



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Masterarbeit

Entwurf vom
4. März 2018

**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 4. März 2018

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
■ Nach Einleitung nochmal umschreiben	11
■ Gesetzestexte als „Quellen,, in Fußnoten?	11
■ Quelle	12
■ Quelle	12
■ Beispiele für Überwachungsskandale	13
■ Evtl. auch Telemediengesetz erwähnen - Nutzung eines Systems unter einem Pseudonym	15
■ Vielleicht diese Werte direkt nutzen?	17
■ mod p irgendwo erwähnen	17
■ Vmtl.? Mehr Details...	18
■ EZ: c' ist ungeeignet fuer die Benennung eines Entschluesselungsergebnisses, wenn du den Schluesseleltext mit c bezeichnest. Wo kommt das j her?	18
■ Erweitern, um später darauf zurückgreifen zu können	21
■ Performance?	25
■ EZ: Hier interessieren insbesondere auch die Auswirkungen auf den Rest des Systems? Wie muss das Schwellwertschema angepasst werden? Welche Akteure wissen zu welchem Zeitpunkt was? Wer bekommt welche Schluesssel? Usw... Der Abschnitt hat also wesentlich mehr Potential!	26
■ Auch auf Angreifermodell beziehen?	28
■ Verhindert verteiltes System - diesen Aspekt betonen (mehrere Proxies, ...)	42
■ Hash-Kollisionen und Geburtstags-Paradoxon hier erwähnen?	42

■ Um Implementation erweitern (welche Parameter in welchen Systemteilen, Setup, ...)	46
■ ergänzen	48
■ evtl. auf Evaluation beziehen.	49
■ TBW	52
■ Beispiel, Diagramm?	55
■ Beispiel	56
■ TBW	58

Zusammenfassung

TBW

Inhaltsverzeichnis

1	Einführung	8
1.1	Related work	9
2	Grundlagen	11
2.1	Definitionen und Notationen	11
2.1.1	Mathematische Notationen	11
2.2	Arbeitnehmerdatenschutz	11
2.2.1	Das Recht auf informationelle Selbstbestimmung	12
2.2.2	Datenschutz im Beschäftigungsverhältnis	12
2.3	SIEM-Systeme	13
2.3.1	Das Syslog-Protokoll	15
2.4	Pseudonymisierung	15
2.5	Schwellwertschemata	16
2.5.1	Shamir's Secret Sharing	17
2.5.2	Threshold Decryption	18
2.6	Weitere kryptographische Verfahren und Techniken	19
2.6.1	Hashfunktion	19
2.6.2	Message Authentication Code	20
2.6.3	Hybride Kryptosysteme	20
2.6.4	Authenticated Encryption Schemes	20
2.6.5	ElGamal-Kryptosystem	20
2.7	Searchable Symmetric Encryption	21
3	Ueberblick	22
3.1	Anforderungen	23
3.1.1	Anforderungen an die Integration in das SIEM-System	23
3.1.2	Anforderungen an die Pseudonymisierung	23
3.1.3	Anforderungen im Bezug auf den Einsatz eines kryptographischen Schwellwertschemas	24
3.1.4	Benutzerinteraktion	24
3.1.5	Erweiterbarkeit um neue Datenquellen	25
3.1.6	Erweiterbarkeit um neue Datenschutztechniken	25
3.1.7	Übersicht	25
3.2	Entwurf	26
3.2.1	Eingriff in den Datenfluss des SIEM-Systems	26
3.2.2	Architektur	28
3.3	Angreifermodell	29
4	Stand der Wissenschaft und Auswahl von Verfahren	32
4.1	SIEM-Systeme	32
4.1.1	OSSIM-Überblick	33
4.1.2	Parsen von Logdaten in OSSIM	34

4.2	Pseudonymisierung	34
4.2.1	Mobilfunknetze	35
4.2.2	VANets	35
4.2.3	Pseudonymisierung im zu entwickelnden System	36
4.3	Schwellwertschemata	37
4.3.1	Übersicht	37
4.3.2	ElGamal-basiertes Schwellwertschema	38
4.3.3	Verteilte Schlüsselgenerierung	39
4.3.4	Existierende Implementierungen	40
4.4	Searchable Encryption	40
4.4.1	Entschlüsseln aller Datensätze	41
4.4.2	Deterministische Verschlüsselung	41
4.4.3	Nutzung von Hashfunktionen	41
4.4.4	Lokale Zuordnung	41
4.4.5	Message Authentication Codes	42
4.4.6	Weitere Möglichkeiten der Searchable Encryption	42
5	Implementierung	44
5.1	Einbindung in OSSIM	44
5.1.1	Konfiguration von Datenquellen	45
5.1.2	Umsetzung von Datenschutztechniken durch Plugins	46
5.2	Umsetzung der Pseudonymisierung	46
5.2.1	Setup-Phase	47
5.2.2	Proxy	47
5.2.3	Service	47
5.2.4	Perfect Forward Privacy	47
5.3	Implementierung und Integration des Schwellwertschemas	48
5.3.1	ThresholdCrypto „Lib“	48
5.3.2	Service und Setup-Verfahren	51
5.3.3	Client-Anwendung	51
5.3.4	Proxy	52
5.4	Evaluation	52
5.4.1	Erfüllung der Anforderungen	52
5.4.2	Angriffsmöglichkeiten	52
5.4.3	Performance	53
6	Ergänzende und alternative Datenschutztechniken	55
6.1	Unterdrückung	55
6.2	Generalisierung	55
6.3	Verrauschen	56
6.4	Nutzung von Hashverfahren	56
6.5	Vorgehen zur Integration	56
7	Fazit	58
	Literatur	59

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 tatsächlichen 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

Im Bereich der Erkennung von Insiderangriffen fand bereits einige Forschungsarbeit statt. In [SHS08] bieten die Autoren einen Überblick über Forschungsergebnisse basierend auf unterschiedlichen Verfahren aus der Statistik und des maschinellen Lernens sowohl auf Host- als auch auf Netzwerkebene. Hier wird auch die Frage nach Erhalt der Privatsphäre eines Nutzers als Feld weiterer notwendiger Forschung dargestellt:

Hence, we also believe that any technologies developed to detect insider attack have to include strong privacy-preserving guarantees to avoid making false claims that could harm the reputation of individuals whenever errors occur. [...] How might a system alert a supervisor of a possible attack without disclosing an employee's true identity unless and until an attack has been validated? [SHS08]

Mit dieser Fragestellung beschäftigen sich weitere Veröffentlichungen im Bereich der Intrusion-Detection-Systeme. Oftmals wird – wie in dieser Arbeit auch – Pseudonymisierung als Verfahren zum Erhalt der Privatsphäre genutzt.

In [SFHR97] werden zwei Ansätze zur *Privacy Enhanced Intrusion Detection* vorgestellt. Die Pseudonymisierung wird jeweils bereits im Betriebssystem-Kernel vorgenommen und mithilfe symmetrischer Verschlüsselung erreicht. Auch die Nutzung des 4-Augen-Prinzips wird bei der Pseudonymaufdeckung bereits erwähnt – hier angestrebt durch eine Aufteilung des symmetrischen Schlüssels.

In [BK99] stellt der Autor einen Architekturansatz für Intrusion-Detection-Systeme vor, der ebenfalls auf der Nutzung von Pseudonymen beruht. Es werden zwei Ansätze basierend auf Kerberos-Tickets bzw. dem MIX-Konzept vorgestellt. Für die Generierung bzw. Aufdeckung eines Pseudonyms wird eine *Trusted Third Party* benötigt.

In [LJ00] wird von den Autoren ein System zur Anomalieerkennung auf Basis von Pseudonymen entwickelt und anhand von Logdaten einer Firewall überprüft. Das Aufdecken von Pseudonymen wird hier als durch organisatorische Maßnahmen zu regelnder Prozess verstanden.

In [BF00] und [BF01] nutzen die Verfasser Shamir's Secret Sharing zur Erzeugung von Pseudonymen. Jeder Share bildet ein Pseudonym. Hierdurch wird sichergestellt, dass ein Pseudonym erst aufgedeckt werden kann, wenn eine einen Schwellwert übertreffende Anzahl von Warnmeldungen zu einem Nutzer im System eingetroffen ist.

Neben den Pseudonym-basierten Lösungen gibt es weitere auf anderen Verfahren basierende Forschungsergebnisse. In [Par+07] beispielsweise werden von den Autoren die Eigenschaften homomorpher Verschlüsselung zur Privatsphäre-erhaltenden Angriffserkennung eingesetzt, in [Nik+13] wird eine Art der Mehrparteienberechnung genutzt, um mehrseitigen Datenschutz insbesondere im Hinblick auf Zero-Day-Lücken in einem Intrusion-Detection-System zu garantieren.

In [NKS17] bieten die Autoren einen Überblick über weitere existierende Lösungen im Bereich der *Privacy Enhanced Intrusion Detection*.

2 Grundlagen

Nach Einleitung
nochmal umschrei-
ben

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

2.1.1 Mathematische Notationen

p, q

\mathbb{Z}_p

\mathbb{Z}_p^*

2.2 Arbeitnehmerdatenschutz

Gesetzestexte als
„Quellen“, in Fuß-
noten?

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.

1. In manchen Veröffentlichungen wird der Arbeitnehmerdatenschutz auch als Mitarbeiterdatenschutz, Beschäftig-tendatenschutz, 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?

2.3.1 Das Syslog-Protokoll

Das Syslog-Protokoll wird in dieser Arbeit als Grundlage für den Empfang von Logdaten im SIEM-System genutzt und soll daher hier kurz dargestellt werden. Das Protokoll wird weitverbreitet für die Übertragung von Logdaten in Rechnernetzen genutzt. Nachrichten enthalten drei Felder: die *Priority*, die die Schwere des gemeldeten Vorfalls beschreibt, die *Facility*, die die Komponente beschreibt, in der der Vorfall auftrat, und die eigentliche Nachricht in textueller Form [Ger09].

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 (auch bekannt als Pseudonymhalter) 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 öffentlichen Telefonbuch mit der Identität der Person verknüpft ist. Eine Kontonummer zu einem Bankkonto, dessen Inhaber bei der Kontoeröffnung nur der Bank und ihm selbst bekannt ist, bildet ein nicht-öffentliches Pseudonym – die Zuordnung ist nur wenigen Stellen bekannt. Ein anonymes Pseudonym ist zu Beginn nur dem Besitzer bekannt. Ein Beispiel bildet ein selbstgewählter Benutzername in einem Webforum ohne Hinterlegung identifizierender Merkmale wie Klarname oder Adresse.

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.

Personenpseudonyme beschreiben Pseudonyme wie beispielsweise die Sozialversicherungsnummer, die stellvertretend für die eindeutige Identität des Subjekts in der Gesellschaft stehen.

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

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

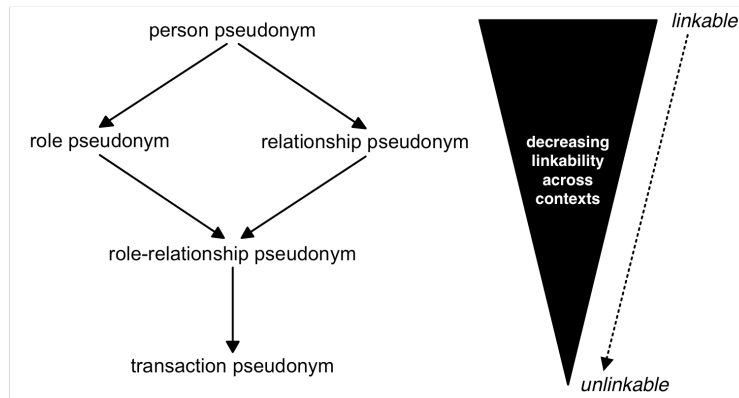


Abbildung 2.1: Pseudonym-Verband entsprechend ihrer Nutzung in verschiedenen Kontexten. Entnommen aus [PH10].

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 z.B. 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 (t, n) -Schwellwertschema: Ein Geheimnis D wird so in n Teile D_1, \dots, D_n (engl. *Shares*) zerlegt, dass durch Kenntnis von mindestens t Teilen das Geheimnis wieder aufgedeckt werden kann, aber jede Kombination aus höchstens $t - 1$ Teilen keine Informationen über D liefert [Sha79].⁷ Keine Information meint hier, dass jedes mögliche Geheimnis gleich wahrscheinlich D darstellt und die Kenntnis von weniger Shares als nötig nichts an diesen Wahrscheinlichkeiten ändert. Man spricht hierbei auch von informationstheoretischer Sicherheit.

Auf Basis dieses Verfahrens kann also die Integrität eines Schlüssels erhöht werden, da nun selbst bei Verlust von $n - t$ 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 t Teile des Schlüssels zur Wiederherstellung vorliegen müssen.

Shamirs Verfahren wird nachfolgend 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 $p > n$, wobei n die Anzahl an späteren Share-Besitzern bezeichnet, gewählt. Weiterhin wird ein Polynom

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

derart gewählt, dass a_1, \dots, a_{t-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, die auch als Lagrange-Koeffizienten bezeichnet werden. Das gesuchte Geheimnis D kann nun als

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

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

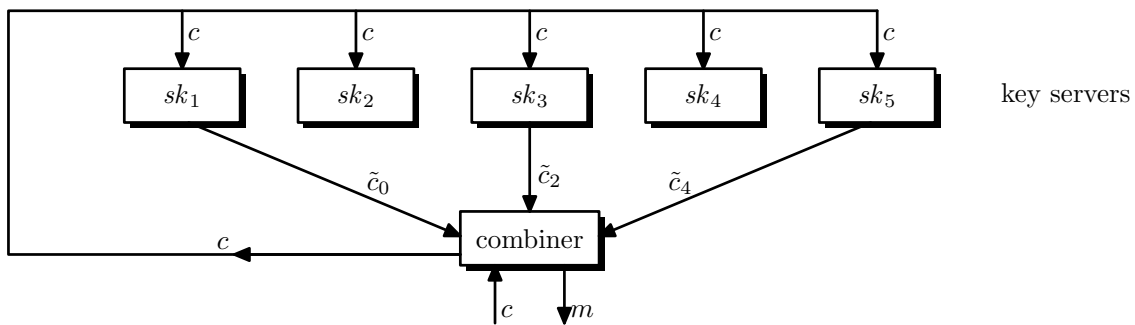


Abbildung 2.2: Übersicht über den Entschlüsselungsvorgang bei der Nutzung eines (3,5)-Schwellwertschemas. Entnommen aus [BS16].

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.

2.5.2 Threshold Decryption

In [Des87] wird das Verfahren der Schwellwertschemata 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 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 einsatzbare 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*. r ist als stellvertretend für die einfließenden Zufallswerte zu betrachten.
- $E(pk, m, r)$ steht für den Algorithmus, der der Verschlüsselung eines Klartexts 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'_j liefert.

Vmtl.? Mehr Details...

EZ: c' ist ungeeignet fuer die Benennung eines Entschlueßelungsergebnisses, wenn du den Schluesselformat mit c bezeichnest. Wo kommt das j her?

- $C(c, c'_1, \dots, c'_t)$ ist der Algorithmus, der aus dem Schlüsseltext c und aus t durch D generierten partiellen Entschlüsselungen wieder den Klartext m liefert. Dieser Algorithmus wird auch *Combiner* genannt.

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 $c'_j = D(sk_j, c)$, wobei sk_i der i -te von G erstellte *Share* ist, gelte: $V(pk, vk, c, c'_j)$ liefert ein valides Ergebnis.

2.6 Weitere kryptographische Verfahren und Techniken

In diesem Abschnitt werden die Grundlagen weiterer kryptographischer Verfahren und Techniken vorgestellt, die in dieser Arbeit verwendet werden.

2.6.1 Hashfunktion

Eine Hashfunktion ist eine Funktion, die eine Eingabe variabler Länge auf eine Ausgabe fester Länge (den Hashwert) abbildet.

In der Kryptographie werden meist kryptographisch sichere Hashfunktionen eingesetzt. Bei dieser Art von Hashfunktionen handelt es sich um Einwegfunktionen, d.h. es ist leicht, aus einer Eingabe den Hashwert zu berechnen, jedoch nicht mit vertretbarem Aufwand möglich, zu einem gegebenen Hashwert eine Eingabe zu finden, die auf diesen Wert abgebildet wird. Zusätzlich müssen die Hashfunktionen kollisionsresistent sein: Für einen gegebenen Wert ist es praktisch nicht möglich einen zweiten Wert zu finden, der den gleichen Hashwert besitzt [Sch06].

2.6.2 Message Authentication Code

Ein Message Authentication Code (MAC) ist ein symmetrisches Verfahren, das dazu dient, die Authentizität und die Integrität einer Nachricht sicherzustellen. Dazu wird vom Sender aus einem geheimen Schlüssel k und der Nachricht m eine Art Prüfsumme generiert und zusammen mit der Nachricht versendet. Ein Empfänger kann den MAC überprüfen, wenn er im Besitz des gleichen geheimen Schlüssels ist, und so sicher sein, dass die Nachricht nicht verändert wurde [Sch06].

