Wasch mich, aber mach mich nicht nass – Anonymisierungsverfahren als Schlüssel zur datenschutzkonformen E-Mail-Filterung

Nils Gruschka, Meiko Jensen

Fachhochschule Kiel
Grenzstr. 5
24149 Kiel
Nils.Gruschka@fh-kiel.de

Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein
Holstenstr. 98
24103 Kiel
Meiko.Jensen@rub.de

Abstract: Die rechtskonforme Verarbeitung personenbezogener Daten erweist sich in heutigen Anwendungssystemen häufig als schwer zu realisierende Anforderung. Beispielsweise liegt in der Filterung von Spam-Nachrichten aus E-Mail-Strömen eine der größten diesbezüglichen Herausforderungen für die heutigen E-Mail-Systeme, da hier der Personenbezug besonders evident ist. Ein oft angestrebter Ansatz beruht auf der zentralen Sammlung von Informationen über möglichst viele versandte E-Mails, um in diesen Datenbergen Spam-Wellen zu erkennen. Dies erfordert aber meist die Weitergabe von Informationen aus E-Mail-Daten an Dritte, die in zentraler Position aus diesen Daten Signaturen für Spam-Nachrichten errechnen. Das grundlegende Problem hierbei besteht darin, dass diese zentralen Analyse- und Erkennungssysteme nicht in den Besitz personenbezogener Daten, welche in E-Mail enthalten sind, gelangen dürfen.

In diesem Artikel analysieren wir die bestehenden rechtlichen und technischen Problemfelder rund um die zentralisierte Detektion von Spam-Nachrichten. Basierend auf dem Konzept der zielgerichteten Anonymisierung elaborieren wir eine mögliche Verarbeitungsmethodik für E-Mails, die eine zentrale Verarbeitung in datenschutzkonformer Art und Weise ermöglichen kann. Durch die Art der von uns vorgeschlagenen Anonymisierung wird der Personenbezug weitgehend aus den einzelnen E-Mails herausgelöst, dennoch bleibt die anonymisierte E-Mail hinreichend geeignet zur Identifikation und Extraktion spezifischer Charakteristika von Spam-Nachrichten.

1 Motivation

Von einem technischen Standpunkt aus betrachtet unterscheidet sich die Verarbeitung personenbezogener Daten nicht von der Verarbeitung jedweder anderer Form von Daten. Es werden logische Operationen ausgeführt, die Daten miteinander verknüpfen, verändern, aggregieren, in Bezug zueinander setzen, und daraus Rückschlüsse ziehen. Die Zwecke

der Verarbeitung werden dabei auch eher zweitrangig betrachtet; es macht technisch gesehen einfach keinen Unterschied, ob das aus Geburtstag und aktuellem Datum errechnete Alter einer Person später lediglich als Textfeld in einem Brief auftaucht, oder ob es zur Altersverifikation beim Kauf von Zigaretten am Automaten dient. Die Vorgehensweise zur Ermittlung des Alters aus Geburtsdatum und aktuellem Datum ist technisch identisch.

Von einem rechtlichen, speziell vom datenschutzrechtlichen Standpunkt (z.B. nach Bundesdatenschutzgesetz [Bun78] oder europäischer Datenschutzrichtlinie [uE]) aus ergeben sich hier aber geradezu grotesk unterschiedliche Rahmenbedingungen, die oft genug aufgrund minimaler kontextueller Unterschiede zu völlig unterschiedlicher rechtlicher Würdigung eines Vorganges kommen. Am Beispiel der Verarbeitung personenbezogener Daten ist dies besonders deutlich zu veranschaulichen. Für obiges Beispiel ist etwa die Verarbeitung des aktuellen Datums im Rahmen der erforderlichen Berechnung rechtlich völlig unbedenklich: diese Information ist frei verfügbar, unterliegt keinerlei rechtlichen Nutzungseinschränkungen, und darf somit jederzeit und zu jedem beliebigen Zweck verarbeitet werden.

Ganz anders die Verarbeitung des Geburtsdatums einer Person. Obwohl es sich hier (wie beim aktuellen Datum) technisch um eine tagesgenaue Zeit-Information handelt, die vermutlich innerhalb eines technischen Systems mittels eines identischen Datentyps abgebildet wird, gibt es hier vom rechtlichen Standpunkt aus gesehen bereits eine Fülle von gesetzlichen Normen zu beachten. Diese resultieren allein aus der Tatsache, dass ein Geburtstag stets eine personenbezogene Information darstellt.

Aus diesem rein kontextuellen Unterschied ergibt sich anschließend eine gravierende Abweichung hinsichtlich der rechtlich erlaubten Menge an technischen Verarbeitungsschritten, die mit dieser Information durchgeführt werden dürfen. So muss die Verarbeitung eines Geburtstages grundsätzlich auf Basis einer geeigneten Rechtsgrundlage erfolgen, beispielsweise aufgrund einer expliziten Einwilligung der betreffenden Person in die Verarbeitung seiner personenbezogenen Daten. Diese Einwilligung wiederum kann üblicherweise nur unter bestimmten Umständen erfolgen.

