Objects*, F. de Boer, M. Bonsangue, S. Graf, and W.-P. de Roever, Eds., vol. 3188 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004, pp. 344–371.

[56] Sweeney, L. K-anonymity: A Model for Protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst 10*, 5 (October 2002), 557–570.

[57] Swift, D. A Practical Application of the SIM/SEM/SIEM automating Threat Identification, 2007. `http://www.sans.org/reading-room/whitepapers/malicious/worm-propagation-countermeasures-1410`. Accessed online 1 August 2014.

[58] Veness, C. Calculate distance, bearing and more between latitude/longitude points. `http://www.movable-type.co.uk/scripts/latlong.html` . Accessed online 15 July 2014.

[59] Vert, G., Frincke, D. A., and McConnell, J. C. A visual mathematical model for intrusion detection. In *Proceedings of the 21st National Information Systems Security Conference* (1998), pp. 329–337.

[60] Walden, I. Accessing Data in the Cloud: The Long Arm of the Law Enforcement Agent. *Queen Mary School of Law Legal Studies Research Paper No. 74/2011* (November 2011).

[61] Wang, Y., Burgener, D., Flores, M., Kuzmanovic, A., and Huang, C. Towards Street-level Client-independent IP Geolocation. In *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation* (Berkeley, CA, USA, 2011), NSDI'11, USENIX Association, pp. 365–379.

[62] Whimsley. Data Anonymisation and Re-identification, June 2013. `http://whimsley.typepad.com/whimsley/2011/09/data-anonymization-and-re-identification-some-basics-of-data-privacy.html`. Accessed online 5 May 2014.

[63] Whuber. How to Measure the Accuracy of Latitude and Longitude?, 2014. `http://gis.stackexchange.com/questions/8650/how-to-measure-the-accuracy-of-latitude-and-longitude`. Accessed online 17 August 2014.

[64] Wolthusen, S. GIS-based Command and Control Infrastructure for Critical Infrastructure Protection. *First IEEE International Workshop on Critical Infrastructure Protection (IWCIP'05)* (2005), 40–50.

Appendix A

Event Rules

Event Rules script for OSSIM, so that it can recognise log events through event IDs and prioritise accordingly.

```
1  -- Herah Test TSOM logs
2  -- Plugin id: 9010
3
4  DELETE FROM plugin WHERE id = "9010";
5  DELETE FROM plugin_sid where plugin_id = "9010";
6
7  INSERT INTO plugin (id, type, name, description) VALUES (9010, 1, 'herah-tsom', 'Herahs TSOM log
       testing');
8  INSERT INTO plugin_sid (plugin_id, si
9  d, category_id, class_id, name, priority, reliability) VALUES (9010, 528, NULL, NULL, 'Successful logon', 1,
       3);
10 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       4624, NULL, NULL, 'Successful logon', 1, 3);
11 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       540, NULL, NULL, 'Successful Network logon', 1, 3);
12 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       529, NULL, NULL, 'Logon Failure: Unknown username or bad password', 1, 3);
13 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       4625, NULL, NULL, 'Logon Failure', 1, 3);
14 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       4672, NULL, NULL, 'Special Privileges assigned to new logon', 1, 3);
15 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       4634, NULL, NULL, 'User logoff', 1, 3);
16 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       530, NULL, NULL, 'Logon Failure: Account logon time restriction violation', 1, 3);
17 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       531, NULL, NULL, 'Logon Failure: Account currently disabled', 1, 3);
18 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       532, NULL, NULL, 'Logon Failure: The specified user account has expired', 1, 3);
19 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       533, NULL, NULL, 'Logon Failure: User not allowd to logon at this computer', 1, 3);
20 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       534, NULL, NULL, 'Logon Failure: The user has not been granted the requested logon', 1, 3);
21 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       537, NULL, NULL, 'Logon Failure: An unexpected error occurred during logon', 1, 3);
22 INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010,
       539, NULL, NULL, 'Logon Failure: Account locked out', 1, 3);
```

```
INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010, 576, NULL, NULL, 'Special privileges assigned to new logon', 1, 3);
INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010, 538, NULL, NULL, 'User logoff', 1, 3);
INSERT INTO plugin_sid (plugin_id, sid, category_id, class_id, name, priority, reliability) VALUES (9010, 551, NULL, NULL, 'User initiated logoff', 1, 3);
```

