



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Entwurf vom
8. November 2017

Masterarbeit

**Datenschutzfreundliche Speicherung
unternehmensinterner Überwachungsdaten mittels
Pseudonymisierung und kryptographischer
Schwellwertschemata**

vorgelegt von

Tom Petersen

geb. am 13. Dezember 1990 in Hannover

Matrikelnummer 3659640

Studiengang Informatik

eingereicht am 8. November 2017

Betreuer: Dipl.-Inf. Ephraim Zimmer

Erstgutachter: Prof. Dr. Hannes Federrath

Zweitgutachter: Prof. Dr. Mathias Fischer

Aufgabenstellung

Die technologiegestützte Bekämpfung von Insider-Angriffen im Unternehmenskontext basiert aktuell häufig auf der Analyse des Nutzerverhaltens einzelner Mitarbeiter und der Erkennung von Abweichungen zum erwarteten Normalverhalten. Diese sogenannte Anomalieerkennung benötigt umfassende Überwachungsdaten aller digitalen Endgeräte und Datenkommunikationssysteme zur Erstellung und eindeutigen Zuordnung von Nutzerprofilen. Dabei entsteht ein Konflikt mit dem Datenschutz der Mitarbeiter, da die Erhebung, Verarbeitung und Speicherung von Überwachungsdaten einen schweren Eingriff in die Privatsphäre und die informationelle Selbstbestimmung der Mitarbeiter darstellt. Um diesen Konflikt zu lösen, können auf der einen Seite Datenschutztechniken eingesetzt werden, die den unmittelbaren Personenbezug gesammelter Daten entfernen. Auf der anderen Seite kann mithilfe von Kryptographie die Rückgewinnung des Personenbezugs im Verdachtsfall und unter der Voraussetzung einer mehrseitigen Kollaboration ermöglicht werden.

Das Ziel der Masterarbeit ist die konzeptionelle Erarbeitung einer solchen datenschutzfreundlichen und mehrseitig sicheren Erhebung, Verarbeitung und Speicherung von Überwachungsdaten sowie die prototypische Implementierung auf Basis eines Security Information Event Management Systems. Dabei sollen insbesondere die folgenden Punkte bearbeitet werden:

- Wie ist der aktuelle Stand sowohl der Technik als auch der Wissenschaft im Bereich der Pseudonymisierung und der kryptographischen Schwellwertschemata?
- An welcher Stelle des konzipierten Systems können die Überwachungsdaten entsprechend des Datenschutzes und der späteren möglicherweise erforderlichen Rückgewinnung des Personenbezugs verarbeitet werden und welche Auswirkungen können entstehen?
- Wie und an welcher Stelle muss das Schlüsselmanagement der benötigten kryptographischen Funktionen erfolgen?
- Welche Alternativen gibt es neben der Pseudonymisierung und den kryptographischen Schwellwertschemata zur Lösung des genannten Zielkonflikts und wie können diese in das Konzept und die prototypische Implementierung integriert werden?

Todo list

■ TBW	5
■ Wo passen Abschnitte zu folgenden Stichworten hin? - Verschiedene Datenarten: Identifizierend, Traffic, nicht relevant, ... - Grundlegende Definition Insiderangriff .	8
■ Schadenshöhe (siehe Antrag)?	8
■ EZ: Warum nicht? Einige werben damit, Innentaeter erkennen zu koennen, z.B. IBM QRadar	8
■ EZ: Warum? Welche?	8
■ EZ: Was genau ist das Szenario? Liegt dein Fokus nun auf den zusaetzlichen Daten- quellen und Erkennungslogiken oder auf den datenschutzrechtlichen Bedenken? . .	9
■ Zu ergaenzen: Aufbau der Arbeit, Literatur	9
■ Gesetzestexte als „Quellen,, in Fußnoten?	10
■ Quelle	11
■ Quelle	11
■ Beispiele für Überwachungsskandale	12
■ Evtl. auch Telemediengesetz erwähnen - Nutzung eines Systems unter einem Pseudonym	14
■ Vielleicht diese Werte direkt nutzen?	16
■ mod p irgendwo erwähnen	16
■ Vmtl.? Mehr Details...	17
■ Übersichtsdiagramm angelehnt an overview-sse, p.3	19
■ Naja SIEM eher weniger Wissenschaft	20
■ Details	22
■ TBW	23
■ To be written...	23
■ Einführender Absatz	24
■ Angriffsmöglichkeiten erläutern: Mitschneiden aller Daten und damit Aufdecken von Pseudonymen,	24
■ ? außerdem erweitern um aktive Angreifer?	24
■ Denial of Service etc?	24
■ Anpassen nach Kapitel 3 - Threshold-Abschnitt?	26

■ Performance?	26
■ Einführender Absatz	27
■ Hier oder in Anforderungen?	27
■ Beschreibung erweitern - analog zu Plugins in OSSIM	27
■ Syslog-Problematik erwähnen	28
■ Auch auf Angreifermodell beziehen	28
■ Change Store to Service, Webinterface, Log-Proxy	29
■ TBW	32

Zusammenfassung

TBW

Inhaltsverzeichnis

1	Einführung	8
1.1	Related work	9
2	Grundlagen	10
2.1	Definitionen und Notationen	10
2.2	Arbeitnehmerdatenschutz	10
2.2.1	Das Recht auf informationelle Selbstbestimmung	11
2.2.2	Datenschutz im Beschäftigungsverhältnis	11
2.3	SIEM-Systeme	12
2.4	Pseudonymisierung	14
2.5	Schwellwertschemata	15
2.5.1	Shamir's Secret Sharing	16
2.5.2	Threshold Decryption	17
2.6	Searchable Encryption	18
3	Stand der Wissenschaft und Auswahl von Verfahren	20
3.1	SIEM-Systeme	20
3.1.1	OSSIM-Überblick	20
3.1.2	Parsen von Logdaten in OSSIM	21
3.2	Pseudonymisierung	22
3.3	Schwellwertschemata	22
3.4	Searchable Encryption	23
4	Implementierung	24
4.1	Anforderungen	24
4.1.1	Zugrundeliegendes Angreifermodell	24
4.1.2	Anforderungen an die Integration in OSSIM	25
4.1.3	Anforderungen an die Pseudonymisierung	25
4.1.4	Anforderungen im Bezug auf den Einsatz eines kryptographischen Schwellwertschemas	26
4.1.5	Benutzerinteraktion	26
4.1.6	Erweiterbarkeit um neue Datenquellen	27
4.1.7	Erweiterbarkeit um neue Datenschutztechniken	27
4.2	Entwurf	27
4.2.1	Eingriff in den OSSIM-Datenfluss	27
4.2.2	Architektur	29
4.3	Einbindung in OSSIM	30
4.4	Umsetzung der Pseudonymisierung	30
4.5	Implementierung und Integration des Schwellwertschemas	30
4.6	Evaluation	30

5 Alternativen	31
5.1 Alternative1	31
5.2	31
5.3 Vorgehen zur Integration	31
6 Fazit	32
Literatur	33

1 Einführung

Liest man von erfolgreichen Angriffen auf Unternehmensnetzwerke, so ist die implizite Annahme von außenstehenden, unternehmensfremden Angreifern weit verbreitet. Doch häufig sind die Angreifer bereits im Netzwerk ansässig. Es handelt sich um (ehemalige) Mitarbeiter oder zumindest Personen mit legitimem Zugriff auf das Netzwerk, wie Geschäftspartnern oder Kunden. Hierbei geht es keineswegs lediglich um Einzelfälle.

Wo passen Abschnitte zu folgenden Stichworten hin? - Verschiedene Datenarten: Identifizierend, Traffic, nicht relevant, ... - Grundlegende Definition Insiderangriff

In dem *IBM Cyber Security Intelligence Report* von 2015 werden 55% der Angriffe als aus dem internen Netz stammend angegeben [Bra+15]. Zu beachten ist allerdings, dass nicht nur mit Absicht ausgeführte Angriffe hierunter erfasst wurden, sondern auch unbeabsichtigte wie das versehentliche Veröffentlichen schützenswerter Kundendaten.

Auch der Branchenverband bitkom führt in seiner *Spezialstudie Wirtschaftsschutz* aus dem Jahr 2016 nach einer Befragung von über 1000 Unternehmen aus, dass etwa 60% der erfolgten Handlungen aus dem Bereich Datendiebstahl, Industriespionage oder Sabotage durch (ehemalige) Mitarbeiter erfolgten [bit16].

Schadenshöhe (siehe Antrag?)

Auch wenn die genauen Zahlen aufgrund von unterschiedlichen Annahmen und der in diesem Bereich nicht zu vernachlässigenden Dunkelziffer¹ mit Vorsicht zu betrachten sind, so geben sie doch Hinweise darauf, dass Angriffe von Innentätern weit verbreitet sind und ein hohes Schadenspotenzial aufweisen. Die Erkennung und Verhinderung solcher Angriffe sollte daher ein wichtiger Teil des IT-Sicherheitskonzepts eines Unternehmens sein.

Zur Erkennung von Angriffen in Netzwerken können SIEM-Systeme eingesetzt werden (siehe Abschnitt ??). Diese sind jedoch in erster Linie auf das Erkennen von externen Angriffen ausgelegt und in ihrer derzeitigen Form kaum sinnvoll für das Erkennen von Innentätern zu nutzen.

Hierfür würden zusätzliche Datenquellen und Erkennungslogiken nötig sein. Zusätzlich spielen auch datenschutzrechtliche Bedenken im Bezug auf das Sammeln von großen Datenmengen über Mitarbeiter des eigenen Unternehmens hier eine entscheidende Rolle. Details hierzu sind im folgenden Kapitel ?? zu finden.

EZ: Warum nicht? Einige werben damit, Innentäter erkennen zu können, z.B. IBM QRadar

EZ: Warum? Welche?

Ein Ansatz, der diese Bedenken ausräumen oder zumindest lindern könnte, ist die Nutzung von Pseudonymen bei der Datenerfassung (siehe Abschnitt ??). Anstatt direkt identifizierende Merkmale eines Arbeitnehmers abzuspeichern, werden diese Merkmale durch ein Pseudonym ersetzt. Eine Liste dieser Ersetzungen wird verschlüsselt abgelegt. Im Fall eines Angriffs durch einen Innentäter kann die Liste entschlüsselt werden und relevante Ereignisse de-pseudonymisiert, also ihrem ursprünglichen Verursacher wieder zweifelsfrei zugeordnet, werden.