Zum Einen muss die betroffene Person, deren Daten verarbeitet werde sollen, vorab über die Natur der Verarbeitung hinreichend informiert werden, denn nur eine informierte Einwilligung erfüllt die hier gestellten rechtlichen Anforderungen. Für das Zigarettenbeispiel wird diese Einwilligung etwa darüber realisiert, dass der Benutzer darauf hingewiesen wird, dass nur Personen ab 18 Jahren Zugang zu Zigaretten haben dürfen, und dass folglich im Automaten eine Überprüfung des Alters, z.B. anhand des Geburtstages auf einer Chipkarte, erfolgt. Stellt die betroffene Person ihre Chipkarte dann aktiv dem Automaten zur Verfügung, ist damit die explizite, informierte Einwilligung als gegeben anzusehen.

Zum Anderen muss eine Einwilligungserklärung stets auf einen bestimmten, klar definierten Anwendungszweck hin erteilt werden. Eine pauschale Bewilligung der Verarbeitung des eigenen Geburtstages zu *beliebigen* Zwecken beispielsweise ist rechtlich nicht tragbar. Würde der Zigarettenautomat etwa das konkrete Geburtsdatum zur Profilierung der Zigarettenkäufer nutzen (beispielsweise zur Identifikation von Jahrgängen, die besonders häufige Konsumenten eines Zigarettenautomaten sind), wäre dies allein schon deshalb unzulässig, weil die Einwilligung nicht informiert gewesen ist. Der Nutzer wurde nicht hin-

reichend (weil gar nicht) über die Existenz und Durchführung der Profilierung aufgeklärt.

Aus diesen rechtlich motivierten unterschiedlichen Betrachtungen lässt sich bereits ablesen, dass die Verarbeitung bestimmter, technisch äquivalenter Daten aufgrund rechtlicher Unterscheidungen auch technisch unterschiedlich gehandhabt werden muss. Ein entscheidender Faktor hierbei ist die Wertung bestimmter Arten von Daten als *personenbezogen* bzw. *personenbeziehbar*. Für obiges Beispiel ist die Zuordnung diesbezüglich relativ einfach. Das aktuelle Datum lässt sich (ohne weitere Kontextinformation) niemals einer natürlichen Person zuordnen, ist folgerichtig definitiv keine personenbezogene Information. Ein Geburtstag dagegen bezieht sich explizit auf die Person, der dieses Datum als Geburtsdatum zugeordnet wird. Folgerichtig handelt es sich hier in den meisten Fällen um eine personenbezogene oder zumindest personenbeziehbare Information. So wird der Zigarettenautomat aus obigem Beispiel potentiell jedwede Chipkarte akzeptieren, die ein Alter von mindestens 18 Jahren ausweist, selbst wenn die Chipkarte keinerlei weitere Informationen über die Person enthält.

Nichtsdestotrotz lässt sich hinsichtlich der Verarbeitung personenbezogener Daten grundsätzlich feststellen, dass der explizite Verzicht auf bestimmte Kontextinformationen die rechtliche Bewertung eines Datums bezüglich ihrer Personenbeziehbarkeit durchaus gravierend zu ändern vermag. Derartige Verfahren, die den Personenbezug aus Daten explizit löschen (*Anonymisierung*) bzw. sehr stark verschleiern (*Pseudonymisierung*), spielen folglich bei der technischen Verarbeitung personenbezogener Daten eine große Rolle, wenn es um die rechtliche Bewertung der Gesetzeskonformität einer Verarbeitung geht.

In diesem Beitrag werden derartige Verfahren zur Anonymisierung und Pseudonymisierung am Beispiel automatischer Filterung von E-Mails untersucht. Konkret wird dabei analysiert, welche Verfahren bezüglich welcher Kontextinformationen eine Löschung bzw. Verschleierung vornehmen, und welchen Einfluss dies auf die rechtliche Würdigung der Implementierung eines solchen Verfahrens folglich haben könnte.

Im nächsten Abschnitt wird dazu zunächst das System zur E-Mail-Filterung vorgestellt. Abschnitt 3 diskutiert, welche Personenbezüge sich in einer E-Mail zu finden ist. Danach wird im Abschnitt 4 allgemeine Konzepte und Gefahren von Anonymisierung präsentiert. Abschnitt 5 zeigt dann verschiedene Verfahren zu Anonymisierung von E-Mails. Im Kapitel 6 vergleichbare Arbeiten vorgestellt und schließlich im Kapitel 7 die Beiträge dieses Artikels zusammengefasst und bewertet.