A.1

Custom OSSIM Plugin defined to read in events in this case, the plugin was created to test windows event data and pull through its geographical information.

```
 [DEFAULT]
plugin_id=9010


[config]
type=detector
enable=yes

#process=
#start=no
#stop=no
#startup=
#shutdown=

source=log
location=/var/log/herahlogs/brutenewwithgis.txt

create_file=false


[ossim-herahk-format]
event_type=event

# event log format: TBS;123423434;;3345454534;223432;...etc basically a semi-colon csvd version of logs
regexp="^([^;]*);([^;]*);([^;]*);(?P<datee>\d+);([^;]*);([^;]*);([^;]*);
([^;]*);([^;]*);([^;]*);([^;]*);(?P<user>[^;]*);([^;]*);(?P<src>[^;]*);
([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);(?P<dst>[^;]*);([^;]*);([^;]*);
([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);([^;]*);(?P<eventid>[^;]*)
;(.*)$"

#------------------------ WORKS --------------------
plugin_id=9010
plugin_sid={$eventid}
username={$user}
src_ip={$src}
dst_ip={$dst}
date={normalize_date($datee)}
#geo_lat={$lat}
#geo_long={$long}
```

A.2

Script for timedate adjustment for log data to simulate real-time, enabling detection in OSSIM based on the three day limitation.

```
#!/usr/bin/env python
# −∗− coding: utf−8 −∗−

import argparse
import datetime
import time
import os
import sys
import socket
from os import path
import re
import urllib.request
import pygeoip

lat='lat'
lon='lon'
#pattern = '<<([^>]∗)>>'
pattern = dict(tsom='^TBS;[^;]∗;[^;]∗;([0−9]+);∗$',mcafee='(herah)',apache='will be some regular
    expression')
mcaffee_pattern = ''
win_tsom_pattern = ''
uct_apache_pattern = ''
datematchto = 'nodate'
datachangeto = 'nodate'
old_date = ''
timestamp = False
geo_line = ""

def user_replace(match):
  print('in the replace function')
  if (isinstance(match, str)):
    matchfrom = match
  else:
    matchfrom = match.group(1);

  print(matchfrom)

  global datematchto
  global datechangeto
  global timestamp

  if(matchfrom != datematchto):
    datematchto = matchfrom

    if(timestamp == True):
      datechangeto = input('Change the current field from %s to? ' % time.strftime('%Y−%m−%d %H:%M
          :%S', time.localtime(int(matchfrom))))
      return datetime.datetime.strptime(datechangeto, '%Y−%m−%d %H:%M:%S').strftime('%s')
    else:
      datechangeto = input('Change the current field from %s to?' % matchfrom)
      return datechangeto
```

```
50    else:
51      if(timestamp == True):
52        return datetime.datetime.strptime(datechangeto,'%Y-%m-%d %H:%M:%S').strftime('%s')
53      else:
54        return datematchto
55
56  def main():
57
58
59
60      print('Event Generator for events from the following sources - CSV-ed McAffee, CSV-ed TSOM
            Windows and Apache Logs:')
61      print(' The following arguments are required - <inputfile> <outputfile> <log source (mcaffee/tsom/
            apache)>')
62      parser = argparse.ArgumentParser(description="***** \n \
63        This script regenerates the events as if they were real time so they can work with Alienvaults real time
              analysis",
64      epilog="Many thanks, \n created by Herah A Khan. *****")
65      parser.add_argument('infile', type=argparse.FileType('r'))
66      parser.add_argument('outfile', type=argparse.FileType('w'))
67      parser.add_argument('logsource' , type=str)
68      args = parser.parse_args()
69      print('Working through ' + args.logsource + ' logs...')
70      gi = pygeoip.GeoIP('c:/Users/Herah/Documents/Mine/GeoIP/GeoLiteCity.dat')
71
72
73      matcher = re.compile(pattern[args.logsource])
74
75      count = 0;
76      for line in args.infile:
77        if (args.logsource == 'tsom'):
78          timestamp = True
79          old_date_array = re.split(';', line)
80          arraylen = len(old_date_array)
81          if(arraylen > 3):
82            print(old_date_array)
83            if (len(old_date_array[15]) > 1) :
84              geosrcinfo = gi.record_by_addr(old_date_array[15])
85              print(geosrcinfo['latitude'])
86              old_date_array[19] = str(geosrcinfo['latitude'])
87              old_date_array[20] = str(geosrcinfo['longitude'])
88
89            if (len(old_date_array[22]) > 1):
90              geodstinfo = gi.record_by_addr(old_date_array[22])
91              print(geodstinfo['latitude'])
92              old_date_array[26] = str(geodstinfo['latitude'])
93              old_date_array[27] = str(geodstinfo['longitude'])
94            print(old_date_array[3])
95            old_date_array[arraylen -1] = ";\n"
96            geoline = ""
97            geoline = ";".join(old_date_array)
98            #new_line = re.sub( old_date_array[3], user_replace(old_date_array[3]), geoline)
99            #print('The new tsom line is' + new_line)
100           args.outfile.write(geoline)
101       if (args.logsource == 'apache'):
102         new_line = matcher.sub(user_replace, line)
103         print('The new line is' + new_line)
```

```
            args.outfile.write(new_line)
        if (args.logsource == 'mcafee'):
            old_date_array = re.split(';', line)
            if(len(old_date_array) > 5):
              print(old_date_array)
              print(old_date_array[4])
              print(old_date_array[27])
              response = urllib.request.urlopen('http://api.hostip.info/get_html.php?ip=' + old_date_array[27]
                    + '&position=true').read()
              data = str(response.decode('utf-8'))
              print(data)
              m = re.match('^.*Latitude:(.*)Longitude:(.*)\nIP.*$', data)
              print(m)
              old_date_array[32] = lon + ',' + lat
              geoline = ""
              geoline = ";".join(old_date_array)
              new_line = re.sub( old_date_array[4], user_replace(old_date_array[4]), geoline)
              print('The new line is' + new_line)

              args.outfile.write(new_line)

    print('End of regeneration. Goodbye. ')

    args.infile.close()
    args.outfile.close()

if __name__ == '__main__':
    main()
```