Um die Entschlüsselung nicht einzelnen (möglicherweise bösartig agierenden) Personen zu

1. Insbesondere die Angst vor Imageschäden, die auch in der *Spezialstudie Wirtschaftsschutz* erwähnt wird, könnte ein Grund für das Geheimhalten von Vorfällen sein.

ermöglichen, können sogenannte Schwellwertschemata eingesetzt werden (siehe Abschnitt ??). Durch sie wird die Entschlüsselung erst durch die Kooperation mehrerer Parteien möglich gemacht.

Bei diesem Ansatz muss jedoch auch beachtet werden, dass durch den Einsatz von Pseudonymen die Erkennung von Angriffen erschwert werden könnte. Beispielsweise könnte das Ändern von Pseudonymen in regelmäßigen Zeitintervallen und die dadurch entstehende Nicht-Verkettbarkeit von Ereignissen dafür sorgen, dass langfristig angelegte Angriffe nicht aufgedeckt werden.

Ziele der Arbeit

In dieser Arbeit soll es darum gehen, prototypisch ein solches Szenario auf Basis eines Open-Source-SIEM-Systems umzusetzen. Hierbei müssen einige Fragen betrachtet werden:

- An welcher Stelle des Systems kann eingegriffen werden, um die erfassten Daten zu verändern, und welche Auswirkungen hat dies?
- Wie erfolgt die angesprochene Pseudonymisierung technisch?
- Welche kryptographischen Schwellwertschemata können genutzt werden? Gibt es bereits quelloffene Implementierungen? Was muss selbst entwickelt werden? Wie erfolgt das Schlüsselmanagement?
- Können neben der Pseudonymisierung noch weitere Funktionen zur Veränderung von Daten sinnvoll sein und wie könnten diese umgesetzt werden?

EZ: Was genau ist das Szenario? Liegt dein Fokus nun auf den zusätzlichen Datenquellen und Erkennungslogiken oder auf den datenschutzrechtlichen Bedenken?

Gerade die letzte Frage sorgt dafür, dass zusätzliche Anforderungen an den zu entwickelnden Prototypen gestellt werden. Es sollte möglich sein, abhängig von den eingehenden Daten die entsprechend gewünschten Funktionen konfigurieren und den Prototypen in eventuell aufbauenden Arbeiten auch um zusätzliche Funktionen ergänzen zu können.

Zu ergänzen: Aufbau der Arbeit, Literatur

1.1 Related work

2 Grundlagen

Dieses Kapitel widmet sich den Grundlagen der in dieser Arbeit verwendeten Konzepte, Verfahren und Systeme. Zu Beginn werden für den Verlauf der Arbeit notwendige Definitionen gegeben. Anschließend werden die juristischen Hintergründe des Arbeitnehmerdatenschutzes erläutert, die den rechtlichen Rahmen für das Thema dieser Arbeit bilden.

Es folgen Erläuterungen zu SIEM-Systemen, in die - wie bereits in der Einleitung erläutert - die prototypische Umsetzung der datenschutzfreundlichen Speicherung erfolgen soll, sowie zu den zu verwendenden Datenschutztechniken Pseudonymisierung, kryptographische Schwellwert-schemata und Searchable Encryption.

2.1 Definitionen und Notationen

- Insider-Angriff
- Datenarten
- ...

2.2 Arbeitnehmerdatenschutz

Der Begriff des Arbeitnehmerdatenschutzes¹ beschreibt die Rechte von Arbeitnehmern im Beschäftigungsverhältnis im Bezug auf den Umgang mit personenbezogenen Daten. In diesem Abschnitt soll ein kompakter Überblick über aktuell geltende und in nächster Zeit in Kraft tretende gesetzliche Regelungen im Bezug hierauf gegeben werden, wobei der Fokus auf zum Thema der Arbeit passenden Regelungen liegt.

Zu Beginn soll kurz auf das Recht auf informationelle Selbstbestimmung eingegangen werden, das die Grundlage für alle folgenden Betrachtungen zum Arbeitnehmerdatenschutz bildet.

Gesetzestexte als
„Quellen“, in Fuß-
noten?

1. In manchen Veröffentlichungen wird der Arbeitnehmerdatenschutz auch als Mitarbeiterdatenschutz, Beschäftigtendatenschutz, Personaldatenschutz oder Betriebsdatenschutz bezeichnet.

2.2.1 Das Recht auf informationelle Selbstbestimmung

Im sogenannten Volkszählungsurteil aus dem Jahr 1983 wurde das Recht auf informationelle Selbstbestimmung als Grundrecht anerkannt [TODO]. Es handelt sich um eine Ausprägung des allgemeinen Persönlichkeitsrechts² nach Artikel 2, Absatz 1 in Verbindung mit Artikel 1, Absatz 1 des Grundgesetzes. Es beschreibt das Recht des Einzelnen, selber über den Umgang mit seinen personenbezogenen Daten entscheiden zu können.

Mit dem vermehrten Aufkommen automatisierter Datenverarbeitung stellten die Richter des Bundesverfassungsgerichts damals die besondere Schutzbedürftigkeit der Selbstbestimmung des Einzelnen im Bezug auf die Offenbarung von Lebenssachverhalten heraus. Sie betonten die Notwendigkeit dieser Selbstbestimmung als Voraussetzung für eine freie Entfaltung der Persönlichkeit und auch für die Ausübung bestimmter Grundrechte wie der Versammlungsfreiheit. Damit sei das Recht auf informationelle Selbstbestimmung auch „eine elementare Funktionsbedingung eines auf Handlungs- und Mitwirkungsfähigkeit seiner Bürger begründeten freiheitlichen demokratischen Gemeinwesens“ [TODO] .

Einschränkungen dieses Rechts sind dem Urteil nach möglich, jedoch in Gesetzen festzuhalten. Hierbei müssen das Geheimhaltungsinteresse des Betroffenen und das öffentliche Informationsinteresse verarbeitender Stellen gegeneinander abgewogen werden.

Auch wenn sich das Urteil des Bundesverfassungsgerichts nur auf die Rechte des Betroffenen gegenüber staatlichen Akteuren bezieht, so bildet die Intention des Rechts auf informationelle Selbstbestimmung doch die Grundlage für das heutige Bundesdatenschutzgesetz und auch die Datenschutzgrundverordnung der EU, die auch für nicht-staatliche Akteure Gültigkeit besitzen.

Zusätzlich findet sich das Recht auf informationelle Selbstbestimmung auch in der Grundrechtcharta der EU: „Jede Person hat das Recht auf Schutz der sie betreffenden personenbezogenen Daten.“³

Quelle

2.2.2 Datenschutz im Beschäftigungsverhältnis

Eine besondere Situation ergibt sich im Unternehmenskontext. Hier muss das Recht des Arbeitnehmers auf informationelle Selbstbestimmung gegenüber dem berechtigten Interesse des Arbeitgebers an der Aufklärung von Straftaten im Beschäftigungsverhältnis abgewogen werden.

Im zur Zeit gültigen Bundesdatenschutzgesetz (BDSG) wird in § 4 die Erhebung, Verarbeitung und Nutzung personenbezogener Daten nur als zulässig angesehen, falls der Betroffene einwilligt oder ein Gesetz dieses erlaubt. Personenbezogene Daten werden in § 3 hierbei als „Einzelangaben über [...] Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person“ definiert.

Quelle

§ 32 beschreibt die Datenerhebung, -verarbeitung und -nutzung für Zwecke des Beschäftigungsverhältnisses:

-
2. Das allgemeine Persönlichkeitsrecht beschreibt den Schutz der Persönlichkeit einer Person vor Eingriffen in ihren Lebens- und Freiheitsbereich.
 3. Artikel 8, Absatz 1

Personenbezogene Daten eines Beschäftigten dürfen für Zwecke des Beschäftigungsverhältnisses erhoben, verarbeitet oder genutzt werden, wenn dies für die Entscheidung über die Begründung eines Beschäftigungsverhältnisses oder nach Begründung des Beschäftigungsverhältnisses für dessen Durchführung oder Beendigung erforderlich ist.

Zur Aufdeckung von Straftaten dürfen personenbezogene Daten eines Beschäftigten nur dann erhoben, verarbeitet oder genutzt werden, wenn zu dokumentierende tatsächliche Anhaltspunkte den Verdacht begründen, dass der Betroffene im Beschäftigungsverhältnis eine Straftat begangen hat, die Erhebung, Verarbeitung oder Nutzung zur Aufdeckung erforderlich ist und das schutzwürdige Interesse des Beschäftigten an dem Ausschluss der Erhebung, Verarbeitung oder Nutzung nicht überwiegt, insbesondere Art und Ausmaß im Hinblick auf den Anlass nicht unverhältnismäßig sind.⁴

Während sich der erste Satz auf den Umgang mit personenbezogenen Daten in einem normalen Beschäftigungsverhältnis befasst und bezogen auf das Thema dieser Arbeit beispielsweise den Rahmen für erforderliche Datenverarbeitung zur Aufdeckung von Vertragsbrüchen unterhalb der Straftatgrenze darstellt, behandelt der zweite Satz den Straftatfall. Hier sind insbesondere der notwendige Anfangsverdacht als Voraussetzung und die Verhältnismäßigkeit der Datennutzung zu beachten.

Weiterhin insbesondere im Rahmen dieser Arbeit entscheidend ist die Ausrichtung des BDSG auf personenbezogene Daten, die wie bereits definiert einer bestimmbar Person zugeordnet werden können müssen. Das in dieser Arbeit angestrebte System wird durch Pseudonymisierung und erst durch Kollaboration ermöglichte De-Pseudonymisierung den direkten Personenbezug verhindern und erst im durch mehrere Instanzen bestätigten Straftatverdacht ermöglichen⁵.

2018 tritt die EU-Verordnung 2016/679, besser bekannt als Datenschutzgrundverordnung, in Kraft. In Deutschland wird das bestehende BDSG durch das Datenschutz-Anpassungs- und Umsetzungsgesetz grundlegend überarbeitet und an die Verordnung angepasst, um diese zu ergänzen. Hier finden sich in § 26 die Bestimmungen zur Datenverarbeitung für Zwecke des Beschäftigungsverhältnisses. Der zitierte Absatz aus dem BDSG ist dort in ähnlicher Form zu finden, wird also auch weiterhin seine Gültigkeit behalten.