2.6.3 Hybride Kryptosysteme

Als hybrides Kryptosystem wird die Kombination von symmetrischen und asymmetrischen Kryptoverfahren zur Verschlüsselung bzw. Entschlüsselung einer Nachricht bezeichnet. Ein Schlüssel k_{symm} für die Verwendung im symmetrischen Verfahren wird zufällig erzeugt und mithilfe des öffentlichen Schlüssels eines asymmetrischen Verfahrens als c_{public} verschlüsselt. Der zu verschlüsselnde Klartext m wird anschließend mithilfe des symmetrischen Verfahrens und des erzeugten Schlüssels k_{symm} als Chiffretext c_{symm} verschlüsselt.

Zur Entschlüsselung wird c_{public} mit dem geheimen Schlüssel des asymmetrischen Verfahrens entschlüsselt. Der hieraus erhaltene Schlüssel k_{symm} kann nun zur Entschlüsselung von c_{symm} genutzt werden, um m zu erhalten [KL14].

Der Vorteil dieser Lösung beruht darin, dass Vorteile symmetrischer und asymmetrischer Verfahren kombiniert werden: Symmetrische Verfahren sind im Allgemeinen deutlich schneller als asymmetrische, die jedoch das bei symmetrischen Systemen bestehende Problem des Schlüsselaustauschs lösen.

2.6.4 Authenticated Encryption Schemes

Symmetrische Kryptosysteme sorgen erst einmal nur für den Schutz der Vertraulichkeit einer Nachricht. Wird zusätzlich die Integrität einer Nachricht durch das System geschützt, so spricht man von einem *Authenticated Encryption Scheme*. Hierdurch wird erreicht, dass Änderungen am Schlüsseltext bei der Entschlüsselung erkannt werden und der Vorgang abgebrochen werden kann.

Ein solches System kann durch Berechnung eines MACs zusätzlich zur Verschlüsselung erreicht werden. Alternativ dazu gibt es Schemata, die direkt auf einer Blockchiffre aufbauen [BS16]. Ein Beispiel hierzu ist der GCM-Betriebsmodus, der in Kombination mit AES in vielen verbreiteten Protokollen wie TLS zu finden ist.

2.6.5 ElGamal-Kryptosystem

Das ElGamal-Kryptosystem ist ein von Taher ElGamal entwickeltes asymmetrisches Kryptosystem, das zur Verschlüsselung und der Erstellung von Signaturen genutzt werden kann [ElG85]. Im Folgenden wird die Ver- und Entschlüsselung beschrieben, die für das in dieser Arbeit verwendete Skryptographische Schwellwertschema relevant ist.

Im Folgenden sei \mathbb{G} eine zyklische Gruppe der primen Ordnung p und g ein Generator dieser Gruppe. Diese Parameter können öffentlich bekannt gegeben werden. Alle folgenden Berechnungen werden in \mathbb{G} (also modulo p) ausgeführt.

Ein Teilnehmer wählt nun ein zufälliges Element $x \in \mathbb{Z}_p$. Dies ist der private Schlüssel des Teilnehmers. Er berechnet zusätzlich seinen öffentlichen Schlüssel $h = g^x$.

Um eine Nachricht m , die an den Teilnehmer geschickt werden soll, zu verschlüsseln, wird zuerst ein zufälliges Element $y \in \mathbb{Z}_p$ gewählt. Anschließend kann die Nachricht verschlüsselt als $(v, c) = (g^y, h^y \cdot m)$ versendet werden.

Zur Entschlüsselung berechnet der Empfänger $k' = (v^x)^{(-1)}$ und kann die Nachricht $m = c \cdot k'$ entschlüsseln. Dies funktioniert, da

$$c \cdot k' = (h^y \cdot m) \cdot (v^x)^{(-1)} = g^{xy} \cdot m \cdot g^{(-yx)} = m$$

gilt.

Die Sicherheit des Verfahrens beruht auf dem Diskreten-Logarithmus-Problem. Details und Beweise hierzu sind beispielsweise in [KL14] zu finden.

Erweitern, um später darauf zurückgreifen zu können

2.7 Searchable Symmetric Encryption

Searchable Symmetric Encryption (SSE) ist ein Konzept, das es ermöglicht, Daten in verschlüsselter Form auf einen Server auszulagern und trotzdem Suchanfragen auf den Daten ausführen zu können. 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.

3 Ueberblick

Das Ziel dieser Arbeit ist es ein System zu entwickeln, das mithilfe von Pseudonymisierung die datenschutzgerechte Speicherung von Überwachungsdaten ermöglicht, wobei die Identität eines Pseudonymhalters im Bedarfsfall durch die Kollaboration verschiedener Akteure unter Nutzung eines kryptographischen Schwellwertschemas aufdeckbar sein muss. Um eine Basis für die Erarbeitung von Anforderungen und eines Systementwurfs zu erhalten, soll nun kurz dargelegt werden, wie die verschiedenen Verfahren ineinander greifen.

Aus Datenquellen wie Firewalls, Zugriffsprotokollen von Dateisystemen oder auch elektronischen Türschlössern werden personenbeziehbare Logdaten¹ an ein SIEM-System gesendet und dort gespeichert. In diesen Datenfluss wird nun durch ein zu entwickelndes System eingegriffen, das die personenbeziehbaren Informationen des Logdatums durch ein Pseudonym ersetzt. Die Zuordnung zwischen dem gesetzten Pseudonym und der personenbeziehbaren Information wird durch ein kryptographisches Schwellwertschema verschlüsselt und in dem System gespeichert. Damit die anschließend auf den pseudonymisierten Logdaten stattfindende Anomalieerkennung Aktionen eines Mitarbeiters verknüpfen kann, muss sichergestellt werden, dass für einen speziellen Benutzer das gleiche Pseudonym verwendet wird. Wird nun durch die Anomalieerkennung ein Angriff erkannt, so können berechtigte Benutzer durch kollaborative Entschlüsselung der Pseudonymzuordnung den hinter dem Pseudonym stehenden Benutzer wieder aufdecken.

Anschaulich wird der Vorgang noch einmal in Abbildung 3.1 dargestellt: Die Benutzerin Eve agiert in dem Unternehmensnetzwerk. Ihre Aktionen werden protokolliert und Logdaten, die ihren Benutzernamen enthalten, werden versendet. Dieser Benutzername wird durch das Pseudonym ps01 ersetzt. Die Zuordnung des Pseudonyms wird nun mithilfe eines kryptographischen Schwellwertschemas verschlüsselt und in einer Datenbank abgelegt. Das pseudonymisierte Logdatum wird nun im SIEM-System gespeichert.

Eingesetzte Anomalieerkennungsverfahren können nun auf die Daten des SIEM-Systems zugreifen. Wird ein möglicher Insider-Angriff durch einen Benutzer mit dem Pseudonym ps01 erkannt, so kann die Zuordnung zu dem ursprünglichen Benutzernamen mithilfe des Schwellwertschemas wieder aufgedeckt werden. Hierzu ist jedoch die Mitarbeit Alice, Bob und Carol notwendig, die jeweils im Besitz eines Teils des Entschlüsselungsschlüssels sind. Stimmen Sie der Aufdeckung zu, so wird Eve als Halter des Pseudonyms ps01 und damit als möglicher Innentäter aufgedeckt.

In diesem Kapitel werden nun zentrale Anforderungen an ein solches System entwickelt und eine abstrakte Architektur für ein solches entworfen. Anschließend wird aufbauend darauf ein Angreifermodell für das System definiert.

1. Da im Bereich technischer Systeme eher von Logdaten oder Protokolldaten im Gegensatz zu Überwachungsdaten gesprochen wird, wird diese Terminologie hier verwendet. Im Rahmen dieser Arbeit sind die Begriffe jedoch synonym zu verstehen.

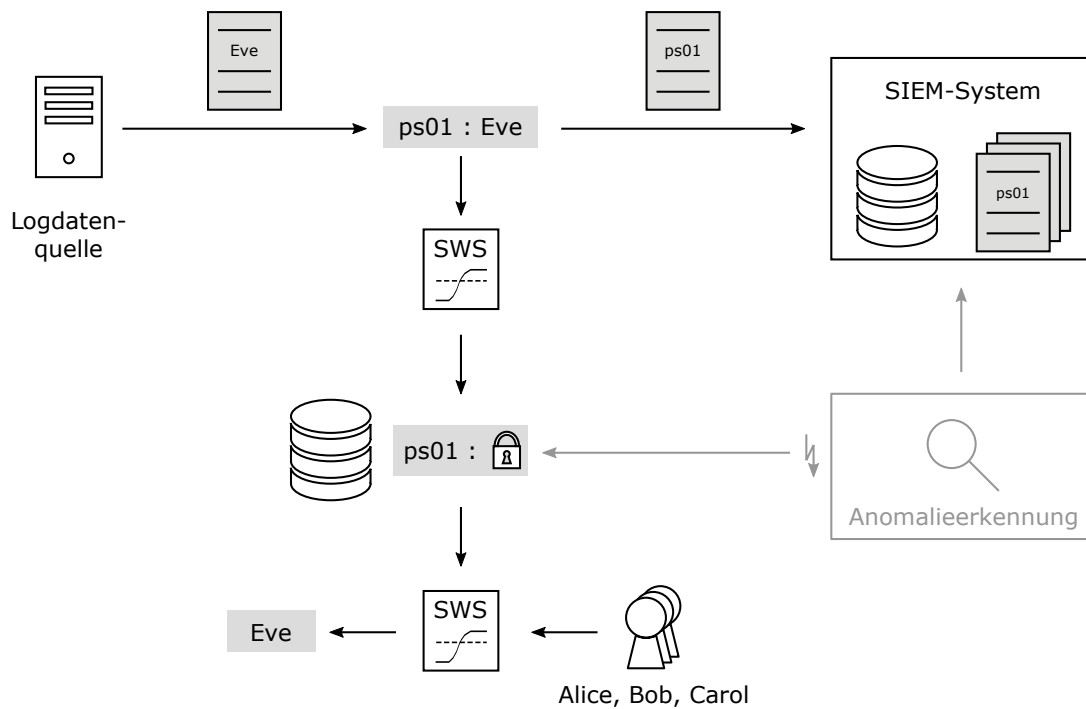


Abbildung 3.1: Übersicht über das angestrebte Verfahren.

3.1 Anforderungen

Neben den primären Anforderungen, die sich direkt aus der Funktionsbeschreibung des Systems und dem Zusammenspiel der enthaltenen Verfahren ergeben, sollte das System noch weitere Eigenschaften wie beispielsweise die Erweiterbarkeit um weitere Datenschutztechniken erfüllen. Alle diese Anforderungen sollen im folgenden Abschnitt aufgestellt und näher erläutert werden.

3.1.1 Anforderungen an die Integration in das SIEM-System

Es muss eine geeignete Stelle für den Eingriff in den Datenfluss der Logdaten in dem verwendeten SIEM-System gefunden werden. Hierzu müssen Auswirkungen des Eingriffs betrachtet und Vor- bzw. Nachteile der verschiedenen Möglichkeiten gegeneinander abgewogen werden.

3.1.2 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 mithilfe der Verknüpfung von Einträgen Anomalieerkennungsverfahren

sinnvoll einsetzen zu können. Auf diese Anforderung wird in Abschnitt 4.4 noch genauer eingegangen.

Außerdem muss es eine Möglichkeit geben, die Parameter der Pseudonymisierung wie den Zeitraum ihrer Verwendung konfigurierbar zu machen (siehe Abschnitt 4.2).

3.1.3 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, in Abschnitt 2.5 beschriebenen Parameter t und n sowie die beteiligten Share-Besitzer müssen in dem System initial konfigurierbar 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 Stellen sicher gespeichert und abrufbar sein. Es existieren zwei Möglichkeiten für diese Phase:

- **Zentrale Generierung von öffentlichem Schlüssel und Shares:** Eine vertrauenswürdige Komponente generiert ein Schlüsselpaar und zerlegt den geheimen Schlüssel in die einzelnen Shares, die anschließend verteilt werden können.
- **Verteilte Schlüsselgenerierung:** Hierbei generieren die einzelnen Share-Besitzer jeweils ihre Shares. Durch verteilte Berechnungen kann hieraus der gemeinsame öffentliche Schlüssel generiert werden. Der geheime Schlüssel liegt dadurch niemals an einer Stelle vor und ein vertrauenswürdiger Dritter ist nicht notwendig. Aus diesem Grund ist diese Lösung zu bevorzugen.

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. Anschließend muss eine Komponente die Rolle des *Combiners* übernehmen, so dass anschließend der den Pseudonymhalter beschreibende entschlüsselte Datensatz im System vorliegt.

3.1.4 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 am Prozess des Aufdeckens mithilfe des Schwellwertschemas zu beteiligen.

3.1.5 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 dem verwendeten SIEM-System weiterhin zu ermöglichen.

3.1.6 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 6). 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.

Performance?

3.1.7 Übersicht

Zusammenfassend sollte ein System, wie es in dieser Arbeit angestrebt wird, also folgende Anforderungen erfüllen:

- Geeignete Stelle zum Eingriff in den Datenfluss zwischen Logdatenquelle und SIEM-System
- Parameterabhängige Generierung eindeutiger, aber in gewissem Rahmen verknüpfbarer Pseudonyme
- Sicherer, verteilter Einsatz eines anpassbaren kryptographischen Schwellwertschemas – vorzugsweise mit verteilter Schlüsselgenerierung
- Geeignete Benutzerinteraktion mit dem System an notwendigen Stellen
- Erweiterbarkeit um unbekannte Datenquellen
- Erweiterbarkeit um weitere Datenschutztechniken

3.2 Entwurf

In diesem Abschnitt wird basierend auf den Anforderungen aus Abschnitt 3.1 eine von eingesetzten Verfahren unabhängige Architektur für das System entworfen. Der erste Unterabschnitt beschäftigt sich mit der Frage, an welcher Stelle in den Datenfluss zwischen Quelle der Logdaten und SIEM-System eingriffen werden kann. Anschließend wird basierend hierauf die Systemarchitektur erstellt.

3.2.1 Eingriff in den Datenfluss des SIEM-Systems

Für den Eingriff zur Pseudonymisierung der Logdaten bieten sich verschiedene Stellen im Datenfluss eines SIEM-Systems an. Im Folgenden sollen diese Möglichkeiten dargestellt und bezogen auf die folgenden Eigenschaften einer Möglichkeit bewertet werden:

- **Veränderung des SIEM-Systems:** Muss das eingesetzte SIEM-System für die Umsetzung der Lösung angepasst werden? Dies wäre im Hinblick auf zukünftige Updates, die das SIEM-System durch seinen Entwickler erfährt, nicht wünschenswert, da jedes dieser Updates dafür sorgen könnte, dass die umgesetzte Lösung angepasst werden muss. Weiterhin würde dieser Ansatz ein SIEM-System erfordern, dass entweder quelloffen vorliegt und verändert werden darf oder die gewünschte Verhaltensänderung zumindest durch Erweiterungen zulässt.
- **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 des SIEM-Systems nicht beeinflusst werden kann, wäre es wünschenswert, dass die Logdaten bereits in pseudonymisierter Form das SIEM-System erreichen.
Die Relevanz dieser Eigenschaft lässt sich am Beispiel des später in dieser Arbeit genutzten SIEM-Systems OSSIM erkennen: Wird das Syslog-Protokoll genutzt, um Logdaten in OSSIM aufzunehmen, so werden die Einträge erst in einer Logdatei abgelegt 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 oder verändert werden.
- **Mehrfaches Parsen von Logdaten:** Durch das SIEM-System werden die Logdaten - wie in Abschnitt 2.3 beschrieben - geparkt und normalisiert. Aus Performancegründen ist eine Lösung zu bevorzugen, die diesen Vorgang oder Teile davon nicht mehrfach voraussetzt.
- **Abhängigkeit von Besonderheiten des SIEM-Systems:** Einige SIEM-Systeme bieten Möglichkeiten der verteilten Installation oder andere spezifische Eigenschaften. Eine Lösung, die unabhängig von dem verwendeten SIEM-System funktioniert, ist zu bevorzugen, da sie universell einsetzbar ist.

Eine Übersicht über die verschiedenen Stellen des Eingriffs bietet Abbildung 3.2. Die Ziffern der Möglichkeiten beziehen sich auf die in der Abbildung gekennzeichneten Stellen.

EZ: Hier interessieren insbesondere auch die Auswirkungen auf den Rest des Systems? Wie muss das Schwellwertschema angepasst werden? Welche Akteure wissen zu welchem Zeitpunkt was? Wer bekommt welche Schlüssel? Usw... Der Abschnitt hat also wesentlich mehr Potential!

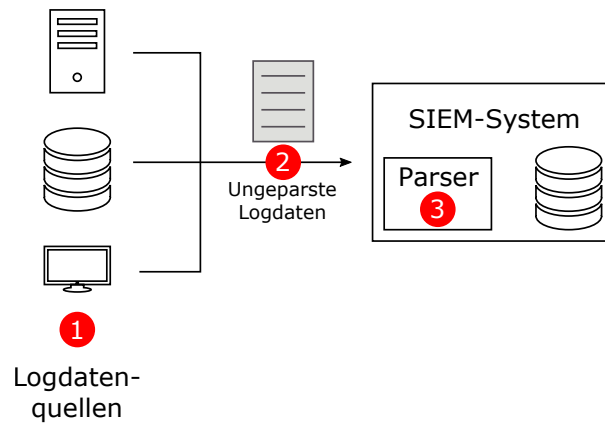


Abbildung 3.2: Mögliche Eingriffspunkte in den Datenfluss eines SIEM-Systems.