A.3

Script to calculate the Haversine distance for data points retrieved from all countries. Calculations for each data point are collected and stored in a CSV file for download.

```
 geocoder = new google.maps.Geocoder();

       function measure(lat1, lon1, lat2, lon2){ // generally used geo measurement function
    var R = 6378.137; // Radius of earth in KM
    var dLat = (lat2 - lat1) * Math.PI / 180;
    var dLon = (lon2 - lon1) * Math.PI / 180;
    var a = Math.sin(dLat/2) * Math.sin(dLat/2) +
    Math.cos(lat1 * Math.PI / 180) * Math.cos(lat2 * Math.PI / 180) *
    Math.sin(dLon/2) * Math.sin(dLon/2);
    var c = 2 * Math.atan2(Math.sqrt(a), Math.sqrt(1-a));
    var d = R * c;
    return d * 1000; // meters
}


//download results in CSV format file for parsing and graphing
function getCSV(distarray)
{
                    var csvRows = [];

```

```
21                for(var i=0;i < distarray.length ; ++i){
22                    csvRows.push(distarray[i].join(','));
23                }
24
25                var csvString = csvRows.join("%0A");
26 var a = document.createElement('a');
27 a.href = 'data:attachment/csv,' + csvString;
28 a.target = '_blank';
29 a.download = 'myFile.csv';
30
31 document.body.appendChild(a);
32 a.click();
33 }
34
35 //accuracy for a location in each country
36 function getCountry(country, map, distarray) {
37     geocoder.geocode( { 'address': country }, function(results, status) {
38         if (status == google.maps.GeocoderStatus.OK) {
39             var ln;
40             var lt;
41             var ltg = results[0].geometry.location.lat();
42             var lng = results[0].geometry.location.lng();
43           // map.setCenter(results[0].geometry.location);
44            var marker = new google.maps.Marker({
45                map: map,
46                position: results[0].geometry.location
47            });
48       // alert("Geocode was successful for the following reason: " + results[0].geometry.location.lat() );
49             //accuracy of 1
50             ln = Math.floor(lng);
51             lt = Math.floor(ltg);
52             var lnb = ln + 0.999999;
53             var ltb = lt + 0.999999;
54             var accuracy = measure(lt,ln,ltb,lnb);
55          // alert('Accuracy of ... ' + accuracy + 'check:' + lnb);
56             //accuracy to .1
57             ln = Math.floor(10 * lng) / 10;
58             lt = Math.floor(10 * ltg) / 10;
59             var lnb = ln + 0.099999;
60             var ltb = lt + 0.099999;
61             var accuracy1 = measure(lt,ln,ltb,lnb);
62             //accuracy to .2
63             ln = Math.floor(100 * lng) / 100;
64             lt = Math.floor(100 * ltg) / 100;
65             var lnb = ln + 0.009999;
66             var ltb = lt + 0.009999;
67             var accuracy2 = measure(lt,ln,ltb,lnb);
68               //accuracy to .3
69             ln = Math.floor(1000 * lng) / 1000;
70             lt = Math.floor(1000 * ltg) / 1000;
71             var lnb = ln + 0.000999;
72             var ltb = lt + 0.000999;
73             var accuracy3 = measure(lt,ln,ltb,lnb);
74               //accuracy to .4
75             ln = Math.floor(10000 * lng) / 10000;
76             lt = Math.floor(10000 * ltg) / 10000;
77             var lnb = ln + 0.000099;
```

```
78                  var ltb = lt + 0.000099;
79                  var accuracy4 = measure(lt,ln,ltb,lnb);
80                  distarray.push([country, accuracy, accuracy1, accuracy2, accuracy3, accuracy4]);
81  //wait = true;
82    // setTimeout('wait=true',2000);
83              // alert(country + ' acc: ' + accuracy +' acc1: ' + accuracy1+ 'acc2: ' + accuracy2+ ' acc3: ' +
                    accuracy3 + 'acc4: ' + accuracy4);
84            } else {
85              alert("Geocode was not successful for the following reason: " + status);
86            }
87      });
88  }
```