Wie sollten Gesetzestexte zitiert werden?

Beispiele für Überwachungskandale

2.3 SIEM-Systeme

SIEM-Systeme dienen dazu Daten in Netzwerken zu sammeln, um so einen zentralisierten Überblick über das Netzwerk zu erhalten und Bedrohungen erkennen und verhindern zu können.

Der Begriff *Security Information and Event Management* (SIEM) wurde von zwei Analysten des IT-Marktforschungsunternehmens Gartner geprägt, das auch jährlich einen Bericht über aktuelle

4. § 32, Absatz 1, Bundesdatenschutzgesetz

5. Der Autor maßt sich an dieser Stelle allerdings keine Beurteilung über die tatsächliche rechtliche Bewertung dieser Lösung an.

Trends im Bereich der SIEM-Systeme veröffentlicht. Er setzt sich zusammen aus *Security Event Management* (SEM), das sich mit Echtzeitüberwachung und Ereigniskorrelation befasst, sowie *Security Information Management* (SIM), in dessen Fokus Langzeiterfassung und Analyse von Log-Daten steht [NK11].

Ein SIEM-System sollte nach [Det+15] die folgenden Aufgaben erfüllen können:

- **Network Behaviour Anomaly Detection:** Beschreibt die Erkennung von Anomalien auf Netzwerkebene durch die Erkennung von vom Normalzustand abweichenden Kommunikationsverhalten.
- **Identity Mapping:** Abbildung von Netzwerkadressen auf Nutzeridentitäten.
- **Key Performance Indication:** Zentrale Analyse sicherheitsrelevanter Informationen und Netzwerkdetails.
- **Compliance Reporting:** Überprüfung der Einhaltung von durch Regelungen vorgeschriebenen Anforderungen wie Integrität, Risiko und Effektivität.
- **API:** Bereitstellung von Schnittstellen zur Integration heterogener Systeme im Netzwerk.
- **Role based access control:** Zuständigkeitsabhängige Sichten auf sicherheitsrelevante Ereignisse.
- **Event Correlation:** im Folgenden näher erläutert.

Eine besondere Bedeutung im Kontext dieser Arbeit kommt hier der Behandlung von sicherheitsrelevanten Ereignissen (Events) zu, die beispielsweise von Intrusion-Detection-Systemen oder aus den Log-Daten von Firewalls, Switches oder anderen Netzwerkgeräten stammen können.

Um diese Ereignisse zu erhalten, muss ein SIEM-System nach [DRS14] vor ihrer Speicherung insbesondere drei Aufgaben wahrnehmen. Zu Beginn werden die Daten aus Logeinträgen oder empfangenen Systemmeldungen herausgelesen (Extraktion).

Anschließend müssen die extrahierten Daten in ein SIEM-spezifisches Format übersetzt werden, um eine sinnvolle Weiterverarbeitung zu gewährleisten (Homogenisierung). Hierbei werden relevante Felder eines SIEM-Events wie Datumsangaben, Adressen oder Aktionen aus den empfangenen Daten befüllt. Dieser Schritt wird in anderen Quellen auch als Normalisierung oder Mapping bezeichnet.

Optional können darauf folgend gleichartige Events in bestimmten Fällen anschließend zusammengefasst werden, um aussagekräftigere Informationen zu erhalten (Aggregation).

Liegen die Events nun in einem vorgegebenen Format im System vor, so können sie weiterhin mit dem System bekannten Umgebungsdaten über Benutzer, Geräte oder Bedrohungen verknüpft werden, um ihre Relevanz besser einschätzen zu können.

Anschließend lassen sich vorgegebene Regeln anwenden, um aus der Korrelation von Ereignissen aus verschiedenen Datenquellen auf eine Bedrohung schließen zu können, die in den einzelnen Events nicht erkennbar wäre (Event Correlation).

Wäre hier ein Beispiel für Event Correlation nötig/hilfreich?

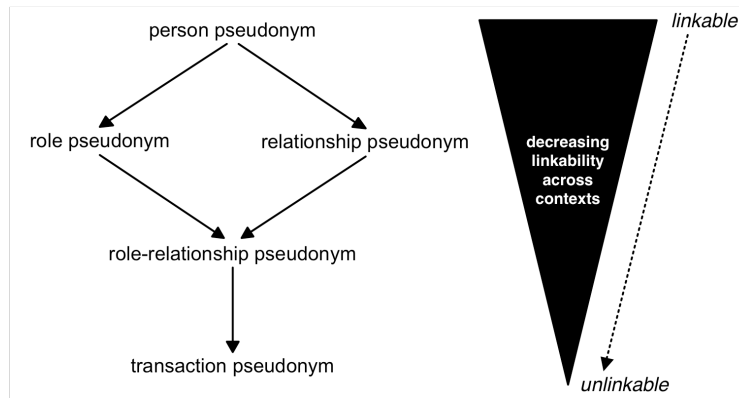


Abbildung 2.1: Pseudonym-Verband entsprechend ihrer Nutzung in verschiedenen Kontexten. Entnommen aus [PH10].

2.4 Pseudonymisierung

Ein Pseudonym⁶ bezeichnet nach [PH10] einen Identifikator eines Subjekts ungleich seinem echten Bezeichner. In §3, Absatz 6a des BDSG wird zusätzlich noch der „Zweck [eines Pseudonyms], die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren“**[TODO]** ergänzt. Beispiele für Pseudonyme lassen sich in verschiedensten Bereichen finden: E-Mail-Adressen, Sozialversicherungsnummern oder auch Autoren, die unter einem Pseudonym ihre Schriften veröffentlichen.

Es lassen sich nach [PWP90] unterschiedliche Arten von Pseudonymen unterscheiden. Eine Eigenschaft, die zur Unterscheidung herangezogen werden kann, ist die (Un-)Kenntnis des Zusammenhangs zwischen dem Pseudonym und dem zugehörigen Subjekt zu Beginn seiner Verwendung. Dieser Zusammenhang wird auch als Zuordnungsvorschrift bezeichnet und kann beispielsweise als Funktion oder in Tabellenform vorliegen.

Hier kann zwischen öffentlichen, nicht-öffentlichen und anonymen Pseudonymen unterschieden werden. Ein öffentliches Pseudonym ist beispielsweise die Telefonnummer einer Person, die von Beginn an im Telefonbuch mit der Identität der Person verknüpft ist. Eine Kontonummer zu einem Bankkonto, dessen Inhaber bei der Kontoeröffnung nur der Bank bekannt ist, bildet ein nicht-öffentliches Pseudonym. Ein anonymes Pseudonym bildet beispielsweise die DNA-Sequenz eines Menschen, die oftmals nicht einmal ihm selbst bekannt ist. Diese Eigenschaft eines Pseudonyms kann sich im Laufe der Zeit ändern, wenn Informationen über den Zusammenhang zwischen Pseudonym und zugehörigem Subjekt veröffentlicht werden.

Eng in Verbindung mit Pseudonymen steht auch der Begriff der Verkettbarkeit. Verkettbarkeit bezeichnet dabei die Eigenschaft, dass ein Außenstehender mit hoher Wahrscheinlichkeit entscheiden kann, ob zwei Objekte in einem System unabhängig sind[PH10].

Auf Basis dieser Eigenschaft lassen sich nun ebenfalls verschiedene Arten von Pseudonymen ausmachen: Diese Arten unterscheiden sich durch Nutzung des Pseudonyms in verschiedenen Kontexten. Personen-, Rollen-, Beziehungs-, Rollen-Beziehungs- und Transaktionspseudonym sind mögliche Ausprägungen dieser Eigenschaft. Eine Übersicht über diese Pseudonymarten und deren Zusammenhang zur Verkettbarkeit ist Abbildung 2.1 zu entnehmen.

6. ursprünglich aus dem Griechischen stammend: *pseudonumon* - falsch benannt

Evtl. auch Telemediengesetz erwähnen - Nutzung eines Systems unter einem Pseudonym

Personenpseudonyme beschreiben Pseudonyme wie beispielsweise die Sozialversicherungsnummer, die stellvertretend für die eindeutige Identität des Subjekts in der Gesellschaft stehen. Ereignisse, die mit solchen Pseudonymen verbunden sind, lassen sich über die gesamte Gültigkeitsspanne des Pseudonyms verketteten.

Rollen-, Beziehungs- und Rollen-Beziehungspseudonyme stellen Pseudonyme da, die das Subjekt nur in einer besonderen Funktion oder mit einem bestimmten Kommunikationspartner bzw. in der Kombination beider Möglichkeiten nutzt. Ein Beispiel hierfür wäre eine E-Mail-Adresse wie *administrator@unternehmen.de* in einem Unternehmen, die nicht das Subjekt als solches sondern nur in seiner Rolle in dem Unternehmen beschreibt. Ereignisse verbunden mit diesen Pseudonymarten lassen sich nur in bestimmten Kontexten, jedoch nicht über Kontextgrenzen hinweg, verknüpfen. Beispielsweise könnten zwei Kommunikationspartner mit derselben Person kommunizieren, ohne dies feststellen zu können, wenn die Person ihnen gegenüber unter unterschiedlichen Beziehungspseudonymen auftritt.

Transaktionspseudonyme stellen die stärkste Form der Unverkettbarkeit da. Bei dieser Art von Pseudonymen wird für jedes Ereignis ein neues Pseudonym verwendet, das daher nur ein einziges Mal auftritt und Verkettbarkeit verhindert.

Abgrenzung zur Anonymität sinnvoll?

Ist der „Schnitt“ zwischen Kapitel 2 und 3 hier sinnvoll?