1. In der Quelle der Logdaten

Bei diesem Ansatz werden die Daten bereits pseudonymisiert, bevor sie die Datenquelle verlassen. Dieser Ansatz sorgt dafür, dass die Daten bereits pseudonymisiert auf der Übertragungsstrecke und im SIEM-System vorliegen. Es ist kein mehrfaches Parsen der Daten notwendig und der Ansatz ist unabhängig vom verwendeten SIEM-System zu realisieren. Auf der anderen Seite macht der Ansatz die universelle Veränderung jeder Datenquelle notwendig. Dies kann bei Datenquellen, die auf ähnlichen gut erweiterbaren Plattformen beruhen, relativ einfach umzusetzen sein. Beispielsweise könnte die im nächsten Ansatz vorgestellte Proxy-Komponente lokal auf der Datenquelle eingesetzt werden. Schwierigkeiten würde dieser Ansatz hingegen bei Datenquellen bereithalten, die beispielsweise aus Gründen abgespeckter zugrundeliegender Betriebssysteme oder geringer Rechenleistung nur schwer erweiterbar sind. Außerdem würde er in vielen Fällen die Kooperation des Herstellers voraussetzen, wenn es sich um nicht quelloffene „Box“-Lösungen handelt.

2. Proxy-basierter Ansatz

Dieser Ansatz pseudonymisiert die Daten vor dem ersten Kontakt mit dem SIEM-System, indem Datenquellen ihre Logdaten an einen Proxy senden, der die Daten pseudonymisiert und erst anschließend an das SIEM-System weiterreicht. Hierdurch wird erreicht, dass die Daten zu keiner Zeit nicht-pseudonymisiert in dem SIEM-System vorliegen. Außerdem ist er unabhängig von Datenquellen und SIEM-System und erfordert keinen direkten Eingriff in diese (abgesehen von geringen Konfigurationsanpassungen). Ein Nachteil dieser Lösung ist, dass sie das Parsen und Neuzusammensetzen der Logdaten im Proxy zusätzlich zu deren anschließender Behandlung im SIEM-System erfordert. Außerdem müssen für verschiedene Arten der Logdatenübermittlung (Protokolle wie Syslog oder SNMP) unterschiedliche Proxys entwickelt werden.

3. Patchen des SIEM-Systems

Die letzte Möglichkeit ist das Verändern des SIEM-Systems selbst. Hierzu wird in die Logdaten parsende Komponente eingegriffen, um vor, während oder nach diesem Vorgang die Logdaten zu pseudonymisieren. Dieser Ansatz erfordert kein mehrfaches Bearbeiten von Logdaten wie in dem Proxy-basierten Ansatz. Auf der anderen Seite ist er abhängig vom eingesetzten SIEM-System und erfordert seine Veränderung. Zusätzlich liegen die Daten erst einmal in nicht veränderter Form im SIEM-System vor, was die in Abschnitt 3.1.1 erwähnten Nachteile mit sich bringt.

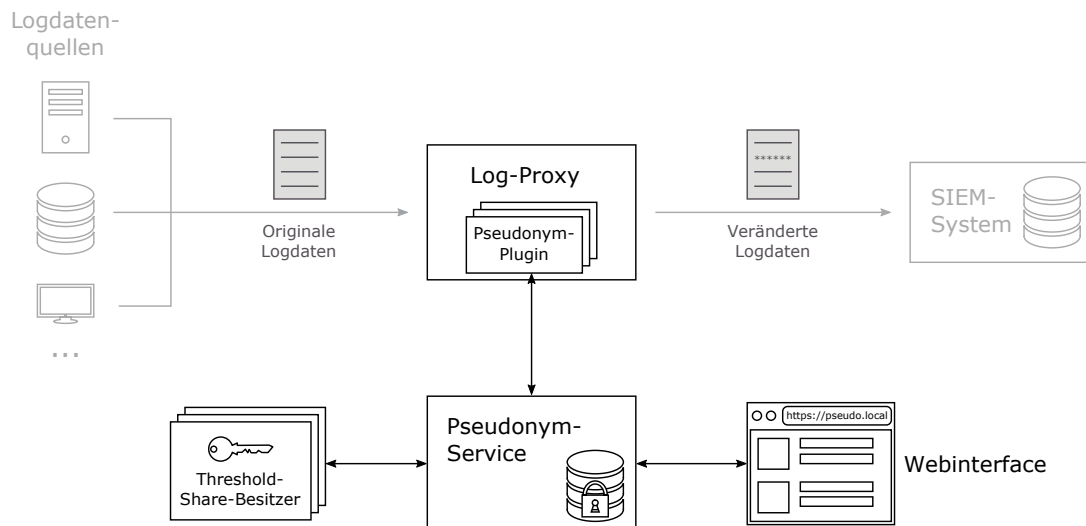


Abbildung 3.3: Ein Überblick über die entworfene Architektur.

Aus datenschutztechnischer Sicht ist eine frühestmögliche Pseudonymisierung zu bevorzugen, wie sie auch in [SW17] empfohlen wird: „Die Pseudonymisierung ist im Verarbeitungsprozess so früh wie möglich durchzuführen.“ Daher wäre eine Pseudonymisierung bereits in der Datenquelle der Optimalfall. Demgegenüber stehen jedoch die erwähnten Nachteile des ersten Ansatzes im Bezug auf die Umsetzbarkeit, da hierzu jede mögliche Quelle von Logdaten universell verändert werden müsste. Eine erst im SIEM-System stattfindende Pseudonymisierung bringt jedoch die beschriebenen Risiken im Bezug auf das Vorliegen des pseudonymisierten Daten im Originalformat mit.

Dies ließ die Entscheidung auf den Proxy-basierten Ansatz fallen. Dass die Lösung außerdem noch keine Anpassungen an dem SIEM-System selbst erfordert, wiegt den Nachteil des zusätzlichen Parsens und wieder Zusammensetzens der Lognachricht bei Weitem auf.

Auch auf Angreifermodell beziehen?

3.2.2 Architektur

Ausgehend von diesen Überlegungen wurde ein System entworfen, dass die Anforderungen aus Abschnitt 3.1 erfüllt und an der beschriebenen Stelle in den Datenfluss eingreift.

Bei dem Entwurf handelt es sich um ein verteiltes System, bei der die Verarbeitung der Logdaten und die Speicherung der Pseudonymzuordnung an unterschiedlichen Stellen geschieht. Hierfür sprechen verschiedene Gründe. Die Kompromittierung der speichernden Komponente schützt die erstellten Pseudonyme vor Aufdeckung durch die Verschlüsselung der Datensätze mit einem kryptographischen Schwellwertschema. Die Kompromittierung der verarbeitenden Komponente lässt zwar eine Verknüpfung neu erstellter Pseudonyme mit eintreffenden Daten zu, sorgt aber nicht für eine Aufdeckung bereits erstellter Pseudonyme, da diese in der anderen Komponente vorliegen. Weiterhin sorgt dieser Ansatz auch für eine zusätzliche Erweiterbarkeit des Systems. Eine speichernde Komponente kann so als Datenspeicher für mehrere verarbeitende Komponenten agieren, was beispielsweise die Erweiterung um zusätzliche Protokolle (vgl. Abschnitt 5.1) oder Pseudonyme über verschiedene Datenarten (vgl. Abschnitt 4.4.6) ermöglicht. Einen Überblick über den Entwurf bietet Abbildung 3.3. Die verschiedenen Komponenten des Systems werden im Folgenden näher beschrieben.

Die erste Komponente ist der **Log-Proxy**, der die Daten entgegennimmt, verändert und anschließend an das SIEM-System 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 3.1.6 ermöglicht. Der Proxy leistet die Behandlung von Logdaten aus verschiedenen Quellen wie in Abschnitt 3.1.5 beschrieben. Das für diese Art des Dateneingriffs erforderliche Parsen und Wiederzusammensetzen der Daten muss hier Datenquellen-abhängig konfigurierbar geschehen.

Ein in dem Proxy enthaltenes Plugin ist 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 der Logdatenverarbeitung verwenden.

Der **Pseudonym-Service** erfüllt zwei Aufgaben: das Speichern und Verwalten der Pseudonyme sowie die Integrierung des kryptographischen Schwellwertschemas. Initial muss die Schlüsselgenerierung des Schwellwertschemas durch den Service geleistet werden. Dies kann wie bereits im vorhergehenden Abschnitt beschrieben zentral oder verteilt geschehen.

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 (dieses Problem wird in Abschnitt 4.4 genauer dargestellt). Über ein Webinterface kann ein berechtigter Benutzer die Aufdeckung eines bestimmten Pseudonyms fordern und den Status seiner Forderung bzw. im Erfolgsfall das aufgedeckte Datum betrachten. Dieses Datum wird durch das Kombinieren der partiellen Entschlüsselungen erhalten, die von den entsprechenden *Share*-Besitzern berechnet werden. Weiterhin kann über dieses Webinterface auch die initiale Konfiguration des Systems im Bezug auf Eigenschaften des Pseudonymisierung und des kryptographischen Schwellwertschemas vorgenommen werden.

Benutzer, die über das Webinterface oder den Webservice mit dem Pseudonym-Service agieren möchten, werden durch geeignete Maßnahmen authentifiziert und ihre Autorisierung überprüft.

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 Anwendung verwaltet den *Share* des Benutzers für das kryptographische Schwellwertschema und kann nach der Bestätigung des Benutzers zu der Aufdeckung eines Pseudonyms die partielle Entschlüsselung eines Datensatzes erstellen und an den Pseudonym-Service senden.

Für alle Übertragungsstrecken wird angenommen, dass ein Angreifer keinen Zugriff auf Kommunikationsinhalte erhält oder diese verändern kann. Dies ist durch Transportverschlüsselung mittels des weitverbreiteten TLS-Protokolls zu erreichen, muss jedoch bei der Verwendung des Systems beachtet werden.

3.3 Angreifermodell

Das Ziel des Systems bezogen auf seine Sicherheit lässt sich folgendermaßen definieren: Das Pseudonym eines Nutzers erlaubt (ohne Anwendung von Hintergrundwissen) keinen Rück-

schluss auf die Identität eines Nutzers. Erst die Kollaboration berechtigter Akteure ermöglicht das Aufdecken eines Pseudonyms. Aus diesem Grund soll sich auch das nachfolgend aufgestellte Angreifermodell auf dieses Ziel fokussieren. Andere Angriffsarten beispielsweise auf die Verfügbarkeit des Systems werden daher nicht betrachtet.

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/passives und beobachtendes/veränderndes Verhalten und die Rechenkapazität, die der Angreifer zum Überwinden der eingesetzten Schutzmaßnahmen aufbringen kann. [BFP14]

Bezogen auf die verfügbare Rechenleistung des Angreifers sollen verbreitete und nach heutigem Wissensstand für sicher befundene kryptographische Algorithmen als nicht mit vertretbarem Aufwand zu brechen angesehen werden. Es handelt sich daher um die Annahme von Komplexitätstheoretischer Sicherheit. Das in Abschnitt 2.6.5 erwähnte Diskrete-Logarithmus-Problem wird also beispielsweise für ausreichend große Primzahlen als nicht mit vertretbarem Aufwand zu brechen angesehen.

Im Bezug auf die Verbreitung eines Angreifers 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 Schutz ihrer Integrität über das Netzwerk. Hierdurch könnte ein Angreifer bereits vor dem Eintreffen der Daten im Proxy passiv alle Daten mitlesen und die anschließend stattfindenden Pseudonymisierung würde nichts an dem gewonnenen Wissen verändern können. Es ist kein Ziel dieser Arbeit, die sich mit der datenschutzfreundlichen **Speicherung** von Überwachungsdaten beschäftigt, diese Angriffsmöglichkeiten zu verhindern. Daher werden bei der Verbreitung eines Angreifers die Datenquellen und Übertragungswege zum Log-Proxy nicht betrachtet.

Mit dieser Einschränkung ergeben sich verschiedene Rollen und darauf basierend Verbreitungen, die ein Angreifer annehmen kann:

Außenstehender Als Außenstehender wird hier jeder Akteur verstanden, der keinen legitimen Zugriff auf Teile des Systems besitzt. Gemeint sind also genauso Mitarbeiter in einem Unternehmensnetzwerk, in dem das System genutzt wird, wie externe Angreifer. Sind die Zugriffsmechanismen im Pseudonym-Service korrekt umgesetzt, bleibt Außenstehenden nur das passive Beobachten von Nachrichten. Durch den Einsatz der in Abschnitt 3.2.2 erwähnten notwendigen Transportverschlüsselung erfahren passive Angreifer jedoch keine brauchbaren Informationen. Diese verhindert auch das Verändern der Nachrichten auf der Übertragungsstrecke durch aktive Angreifer.

Benutzer mit SIEM-Zugriff Ein Benutzer, der Zugriff auf das SIEM-System besitzt, sieht nur die pseudonymisiert gespeicherten Logdaten. Hieraus erfährt er erst einmal nichts über den Benutzer hinter dem Pseudonym. Durch die Anwendung von Hintergrundwissen und die Verknüpfung von Datenbankeinträgen kann er bestimmte Pseudonyme eventuell aufdecken – ein Angriff, der nicht zu verhindern ist, durch regelmäßige Pseudonymwechsel jedoch zumindest in seiner Reichweite beschränkt werden kann.

Benutzer mit Recht auf Pseudonymaufdeckung Legitimierte Benutzer können Anfragen zur Aufdeckung eines Pseudonyms stellen. Durch den Einsatz des kryptographischen Schwellwertschemas führt jedoch erst die Kollaboration von ausreichend Share-Besitzern zur wirklichen Aufdeckung. Vorher erfährt der Benutzer nichts über den Pseudonymhalter.

Share-Besitzer Besitzer eines Shares erhalten für aufzudeckende Pseudonymzuordnungen die verschlüsselten Daten und berechnen aus ihrem Share und dem Schlüsseltext eine partielle Entschlüsselung. Ausgehend von den Eigenschaften des kryptographischen Schwellwertschemas erfahren Sie hieraus jedoch nichts über das verschlüsselte Datum, solange nicht eine Mindestanzahl t an partiellen Entschlüsselungen vorliegt. Erst eine Kollaboration von mindestens t als aktive Angreifer handelnden Share-Besitzern kann so unberechtigt Pseudonyme aufdecken, wenn sie zusätzlich Zugriff auf die Datenbank des Pseudonym-Service erlangen.

Eine Kollaboration von mindestens $n - t + 1$ böartigen Share-Besitzern könnte auch dazu führen, dass die Aufdeckung eines Pseudonyms durch das Senden fehlerhafter partieller Entschlüsselungen fehlschlägt. Die sinnvollen Aufteilung der Shares und damit die Modellierung von verteiltem Vertrauen spielt also in dem System eine wichtige Rolle. Zusätzlich könnten im Falle der verteilten Schlüsselgenerierung böartige Share-Besitzer beispielsweise durch das Senden falscher Daten versuchen, die Generierung zu stören und damit das anschließende Aufdecken von Pseudonymen zu verhindern. Dieser Angriff muss bei der Schlüsselgenerierung durch das verwendete Schema verhindert werden.

Administrator mit Proxy-Zugriff Der Zugriff auf den Proxy, an dem die Logdaten im Klartext eintreffen und pseudonymisiert werden, erlaubt die Zuordnung von Pseudonymen zu Benutzern. Hier muss der Zugriff nach der Initialisierung eingeschränkt werden und soweit wie möglich sichergestellt werden, dass auch zugriffsberechtigte Benutzer den Zugriff nicht ausnutzen (z.B. durch Mehraugenprinzip oder zumindest Protokollierung der Handlungen).

Administrator mit Pseudonym-Service-Zugriff Der Administrator des Pseudonym-Service hat Zugriff auf die Datenbank der verschlüsselten Pseudonymzuordnungen. Durch den Einsatz des Schwellwertschemas und die Verschlüsselung der Logdaten-Pseudonym-Zuordnung bereits im Proxy erfährt er (kein Hintergrundwissen vorausgesetzt) nichts über die Pseudonymhalter. Besondere Bedeutung kommt der Schlüsselgenerierung im Pseudonym-Service im Falle der zentralen Schlüsselgenerierung zu. Gelingt es dem Administrator während der Schlüsselgenerierung in den Besitz des temporär erzeugten geheimen Schlüssels oder mindestens t Shares zu kommen, so kann er jederzeit in der Datenbank abgelegte Pseudonymzuordnungen entschlüsseln, ohne dass andere Benutzer dies mitbekommen. Aus diesem Grund ist die verteilte Schlüsselgenerierung zu bevorzugen.

4 Stand der Wissenschaft und Auswahl von Verfahren

In diesem Kapitel soll der aktuelle Stand der Forschung und Entwicklung bezogen auf die in dieser Arbeit verwendeten Klassen von Systemen bzw. Verfahren betrachtet sowie ausgehend von den Anforderungen an das zu entwickelnde System passende Lösungen ausgewählt und in für folgende Kapitel notwendiger Detailtiefe beschrieben werden.