Appendix B

Component Configuration

GET	REB	RES
GET - version D3.4.4 Dispatcher - version 12/06/2013 GET Access Point (GAP) version 12/06/2013 GET Manager - version 12/06/2013 MASSIF Event Handler (MEH) version 12/06/2013 RegistryService version 12/06/2013 SenderAgent - version 12/06/2013	REB - version V1.0.1 REBreceiverStubCINI.zip version 29/04/2013	RES- version V1.0

B.0.1 GET Component

Installing GET consists in extracting all the files inside the archive to a directory. The Sender Agent module needs to be flagged to use the REB with the appropriate flags. The GET does not require any further configuration besides the storage of the parser JAR files, Macafee_v1.jar and Winser_v1.jar to be used by its GAP module, in the relevant folder of the GET placed in the relevant virtual machine.

B.0.2 REB Component

1. Since REB runs over UDP/IP, it is necessary to allow UDP traffic between the hosts where the GET and CEP are running.

2. In order to allow Java applications (GET) to use the REB, it is necessary to include the directory where the REB.jar file is located in the environment variable CLASSPATH.

3. Finally, it is required to establish the following kernel parameters:

```
1 − Sets the maximum socket receive buffer size
2 net.core.rmem\_max=2904000
3 − Sets the maximum socket send buffer size
4 net.core.wmem\_max=2904000
```

Since certain operating systems impose a small limit on the maximum socket read/write buffer sizes, REB might not be able to set the ideal buffer size for the local UDP sockets during initialisation. As of the latest version, REB uses an ideal socket buffer size of 29,040,00 bytes (2.9 MB). Failure to set the buffer size to this value may result in increased packet dropping at receiver nodes. It is thus necessary to configure the maximum limit imposed by the local OS to a value of at least the ideal size[22].

Installing REB consists in extracting all the files inside the archive to a directory. The files include: a Jar file REB.jar which contains the REB API classes necessary for interaction with other Java programs (such as GET); a default configuration file nodes.cfg to be completed with information about the REB nodes; script files for generating shared cryptographic keys (genkeys) and validating the configuration (checkreb). The script files should be given permission to execute[22]. Optionally, the environment variable REB_HOME may be set to the directory containing the Jar file.

In order to verify the REB installation, the script checkreb can be used with each of the REB node ids using the command at terminal './checkreb x' where x is the node id[22].

In the simulation, the REB is used for transmitting events through a resilient channel. It does not modify the events or contribute further information for this scenario but rather facilitates the reliability and integrity of events as it is necessary especially in situations of high load data. The REB is installed and resides in the same machine as the GET, the massif-get machine. It is used in conjunction with the Sender Agent component of the GET as it is part of the transmission phase of events. The Sender Agent module needs to be flagged to use the REB with the appropriate flags, as show in the GET section of the MESI scenario.

B.0.3 RES Component

Installing the RES consists in extracting all the files inside the archive to a directory and does not require any further configurations besides running the jar through script specifying the relevant port and address.