2.5 Schwellwertschemata

Mit der Verbreitung technischer Systeme, die kryptographische Verfahren nutzen, in den 70er Jahren musste auch das Problem der sicheren Aufbewahrung und Verteilung kryptographischer Schlüssel betrachtet werden. Die Sicherheit dieser Schlüssel ist essentiell für den Betrieb solcher Systemen. Das einfache Speichern eines Schlüssels an einem einzigen Ort resultiert in einer hohen Verlustwahrscheinlichkeit, da ein einzelner Fehler wie unbeabsichtigtes Löschen oder Speichermedienausfall den Schlüssel unwiederbringlich verloren gehen lassen kann. Das mehrfache Speichern eines Schlüssels an verschiedenen Orten erhöht hingegen die Gefahr eines Schlüsseldiebstahls oder -missbrauchs, da auch die Angriffsfläche vergrößert wird. Bei möglichen Lösungen dieses Problems müssen also immer die Integrität und die Vertraulichkeit eines Schlüssels gegeneinander abgewogen werden. [Gem97]

Ausgehend von diesen Überlegungen entwickelte Shamir das erste (k, n) -Schwellwertschema: Ein Geheimnis D wird so in n Teile D_1, \dots, D_n (*Shares*) zerlegt, dass durch Kenntnis von mindestens k Teilen das Geheimnis wieder aufgedeckt werden kann, aber jede Kombination aus höchstens $k - 1$ Teilen keine Informationen über D liefert [Sha79].⁷

Auf Basis dieses Verfahrens kann also die Integrität eines Schlüssels erhöht werden, da nun selbst bei Verlust von $n - k$ Teilen der Schlüssel noch wiederhergestellt werden kann. Auf der anderen Seite ist die Vertraulichkeit jedoch höher als bei der mehrfachen Speicherung des Schlüssels im Original, da mindestens k Teile des Schlüssels zur Wiederherstellung vorliegen müssen.

7. Im selben Jahr veröffentlichte auch Blakley eine Lösung dieses Problems, die auf den Schnittpunkten von Hyperebenen über endlichen Feldern beruht [Bla79].

Shamirs Verfahren wird im Folgenden im Detail beschrieben, da es auch im später erläuterten und verwendeten Schwellwertschema eine wichtige Rolle spielt.

2.5.1 Shamir's Secret Sharing

Die Menge aller Ganzzahlen modulo einer Primzahl p bilden den (endlichen) Körper \mathbb{Z}_p , dessen Eigenschaften für das Verfahren entscheidend sind. Soll das Geheimnis D (das o.B.d.A. als Ganzzahl angenommen wird) aufgeteilt werden, so wird eine Primzahl p mit $p > D$ und $p > n$ gewählt. Weiterhin wird ein Polynom

$$q(x) = a_0 + a_1x + \dots + a_{k-1}x^{k-1} \text{ mit } a_0 = D$$

derart gewählt, dass a_1, \dots, a_{k-1} zufällig gleichverteilt aus der Menge $\mathbb{Z}_p \setminus \{0\}$ stammen. Die einzelnen *Shares* werden nun als

$$D_1 = (x_1, q(x_1)), \dots, D_i = (x_i, q(x_i)), \dots, D_n = (x_n, q(x_n))$$

jeweils modulo p berechnet, wobei die x_i paarweise unterschiedlich aus \mathbb{Z}_p gewählt werden können. Beispielsweise kann schlicht $x_i = i$ gelten.

Vielleicht diese Werte direkt nutzen?

Um nun aus diesen einzelnen Teilen wieder das ursprüngliche Geheimnis zu erhalten, wird das Verfahren der Lagrange'schen Polynominterpolation verwendet, das ausgehend von einer Menge von Punkten ein Polynom findet, das durch diese Punkte verläuft. Hierbei wird die Tatsache ausgenutzt, dass jedes Polynom vom maximalen Grad $t - 1$ in einem mathematischen Körper durch t Punkte exakt bestimmt wird.

Für die zur Rekonstruktion verwendeten t Teile $D'_1 = (x'_1, q(x'_1)), \dots, D'_t = (x'_t, q(x'_t))$ werden t Werte

mod p irgendwo erwähnen

$$\lambda_i := \prod_{j=1, j \neq i}^t \frac{-x'_j}{x'_i - x'_j} \text{ für } i \in \{1, \dots, t\}$$

definiert. Das gesuchte Geheimnis D kann nun als

$$D = \sum_{i=1}^t \lambda_i \cdot q(x'_i)$$

berechnet werden. Da λ_i nicht von $q(x_i)$ abhängt, können diese Werte in der Praxis bereits vorberechnet werden. Details zu der Korrektheit dieses Verfahrens sind [BS16] zu entnehmen.

Das Problem dieser Lösung bezogen auf den in dieser Arbeit behandelten Anwendungsfall ist jedoch, dass das Geheimnis nach erstmaligem Aufdecken bekannt ist. Wünschenswert wäre ein Verfahren, bei dem nur ein entsprechend verschlüsseltes Datum (bspw. der gesuchte Eintrag in einer Pseudonym-Tabelle) aufgedeckt werden kann, ohne dass der kombinierte Schlüssel selbst bekannt wird.

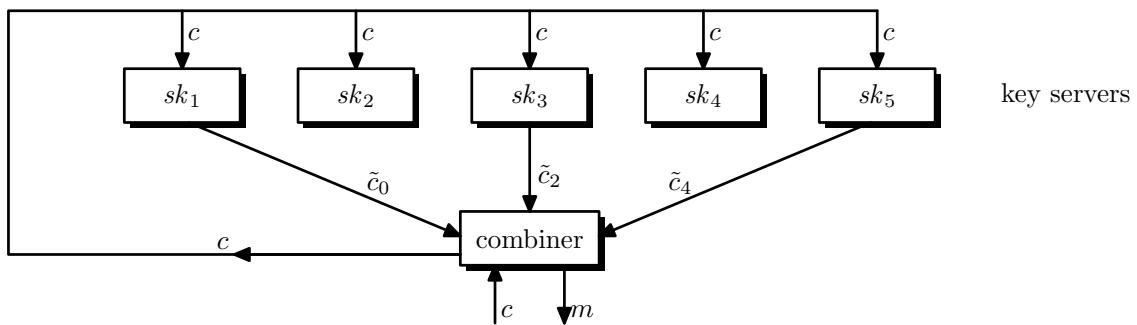


Abbildung 2.2: Übersicht über den Entschlüsselungsvorgang bei der Nutzung eines (3,5)-Schwellwertschemas. Entnommen aus [BS16].

2.5.2 Threshold Decryption

In [Des87] wird das erwähnte Problem das erste Mal im Kontext von verschlüsselten Nachrichten an Gruppen betrachtet: Ein Sender möchte eine Nachricht an eine Gruppe von Empfängern senden, die nur in Zusammenarbeit die Nachricht entschlüsseln können sollen. Hier wird auch die zentrale Forderung an mögliche Lösungen des Problems aufgestellt, den mehrfachen Nachrichtenaustausch zwischen Sender und Empfänger(n) bei der Entschlüsselung (sogenannte Ping-Pong-Protokolle) zu vermeiden.

In [Des93] spricht der Autor bei dieser Klasse von Verfahren von *Threshold Decryption* und fordert weiterhin, dass praktisch einsetzbare Systeme auch *non-interactive* sein sollten, also bei der Entschlüsselung keinen aufwendigen Datenaustausch zwischen den Besitzern der *Shares* notwendig machen.

In [BS16] werden diese Systeme formalisiert. Ein *Threshold-Public-Key-Decryption*-Schema $\varepsilon = (G, E, D, C)$ besteht aus vier Algorithmen:

- $G(t, n, r)$ ist der Algorithmus zur Generation des öffentlichen Schlüssels pk und der n *Shares* des geheimen Schlüssels $\{sk_1, \dots, sk_n\}$. t steht für die Anzahl der zur Entschlüsselung benötigten *Shares*, wie bereits beschrieben. r ist als stellvertretend für die einfließenden Zufallswerte zu betrachten.
- $E(pk, m, r)$ steht für den Algorithmus, der der Verschlüsselung einer Nachricht m mit dem öffentlichen Schlüssel pk dient. r dient der Vermeidung von *Known-Ciphertext-Angriffen*.
- $D(sk_i, c)$ ist der Algorithmus der für einen bestimmten *Share* und einen Schlüsseltext c eine partielle Entschlüsselung c'_i liefert.
- $C(c, c'_1, \dots, c'_t)$ ist der Algorithmus, der aus dem Schlüsseltext c und aus t durch D generierten partiellen Verschlüsselungen wieder die Nachricht m liefert. Dieser Algorithmus wird auch *Combiner* genannt.

Vmtl.? Mehr Details...

Zusätzlich wird von diesen Algorithmen die folgende Eigenschaft verlangt. Sie beschreibt die korrekte Entschlüsselung von validen Schlüsseltexten im Kontext eines Schwellwertschemas: Für alle möglichen Ergebnisse $(pk, \{sk_1, \dots, sk_n\})$ von G , alle möglichen Nachrichten m und alle t -elementigen Teilmengen der *Shares* $\{sk'_1, \dots, sk'_t\}$ soll für alle möglichen Schlüsseltexte $c = E(pk, m, r)$ gelten: $C(c, D(sk'_1, c), \dots, D(sk'_t, c)) = m$.

Eine Übersicht über den Entschlüsselungsvorgang ist in Abbildung 2.2 zu finden. Dort sind die partiellen Entschlüsselungen und der *Combine*-Vorgang eines $(3,5)$ -Schwellwertschemas dargestellt. Der Algorithmus D für die partielle Entschlüsselung läuft dabei auf den einzelnen *Key-Servern* ab.

In [BBH06] werden diese Algorithmen noch um einen fünften erweitert, der dazu dient, einzelne partielle Entschlüsselungen auf Validität zu überprüfen. Hierdurch können fehlerhaft handelnde *Key-Server* aufgedeckt werden. Hierzu wird auch der Algorithmus G verändert, der zusätzlich einen Validierungsschlüssel vk liefert:

1. $G(t, n, r)$ liefert nun $(pk, vk, \{sk_1, \dots, sk_n\})$.
- ...
5. $V(pk, vk, c, c'_j)$ überprüft die j -te partielle Entschlüsselungen auf Validität.

Weiterhin wird für den neuen Algorithmus eine weitere Eigenschaft verlangt. Für jeden Schlüsseltext c und $c'_j = D(sk_i, c)$, wobei sk_i der i -te von G erstellte *Share* ist, gelte: $V(pk, vk, c, c'_j)$ liefert ein valides Ergebnis.

2.6 Searchable Encryption

Ist es nötig Searchable Encryption zu verwenden?

Reicht eine lokale Zuordnungstabelle (User \rightarrow Pseudonym)? Kompromittierung des Systems würde Schlüssel für Tokengenerierung oder äquivalent dazu Tabelle betreffen, oder?