Der erste Abschnitt befasst sich mit zum Zeitpunkt der Erstellung dieser Arbeit verfügbaren SIEM-Systemen und beschreibt Eigenschaften des ausgewählten Systems näher, die für die Integration des in dieser Arbeit zu entwickelnden Prototypen relevant sind.

Im zweiten Abschnitt werden Eigenschaften von Pseudonymen und Anforderungen an die Verwendung von Pseudonymisierung herausgearbeitet, die bei der Entwicklung eines Systems beachtet werden müssen.

Der dritte Abschnitt stellt den Forschungsstand im Bereich der kryptographischen Schwellwert-schemata dar und beschreibt das ausgewählte Verfahren im Detail.

Der letzte Abschnitt befasst sich mit verschiedenen Ansätzen zur Lösung des in Abschnitt 3.2.2 erwähnten Problems des Suchens und Wiederverwendens bereits vergebener Pseudonyme für ein eintreffendes Datum. Basierend auf Vor- und Nachteilen der verschiedenen Lösungen wird eine Auswahl für ein zu implementierendes Verfahren getroffen.

4.1 SIEM-Systeme

Zur Zeit gibt es eine vielfältige Auswahl an SIEM-Systemen auf dem Markt: Splunk¹, QRadar von IBM² oder ArcSight von Micro Focus³ sind nur einige Beispiele aus diesem Bereich. Neben den in Abschnitt 2.3 beschriebenen grundlegenden Funktionen eines SIEM-Systems, die von allen Kandidaten in unterschiedlichem Maße bereitgestellt werden, unterscheiden sie sich insbesondere in darüber hinausgehenden Techniken: Die Nutzung von Machine Learning zur Erkennung ungewöhnlichem Verhaltens oder die Automatisierung von Handlungen im Bedrohungsfall sind hier beispielhafte Möglichkeiten.

Die Auswahl an quelloffener Software in diesem Bereich ist jedoch sehr gering. Eine Ausnahme 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

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>

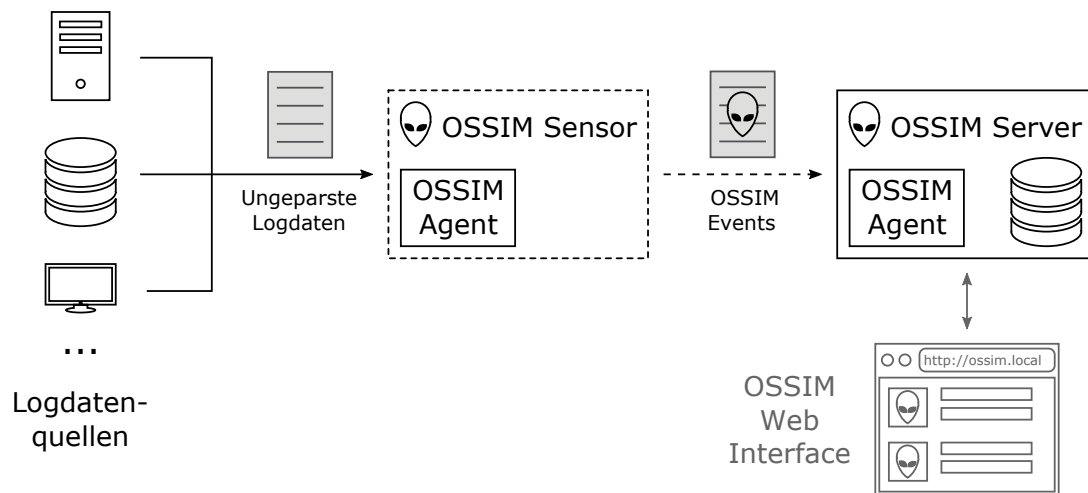


Abbildung 4.1: High-Level-Übersicht über die OSSIM-Architektur und den Datenfluss.

die in USM fließt, profitiert jedoch auch OSSIM, beispielsweise durch die Aktualisierung von Plugins für die Einbindung von aktuellen Netzwerkgeräten.

Da die Quelloffenheit auch die Möglichkeit bietet, Komponenten des SIEM-Systems direkt zu verändern, und außerdem gerade im Sicherheitsbereich generell zu bevorzugen ist, fiel die Entscheidung des in dieser Arbeit verwendeten SIEM-Systems auf OSSIM.

4.1.1 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 4.1 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 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.

4.1.2 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.

4.2 Pseudonymisierung

Der Begriff der Pseudonymisierung beschreibt die Benutzung von Pseudonymen zur Identifizierung von Subjekten. Ein Pseudonym (im technischen Sinne) kann nach [PH10] als einfache Bitkette betrachtet werden. Es sollte zufällig generiert werden, d.h. vollkommen unabhängig von dem zugehörigen Subjekt oder es betreffenden Eigenschaften sein, um keine Rückschlüsse aus dem Pseudonym selbst zu ermöglichen. Ein Negativbeispiel wäre ein nutzervergebenes Pseudonym, das den Namen des Haustiers enthält. Aber beispielsweise auch eine aufsteigende Nummerierung als Pseudonym könnte durch den hierdurch genauer spezifizierten Erstellungszeitpunkt Rückschlüsse auf das Subjekt hinter dem Pseudonym ermöglichen.

Pseudonymisierung sagt 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 nach [PK01] 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

Um die Auswirkungen dieser Eigenschaften einordnen zu können, soll im Folgenden kurz die Pseudonymisierung in zwei Systemen betrachtet und die Relevanz der eben genannten Eigenschaften verdeutlicht werden: Pseudonyme in Mobilfunknetzen und in der Fahrzeug-zu-Fahrzeug-Kommunikation.

4.2.1 Mobilfunknetze

In Mobilfunknetzen wird zur Identifikation eines Teilnehmers anstelle seiner identifizierenden *International Mobile Subscriber Identity* meist ein Pseudonym – die *Temporary Mobile Subscriber Identity* (TMSI) – genutzt, das in bestimmten Situationen gewechselt wird und so Ortung und Bewegungsprofile der Teilnehmer verhindern soll. In [Ara+14] beschreiben die Autoren Schwächen der Implementierungen von Mobilfunkstandards in Netzen bei der (Neu-)Vergabe einer TMSI. Bestimmte Eigenschaften im Bezug auf die Unverkettbarkeit von Pseudonymen und damit auf die Privatsphäre der Nutzer werden in vielen Netzen aufgrund einiger Schwächen nicht erreicht:

- Zu seltene Änderung der Pseudonyme
- Keine nutzungsabhängige Änderung von Pseudonymen
- Pseudonyme werden in verschiedenen Bereichen beibehalten
- Anfälligkeit der Neuvergabe für Replay-Angriffe

Die letzten beiden Schwächen sind für den Anwendungskontext dieser Arbeit nicht relevant, aber die zeit- und aktivitätsabhängige Neuvergabe von Pseudonymen müssen auch hier beachtet und umgesetzt werden.

4.2.2 VANets

Ein anderer Bereich, der sich besonders mit der Nutzung von Pseudonymen beschäftigt hat, ist die Forschung an Vehicular Ad Hoc Networks (VANets). Hierbei handelt es sich um Netzwerke für die Kommunikation zwischen Fahrzeugen, die beispielsweise für die Datenübermittlung zur Bremserkennung naher Fahrzeuge oder die Stauerkennung genutzt werden können. Um die Privatsphäre der Fahrzeughalter zu schützen, wird für die Kommunikation in vielen Ansätzen auf die Verwendung von Pseudonymen gesetzt. So soll sich beispielsweise das Anlegen von Bewegungsprofilen verhindern lassen.

Unter anderem in [Dö05] und [Pet+15] widmen sich die Autoren der Nutzung von Pseudonymen in VANets und den besonderen Anforderungen, die diese erfüllen müssen – insbesondere auch im Hinblick auf die Häufigkeit von Pseudonymwechseln. Es ergibt sich, dass die Häufigkeit und Situation⁵, in der Pseudonymwechsel stattfinden sollten, abhängig vom gewünschten Grad an Anonymität bzw. Angreifermodell sind und außerdem gegenüber Sicherheitsanwendungen⁶ abgewogen werden müssen.

5. Es werden beispielsweise Lösungen vorgeschlagen, die abhängig von Geschwindigkeitsänderungen, einer gewissen Anzahl anderer Fahrzeuge oder besonderen Verkehrssituationen wie Kreuzungen die Pseudonymänderung vornehmen. Das Ziel ist hier immer die Möglichkeit der Pseudonymverkettung bzw. der Bewegungsprofilerstellung durch die äußere Situation der Pseudonymänderung zu erschweren.

6. Beispielsweise wäre zur VANet-basierten Kollisionsvermeidung eine Verkettung von Orten, an denen sich ein Fahrzeug zu verschiedenen Zeitpunkten befindet, erstrebenswert.

Bei der Nutzung von Pseudonymen in VANets handelt es sich natürlich um eine Anwendung mit anders gelagerten Prioritäten im Vergleich zu dem Kontext dieser Arbeit. Dennoch wird deutlich, dass die Strategie zum Pseudonymwechsel stark von der Anwendungssituation abhängig ist. Bezogen auf den hier vorliegenden Anwendungsfall werden insbesondere Besonderheiten der Datenquelle, wie die Häufigkeit von auftretenden Überwachungsdaten, und Anforderungen an die Verknüpfbarkeit von Ereignissen der auf den Daten beruhenden Anomalieerkennung zu beachten sein.

In [SMK09] stellt der Autor eine weitere Anforderung an die Nutzung von Pseudonymen in VANets, die jedoch nicht nur für diesen speziellen Anwendungsfall relevant ist: Er verlangt, dass die Aufdeckung eines Pseudonyms keine Informationen über die Identität eines Nutzers im Bezug auf andere Pseudonyme ermöglichen sollte. Diese Eigenschaft bezeichnet er als *Perfect Forward Privacy*⁷.

4.2.3 Pseudonymisierung im zu entwickelnden System

Aus diesen Vorüberlegungen können nun die Rahmenbedingungen der in dieser Arbeit verwendeten Pseudonymisierung aufgestellt werden. Pseudonyme sollten als zufällig gewählte Bitketten hinreichender Länge gewählt werden. Ihre Eindeutigkeit muss sichergestellt werden.

Wie auch in den Beispielen deutlich wurde, müssen Pseudonyme abhängig von dem Anwendungsszenario in bestimmten Fällen für einen Benutzer gewechselt werden. In dem hier vorliegenden Anwendungsfall, in dem Pseudonyme für die Zuordnung von eintreffenden Überwachungsdaten in Unternehmensnetzen genutzt werden, sind insbesondere die Zeitabhängigkeit sowie die Abhängigkeit von der Nutzungshäufigkeit für die Pseudonymwechselstrategie ausschlaggebend. Verschiedene Nutzeraktionen sollten nur in einem gewissen zeitlichen Rahmen und nur in einer gewissen Häufigkeit verkettbar sein. Es handelt sich also um eine schwächere Form der Transaktionspseudonyme, bei der ein Pseudonym je nach Pseudonymwechselstrategie nur für eine bestimmte Anzahl an Ereignissen verwendet wird.

Eine über diese generelle Aussage hinausgehende Bewertung davon, wie diese Abhängigkeiten konkret zu implementieren sind, ist jedoch im Rahmen dieser Arbeit nicht zu leisten. Hierfür sind zwei Gründe ausschlaggebend:

- Sie hängen stark von den Eigenschaften der Datenquellen ab, die die Überwachungsdaten liefern. Beispielsweise wäre das Datenprofil, das von einem elektrischen Türschließsystem geliefert wird, sehr unterschiedlich zu dem, das Zugriffe auf einen Netzwerkspeicher protokolliert. Im ersten Fall würden im Allgemeinen selten Daten anfallen, die zudem durch die Anwendung von Hintergrundwissen (Benutzer wird beim Betreten eines Raumes beobachtet) eher zur Aufdeckung eines Pseudonyms führen könnten. Hier wären wahrscheinlich häufige nutzungsabhängige Wechsel angebracht. Eventuell wäre sogar der Extremfall einer einmaligen Pseudonymvergabe pro Aktion in Erwägung zu ziehen. Im zweiten Fall hingegen würden im Allgemeinen häufig Daten anfallen und erst die Verkettung dieser Daten könnte hilfreiche Rückschlüsse auf vorliegende Anomalien liefern.

7. Die Bezeichnung ist an *Perfect Forward Secrecy* angelehnt. Diese Eigenschaft beschreibt ein ähnliches Verhalten bei der verschlüsselten Kommunikation: Ein Angreifer, der in den Besitz des Langzeitschlüssels eines Kommunikationspartners kommt, sollte trotzdem nicht in der Lage sein, bereits aufgezeichnete Nachrichten entschlüsseln zu können.

Ein einzelner Datenzugriff hätte meist wenig Aussagekraft, wohingegen ein massenhafter Zugriff beispielsweise auf die Kundendatenbank eines Unternehmens durch einen gekündigten Mitarbeiter möglicherweise auf Datendiebstahl schließen lassen könnte.

- Weiterhin muss die Pseudonymwechselstrategie auch abhängig von der später auf den pseudonymisierten Überwachungsdaten auszuführenden automatisierten Anomalieerkennung sein. Je nachdem welche Verfahren auf Daten aus welchen Datenquellen eingesetzt werden sollen, könnte auch hier unterschiedliche Verknüpfbarkeit der Daten erforderlich sein. Hieraus ergibt sich auch ein Spannungsfeld zwischen den Anforderungen der Anomalieerkennung gegenüber der Verknüpfbarkeit der Daten und damit der Privatsphäre der Arbeitnehmer.

Aus diesen Gründen wird eine parameterabhängige Pseudonymwechselstrategie implementiert, die sowohl zeit- als auch die nutzungsabhängige Wechsel ermöglicht. Wie lange bzw. häufig ein Pseudonym verwendet wird, kann so in konkreten Anwendungen mit gesetzten Rahmenbedingungen beurteilt und gesetzt werden. Dieses Vorgehen wird auch in den *Leitlinien für die rechtssichere Nutzung von Pseudonymisierungslösungen unter Berücksichtigung der Datenschutz-Grundverordnung* beschrieben: „Abhängig vom Anwendungsfall sind – zeit- oder datenvolumenabhängig – geeignete Intervalle zu definieren, in denen ein Wechsel [...] erfolgt.“[SW17]

Weiterhin wird angestrebt für die Pseudonyme bzw. ihre Aufdeckung die erwähnte Perfect Forward Privacy zu ermöglichen. Die konkrete Umsetzung dieser Eigenschaft wird in einem späteren Abschnitt beschrieben werden.

4.3 Schwellwertschemata

Aufbauend auf den Ideen von Shamir und Blakley und den ersten Ideen zu kryptographischen Schwellwertschemata wurden für verschiedene Algorithmen und Anwendungsfälle Schemata mit unterschiedlichen Eigenschaften entwickelt.

4.3.1 Übersicht

Eine Vielzahl von Veröffentlichungen behandeln das Problem der verteilten Erstellung von Signaturen: Die in [Sho00] entwickelte Lösung basiert auf dem RSA-Verfahren, [Gen+96b] erweitert den DSS-Standard um ein Schwellwertschema und [SS01] entwickelt ein Schema zur verteilten Signatur mittels Schnorr-Signaturen.

Weitere Forschungen haben sich mit der Entwicklung von RSA-basierten Schwellwertschemata zur verteilten Entschlüsselung beschäftigt, die im Kontext dieser Arbeit genutzt werden [Fra+97; Gen+96a; Rab98].

Ein zusätzliches Verfahren, das im Zusammenhang mit verteilter Entschlüsselung Aufmerksamkeit erfuhr, ist das Paillier-Kryptosystem. In [DJ01] und [FPS00] entwickelten die Autoren auf diesem System basierte Schwellwertschemata, die insbesondere durch ihre homomorphe Eigenschaft hervorstechen und dadurch im Bereich der elektronischen Wahlsysteme genutzt werden können.

Einen Überblick über weitere Veröffentlichungen in diesem Bereich bieten beispielsweise [Des97], [Gem97] und [Des93].

4.3.2 ElGamal-basiertes Schwellwertschema

Ein Verfahren zur *Threshold Decryption*, das auf einer geschickten Kombination von Shamir's Secret Sharing (Abschnitt 2.5.1) und des ElGamal-Kryptosystems (Abschnitt 2.6.5) basiert, veröffentlichten die Autoren in [DF90]. Aufbereitete Darstellungen lassen sich in [KL14] und [BS16] finden.

Es ist eines der ersten veröffentlichten Schwellwertschemas und erfuhr dadurch viel Beachtung und entsprechend viele aufbauende Arbeiten, die Verbesserungen vorschlugen. Durch die hinterliegende Mathematik bietet das Schema einfachere Umsetzbarkeit (auch von Erweiterungen wie dezentraler Schlüsselgenerierung) gegenüber RSA⁸. Dies gilt ebenso gegenüber den Paillier-basierten Schemata – deren homomorphe Eigenschaften in dieser Arbeit nicht benötigt werden. Aus diesen Gründen fiel die Wahl des in dieser Arbeit umzusetzenden Schemas auf das genannte Verfahren.

Der Rest dieses Abschnitts stellt das Verfahren nun entsprechend den in Abschnitt 2.5.2 aufgeführten Algorithmen eines Threshold-Public-Key-Decryption-Systems im Detail vor.