Möglicher Vorteil: Eindeutige Pseudonymgenerierung kann in Store stattfinden, aber das wäre auch durch $\text{Create}(\text{enc-user})$, das Pseudonym liefert, möglich?

- Searchable Symmetric encryption vs Public Key Encryption With Keyword Search
- Hier relevant SSE

Ein allgemeines SSE-Schema besteht aus vier effizient berechenbaren Algorithmen [WWC16]:

- **GenerateKey:** (k) generiert einen geheimen Schlüssel K anhand eines (verfahrensabhängigen) Sicherheitsparameters k .
- **BuildIndex:** (K, D) erstellt einen Suchwort-Index I aus dem generierten Schlüssel K und einer Dokumentenmenge D .
- **GenerateTrapdoor:** (K, w) erstellt für ein spezielles Suchwort w mithilfe des Schlüssels K das Trapdoor-Element T_w für die Suche nach w .
- **Search:** (I, T_w) liefert eine Menge von Dokumenten basierend auf einem Suchwort-Index I und einem Trapdoor-Element T_w .

Der Besitzer der Daten erstellt sich einen Schlüssel mithilfe von **GenerateKey** und generiert durch **BuildIndex** einen Suchwort-Index für seine Dokumente. Anschließend lädt er diese Dokumente in verschlüsselter Form zusammen mit dem Index auf den Server. Möchte der Besitzer nun alle Dokumente erhalten, auf die ein spezielles Suchwort zutrifft, so erstellt er für dieses Suchwort mithilfe von **GenerateTrapdoor** ein Trapdoor-Element und sendet dieses an den Server. Dort wird nun auf dem Suchwort-Index durch **Search** die Suche nach dem Trapdoor-Element ausgeführt, die eine Menge von verschlüsselten Dokumenten liefert, auf die das Suchwort zutrifft. Diese können zurück an den Besitzer gesendet werden, der sie lokal entschlüsseln kann.

Übersichtsdiagramm
angelehnt an
overview-sse, p.3

3 Stand der Wissenschaft und Auswahl von Verfahren

In diesem Kapitel soll der aktuelle Stand der Wissenschaft bezogen auf die in dieser Arbeit verwendeten Klassen von Verfahren betrachtet werden sowie ausgehend von den Anforderungen an das zu entwickelnde System ein passendes Verfahren ausgewählt und im Detail beschrieben werden. An geeigneten Stellen wird auf die im letzten Kapitel dargelegten Grundlagen zurückgegriffen.

Naja SIEM eher weniger Wissenschaft

3.1 SIEM-Systeme

Zur Zeit gibt es eine vielfältige Auswahl an SIEM-Systemen auf dem Markt, die die grundsätzlichen Aufgaben eines SIEM-Systems erfüllen und über diese hinausgehen: Splunk¹, QRadar von IBM² oder ArcSight von Micro Focus³ sind nur einige Beispiele aus diesem Bereich.

Die Auswahl an Open-Source-Software in diesem Bereich ist jedoch sehr gering. Eine der wenigen Ausnahmen stellt OSSIM - ein SIEM-System der Firma AlienVault⁴ - dar, das auf Basis weiterer quelloffener Lösungen aus dem Netzwerksicherheits-Bereich unter anderem die in Abschnitt 2.3 beschriebenen Funktionen bereitstellt. AlienVault bietet zusätzlich eine kommerzielle Variante seines Produkts namens USM an, das insbesondere in den Bereichen Event-Korrelation und Compliance-Reporting die Funktionalität von OSSIM übersteigt. Von der Entwicklungsarbeit die in USM fließt, profitiert jedoch auch OSSIM, beispielsweise durch die Aktualisierung von Plugins für die Einbindung von aktuellen Netzwerkgeräten.

3.1.1 OSSIM-Überblick

Im Folgenden soll eine Übersicht über die für diese Arbeit relevanten Komponenten von OSSIM und deren Zusammenspiel gegeben werden. Diese ist auch in Abbildung 3.1 dargestellt.

Den Kern des SIEM-Systems bildet der OSSIM-Server. Hier werden Events gespeichert sowie aggregiert und es findet die Korrelation von Events statt, die der Erkennung von Angriffen oder ungewöhnlichem Netzverhalten dient. Events und generierte Meldungen können über ein Web Interface betrachtet werden. Weiterhin können hier unter anderem Angaben zur Netzinfrastruktur bereitgestellt, Netzwerk- und Schwachstellenscanner bedient und sämtliche Informationen über den Netzwerkstatus eingesehen werden.

Der OSSIM-Agent ist dafür zuständig, vorliegende Logdaten zu parsen und in ein OSSIM-spezifisches Event-Format zu übersetzen. Auf diesen Vorgang wird im nächsten Abschnitt

1. <https://www.splunk.com>

2. <https://www.ibm.com/us-en/marketplace/ibm-qradar-siem>

3. <https://software.microfocus.com/en-us/software/siem-security-information-event-management>

4. AlienVault OSSIM: The World's Most Widely Used Open Source SIEM
<https://www.alienvault.com/products/ossim>

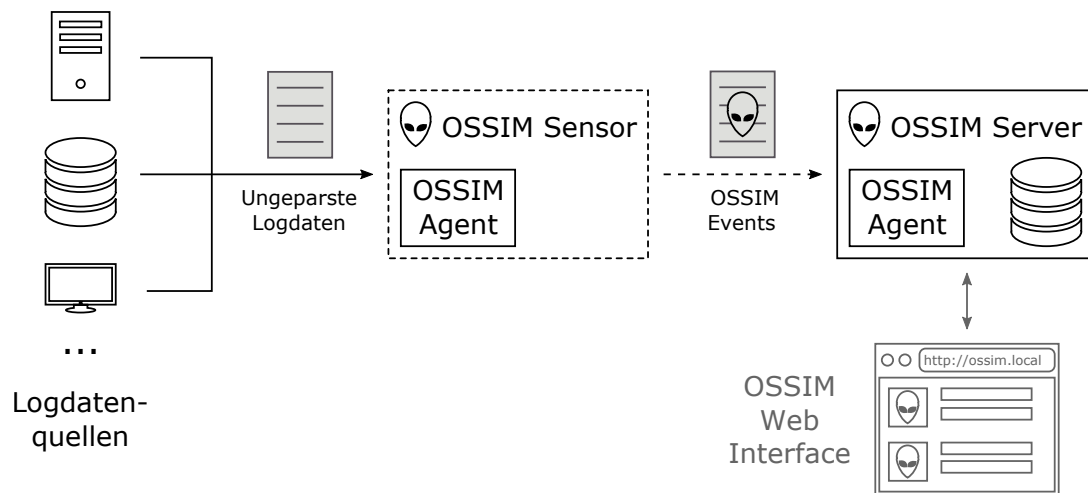


Abbildung 3.1: High-Level-Übersicht über die OSSIM-Architektur und den Datenfluss.

genauer eingegangen. Die erzeugten Events werden anschließend an den Server weitergeleitet. Der Agent befindet sich sowohl direkt auf dem Server als auch auf jedem installierten Sensor.

Eine OSSIM-Umgebung kann optional ein oder mehrere Sensoren nutzen, auf denen jeweils ein Agent seine Arbeit verrichtet. Dies wird im Folgenden verteilte Installation genannt. Der Vorteil dieser Lösung besteht darin, dass das aufwendige Parsen und Normalisieren von Logdaten verteilt stattfinden und dadurch die Serverlast in großen Umgebungen reduziert werden kann. Kommt kein externer Sensor zum Einsatz, so spricht man von einer All-In-One-Installation.

3.1.2 Parsen von Logdaten in OSSIM

Besonders von Bedeutung für diese Arbeit ist die Verarbeitung von Logdaten. OSSIM ermöglicht es, Logdaten aus unterschiedlichen Quellen entgegenzunehmen bzw. aktiv selber abzurufen und in ein gemeinsames Event-Format zu übersetzen. Hierzu stehen verschiedene Möglichkeiten zur Verfügung:

- Entgegennehmen von Daten über das Syslog-Protokoll
- Beschaffen von Daten über das SNMP-Protokoll
- Entgegennehmen von Daten über proprietäre Protokoll wie SDEE oder WMI
- Beschaffen von Daten durch Datenbankabfragen

Unabhängig von der Datenquelle funktioniert die Verarbeitung der Logdaten nach dem immer gleichen Schema. OSSIM bietet die Möglichkeit mitgelieferte oder selber entwickelte Plugins für verschiedene Datenquellen zu aktivieren. Für eintreffende Logdaten überprüft der Agent anhand von regulären Ausdrücken, ob ein Plugin für das entsprechende Datum zuständig ist. Ist so ein Plugin gefunden, so wird ein neues OSSIM-Event angelegt und anhand der Angaben im Plugin die entsprechenden vorgegebenen Felder des Events gesetzt. Hierbei kann es sich beispielsweise um den Zeitpunkt des Events, IP-Adresse und Port der Datenquelle, einen zu dem Event gehörigen Netzwerkbenutzer oder ereignisabhängige selbstgesetzte Felder handeln. Anschließend folgt die Weiterleitung des Events an den Server.

3.2 Pseudonymisierung

- A name or another bit string. Identifiers, which are generated using random data only, i.e., fully independent of the subject and related attribute values, do not contain side information on the subject they are attached to, whereas non-random identifiers may do. E.g., nicknames chosen by a user may contain information on heroes he admires; a sequence number may contain information on the time the pseudonym was issued; an e-mail address or phone number contains information how to reach the user. [PH10]

- evtl. Common Criteria

- VANET

- TIMSI

Der Begriff der Pseudonymisierung beschreibt die Benutzung von Pseudonymen zur Identifizierung von Subjekten. Pseudonymisierung sagt dabei erst einmal lediglich etwas über die Verwendung eines Verfahrens aus, jedoch nichts über die daraus entstehenden Auswirkungen auf die Identifizierbarkeit eines Subjekts oder die Zurechenbarkeit bestimmter Aktionen. Hierfür spielen weitere Eigenschaften von Pseudonymen wie die folgenden eine Rolle:

- garantierte Eindeutigkeit von Pseudonymen
- Möglichkeit von Pseudonymänderungen
- begrenzt häufige Verwendung von Pseudonymen
- zeitlich begrenzte Verwendung von Pseudonymen
- Art der Pseudonymserstellung

Die Ausprägungen dieser Eigenschaften werden auch im Rahmen dieser Arbeit für das umzusetzende System zu bewerten sein.

3.3 Schwellwertschemata

- **Grundlagen** Von Shamirs Secret Sharing bis heute
- **Konkretes Verfahren** Im Detail erläutern - vmtl. Desmedt auf Basis von ElGamal
- Zusätzlich Erweiterungen dieses Verfahrens wie verteilte Schlüsselgenerierung.

Ein solches System, das auf dem ElGamal-Algorithmus und damit dem Diskreten-Logarithmus-Problem basiert, veröffentlichten die Autoren in [DF90]. Dieser Ansatz setzt in der Setup- Phase auf eine zentrale vertrauenswürdige Stelle zur Erzeugung der Schlüssel und *Shares*. In [Ped91] stellt der Autor basierend auf diesen Ergebnissen ein Verfahren vor, das bei der Schlüsselgenerierung ohne eine vertrauenswürdige Instanz auskommt. Dieses Verfahren wird in [Gen+99] noch einmal verbessert.

Basierend auf dem jetzigen Recherchestand würde sich diese Kombination von Verfahren gut für den angestrebten Anwendungszweck eignen. Konkrete offene Implementierungen wurden jedoch bisher nicht gefunden, so dass möglicherweise eine eigene Implementierung umgesetzt werden muss.

Details

Neben diesem Verfahren gibt es noch weitere Ansätze basierend auf RSA [Des93; Ngu05] oder dem Paillier-Kryptosystem [Pai99; DJ01], die jedoch deutlich komplexer zu sein scheinen.

TBW

3.4 Searchable Encryption

To be written...

4 Implementierung

Einführender Absatz

4.1 Anforderungen

Der umgesetzte Prototyp soll es ermöglichen, aufbauend auf dem bestehenden Open-Source-SIEM-System OSSIM Logdaten mittels Pseudonymisierung und Schwellwertschemata so zu verändern, dass diese erst durch Kollaboration einer bestimmten Anzahl an Teilnehmern wieder aufgedeckt werden können. Neben dieser primären Anforderung sollte er noch weitere wie die Erweiterbarkeit um weitere Datenschutztechniken erfüllen. Alle diese Anforderungen sollen im folgenden Abschnitt näher erläutert werden. Zuerst wird jedoch ein Angreifermodell für die prototypische Umsetzung aufgestellt.

4.1.1 Zugrundeliegendes Angreifermodell

Im Kontext dieser Arbeit, die sich mit der datenschutzfreundlichen Speicherung von Überwachungsdaten beschäftigt, muss zuerst folgende Vorüberlegung getroffen werden: Logdaten erreichen das verwendete SIEM-System abhängig von den verwendeten Protokollen im allgemeinen nicht-pseudonymisiert und oftmals weder verschlüsselt noch mit geschützter Integrität über das Netzwerk. Diese Angriffsmöglichkeiten zu verhindern, ist ausdrücklich kein Ziel dieser Arbeit. Daher bezieht sich das Angreifermodell auf die Bearbeitung und Speicherung von Logdaten, erst sobald sie das zu entwickelnde System erreichen, jedoch nicht vorher.

Angriffsmöglichkeiten erläutern: Mitschneiden aller Daten und damit Aufdecken von Pseudonymen, ...

Das Ziel des Systems bezogen auf seine Sicherheit lässt sich folgendermaßen definieren: Das Pseudonym eines Nutzers erlaubt (ohne Anwendung von Hintergrundwissen) keinen Rückschluss auf die Identität eines Nutzers. Erst die Kollaboration berechtigter Akteure ermöglicht das Aufdecken eines Pseudonyms.

? außerdem erweitern um aktive Angreifer?

Ein Angreifermodell beschreibt die maximale Stärke eines Angreifers im Bezug auf verschiedene Faktoren, gegen die ein System abgesichert ist. Enthalten sind die Rolle eines Angreifers, seine Verbreitung im System, aktives oder passives Verhalten und die Rechenkapazität, die der Angreifer zum Überwinden der eingesetzten Schutzmaßnahmen aufbringen kann. Es bildet die Basis für alle Folgeüberlegungen im Bezug auf die Sicherheit des zu entwickelnden Systems.

Für das Angreifermodell werden folgende Annahmen getroffen: Bei dem Angreifer kann es sich um einen Außenstehenden, um einen Berechtigten mit Zugriff auf das SIEM-System oder sogar um einen Administrator mit physischem Zugriff auf Rechner im Netz handeln. Wie bereits beschrieben, wird die ungesicherte Übertragung nicht-pseudonymisierter Daten zu dem System nicht betrachtet. Insofern wird von einem Angreifer ausgegangen, der erst Zugriff auf das zu entwickelnde System oder das SIEM-System beseitigt oder erlangt. Das zu entwickelnde System soll in der Lage sein, sich auch gegen aktive Angreifer zur Wehr zu setzen. Bezogen auf die verfügbare Rechenleistung des Angreifers sollen verbreitete und nach heutigem Wissensstand für

Denial of Service etc?

sicher befundene kryptographische Algorithmen als nicht mit vertretbarem Aufwand zu brechen angesehen werden. Es handelt sich daher um die Annahme von komplexitätstheoretischer Sicherheit.

4.1.2 Anforderungen an die Integration in OSSIM

Um zu beurteilen, an welcher Stelle in den Datenfluss der Logdaten in OSSIM (vergleiche Abschnitt 3.1) eingegriffen wird, um die Daten zu verändern, müssen Vor- bzw. Nachteile der verschiedenen Möglichkeiten gegeneinander abgewogen werden. Folgende Eigenschaften einer Möglichkeit sollten betrachtet werden:

- **Abhängigkeit von der OSSIM-Konfiguration:** Ist die Lösung sowohl in der verteilten als auch in der All-in-One-Installation von OSSIM umsetzbar? Eine Lösung, die unabhängig von der gewählten Installationsmethode funktioniert, ist zu bevorzugen, da sie universell einsetzbar ist.
- **Veränderung des SIEM-Systems:** Muss OSSIM für die Umsetzung der Lösung verändert werden? Dies wäre im Hinblick auf zukünftige Updates, die OSSIM durch den Entwickler erfährt, nicht wünschenswert, da jedes dieser Updates dafür sorgen könnte, dass die umgesetzte Lösung angepasst werden muss. Hierbei können sowohl direkte Änderungen an dem OSSIM-Quellcode als auch Eingriff in die interne Kommunikation betrachtet werden.
- **Nicht-pseudonymisierte Daten im SIEM-System:** Um das Ziel der Arbeit - die Pseudonymisierung, die nur durch Kollaboration aufgedeckt werden kann - zu erreichen, muss sichergestellt sein, dass Logdaten nirgendwo in nicht pseudonymisierter Form vorliegen. Da insbesondere das zukünftige Verhalten von OSSIM nicht beeinflusst werden kann, wäre es wünschenswert, dass die Logdaten bereits in pseudonymisierter Form das SIEM-System erreichen.
Die Relevanz dieser Eigenschaft lässt sich bereits an einem einfachen Beispiel erkennen: Wird das Syslog-Protokoll genutzt, um Logdaten in OSSIM aufzunehmen, so werden die Einträge im Sensor erst in eine Logdatei gespeichert und von dort aus geparkt, normalisiert und in der Datenbank gespeichert. Das Datum verbleibt in der Logdatei. Kommen die Daten in nicht-pseudonymisierter Form in dem OSSIM-Sensor an, so muss sichergestellt werden, dass verarbeitete Einträge gelöscht/verändert werden.
- **Mehrfaches Parsen von Logdaten:** Durch den OSSIM-Agenten werden die Logdaten - wie in Abschnitt 3.1 beschrieben - geparkt und normalisiert. Aus Performancegründen ist eine Lösung zu bevorzugen, die dieses Parsen nicht mehrfach voraussetzt.

4.1.3 Anforderungen an die Pseudonymisierung

Die Pseudonymisierung muss es ermöglichen nach Aufdecken eines Eintrags wieder auf den ursprünglichen Dateninhalt schließen zu können. Daher müssen die Pseudonyme für die Zeit ihrer Speicherung eindeutig sein, d.h. es darf zu keiner gleichzeitigen Mehrfachverwendung von Pseudonymen kommen.

Weiterhin muss es eine Möglichkeit beim Pseudonymisieren von Logeinträgen geben, zu überprüfen, ob für ein Datum bereits ein Pseudonym vergeben wurde. So kann sichergestellt werden,

dass in einem bestimmten Zeitraum Logeinträge zu einer Person mit dem gleichen Pseudonym versehen werden, um über die Verknüpfung von Einträgen die gewünschte Erkennung von Insider-Angriffen erreichen zu können.

Außerdem muss es eine Möglichkeit geben, die Parameter der Pseudonymisierung wie den Zeitraum ihrer Verwendung (vergleiche Abschnitt 3.2) konfigurierbar zu machen.

4.1.4 Anforderungen im Bezug auf den Einsatz eines kryptographischen Schwellwertschemas

Der Einsatz eines kryptographischen Schwellwertschemas setzt eine verteilte Anwendung voraus, die den Zugriff für die Pseudonymisierungskomponente sowie für die bei der Entschlüsselung eines Eintrags beteiligten Akteure bereitstellt. Die für das Schwellwertschema nötigen Parameter t und n sowie die beteiligten Akteure müssen anpassbar bzw. auswählbar sein.

In der Phase der Schlüsselgenerierung muss das System die Kommunikation und Koordination aller Beteiligten unterstützen. Die hier erstellten Schlüssel und *Shares* müssen an geeigneten Stelle sicher gespeichert und abrufbar sein.

Der für die Verschlüsselung erforderliche öffentliche Schlüssel muss so vorliegen, dass er bei der Verschlüsselung eines Pseudonym-Datensatzes genutzt werden kann.

Bei der Entschlüsselung eines Eintrags, also der Aufdeckung eines Pseudonyms, muss das System wiederum die beteiligten Akteure koordinieren und anschließend die Rolle des *Combiners* übernehmen, so dass anschließend der entschlüsselte Datensatz zentral vorliegt.

4.1.5 Benutzerinteraktion

Die zu entwickelnde verteilte Anwendung wird an verschiedenen Stellen Benutzerinteraktion erfordern.