Algorithmus G: Schlüsselgenerierung

In dem Verfahren wird für die Schlüsselgenerierung eine zentrale, vertrauenswürdige Instanz vorausgesetzt, die den öffentlichen Schlüssel und die später benötigten Shares des geheimen Schlüssels erzeugt und verteilt.

Zur Erzeugung werden zwei Primzahlen p und q mit der Eigenschaft $p = 2q + 1$ - bekannt als sichere Primzahl bzw. Sophie-Germain-Primzahl - benötigt. Weiterhin ist ein Generator der Untergruppe der Ordnung q von \mathbb{Z}_p^* notwendig.

Der (temporär erstellte) geheime Schlüssel $a \in \mathbb{Z}_q$ wird analog zu der Schlüsselgenerierung im ElGamal-Verfahren zufällig gewählt. Aus ihm wird der öffentliche Schlüssel $pk = g^a \bmod p$ berechnet.

Der geheime Schlüssel wird anschließend analog zu Shamirs Secret Sharing in \mathbb{Z}_q in einzelne Shares $(x_i, y_i) = (x_i, q(x_i))$ aufgeteilt und diese an die Teilnehmer verteilt. Anschließend werden diese Werte gelöscht, so dass nur noch die Teilnehmer im Besitz ihrer Shares und damit in der Lage sind, Schlüsseltexte zu entschlüsseln.

Algorithmus E: Verschlüsselung

Anschließend kann ein Klartext mithilfe von pk analog zu dem ElGamal-Verfahren verschlüsselt werden. So erhält man $(v, c) = (g^k, m \cdot g^{ak})$ für ein durch den Sender zufällig gewähltes $k \in \mathbb{Z}_q$.

Algorithmus D: Partielle Entschlüsselung

Jeder Besitzer eines Shares (x_i, y_i) kann nun für den zu entschlüsselnden Schlüsseltext (v, c) seine partielle Entschlüsselung (x_i, v^{y_i}) berechnen und diese an eine zentrale Instanz, den Combiner, senden. Empfängt dieser mindestens t partielle Entschlüsselungen⁹, so kann er den Klartext wiederherstellen.

Algorithmus C: Kombination

8. Das ElGamal-Verfahren nutzt zur Berechnung eine öffentlich bekannter Ordnung (sie ist Teil des öffentlichen Schlüssels). Im Gegensatz dazu werden Berechnungen bei RSA in $\Phi(n)$ ausgeführt, das jedoch nicht öffentlich vorliegen darf [Ngu05].

9. Zur Erinnerung: t beschreibt die Mindestzahl zur Entschlüsselung benötigter Shares des Schwellwertschemas.

Hierzu berechnet der Combiner die Lagrange-Koeffizienten $\lambda_i \in \mathbb{Z}_q$ wie in Shamir's Secret Sharing beschrieben¹⁰. Anschließend kann

$$g^{ak} = \prod_{i=1}^k (v^{y_i})^{\lambda_i}$$

berechnet werden. Dies funktioniert, da

$$\prod_{i=1}^k (v^{y_i})^{\lambda_i} = \prod_{i=1}^k (g^k)^{y_i \cdot \lambda_i} = (g^k)^{\sum_{i=1}^k y_i \cdot \lambda_i} \stackrel{(*)}{=} (g^k)^a$$

gilt. Der letzte Schritt (*) folgt direkt aus dem zugrundeliegenden Secret-Sharing-Schema und ist in dieser Form bereits in Abschnitt 2.5.1 zu finden.

Anschließend kann der Klartext als $m = c \cdot (g^{ak})^{(-1)}$ wiederhergestellt werden.

4.3.3 Verteilte Schlüsselgenerierung

Ein Nachteil dieses Verfahrens in der Phase der Schlüsselgenerierung ist, dass für die Generierung des geheimen Schlüssels und der daraus resultierenden Shares eine zentrale und vertrauenswürdige Instanz notwendig ist. Diese Problematik wurde bereits in Abschnitt 3.1.3 dargestellt und Auswirkungen in Abschnitt 3.3 betrachtet.

In [Ped91] wurde vom Autor eine Möglichkeit der verteilten Schlüsselgenerierung für das dargestellte Verfahren vorgeschlagen, die von den Autoren in [Gen+99] noch verbessert wurde.

Das Verfahren besteht aus zwei Phasen: In der ersten Phase wird von allen potentiellen Share-Besitzern ein *Verifiable Secret Sharing Scheme*¹¹ (VSS) nach Pedersen ausgeführt, das dafür sorgt, dass anschließend alle ehrlichen Beteiligten jeweils im Besitz eines Shares sind, die zusammengenommen den geheimen Schlüssel x bilden (der jedoch nirgendwo vorliegt oder im Laufe des Verfahrens vorlag). In der zweiten Phase wird ein VSS nach Feldman dazu genutzt, den gemeinsamen öffentlichen Schlüssel $y = g^x$ auf eine Weise zu berechnen, die wiederum dafür sorgt, dass der geheime Schlüssel nirgendwo vorliegen muss. Auf diese Weise wird die vertrauenswürdige Instanz vermieden und es ist trotzdem sichergestellt, dass die ehrlichen Beteiligten im Besitz von Shares sind, die die Verwendung des kryptographischen Schwellwertschemas so ermöglichen wie im letzten Abschnitt für das Verfahren mit zentraler Schlüsselgenerierung vorgestellt.

10. In diesem Abschnitt gilt $i \in C$. C stellt dabei die Menge der Indizes der beteiligten Sharebesitzer dar. Es gilt also $C \subseteq \{1, \dots, n\}$ und $|C| \geq t$.

11. Verifiable Secret Sharing Schemes sind Secret Sharing Schemes, die es den Share-Besitzern erlauben zu überprüfen, ob ihre Shares konsistent sind, d.h. ob es möglich ist aus den Shares ein gemeinsames Geheimnis wiederherzustellen. Bei dem in Abschnitt 2.5.1 vorgestellten Secret Sharing nach Shamir ist dies beispielsweise nicht der Fall. Ein bössartiger Erzeuger von Shares könnte für jeden Beteiligten ein anderes Geheimnis benutzen, so dass bei der Rekonstruktion abhängig von beteiligten Share-Besitzern unterschiedliche Geheimnisse erhalten werden.

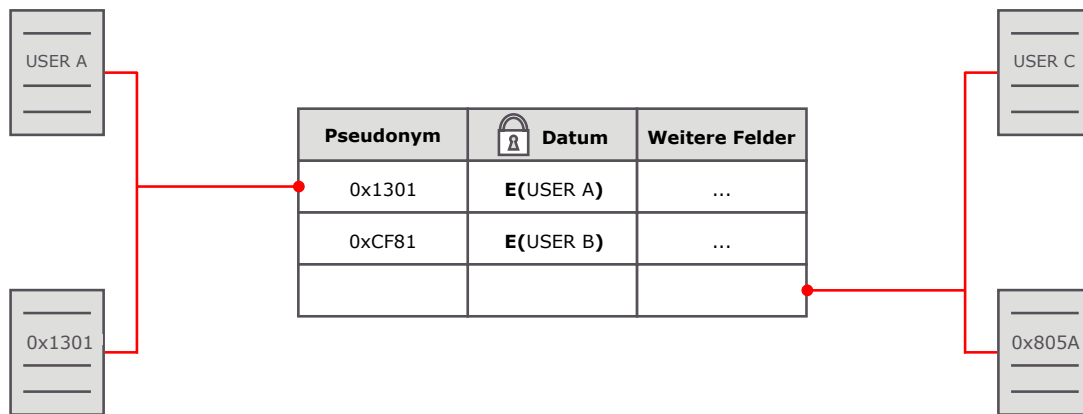


Abbildung 4.2: Erhalt von Pseudonymen aus der Zuordnungstabelle: Links für den Fall eines bereits bekannten Datums (hier Benutzer), rechts für ein unbekanntes Datum.

4.3.4 Existierende Implementierungen

Auch nach umfangreicher Recherche ließ sich keine quelloffene, kryptographisch überprüfte und lizenztlich nutzbare Bibliothek finden, die das gewünschte Schwellwertschema implementiert. Es gab verschiedene verwandte Lösungen wie Civitas¹² oder Helios¹³, die jedoch alle eng mit dem Anwendungskontext der elektronischen Wahl verknüpft waren und dadurch andere Anforderungen erfüllten als sie für diese Arbeit erforderlich sind. Aus diesem Grund wird das beschriebene Schwellwertschema gezwungenermaßen in Teilen selbstständig implementiert.

4.4 Searchable Encryption

Trifft ein neues Datum in dem System ein und soll pseudonymisiert werden, so muss überprüft werden, ob bereits ein Pseudonym für das Datum vergeben wurde. Da die Daten jedoch in verschlüsselter Form vorliegen, stellt sich die Frage, wie diese Überprüfung erreicht werden kann. In Abbildung 4.2 ist dieses Problem noch einmal anhand eines Beispiels dargestellt. Zu einem Zeitpunkt liegen in der Pseudonymtabelle Pseudonyme für zwei Benutzer *User A* und *User B* vor. Da diese Benutzer zu den Pseudonymen gehören, ist jedoch nicht offensichtlich, da die Daten verschlüsselt gespeichert sind. Sollen nun neue Logdaten verarbeitet werden, liegen zwei mögliche Fälle vor:

- Es liegt bereits ein Pseudonym für den Benutzer vor (linker Bereich in der Darstellung). Hier muss das bereits vergebene Pseudonym *0x1301* verwendet werden.
- Es liegt noch kein Pseudonym für den Benutzer vor (rechter Bereich in der Darstellung). Nun wird ein neues Pseudonym *0x805A* angelegt und verwendet werden.

Verschiedene Lösungsmöglichkeiten inklusive ihrer Vor- und Nachteile für das Problem, wie nun trotz der Verschlüsselung der Daten dieses Verhalten erreicht werden kann, werden in diesem Abschnitt beleuchtet.

12. Civitas – A secure voting system. <http://www.cs.cornell.edu/projects/civitas/>

13. Helios Voting. <https://heliosvoting.org/>

4.4.1 Entschlüsseln aller Datensätze

Da für die Verschlüsselung ein kryptographisches Schwellwertschema verwendet wird, scheiden zwei triviale Möglichkeiten aus: Das Entschlüsseln aller Datensätze zur Überprüfung wäre nicht nur unter Performance-Gesichtspunkten nicht wünschenswert. Es darf auch nicht möglich sein, da die *Shares* zur Entschlüsselung nicht am Ort der Verschlüsselung vorliegen dürfen. Dies ist eine der Grundannahmen, die der Sicherheit des Systems zugrundeliegen.

4.4.2 Deterministische Verschlüsselung

Die zweite ausscheidende Möglichkeit wäre das Überprüfen aller Einträge auf Schlüsseltextgleichheit. Bei Gleichheit eines Eintrages könnte das entsprechende Pseudonym zurückgeliefert werden. Hierzu müsste ein deterministisches Verschlüsselungsverfahren genutzt werden, das ein Datum immer auf den gleichen Schlüsseltext abbildet. Diese Möglichkeit scheidet jedoch aus, da es sich bei dem verwendeten Schwellwertschema um ein Public-Key-Verfahren, bei dem bei der Verschlüsselung ein Zufallswert einfließt – folglich ein nicht-deterministisches Verfahren, handelt.

Dieser Nicht-Determinismus ist notwendig, da ansonsten zur Aufdeckung eines Pseudonyms auch ein Wörterbuchangriff mithilfe des öffentlichen Schlüssels des Schwellwertschemas genutzt werden könnte. Ein Angreifer würde alle möglichen Werte, die ein Datum annehmen kann, mit dem öffentlichen Schlüssel verschlüsseln und mit dem aufzudeckenden Eintrag vergleichen. Gleichheit der Schlüsseltexte würde den gesuchten Klartext liefern. Der im Kontext dieser Arbeit vorliegende kleine und bekannte Wertebereich (wie beispielsweise Mitarbeiternamen) würde diesen Angriff relativ effizient machen. Aus diesem Grund muss ein nicht-deterministisches Verschlüsselungsverfahren genutzt werden und damit ist die Überprüfung auf Schlüsseltextgleichheit nicht möglich.

4.4.3 Nutzung von Hashfunktionen

Ein weiterer Ansatz, der ebenfalls anfällig für diese Art von Wörterbuch-Angriff wäre, ist die Verwendung von (kryptographisch sicheren) Hashfunktionen zur Suche: Neben dem Pseudonym und dem verschlüsselten Datum wird ein Hash des Datums abgespeichert. Muss nun für ein neues Datum überprüft werden, ob bereits ein Pseudonym vorliegt, kann der Hash des Datums gebildet und mit allen vorliegenden Hashes verglichen werden. Bei Übereinstimmung wäre das Datum (mit großer Wahrscheinlichkeit) gleich dem verschlüsselten Datum und das Pseudonym könnte genutzt werden. Diese Möglichkeit ist jedoch durch den bereits erwähnten kleinen Wertebereich ebenso anfällig für einen Wörterbuchangriff: Ein Angreifer könnte mithilfe der bekannten Hashfunktion die Hashes aller Werte berechnen und mit dem des aufzudeckenden Pseudonyms vergleichen.

4.4.4 Lokale Zuordnung

Eine andere Möglichkeit ist die Anlage einer vor externem Zugriff geschützten Zuordnungstabelle zwischen Datum und Pseudonym am Ort der Ersetzung. Bei Eintreffen eines neuen Datums kann in der Tabelle das zugehörige Pseudonym ermittelt werden oder falls es noch nicht existiert, ein

neues Pseudonym erstellt und zusammen mit dem verschlüsselten Datum gespeichert werden. Diese Lösung wird auch in [Goh03] erwähnt.

Nachteil bezogen auf das für diese Arbeit zu entwickelnde System ist jedoch die notwendige Generierung von Pseudonymen und Speicherung der Zuordnungstabelle an der Stelle, an der neue Daten eintreffen. Hierdurch wird ein relativ leichtgewichtiger Proxy-Server wie er in der Architektur vorhergesehen ist (siehe Abschnitt 3.2) verhindert. Außerdem würde eine Kompromittierung dieser Komponente direkt zur Aufdeckung des Pseudonymzusammenhangs führen.

Verhindert verteiltes System - diesen Aspekt betonen (mehrere Proxies, ...)

4.4.5 Message Authentication Codes

Aufbauend auf der Hash-basierten Lösung lässt sich jedoch auch eine nicht für einen Wörterbuchangriff anfällige Lösung entwickeln. Dazu wird der verwendete Hash durch einen schlüsselabhängigen MAC (siehe Abschnitt 2.6.2) ersetzt. Beim Speichern eines neuen Eintrags wird dazu unter Zuhilfenahme eines zufällig generierten Schlüssels ein MAC über das Datum berechnet und mit dem Eintrag gespeichert. Für ein Datum kann jetzt durch Überprüfen aller MACs bestimmt werden, ob bereits ein Pseudonym vergeben wurde. Ein Angreifer kann den beschriebenen Wörterbuchangriff jedoch ohne Kenntnis des Schlüssels nicht ausführen.

Bei dieser Lösung handelt es sich um eine einfache Form der Searchable Symmetric Encryption wie sie in Abschnitt 2.7 dargestellt ist. Die durch Pseudonyme zu ersetzenden Daten bilden die zu durchsuchenden Dokumente. Der MAC bildet den Suchwort-Index für jeden verschlüsselten Eintrag und wird so auch als Trapdoor-Element für die Suche nach einem passenden Eintrag genutzt. Ausgehend von den Anforderungen des umzusetzenden Systems eignet sich dieser Ansatz: Er ermöglicht einer Komponente die Zuordnung eines Datums zu einem Pseudonym, ohne dass diese Zuordnung direkt gespeichert werden muss oder der Datenbank bei der Abfrage bekannt wird. Auch ein Wörterbuchangriff, der bei dem erwähnten kleinen Wertebereich geringen Aufwand bedeutete, wird verhindert. Aus diesen Gründen wird dieser Lösungsansatz in einem späteren Schritt im zu entwickelnden System umgesetzt.

Die Nutzung von deterministischer Verschlüsselung (mit der hier beschriebenen Verwendung eines MACs als Sonderfall) wird erstmals in [BBO07] beschrieben. Dort werden auch einige Schwächen dieser Lösung diskutiert: Die Datenbank erfährt durch den Suchindex bereits einiges über die gespeicherten Dokumente, da durch die deterministische Struktur gleiche Suchwörter auf gleiche Trapdoor-Elemente abgebildet werden. Diese Schwäche ist im Bezug auf den besonderen Anwendungsfall dieser Arbeit jedoch zu vernachlässigen, da Dokumente (meint Benutzernamen, ...) nur einmalig vorliegen dürfen. Eine weitere Schwäche, die auch in dieser Arbeit beachtet werden muss, ist, dass die Datenbank etwas über die Häufigkeit verschiedener Suchanfragen erfährt, da die Trapdoor-Elemente für ein bestimmtes Datum immer gleich sind. Durch Korrelation mit Hintergrundwissen könnte so möglicherweise in bestimmten Fällen der Inhaber eines Pseudonyms herausgefunden werden.

Hash-Kollisionen und Geburtstags-Paradoxon hier erwähnen?