Das Konfigurieren des Systems zur Integration verschiedener Datenquellen und deren Beschreibung muss einem berechtigten Nutzer zugänglich gemacht werden. Ebenso sollte es im Hinblick auf die in der Aufgabenstellung geforderte Erweiterbarkeit im Bezug auf weitere Datenschutztechniken relativ leicht sein, diese Techniken im System nutzen zu können.

Für pseudonymisierte Datensätze muss es berechtigten Benutzern ermöglicht werden, Anfragen zur Aufdeckung eines Pseudonyms zu stellen und sich über ihren Status informiert zu halten.

Für die Benutzung eines kryptographischen Schwellwertschemas sollte es einem Administrator des Systems ermöglicht werden, grundlegende Parameter des Systems wie die Schwellwertparameter und die beteiligten Nutzer auszuwählen sowie die Initialisierung des Schemas anzustoßen. Die am Schwellwertschema beteiligten Nutzer müssen die Möglichkeit erhalten, eine Übersicht über sie betreffende Anfragen zur Aufdeckung eines Pseudonym-Datensatzes zu bekommen sowie einzelne Anfragen abzulehnen oder sich an m Prozess des Aufdeckens mithilfe des Schwellwertschemas zu beteiligen.

Anpassen nach Kapitel 3 - Threshold-Abschnitt?

Performance?

4.1.6 Erweiterbarkeit um neue Datenquellen

Das umzusetzende System sollte es ermöglichen, Daten aus verschiedenen Quellen und (abhängig vom gewählten Eingriffspunkt in OSSIM auch) in verschiedenen Formaten entgegenzunehmen und mithilfe der umgesetzten Datenschutztechniken verändern zu können. Dabei muss das Format der Logdaten grundsätzlich beibehalten werden, um die Behandlung der Daten in OSSIM weiterhin zu ermöglichen.

4.1.7 Erweiterbarkeit um neue Datenschutztechniken

Neben der im Fokus dieser Arbeit stehenden Pseudonymisierung und dem Einsatz von kryptographischen Schwellwertschemata zum Schutz der Logdaten gibt es weitere Datenschutztechniken, die für den Anwendungsfall genutzt werden könnten (siehe Kapitel 5). Der umgesetzte Prototyp sollte leicht um diese Techniken erweiterbar sein, d.h. so gestaltet sein, dass andere Techniken ohne große Änderungen am System integriert und auf eingehende Logdaten angewendet werden können.

4.2 Entwurf

Einführender Absatz

4.2.1 Eingriff in den OSSIM-Datenfluss

Hier oder in Anforderungen?

Für den Eingriff zur Pseudonymisierung der Logdaten bieten sich verschiedene Stellen im Datenfluss von OSSIM an. In diesem Abschnitt sollen die verschiedenen Möglichkeiten dargestellt und gegeneinander abgewogen werden. Eine Übersicht über die verschiedenen Stellen bietet Abbildung 4.1. Im Folgenden sollen die verschiedenen Möglichkeiten bezogen auf die in Abschnitt 4.1.2 dargestellten Eigenschaften bewertet werden.

1. In der Quelle der Logdaten

Bei diesem Ansatz werden die Daten bereits pseudonymisiert, bevor sie die Datenquelle verlassen. Auch wenn dieser Ansatz aus datenschutztechnischer Sicht die beste Möglichkeit darstellen würde, so ist er doch nicht umsetzbar, da hierzu jede mögliche Quelle von Logdaten universell verändert werden müsste.

2. Syslog-Proxy

Dieser Ansatz pseudonymisiert die Daten vor dem ersten Kontakt mit einer OSSIM-Komponente, indem Datenquellen ihre Logdaten an einen Proxy senden, der die Daten pseudonymisiert und erst anschließend an OSSIM weiterreicht. Hierdurch wird erreicht, dass die Daten zu keiner Zeit nicht-pseudonymisiert in OSSIM vorliegen. Ein Nachteil dieser Lösung ist, dass sie das Parsen und Neuzusammensetzen der Logdaten im Proxy erfordert.

Beschreibung erweitern - analog zu Plugins in OSSIM

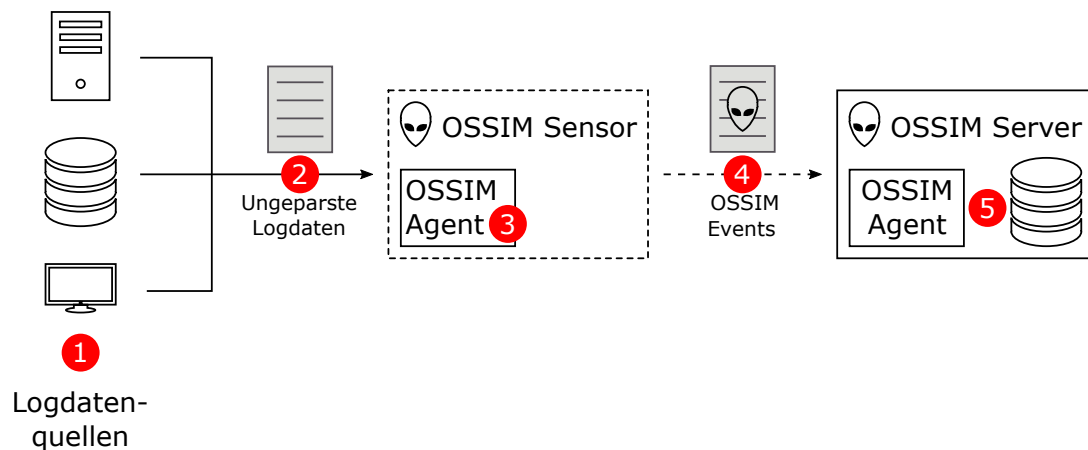


Abbildung 4.1: Mögliche Eingriffspunkte in den OSSIM-Datenfluss.

3. Patchen des OSSIM-Sensor-Agents

Bei dieser Lösung müsste der OSSIM-Agent des Sensors so verändert werden, dass vor dem Senden der Events an den Server die Pseudonymisierung stattfindet. Daten erreichen den OSSIM-Server nur pseudonymisiert und mehrfaches Parsen wie in der zweiten Lösung wird verhindert. Auf der anderen Seite erfordert diese Lösung einen Eingriff in die Funktionsweise von OSSIM, was beispielsweise bei Updates von OSSIM zu Problemen führen kann. Außerdem liegen die Daten zu Beginn in nicht-pseudonymisierter Form im Sensor vor. Zusätzlich erfordert diese Lösung die verteilte Installation von OSSIM-Sensor und -Server, schließt also die All-In-One-Installation aus.

Syslog-
Problematik er-
wähnen

4. Sensor-Server-Proxy

Hier wird ein Proxy zwischen Sensor und Server geschaltet, der bereits geparte Events pseudonymisiert und anschließend an den Server sendet. Dieser Ansatz würde mehrfaches Parsen verhindern und dafür sorgen, dass nur pseudonymisierte Logdaten den OSSIM-Server erreichen. Wie die vorhergehende Lösung würde er jedoch nur in der verteilten Installation funktionieren und zusätzlich in die Kommunikation zwischen Sensor und Server aktiv eingreifen, was im Hinblick auf die Nachrichtenintegrität¹ und auch auf geändertes Verhalten nach Updates von OSSIM einen Nachteil darstellt.

5. Patchen des OSSIM-Servers

Die letzte Möglichkeit ist das Verändern des OSSIM-Servers selbst. Diese Lösung ist vergleichbar mit der dritten Möglichkeit. Zusätzlich würde sie bei der verteilten sowie bei der All-In-One-Installation funktionieren, auf der anderen Seite aber zulassen, dass nicht-pseudonymisierte Events sogar noch direkt auf dem Server vorliegen.

Insbesondere der aus datenschutztechnischer Sicht relevante Vorteil, dass die Daten bereits pseudonymisiert in allen OSSIM-Komponenten eintreffen, ließ die Entscheidung auf die **zweite Möglichkeit** fallen. Dass die Lösung außerdem noch für beide Varianten der OSSIM-Installation möglich ist und keine Anpassungen an OSSIM selbst benötigt, wiegt den Nachteil des zusätzlichen Parsens und wieder Zusammensetzens der Lognachricht bei Weitem auf.

Auch auf Angrei-
fermodell beziehen

1. In der aktuellen Version von OSSIM werden Nachrichten unverschlüsselt und nicht signiert zwischen Sensor und Server versendet, aber zu hoffen ist, dass dieser Zustand sich in zukünftigen Versionen noch ändert.

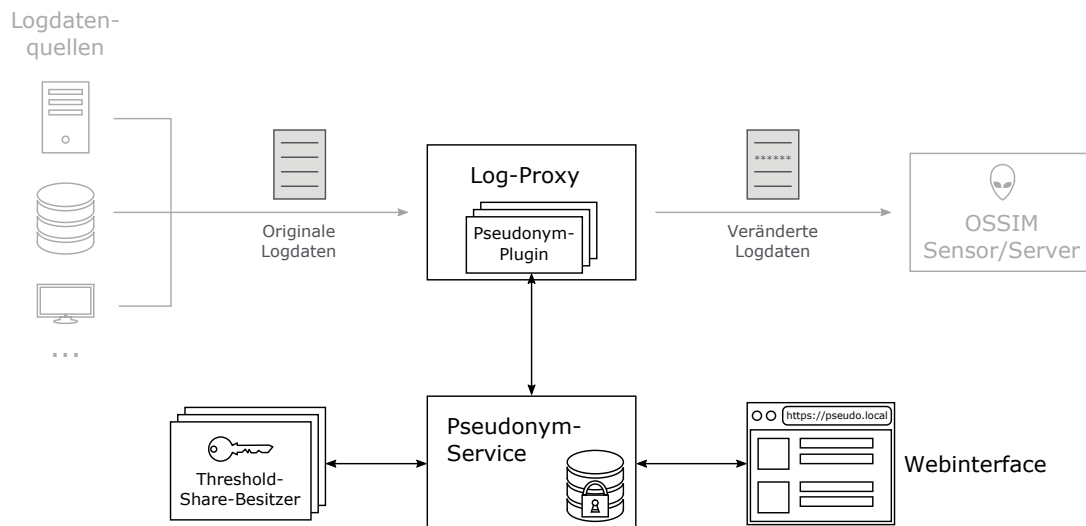


Abbildung 4.2: Ein Überblick über die entworfene Architektur.