4.4.6 Weitere Möglichkeiten der Searchable Encryption

Die im letzten Abschnitt betrachtete MAC-basierte Lösung funktioniert für den in dieser Arbeit behandelten Anwendungsfall. Bei der Abbildung Pseudonym zu Datum (wie den Benutzer-

name) handelt es sich um eine 1:1-Abbildung, die lediglich von einer Komponente – dem zu entwickelnden Log-Proxy – abgefragt wird.

In anderen Umgebungen bzw. Erweiterungen des in dieser Arbeit betrachteten Anwendungsfalls kann es jedoch auch andere Anforderungen geben. Vorstellbar wäre beispielsweise die verteilte Abfrage der Datenbank nach existierenden Pseudonymen. Für den MAC-basierten Ansatz müsste dazu zumindest der genutzte Schlüssel verteilt werden, was Kommunikation innerhalb der verteilten abfragenden Komponenten erfordert. Zusätzlich müssten auch die Sicherheitsauswirkungen dieser Lösung betrachtet werden.

Eine andere Erweiterung könnte die Mehrfachverwendung von Pseudonymen für verschiedene Merkmale eines Benutzers (Name, IP-Adresse, Signaturschlüssel, ...) sein, um die Erkennung von Insiderangriffen zu verbessern. Auch diese Erweiterung wäre mit dem MAC-basierten Ansatz nicht direkt abbildbar.

Andere Lösungsansätze für die in diesen Fällen entstehenden Probleme könnten Forschungsergebnisse aus dem Bereich der *Searchable Encryption* bieten. In [SWP00] wurde von den Autoren das erste Searchable-Symmetric-Encryption-Schema entwickelt (siehe auch Abschnitt 2.7). Verbesserte Schemata folgten in [Goh03] und [CM05]. Durch diese sind variable Dokumentenlängen mit beliebig vielen Suchwörtern möglich, die auch nachträglich erweiterbar sind. Für den Aspekt der verteilten Abfrage könnten die Ergebnisse aus [Bon+04] genutzt werden, in dem sich die Autoren mit der Suche in mit asymmetrischen Verfahren verschlüsselten Daten befassen.

Eine Übersicht über weitere Ergebnisse in diesem Bereich lässt sich in [WWC16] finden.

5 Implementierung

In diesem Kapitel wird die Implementation des Prototypen aufbauend auf dem in Kapitel 3 erstellten Entwurf und den in Kapitel 4 erarbeiteten konkreten Verfahren dargestellt. Neben der Einbindung in das SIEM-System und der Pseudonymisierung wird insbesondere die für die Nutzung eines kryptographischen Schwellwertschemas entwickelte Bibliothek im Detail erörtert. Anschließend erfolgt eine Evaluation des entwickelten Prototypen.

5.1 Einbindung in OSSIM

Wie in Abschnitt 4.1 beschrieben erlaubt OSSIM eine verteilte Installation am Empfang von Logdaten beteiligter Komponenten durch die Einführung optionaler Sensoren, die den Empfang und das Parsen von Logdaten übernehmen. Dadurch bieten sich gegenüber den in Abschnitt 3.2.1 beschriebenen Möglichkeiten zum Eingriff in den Datenfluss eines SIEM-Systems zwei weitere OSSIM-spezifische Möglichkeiten, die folgend als mögliche Alternativen beschrieben werden. Eine Übersicht über den entstehenden Datenfluss bietet Abbildung 5.1. Die Ziffern beschreiben die Stelle des Eingriffs wie in der Abbildung gekennzeichnet:

4. 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 Parsender Logdaten 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 – ein Nachteil der bereits in Abschnitt 4.1 ausführlich diskutiert wurde.

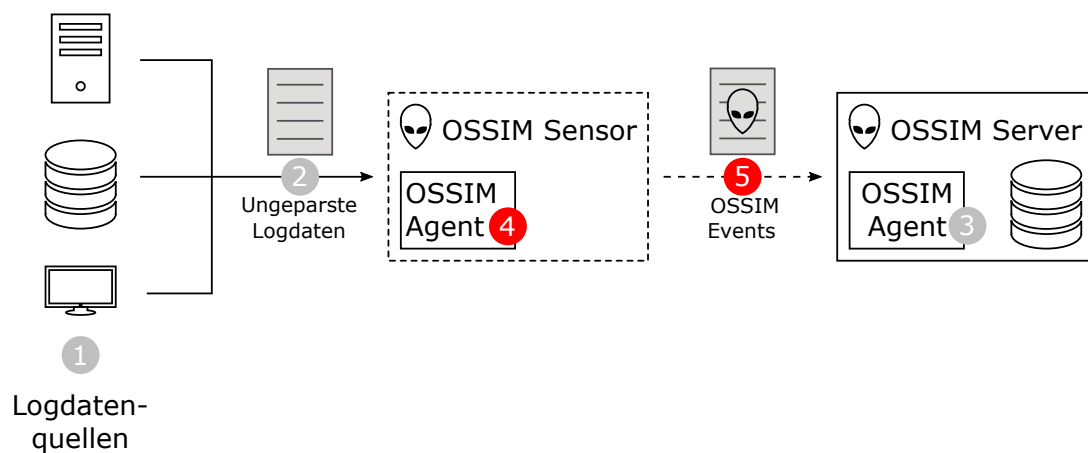


Abbildung 5.1: Mögliche Eingriffspunkte in den OSSIM-Datenfluss.

Zusätzlich erfordert diese Lösung die verteilte Installation von OSSIM-Sensor und -Server, schließt also die All-In-One-Installation aus.

5. Sensor-Server-Proxy

Hier wird ein Proxy zwischen Sensor und Server geschaltet, der bereits geparsete 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.

Die beiden weiteren Möglichkeiten leiden also unter denselben Nachteilen wie der Ansatz einer Veränderung des SIEM-Servers und sind zusätzlich nur bei der Nutzung von OSSIM einsatzbar. Deshalb wurde sich für den zu entwickelnden Prototypen dazu entschieden, die bereits in Abschnitt 3.2.1 präferierte Lösung eines Proxys zwischen Datenquelle und SIEM-System umzusetzen.

Weiterhin wird sich auf den am häufigsten² genutzten Weg des Datenerhalts in OSSIM beschränkt: das Entgegennehmen der Daten über das Syslog-Protokoll (siehe Abschnitt 2.3.1). Lösungen für weitere Wege (siehe Abschnitt 4.1.2) sind bei Bedarf jedoch auf vergleichbarem Weg zu implementieren.

Der entwickelte Log-Proxy nimmt Logdaten über das Syslog-Protokoll entgegen. Verschiedene Datenquellen und Eventformate können mithilfe von Konfigurationsdateien angelegt werden wie im nächsten Abschnitt beschrieben. Für einzelne Felder eines Logdatums können dabei Plugins angegeben werden, die die Behandlung des Feldes übernehmen. Der Proxy-Server kann durch einfach zu entwickelnde Plugins leicht um weitere Datenschutztechniken erweitert werden. Dies wird im übernächsten Abschnitt beschrieben. Nach der Behandlung eines Logdatums wird es anschließend über das Syslog-Protokoll an den OSSIM-Server oder -Sensor geschickt.

5.1.1 Konfiguration von Datenquellen

Die Behandlung von verschiedenen Datenquellen wird durch Konfigurationsdateien ermöglicht:

```
1 | [general]
2 | active=True
3 |
4 | [group1]
5 | pattern=^(?P<time>\w+ *\d{1,2} \d{2}:\d{2}:\d{2}) (?P<device>[^:]+):
   |     Testing my device USER=(?P<user>.+)$
6 | time=Substitute(substitute = 'somevalue_time')
7 | device=Substitute(substitute = 'somevalue_device')
8 | user=Pseudonymize()
```

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.
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.

```

9 ||
10 || [group2]
11 || pattern=^(?P<test>.*)$
12 || test=Pseudonymize()

```

Eine Konfigurationsdatei kann aus mehreren Bereichen bestehen. Der `general`-Bereich enthält allgemeine Angaben über das Plugin. Um unterschiedliche Lognachrichten eines Gerätes bündeln zu können, kann eine Konfigurationsdatei weiterhin mehrere Bereiche enthalten, die jeweils die Verarbeitung einer bestimmten Lognachricht beschreiben. Angegeben werden muss jeweils ein regulärer Ausdruck, der die Nachricht beschreibt und mehrere Gruppen `((?P<name>...))` enthalten kann. Für jede dieser Gruppen muss eine Angabe zu dem Plugin inklusive notwendiger Parameter gemacht werden, das die Gruppe verarbeiten soll. Durch diese Konfigurationsdateien können Nachrichten unbekannter Formate aus neuen Datenquellen leicht in das bestehende System eingebunden werden.

5.1.2 Umsetzung von Datenschutztechniken durch Plugins

Die Erweiterbarkeit um neue Datenschutztechniken wird durch leicht erweiterbare Plugins ermöglicht. Ein Plugin muss lediglich die Methode `handle_data` implementieren, die die originalen Daten und alle in der Konfiguration angegebenen Parameter erhält. Ein einfaches Plugin, das die Daten durch ein in der Konfiguration angegebenen Wert ersetzt, könnte so aussehen:

```

1 || class Substitute(AbstractPlugin):
2 ||
3 ||     def handle_data(self, data: str, **kwargs) -> str:
4 ||         if 'substitute' in kwargs:
5 ||             return kwargs['substitute']
6 ||         else:
7 ||             raise MissingSubstituteError

```

5.2 Umsetzung der Pseudonymisierung

Für die Pseudonymisierung wurde ein Plugin für den Proxy entwickelt, wie in Abschnitt 5.1 beschrieben. Bei jedem eintreffenden Logdatum kann abhängig von dem Datenformat für ein entsprechendes Datenfeld ein Pseudonym als Ersatz für das echte Datum gesetzt werden. Dazu stellt der Pseudonym-Service eine Schnittstelle bereit, über die für ein Datum ein Pseudonym erhalten werden kann. Durch diese Trennung wird eine höhere Sicherheit der Zuordnung zwischen Datum und Pseudonym erreicht: In dem Proxy kommen die Logdaten in unveränderter Form an und werden verändert weitergesendet, daher ist die Zuordnung hier implizit bekannt und muss in Kauf genommen werden. Die Speicherung dieser Zuordnung erfolgt jedoch nur in dem Pseudonym-Service. Durch die Verschlüsselung und die in den folgenden Abschnitten beschriebene MAC-abhängige Pseudonymgenerierung erfährt der Pseudonym-Service nichts über das Datum, das durch das Pseudonym beschrieben wird. So führt unberechtigter Zugriff auf die Datenbank des Pseudonym-Service nicht zu mehr Informationen über das Datum, das ein Pseudonym beschreibt.

Um Implementa-
tion erweitern
(welche Parameter
in welchen Sys-
temteilen, Setup,
...)

5.2.1 Setup-Phase

Vor Verwendung der Pseudonymisierung müssen die Parameter zur Pseudonymgenerierung (vgl. Abschnitt 4.2) dem System bekannt gemacht werden. Diese Parameter können in dem Pseudonym-Service mittels einer Konfigurationsoberfläche durch einen berechtigten Benutzer gesetzt werden. Die für den Proxy relevanten Parameter können anschließend über eine Schnittstelle abgefragt werden. Vorerst handelt es sich hierbei lediglich um das maximale Zeitintervall, in dem ein Pseudonym genutzt werden darf.

5.2.2 Proxy

Beim Start des Proxies wird zuerst der gerade beschriebene Parameter erhalten. Anschließend können eintreffende Logdaten verarbeitet werden. Das eintreffende Datum wird verschlüsselt (siehe dazu Abschnitt 5.3) und anschließend zusammen mit einem über das Datum generierten MAC, der wie in Abschnitt 4.4 beschrieben für die Überprüfung auf bereits bestehende Pseudonyme genutzt wird, an den Pseudonym-Service gesendet. Das ursprüngliche Datum wird nun durch das gelieferte Pseudonym (siehe nächsten Abschnitt) ersetzt und das so veränderte Logdatum wird an das SIEM-System weitergeleitet.

Der Schlüssel, der für die Generierung des MACs verwendet wird, wird abhängig von dem erhaltenen Parameter nach einer bestimmten Zeitspanne neu generiert. Durch diesen Schlüsselwechsel wird wie beschrieben erreicht, dass für gleiche Daten, für die der MAC mit einem neuen Schlüssel erstellt wird, auch neue Pseudonyme erhalten werden. Da der Schlüsselwechsel nicht Pseudonym-abhängig geschieht, ist die Zeitspanne global für alle Pseudonyme gültig und somit als maximale Zeitspanne zu verstehen. Dies kann für die Anomalieerkennung evtl. Probleme bereiten, wenn nicht genügend lange Überwachungsdaten verkettet werden können. Auf der anderen Seite würde eine Verweildauer für einzelne Pseudonyme ein Erfassen des Erstellungszeitpunkts in der Datenbank erfordern, was wie in Abschnitt 4.2 beschrieben Rückschlüsse auf das ursprüngliche Datum des Pseudonyms liefern könnte. Daher wurde sich gegen diesen Ansatz entschieden.

5.2.3 Service

Auf der Service-Seite wird anhand des empfangenen MACs durch Vergleich mit in der Datenbank vorliegenden MACs überprüft, ob bereits ein Pseudonym für das eintreffende Datum vergeben wurde, das noch nicht zu häufig verwendet wurde. Hierzu wird der in der Konfiguration gesetzte Parameter zur maximalen Nutzungshäufigkeit von Pseudonymen verwendet. Liegt kein solches Pseudonym vor, so wird ein noch nicht verwendetes, zufälliges Pseudonym erstellt und zusammen mit dem MAC und dem verschlüsselten Datum in der Datenbank gespeichert. Anderenfalls wird das bereits vergebene Pseudonym zurückgeliefert.

5.2.4 Perfect Forward Privacy

Im Zusammenspiel dieser Parameter kann jedoch noch ein Problem entstehen. Für neu vergebene Pseudonyme, die innerhalb eines Zeitabschnitts durch Überschreiten der maximalen Nutzungsanzahl entstanden sind, liegen in der Datenbank Einträge mit gleichem MAC vor.

Auf diese Weise wird die Verknüpfung verschiedener Pseudonyme ermöglicht, wenn jemand (berechtigt oder unberechtigt) Zugriff auf die Daten erhält. Das Aufdecken eines Pseudonyms deckt auch alle anderen in diesem Zeitintervall erstellten Pseudonyme implizit auf, was dem in Abschnitt 4.2 beschriebenen Prinzip der *Perfect Forward Privacy* widerspricht.

Dieses Problem könnte durch eine Hashwert-Berechnung für den eintreffenden MAC auf der Service-Seite verhindert werden, die einen Pseudonym-abhängigen Zufallswert (eine Art Salt) einbezieht. Hierdurch enthalten Datenbankeinträge, die innerhalb eines Zeitintervalls zu dem gleichen Datum gehören und daher den gleichen MAC besitzen, durch den einfließenden Zufallswert unterschiedliche Hashwerte. Durch die Einweg-Eigenschaft der Hashfunktion wäre die Verkettbarkeit verschiedener Pseudonyme verhindert. Jedoch erfordert dieser Ansatz eine Hash-Berechnung pro Datenbankeintrag für jede Anfrage und ist daher aus Performance-Sicht kritisch zu betrachten. Aus diesem Grund wird diese Möglichkeit vorerst nicht implementiert. Das bestehende Problem ist jedoch für ein konkretes Anwendungsszenario und bei der Wahl der Parameter – insbesondere des Zeitintervalls – zu beachten.

5.3 Implementierung und Integration des Schwellwertschemas

In diesem Abschnitt wird zuerst die entwickelte Bibliothek zur Nutzung eines kryptographischen Schwellwertschemas dargestellt und anschließend ihre Nutzung in den unterschiedlichen Systemkomponenten beschrieben.

5.3.1 ThresholdCrypto „Lib“

Um die Funktionen des Schwellwertschema unterschiedlichen Systemteilen einfach zur Verfügung zu stellen, wurde eine Bibliothek entwickelt, die in den verschiedenen Komponenten genutzt werden kann. Das öffentliche Interface der Bibliothek stellt folgende Funktionen bereit:

- **Parametergenerierung:** Diese Funktion dient dem Erhalt der benötigten sicheren Primzahl bzw. des Generator (vgl.). Hierbei kann zwischen Neugenerierung und Verwendung vorberechneter Parameter verschiedener Schlüsselstärken entschieden werden. Näheres dazu ist im Unterabschnitt *Parametergenerierung* zu finden.
- **Schlüssel- und Sharegenerierung:** Durch diese Funktion wird ein zufälliger geheimer Schlüssel erzeugt und aus ihm der öffentliche Schlüssel sowie die einzelnen *Shares* erzeugt.
- **Verschlüsselung einer Nachricht:** Diese Funktion verschlüsselt mithilfe des öffentlichen Schlüssels eine Nachricht (siehe auch Unterabschnitt *Hybride Kryptographie*). Sie wird im Proxy verwendet.
- **Berechnung einer partiellen Entschlüsselung:** Mithilfe eines *Shares* wird die zugehörige partielle Entschlüsselung einer verschlüsselten Nachricht berechnet. Diese Funktion wird in den Threshold-Clients genutzt.
- **Kombinieren von partiellen Entschlüsselungen:** Aus ausreichend partiellen Entschlüsselungen kann die Nachricht wiederhergestellt bzw. entschlüsselt werden (siehe wiederum Unterabschnitt *Hybride Kryptographie*).

ergänzen

Neben den Funktionen werden noch einige Klassen zur Verfügung gestellt, die das Arbeiten mit den Ergebnissen der Funktionen erleichtern sowie Funktionalität wie Serialisierbarkeit ermöglichen:

- **ThresholdParameters:** Enthalten die Werte t und n des Schwellwertschemas.
- **KeyParameters:** Enthalten die benötigten Primzahlen bzw. den Generator der zugrundeliegenden Gruppe.
- **PublicKey:** Enthält den öffentlichen Schlüssel der zum Verschlüsseln einer Nachricht verwendet werden kann.
- **KeyShare:** Enthält die Werte des Shares eines Teilnehmers am Schwellwertschema.
- **EncryptedMessage:** Enthält die Daten einer verschlüsselten Nachricht (vgl. auch Abschnitt 5.3.1).
- **PartialDecryption:** Enthält die zu einer partiellen Entschlüsselung gehörenden Werte, die zum Entschlüsseln der vollständigen Nachricht genutzt werden.

Im Folgenden soll kurz auf zwei Kernelemente des implementierten Verfahrens eingegangen werden.

Parametergenerierung

Die für das Verfahren benötigten sicheren Primzahlen p und q lassen sich mithilfe eines einfachen Ansatzes finden [MVOV96]: Es wird solange eine Primzahl q im Bereich der vorgegebenen Schlüsselstärke zufällig gewählt, bis $2q + 1$ ebenfalls eine Primzahl ist. Zur Überprüfung der Primalität der entsprechenden Zahlen wird der Miller-Rabin-Test³ genutzt.

Zusätzlich benötigt das Verfahren einen Generator g einer Untergruppe der Ordnung q von \mathbb{Z}_p^* . Hierzu wird lediglich solange ein zufälliges Element g aus \mathbb{Z}_p^* gewählt bis

$$g^q \equiv 1 \pmod{p} \text{ und } g^2 \not\equiv 1 \pmod{p}$$

gelten. Da Untergruppen von \mathbb{Z}_p^* nach dem Satz von Lagrange lediglich die Ordnungen $1, 2, q$ oder $2q$ besitzen, werden durch obenstehende Bedingungen lediglich Untergruppen der Ordnung q zugelassen.

Nach [KL14] beeinträchtigt es nicht die Sicherheit des ElGamal-Verfahrens, wenn vorberechnete Parameter von verschiedenen Benutzern geteilt werden. Auch die Kombination dieses Verfahrens mit Shamirs Secret Sharing dürfte hieran nichts ändern, da das Secret Sharing lediglich für die Aufteilung des geheimen Schlüssels in Shares sorgt, aber an den grundlegenden Eigenschaften der Ver- und Entschlüsselung im ElGamal-Schema nichts ändert. Daher wurden verschiedene Parameter bereits vorberechnet und als statische Parameter zur Verfügung gestellt. Weiterhin ist es jedoch auch möglich eigene Parameter in gewünschter Stärke zu generieren. In den meisten Empfehlungen werden heutzutage Schlüssellängen um 2000 Bit empfohlen⁴.

evtl. auf Evaluation beziehen.

3. Der Miller-Rabin-Primzahltest ist ein Algorithmus, der basierend auf probabilistischen Annahmen die Primalität einer Zahl mit durch Angabe von einer Rundenzahl beliebig hoher Wahrscheinlichkeit bestätigen kann.
4. Für einen Vergleich verschiedener Empfehlungen siehe: <https://www.keylength.com>

Hybride Kryptographie

Die Verschlüsselung bzw. Entschlüsselung wurden in Form eines hybriden Verschlüsselungsverfahrens umgesetzt (siehe Abschnitt 2.6.3): Bei der Verschlüsselung wird ein zufälliger Schlüssel k^{symm} für ein symmetrisches Verfahren E^{symm} erzeugt und dazu genutzt den Klartext m zu verschlüsseln. Das kryptographische Schwellwertschema wird lediglich dazu verwendet, k^{symm} zu verschlüsseln. Ein Schlüsseltext besteht daher aus drei Teilen:

- $v = g^k$: Der erste Teil der ElGamal-Verschlüsselung des symmetrischen Schlüssels (siehe Abschnitte 2.6.5 und 4.3.2)
- $c^{tc} = k^{symm} \cdot g^{ak}$: Der zweite Teil der ElGamal-Verschlüsselung des symmetrischen Schlüssels (siehe Abschnitte 2.6.5 und 4.3.2)
- $c^{symm} = E_{k^{symm}}^{symm}(m)$: Der symmetrisch verschlüsselte Klartext

Bei der Entschlüsselung wird ähnlich vorgegangen: Das kryptographische Schwellwertschema wird dazu genutzt, den symmetrischen Schlüssel wiederherzustellen. Anschließend kann der ursprüngliche Klartext mit diesem Schlüssel wieder entschlüsselt werden. Dieses hybride Vorgehen bietet neben dem im Grundlagenkapitel erwähnten Geschwindigkeitsvorteil weitere Vorteile:

Das Vorgehen ermöglicht es beliebige Nachrichten relativ einfach zu verschlüsseln, da die Beschränkung des asymmetrischen Verfahrens bezogen auf die Nachrichtenlänge (die in kodierter Form kleiner sein muss als der Parameter p) nur noch für den Schlüssel des symmetrischen Verfahrens erfüllt werden muss. Dies stellt kein Problem dar, da symmetrische Verfahren für vergleichbare Sicherheit geringere Schlüssellängen benötigen als asymmetrische Verfahren, die auf dem Diskreten-Logarithmus-Problem beruhen.

Weiterhin kann für die symmetrische Verschlüsselung ein Verfahren genutzt werden, dass neben der Verschlüsselung auch die Validität der Daten überprüft - ein sogenanntes *Authenticated Encryption* (AE) Schema (siehe Abschnitt 2.6.4).

In der Implementierung wird die Standardfunktion zur symmetrischen Verschlüsselung aus der Kryptographie-Bibliothek NaCl⁵ genutzt – ein AE-Schema auf Basis der Algorithmen Salsa20 und Poly1305. Diese verwendete Bibliothek ist weit verbreitet und von erfahrenen Kryptographen entwickelt und überprüft. Dies erhöht das Vertrauen in eine sichere Umsetzung der Verfahren.

Eine direkte Implementierung des bis hierhin beschriebenen hybriden Schemas enthält jedoch noch eine Schwäche: In [BJN00] wird von den Autoren ein Angriff vorgestellt, der bei der direkten Implementierung von auf dem ElGamal-Verfahren basierenden Schemata möglich ist. Der Angriff besteht darin, dass bei der Verschlüsselung symmetrischer Schlüssel durch deren geringere Länge und die Berechnungen in speziellen Untergruppen im ElGamal-Verfahren die Möglichkeit besteht, die symmetrischen Schlüssel mit geringerem Aufwand zu entschlüsseln.

Als Gegenmaßnahme wird die Vorverarbeitung der symmetrischen Schlüssel empfohlen. In [ABR99] wird ein hybrides Schema dargestellt, dass diese Gegenmaßnahme umsetzt. Der entscheidende Schritt im Vorgehen besteht darin, den symmetrischen Schlüssel nicht direkt zufällig zu erzeugen. Stattdessen wird ein zufälliges Untergruppenelement des Nachrichtenraums $r \in \mathbb{Z}_q^*$ gewählt und der symmetrische Schlüssel mithilfe einer kryptographisch sicheren Hashfunktion

5. NaCl: Networking and Cryptography library: <https://nacl.cr.yp.to>

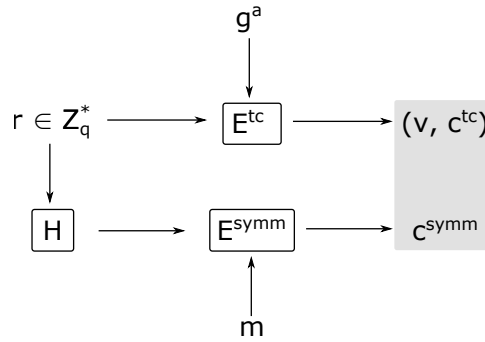


Abbildung 5.2: Übersicht über das umgesetzte hybride Verschlüsselungsschema.

H hieraus abgeleitet. Hierdurch verändert sich die Zusammensetzung eines Schlüsseltextes leicht:

- $v = g^k$: wie vorhergehend beschrieben
- $c^{tc} = r \cdot g^{ak}$: Anstelle des symmetrischen Schlüssels wird das Untergruppenelement r durch das ElGamal-Verfahren verschlüsselt.
- $c^{symm} = E_{H(r)}^{symm}(m)$: Als symmetrischer Schlüssel wird nun der durch H berechnete Hashwert von r genutzt

Zusätzlich schreibt das Schema nach [ABR99] die Benutzung eines MACs zur Sicherung der Integrität der Nachricht vor. Auf diesen Schritt kann in dem in dieser Arbeit implementierten Schema verzichtet werden, da die Integrität bereits durch das verwendete AE-Schema geschützt ist. Einen Überblick über das letztlich umgesetzte Schema, das auch diesen Angriff verhindert, bietet Abbildung 5.2.

5.3.2 Service und Setup-Verfahren

Um die Parameter für das kryptographische Schwellwertschema setzen zu können, wurde die Konfigurationsoberfläche erweitert. Hier müssen nun Angaben zur Schlüsselstärke, zu beteiligten Teilnehmern und erforderlichen Anzahl an Teilnehmern für die Entschlüsselung eines Datenbankeintrages gemacht werden.

Anschließend können berechtigte Benutzer die Aufdeckung eines Pseudonyms über das Webinterface anfragen. Anschließend werden von den teilnehmenden Clients die partiellen Entschlüsselungen entgegengenommen. Liegen ausreichend partielle Verschlüsselungen vor, so werden diese mithilfe der Bibliothek kombiniert und damit der Halter des Pseudonyms entschlüsselt und aufgedeckt.

5.3.3 Client-Anwendung

Für Teilnehmer, die an dem Schwellwertschema beteiligt sind, wurde eine konsolenbasierte Anwendung entwickelt, die die folgende Funktionalität bereitstellt:

Clients melden sich zu Beginn an dem Pseudonym-Service an und können in der anschließenden Setup-Phase mit in das Verfahren integriert werden. Jeder teilnehmende Client empfängt nach

der Generierung das für ihn bestimmte Share vom Server und speichert dieses verschlüsselt lokal ab. Hierzu enthält die Anwendung einen lokalen Webserver. Dies ermöglicht es dem Pseudonym-Service die *Shares* nicht speichern und nach der Generierung sofort entfernen zu können.

Anschließend können sich Teilnehmer regelmäßig nach neuen Anfragen zur Aufdeckung eines Pseudonyms erkundigen. Wurde eine neue Anfrage empfangen, so kann der Teilnehmer diese annehmen oder ablehnen. Die Annahme führt zur Berechnung einer partiellen Entschlüsselung mithilfe der Bibliotheksfunktion, die anschließend an den Pseudonym-Service gesendet wird.

5.3.4 Proxy

Das im Syslog-Proxy implementierte Plugin zur Pseudonymisierung empfängt während der Initialisierung den während der Schlüsselerzeugung generierten öffentlichen Schlüssel vom Pseudonym-Service. Bei einer eintreffenden Syslog-Nachricht werden die zu pseudonymisierenden Daten mit diesem Schlüssel unter Nutzung der Bibliotheksfunktion verschlüsselt und zusammen mit den bereits in Abschnitt 5.2 beschriebenen weiteren Daten an den Service gesendet.

5.4 Evaluation

5.4.1 Erfüllung der Anforderungen

- Geeignete Stelle zum Eingriff in den Datenfluss zwischen Logdatenquelle und SIEM-System
- Parameterabhängige Generierung eindeutiger, aber in gewissem Rahmen verknüpfbarer Pseudonyme
- Sicherer, verteilter Einsatz eines anpassbaren kryptographischen Schwellwertschemas – vorzugsweise mit verteilter Schlüsselgenerierung
- Geeignete Benutzerinteraktion mit dem System an notwendigen Stellen
- Erweiterbarkeit um unbekannte Datenquellen
- Erweiterbarkeit um weitere Datenschutztechniken

...

5.4.2 Angriffsmöglichkeiten

TBW

Zentrale Schlüsselgenerierung

Ein bereits in Abschnitt 3.1.3 betonter Punkt ist die Präferenz von verteilter gegenüber der zentralen Schlüsselgenerierung. In dem entwickelten Prototypen wurde jedoch aus Zeitgründen bisher nur die zentrale Schlüsselgenerierung umgesetzt. Dies ermöglicht einem Angreifer

mit (legitimen oder nicht legitimen) Zugriff auf den Pseudo-Service während der Schlüsselgenerierung den in Abschnitt 3.3 beschriebenen Angriff zur beliebigen Entschlüsselung von Pseudonymzuordnungen. Abhilfe würde ein Schema – wie das in Abschnitt 4.3.3 beschriebene – schaffen.

Nicht überprüfte kryptographische Funktionen

Die Bibliotheksfunktionen des kryptographischen Schwellwertschemas wurden nach bestem Wissen und Gewissen umgesetzt. Trotzdem wurden sie bisher nicht von erfahrenen Kryptographen überprüft und können daher eine Vielzahl von Schwächen und Angriffsmöglichkeiten aufweisen. Die Nutzung einer quelloffenen und vielfach überprüften Bibliothek wäre wünschenswert, aber wie in Abschnitt 4.3.4 beschrieben, wurde eine solche Bibliothek nicht gefunden.

Schlüsseltexte offenbaren Klartextlänge

Eine Eigenschaft des entwickelten hybriden Kryptoschemas kann dazu führen, dass trotz der Verschlüsselung auf den Halter eines Pseudonyms geschlossen werden kann. Bei dem verwendeten symmetrischen Verschlüsselungsalgorithmus Salsa20 handelt es sich um eine Stromchiffre. Hierdurch hat der Schlüsseltext, der in der Datenbank für ein Pseudonym gespeichert wird, genau die gleiche Länge wie der Klartext. Unterscheiden sich die Klartexte (beispielsweise Benutzernamen) in Ihrer Länge, so kann durch Vergleich dieser Längen auf den Pseudonymhalter geschlossen werden. Dieser Angriff erfordert Zugriff auf den Inhalt der Datenbank des Pseudo-Service.

Dieses Problem lässt sich durch Padding beheben, das alle Klartexte auf die gleiche Länge bringt. Die konkrete Umsetzung ist allerdings abhängig von dem Wertebereich der eintreffenden Daten.

Anwendung von Hintergrundwissen

In [LJ99] wird das Problem der Identifikation eines Pseudonymhalters durch die Anwendung von Hintergrundwissen wie Kenntnis über normales Nutzerverhalten dargestellt. Dies ist auch in dem in dieser Arbeit entwickelten System der Fall wie bereits in Abschnitt 3.3 erwähnt. Dieser Angriff lässt sich aus Sicht des Autors nicht vollständig verhindern, da Beobachtungen in der realen Welt nicht zu vermeiden sind. Regelmäßige, parameterabhängige Pseudonymwechsel, wie sie implementiert wurden, sollten diesen Angriff jedoch in seinen Auswirkungen beschränken.

5.4.3 Performance

- Zeitmessungen für einzelne Systemfunktionen
- Performance bzw. Auswirkungen der Nutzung (theoretische Rechenlast, Zeit-/Lastmessung, zusätzlicher Speicherbedarf, ...)
- Bar Chart Times per Key size
- Pie Chart Times per Computational part
- Not deployed in a standard way (uWSGI, Apache, ...) - Crashes of PostgreSQL database driver
- Unreliability of UDP

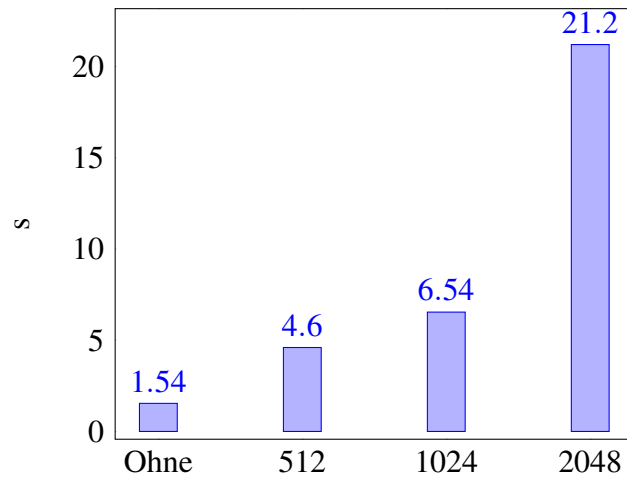


Abbildung 5.3: Dauer der Bearbeitung von 100 Lognachrichten in Sekunden.