4.2.2 Architektur

Ausgehend von diesen Überlegungen wurde ein verteiltes System entworfen, dass die Anforderungen aus Abschnitt 4.1 erfüllt und an der beschriebenen Stelle in den Datenfluss eingreift. Einen Überblick bietet Abbildung 4.2. Das System besteht aus verschiedenen Komponenten, die im Folgenden näher beschrieben werden.

Ein **Log-Proxy**, der die Daten über das Syslog-Protokoll entgegennimmt, verändert und anschließend an OSSIM weiterleitet. Das Verändern der Daten kann mit verschiedenen Plugins geschehen, so dass neben der umzusetzenden Pseudonymisierung auch weitere Datenschutztechniken eingesetzt werden können, was die geforderte Erweiterbarkeit aus Abschnitt 4.1.7 ermöglicht. Der Proxy leistet die Behandlung von Logdaten aus verschiedenen Quellen (siehe Abschnitt 4.1.6), was wie bereits im vorhergehenden Abschnitt beschrieben durch Parsen und Wiederaussetzen der Daten geschehen muss. Die Konfiguration des quellenabhängigen Vorgehens bei der Logdatenverarbeitung erfolgt ebenfalls hier.

Ein in dem Proxy enthaltenes Plugin wird für die Pseudonymisierung von Daten zuständig und kommuniziert dazu mit einer externen Komponente – dem Pseudonym-Service. Die Kommunikation mit dem Proxy erfolgt über einen Webservice-basierten Ansatz. Das Plugin kann für eingehende Daten ein Pseudonym anfordern und dieses anschließend in den Logdaten verwenden.

Der **Pseudonym-Service** erfüllt zwei Aufgaben: das Speichern und Verwalten der Pseudonyme sowie die Integration des kryptographischen Schwellwertschemas. Initial muss die Schlüsselgenerierung des Schwellwertschemas (bei zentraler Schlüsselgenerierung) oder die Koordinierung der teilnehmenden Benutzer (bei dezentraler Schlüsselgenerierung) durch den Service geleistet werden. Es können während des Betriebs neue Pseudonyme angelegt und zusammen mit ihrem durch das Schwellwertschema verschlüsselten Datum abgelegt werden. Sie werden durch geeignete Maßnahmen durchsuchbar gehalten, um für ein Datum überprüfen zu können, ob bereits ein Pseudonym vergeben wurde. Über ein Webinterface kann ein berechtigter Benutzer die Aufdeckung eines bestimmten Pseudonyms fordern und den Status seiner Forderung bzw. im

Change Store to
Service, Webinter-
face, Log-Proxy

Erfolgsfall das aufgedeckte Datum betrachten. Dieses Datum wird durch das Kombinieren der partiellen Entschlüsselungen erhalten, die von den entsprechenden *Share*-Besitzern berechnet werden.

Benutzer, die zuständig für die Bewertung von Anfragen zur Aufdeckung eines Pseudonyms sind, erhalten die Möglichkeit zur Interaktion mit dem System über eine **Client-Anwendung**, für die der Pseudonym-Service ebenfalls als Webservice agiert. Diese Anwendungen leisten im Falle der dezentralen Schlüsselgenerierung die initiale Generierung, verwalten den *Share* des Benutzers und können nach der Bestätigung des Benutzers zu der Aufdeckung eines Pseudonyms partiell beitragen.

4.3 Einbindung in OSSIM

- Vorgehen
- Nur für Syslog, keine anderen Formate betrachtet
- Syslog-Nachrichten werden entgegengenommen, geparkt und Config-basiert einzelne Datenfelder bearbeitet (basierend auf RegEx)
- Konfigurationsdateien für Datenquellen (leichte Erweiterbarkeit, auch im Bezug auf neue Datenquellen, ...) mit Beispiel
- Plugins (leichte Erweiterbarkeit, ...)

Für die in dieser Arbeit zu leistende prototypische Umsetzung des Systems wurde sich auf den am häufigsten² genutzten Weg des Datenerhalts in OSSIM (siehe Abschnitt 3.1.2) beschränkt: Das Entgegennehmen der Daten über das Syslog-Protokoll.

4.4 Umsetzung der Pseudonymisierung

4.5 Implementierung und Integration des Schwellwertschemas

4.6 Evaluation

Inklusive Problemen, ...

2. Die in OSSIM bereits mitgelieferten Plugins bestätigen dies. Unter den Hunderten Plugins ist nur ein einziges, das einen anderen Mechanismus als das Syslog-Protokoll nutzt. Auch wenn beispielsweise Plugins, die eine Datenbankabfrage enthalten, immer dem Anwendungsfall angepasst und daher nicht in OSSIM inkludiert werden können, so unterstützt dies doch die Annahme des Syslog-Protokolls als am häufigsten genutzten Weg und damit als geeignet für den Fokus dieser Arbeit.

5 Alternativen

- **Alternativen** Welche alternativen oder ergänzenden Vorgehensweisen zu Pseudonymisierung + kryptographisches Schwellwertschema gibt es und welche Eigenschaften, Vor- und Nachteile besitzen sie?
- **Umsetzung** Wie könnten diese Alternativen im Prototypen umgesetzt werden?

5.1 Alternative1

5.2 ...

5.3 Vorgehen zur Integration

6 Fazit

TBW

Literatur

- [BBH06] Dan Boneh, Xavier Boyen und Shai Halevi. *Chosen ciphertext secure public key threshold encryption without random oracles*. In: *Topics in Cryptology – CT-RSA 2006*. Hrsg. von David Pointcheval. Springer, 2006, S. 226–243.
- [bit16] bitkom. *Spezialstudie Wirtschaftsschutz*. bitkom. 2016. URL: <https://www.bitkom.org/Bitkom/Publikationen/Spezialstudie-Wirtschaftsschutz.html> (besucht am 24. 06. 2017).
- [Bla79] George Robert Blakley. *Safeguarding cryptographic keys*. In: *Proceedings of the AFIPS 1979 National Computer Conference*. Hrsg. von Richar E. Merwin. AFIPS Press, 1979, S. 313–317.
- [Bra+15] Nicholas Bradley u. a. *IBM 2015 Cyber Security Intelligence Index*. IBM. 2015. URL: <http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=SEW03073USEN> (besucht am 24. 06. 2017).
- [BS16] Dan Boneh und Victor Shoup. *A graduate course in applied cryptography*. In: (2016). URL: https://crypto.stanford.edu/~dabo/cryptobook/draft_0_3.pdf (besucht am 02. 09. 2017).
- [Des87] Yvo Desmedt. *Society and Group Oriented Cryptography: a New Concept*. In: *Conference on the Theory and Application of Cryptographic Techniques 1987*. Hrsg. von Carl Pomerance. Springer, 1987, S. 120–127.
- [Des93] Yvo Desmedt. *Threshold cryptosystems*. In: *Advances in Cryptology — AUS-CRYPT '92*. Hrsg. von Jennifer Seberry und Yuliang Zheng. Springer, 1993, S. 1–14.
- [Det+15] Kai-Oliver Detken u. a. *SIEM approach for a higher level of IT security in enterprise networks*. In: *2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*. IEEE. 2015, S. 322–327.
- [DF90] Yvo Desmedt und Yair Frankel. *Threshold cryptosystems*. In: *Advances in Cryptology - CRYPTO' 89 Proceedings*. Hrsg. von Gilles Brassard. Springer, 1990, S. 307–315.
- [DJ01] Ivan Damgard und Mads Jurik. *A generalisation, a simplification and some applications of Paillier's probabilistic public-key system*. In: *Public Key Cryptography – Proceedings of the 4th International Workshop on Practice and Theory in Public Key Cryptosystems*. Hrsg. von Kwangjo Kim. Bd. 1992. Springer, 2001, S. 119–136.
- [DRS14] Kai-Oliver Detken, Thomas Rossow und Ralf Steuerwald. *SIEM-Ansätze zur Erhöhung der IT-Sicherheit auf Basis von IF-MAP*. In: *Proceedings of the DACH Security 2014: Bestandsaufnahme, Konzepte, Anwendungen und Perspektiven*. Hrsg. von Peter Schartner und Peter Lipp. syssec, 2014.

- [Gem97] Peter Gemmell. *An introduction to threshold cryptography*. In: *CryptoBytes, a technical newsletter of RSA Laboratories* 2.7 (1997), S. 295–310.
- [Gen+99] Rosario Gennaro u. a. *Secure distributed key generation for discrete-log based cryptosystems*. In: *Advances in Cryptology — EUROCRYPT’99*. Hrsg. von Jaques Stern. Springer, 1999, S. 295–310.
- [Ngu05] H.L. Nguyen. *RSA Threshold Cryptography*. 2005. URL: <https://www.cs.ox.ac.uk/files/269/Thesis.pdf> (besucht am 02. 07. 2017).
- [NK11] Mark Nicolett und Kelly M Kavanagh. *Magic quadrant for security information and event management*. In: *Gartner Research Note G00212454* (2011).
- [Pai99] Pascal Paillier. *Public-key cryptosystems based on composite degree residuosity classes*. In: *Advances in Cryptology — EUROCRYPT’99*. Hrsg. von Jaques Stern. Springer, 1999, S. 223–238.
- [Ped91] Torben Pedersen. *A threshold cryptosystem without a trusted party*. In: *Advances in Cryptology — EUROCRYPT’91*. Hrsg. von Donald W. Davies. Springer, 1991, S. 522–526.
- [PH10] Andreas Pfitzmann und Marit Hansen. *A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management*. 2010. URL: http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf (besucht am 27. 06. 2017).
- [PWP90] Birgit Pfitzmann, Michael Waidner und Andreas Pfitzmann. *Rechtssicherheit trotz Anonymität in offenen digitalen Systemen*. In: *Datenschutz und Datensicherung DuD* 14.5-6 (1990), S. 243–253.
- [Sha79] Adi Shamir. *How to share a secret*. In: *Communications of the ACM* 22.11 (1979), S. 612–613.
- [WWC16] Yunling Wang, Jianfeng Wang und Xiaofeng Chen. *Secure searchable encryption: a survey*. In: *Journal of communications and information networks* 1.4 (2016), S. 52–65.