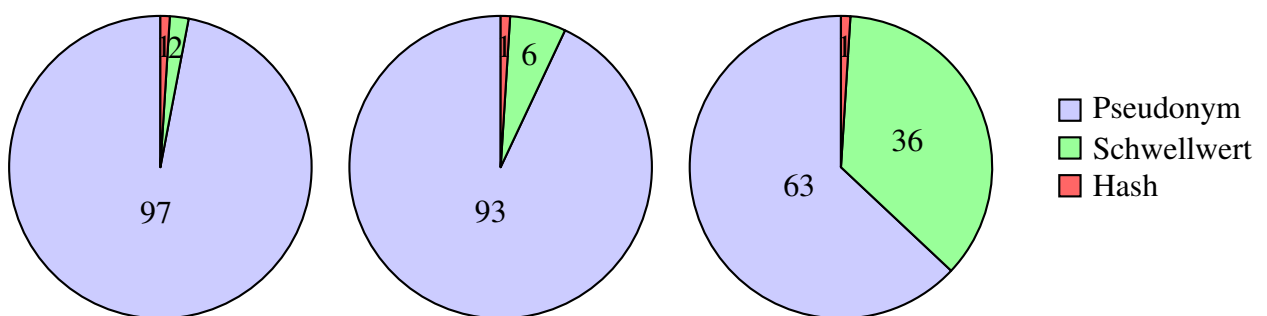


Abbildung 5.4: Zeitlicher Anteil der verschiedenen Berechnungen im Pseudonymisierungs-Plugin bei Schlüssellängen von 512 Bit (links), 1024 Bit (Mitte) und 2048 Bit (rechts).

6 Ergänzende und alternative Datenschutztechniken

Der in dieser Arbeit verfolgte Ansatz der Pseudonymisierung unter Einsatz kryptographischer Schwellwertschemata ist nicht für alle Arten von Logdaten sinnvoll. Sollen beispielsweise Zeitstempel verändert werden, um keine direkten Rückschlüsse auf eine Person durch Kombination mit typischen Verhaltensmustern zu ermöglichen, auf der anderen Seite jedoch zumindest grobe Erkenntnisse aus dem Zeitstempel für die Anomalieerkennung genutzt werden, so kann eine auf Pseudonymisierung basierende Lösung dies nicht leisten.

Daher werden in diesem Abschnitt ergänzende Techniken zum Erhalt der Privatsphäre eines Arbeitnehmers bei der Speicherung von Logdaten dargestellt und erläutert, wie diese in den entwickelten Prototypen eingebunden werden können. Natürlich können auch mehrere kombinierte Techniken für einzelne Logdaten bestehend aus mehreren Feldern sinnvoll sein. Mit Zeitstempeln versehene Logdaten eines Türschließsystems, die die Benutzerkennungen der Mitarbeiter enthalten, könnten so beispielsweise durch Pseudonymisierung der Benutzerkennungen und Verrauschung der Zeitstempel geschützt werden.

Eine Übersicht über weitere in sehr speziellen Einsatzgebieten zu nutzende Datenschutztechniken wie Bloomfilter oder der Einsatz homomorpher Verschlüsselung sind in [NKS17] zu finden.

6.1 Unterdrückung

Als ergänzende Maßnahme für Datenfelder, die für die Anomalieerkennung nicht benötigt werden, aber Rückschlüsse auf den Benutzer zulassen, kommt die Unterdrückung in Frage. Hierbei wird der Wert des Feldes schlicht entfernt oder durch eine Konstante ersetzt. Ein Beispiel für den Kontext dieser Arbeit ist die Unterdrückung von IP-Adressen eines durch einen Arbeitnehmer benutzten Rechners für Logdaten, in denen auch der Benutzername enthalten ist. Bei dem Einsatz von Pseudonymisierung könnte die Zuordnung von Benutzer zu Pseudonym erleichtert werden, wenn der Arbeitnehmer dem physischen Gerät zu einem Zeitpunkt zugeordnet werden kann und die IP-Adresse des Geräts noch in den Logdaten enthalten ist.

6.2 Generalisierung

Bei der Generalisierung wird der Feldinhalt durch einen Wert ersetzt, der das gleiche Konzept beschreiben, jedoch allgemeiner ist. Durch mehrfache Verallgemeinerung entstehen sogenannte Generalisierungshierarchien, wobei die Stufe der höchsten Generalisierung hier gleichbedeutend mit der Unterdrückung ist, da jeder Wert durch den konstanten Wert der höchsten Generalisierungsstufe ersetzt wird.

Ein Beispiel im Unternehmenskontext dieser Arbeit ist die Generalisierung eines Mitarbeiters zu seiner Arbeitsgruppe oder Abteilung – eine Information, die für die Anomalieerkennung ausreichend sein könnte, wenn es beispielsweise um Zugriffe auf Ressourcen geht, die für

Beispiel, Diagramm?

bestimmte Abteilungen üblich, für andere jedoch ungewöhnlich sind. Dieses könnte zusätzlich zur Pseudonymisierung ausgeführt werden, um der Anomalieerkennung zusätzliche Daten zur Verfügung zu stellen, ohne die Identität eines Nutzers direkt offenzulegen.

6.3 Verrauschen

Diese Maßnahme verändert den Wert eines Datenfeldes, indem diesem Werte aus einer Wahrscheinlichkeitsverteilung hinzugefügt werden (statistisches Rauschen). Hierdurch lassen sich die Rückschlüsse auf einen Nutzer aus einem einzelnen Datensatz verringern, aber die Gesamtverteilung bleibt erhalten bzw. lässt sich leicht berechnen. Hierdurch lassen sich zumindest Abweichungen von Durchschnittswerten zur Anomalieerkennung nutzen. Es wird jedoch eine ausreichend große Datenmenge benötigt. Alternativ lassen sich zumindest Aussagen über den Bereich treffen, in dem ein Wert sich befinden muss. Dies könnte beispielsweise bei dem Verrauschen von Ereigniszeitstempeln sinnvoll sein, bei dem zwar nicht auf den konkreten Zeitpunkt geschlossen werden kann, aber zumindest Aussagen darüber getroffen werden können, ob das Ereignis in einem üblichen Intervall wie in den normalen Bürostunden auftrat.

Die Maßnahme ist jedoch nur für bestimmte Felder bzw. Datenarten sinnvoll einsetzbar. Gegenbeispiele sind unter anderem Freitextfelder wie Benutzernamen oder Felder für Aufzählungstypen wie Raumnummern, die sich nicht sinnvoll mit Rauschen verändern lassen.

6.4 Nutzung von Hashverfahren

Neben der zufälligen Generierung von Pseudonymen, wie es in dieser Arbeit genutzt wird, ist auch die Nutzung von Hashwerten als Pseudonym für Daten denkbar. Dies würde die Verknüpfbarkeit von Logdaten ermöglichen, da für gleiche Daten der gleiche Hashwert berechnet wird. Durch den geschickten Einsatz von zusätzlichen zeitabhängig wechselnden Eingaben für das Hashverfahren (sogenannte *Salts*) ließe sich auch die nötige Verknüpfbarkeit für die Anomalieerkennung gegenüber dem Schutz der Privatsphäre eines Benutzers abstimmen. Auf der anderen Seite wäre der Einsatz von Hashverfahren bei einem kleinen Wertebereich für Eingaben wie Benutzernamen wie in Abschnitt 4.4 beschrieben anfällig für Wörterbuchangriffe. Außerdem wären auch Rückschlüsse auf den Pseudonymhalter nicht ohne zusätzlichen Aufwand möglich.

Für Datenfelder, bei denen nur die Verknüpfbarkeit, jedoch nicht der ursprüngliche Wert für die Anomalieerkennung entscheidend ist, und deren Wertebereich ausreichend groß ist, können Hashverfahren sinnvoll sein.

Beispiel

6.5 Vorgehen zur Integration

Die Integration der vorgestellten Datenschutztechniken in den entwickelten Prototypen stellt kein Problem dar. Die jeweiligen Techniken können wie in Abschnitt 5.1.2 beschrieben als Plugins entwickelt werden. Hierbei entsteht je nach Datenschutztechnik unterschiedlicher Aufwand.

Für die Generalisierung und das Verrauschen müssen beispielsweise Eingabedaten-abhängige Plugins entstehen. Die Generalisierung von Zeitstempeln (Generalisierung auf Sekunde, Minute, Stunde, Tag oder selbst gewählte Zeitabschnitte) unterscheidet sich beispielsweise stark von der

Generalisierung der Umgebung eines Mitarbeiters (unternehmensspezifische Generalisierung auf Arbeitsgruppe, Abteilung, ...). Die Unterdrückung oder Nutzung von Hashverfahren kann hingegen unabhängig von den Eingaben entwickelt werden.

Für Datenquellen können diese Plugins nun einzeln oder in Kombination wie in Abschnitt 5.1.1 beschrieben genutzt werden. Durch diesen Ansatz lässt sich für jede Datenquelle abhängig von verwendeten Anomalieerkennungsverfahren eine gute Abwägung zwischen Nutzbarkeit der Daten und Schutz der Privatsphäre eines Arbeitnehmers schaffen.

7 Fazit

TBW

Literatur

- [ABR99] Michel Abdalla, Mihir Bellare und Phillip Rogaway. *DHAES: An Encryption Scheme Based on the Diffie-Hellman Problem*. In: *IACR Cryptology ePrint Archive* 1999 (1999), S. 7.
- [Ara+14] Myrto Arapinis u. a. *Privacy through Pseudonymity in Mobile Telephony Systems*. In: *Network and Distributed System Security Symposium*. 2014.
- [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.
- [BBO07] Mihir Bellare, Alexandra Boldyreva und Adam O’Neill. *Deterministic and efficiently searchable encryption*. In: *Advances in Cryptology-CRYPTO 2007* (2007), S. 535–552.
- [BF00] Joachim Biskup und Ulrich Flegel. *Threshold-based identity recovery for privacy enhanced applications*. In: *Proceedings of the 7th ACM Conference on Computer and Communications Security*. ACM. 2000, S. 71–79.
- [BF01] Joachim Biskup und Ulrich Flegel. *On pseudonymization of audit data for intrusion detection*. In: *Designing Privacy Enhancing Technologies*. Hrsg. von Hannes Federrath. Springer. 2001, S. 161–180.
- [BFP14] Ulrike Baumann, Elke Franz und Andreas Pfitzmann. *Kryptographische Systeme*. Springer, 2014.
- [bit16] bitkom. *Spezialstudie Wirtschaftsschutz*. bitkom. 2016. URL: <https://www.bitkom.org/Bitkom/Publikationen/Spezialstudie-Wirtschaftsschutz.html> (besucht am 24.06.2017).
- [BJN00] Dan Boneh, Antoine Joux und Phong Q Nguyen. *Why textbook ElGamal and RSA encryption are insecure*. In: *ASIACRYPT*. Bd. 1976. Springer. 2000, S. 30–43.
- [BK99] Roland Büschkes und Dogan Kesdogan. *Privacy enhanced intrusion detection*. In: *Proceedings of the Conference on Multilateral Security for Global Communication*. Citeseer. 1999, S. 187–207.
- [Bla79] George Robert Blakley. *Safeguarding cryptographic keys*. In: *Proceedings of the AFIPS 1979 National Computer Conference*. Hrsg. von Richard E. Merwin. AFIPS Press, 1979, S. 313–317.
- [Bon+04] Dan Boneh u. a. *Public key encryption with keyword search*. In: *Eurocrypt*. Bd. 3027. Springer. 2004, S. 506–522.
- [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).
- [CM05] Yan-Cheng Chang und Michael Mitzenmacher. *Privacy preserving keyword searches on remote encrypted data*. In: *ACNS*. Bd. 5. Springer. 2005, S. 442–455.
- [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.
- [Des97] Yvo Desmedt. *Some recent research aspects of threshold cryptography*. In: *International Workshop on Information Security*. Springer. 1997, S. 158–173.
- [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.
- [Dö05] Florian Dötzer. *Privacy issues in vehicular ad hoc networks*. In: *International Workshop on Privacy Enhancing Technologies*. Springer. 2005, S. 197–209.
- [ElG85] Taher ElGamal. *A public key cryptosystem and a signature scheme based on discrete logarithms*. In: *IEEE transactions on information theory* 31.4 (1985), S. 469–472.
- [FPS00] Pierre-Alain Fouque, Guillaume Poupard und Jacques Stern. *Sharing decryption in the context of voting or lotteries*. In: *Financial Cryptography*. Bd. 1962. Springer. 2000, S. 90–104.
- [Fra+97] Yair Frankel u. a. *Proactive rsa*. In: *Annual International Cryptology Conference*. Springer. 1997, S. 440–454.
- [Gem97] Peter Gemmell. *An introduction to threshold cryptography*. In: *CryptoBytes, a technical newsletter of RSA Laboratories* 2.7 (1997), S. 295–310.
- [Gen+96a] Rosario Gennaro u. a. *Robust and efficient sharing of RSA functions*. In: *Advances in Cryptology—CRYPTO'96*. Springer. 1996, S. 157–172.

- [Gen+96b] Rosario Gennaro u. a. *Robust threshold DSS signatures*. In: *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer. 1996, S. 354–371.
- [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.
- [Ger09] Rainer Gerhards. *The syslog protocol*. RFC 5424. 2009.
- [Goh03] Eu-Jin Goh. *Secure indexes*. In: *IACR Cryptology ePrint Archive 2003* (2003), S. 216.
- [KL14] Jonathan Katz und Yehuda Lindell. *Introduction to modern cryptography*. 2. akt. Auflage. CRC press, 2014.
- [LJ00] Emilie Lundin und Erland Jonsson. *Anomaly-based intrusion detection: privacy concerns and other problems*. In: *Computer networks* 34.4 (2000), S. 623–640.
- [LJ99] Emilie Lundin und Erland Jonsson. *Privacy vs. Intrusion Detection Analysis*. 1999. URL: <http://www.raid-symposium.org/raid99/PAPERS/Lundin.pdf>.
- [MVOV96] Alfred J Menezes, Paul C Van Oorschot und Scott A Vanstone. *Handbook of applied cryptography*. CRC press, 1996.
- [Ngu05] H.L. Nguyen. *RSA Threshold Cryptography*. 2005. URL: <https://www.cs.ox.ac.uk/files/269/Thesis.pdf> (besucht am 02. 07. 2017).
- [Nik+13] Salman Niksefat u. a. *ZIDS: a privacy-preserving intrusion detection system using secure two-party computation protocols*. In: *The Computer Journal* 57.4 (2013), S. 494–509.
- [NK11] Mark Nicolett und Kelly M Kavanagh. *Magic quadrant for security information and event management*. In: *Gartner Research Note G00212454* (2011).
- [NKS17] Salman Niksefat, Parisa Kaghazgaran und Babak Sadeghiyan. *Privacy issues in intrusion detection systems: A taxonomy, survey and future directions*. In: *Computer Science Review* 25 (2017), S. 69–78.
- [Par+07] Hyun-A Park u. a. *PPIDS: privacy preserving intrusion detection system*. In: *Pacific-Asia Workshop on Intelligence and Security Informatics*. Springer. 2007, S. 269–274.
- [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.
- [Pet+15] Jonathan Petit u. a. *Pseudonym schemes in vehicular networks: A survey*. In: *IEEE communications surveys & tutorials* 17.1 (2015), S. 228–255.
- [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).
- [PK01] Andreas Pfitzmann und Marit Köhntopp. *Anonymity, unobservability, and pseudonymity — a proposal for terminology*. In: *Designing privacy enhancing technologies*. Hrsg. von Hannes Federrath. Springer, 2001, S. 1–9.

- [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.
- [Rab98] Tal Rabin. *A simplified approach to threshold and proactive RSA*. In: *Advances in Cryptology—CRYPTO’98*. Springer. 1998, S. 89–104.
- [Sch06] Bruce Schneier. *Angewandte Kryptographie - Der Klassiker. Protokolle, Algorithmen und Sourcecode in C*. München: Pearson Studium, 2006.
- [SFHR97] Michael Sobirey, Simone Fischer-Hübner und Kai Rannenberg. *Pseudonymous audit for privacy enhanced intrusion detection*. In: *Information Security in Research and Business*. Springer, 1997, S. 151–163.
- [Sha79] Adi Shamir. *How to share a secret*. In: *Communications of the ACM* 22.11 (1979), S. 612–613.
- [Sho00] Victor Shoup. *Practical threshold signatures*. In: *Advances in Cryptology—EUROCRYPT 2000*. Springer. 2000, S. 207–220.
- [SHS08] Malek Ben Salem, Shlomo HersHKop und Salvatore J Stolfo. *A survey of insider attack detection research*. In: *Insider Attack and Cyber Security*. Springer, 2008, S. 69–90.
- [SMK09] Florian Schaub, Zhendong Ma und Frank Kargl. *Privacy requirements in vehicular communication systems*. In: *International Conference on Computational Science and Engineering*. Bd. 3. IEEE. 2009, S. 139–145.
- [SS01] Douglas R Stinson und Reto Strobil. *Provably secure distributed Schnorr signatures and a (t, n) threshold scheme for implicit certificates*. In: *ACISP*. Bd. 1. Springer. 2001, S. 417–434.
- [SW17] Rolf Schwartzmann und Steffen Weiß. *Whitepaper zur Pseudonymisierung der Fokusgruppe Datenschutz*. Juni 2017. URL: <https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/2017/digital-gipfel-one-pager-fokusgruppe.pdf> (besucht am 15. 11. 2017).
- [SWP00] Dawn Xiaoding Song, David Wagner und Adrian Perrig. *Practical techniques for searches on encrypted data*. In: *Security and Privacy, 2000. S&P 2000. Proceedings. 2000 IEEE Symposium on*. IEEE. 2000, S. 44–55.
- [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.