2 Automatische Filterung von E-Mails

Betrachtet werden soll im Folgenden ein technisches System zur automatischen Filterung von E-Mails, beispielsweise zur Aussortierung unerwünschter E-Mails (*Spam*), oder zur frei konfigurierbaren automatischen Zuordnung eingehender E-Mails zu Postfächern mit passender Semantik (z.B. Trennung privater und beruflicher E-Mails auf Basis der Absenderadresse oder von Schlüsselbegriffen). Da es sich hier ganz klar um die Verarbeitung personenbezogener Daten handelt, ist eine entsprechende rechtliche Grundlage erforderlich. Nun werden solche Filter üblicherweise von den Nutzern eines E-Mail-

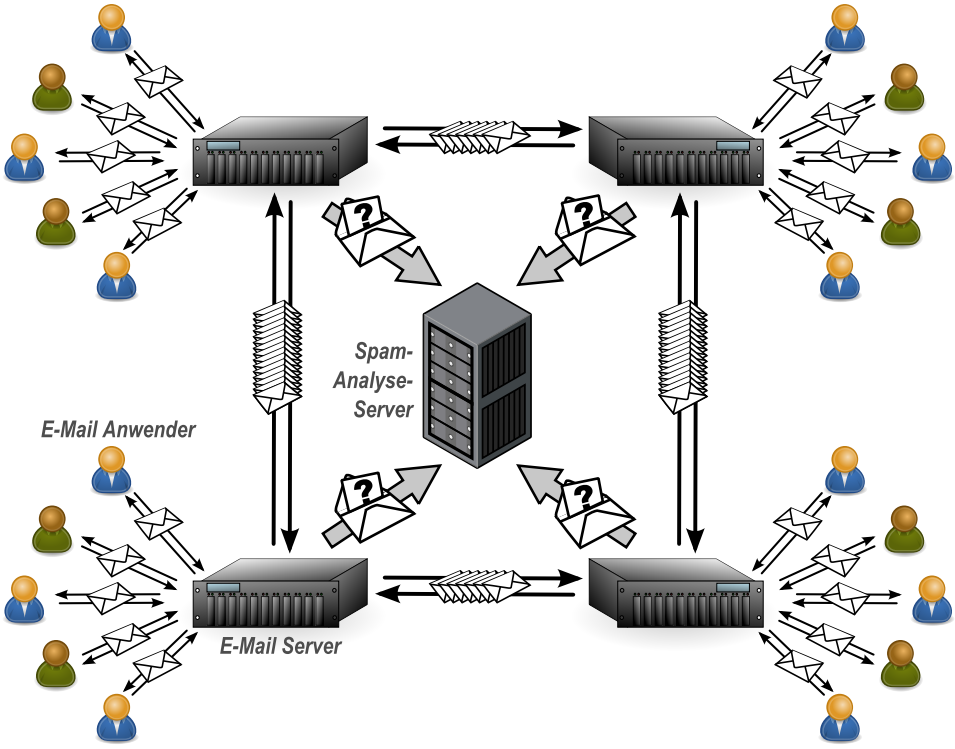

Abbildung 1: Eine verteilte Architektur zur zentralisierten Analyse von E-Mails unter Spamverdacht

Empfangsprogrammes selbsttätig eingerichtet, bzw. wird der automatische Spamfilter selbsttätig und explizit aktiviert. Daraus lässt sich gegebenenfalls eine (informierte?) Einwilligungserklärung in die Filterung ableiten, die die erforderliche Rechtsgrundlage darstellen kann. Diese Einwilligungserklärung ist aber typischerweise auf die lokale Verarbeitung der E-Mails im Empfangsprogramm der betreffenden Person beschränkt, und ist folglich rechtlich nur dort zu verorten. Damit schließt diese Form der Einwilligung nicht automatisch die Zustimmung zu einer automatisierten Vorverarbeitung ein, die etwa bei zustellenden E-Mail-Providern durchgeführt wird. Insbesondere schließt diese Form der Einwilligung auch nicht die Weitergabe der empfangenen E-Mails an Dritte durch den Mail-Provider ein. Letzteres ist technisch aber oft gewünscht, um beispielsweise massenhaft versandte Spam-Nachrichten in größeren Verbünden von Mail-Providern zentral zu identifizieren, und die daraus gewonnenen Erkenntnisse zur Verbesserung der automatischen Spam-Filterung an den einzelnen Mail-Providern zu nutzen (vgl. Abbildung 1).

Derartige zentralisierte Spamerkennungssysteme, wie sie in Abbildung 1 dargestellt sind, bieten den Vorteil, dass durch den Einbezug großer Mengen von Daten präzisere Erkenntnisse darüber gewonnen werden können, welche Merkmale an E-Mails mit hinreichender Sicherheit den Schluss zulassen, dass es sich um eine unerwünschte Spam-Nachricht handelt. Je besser diese Merkmale (*Spam-Signaturen*) definiert werden können, desto präziser

lässt sich die automatische Spam-Erkennung bei den an einen solchen Verbund angeschlossenen Mail-Providern konfigurieren. Dadurch erhöht sich die Rate korrekt erkannter Spam-Nachrichten, und gleichzeitig sinkt folglich die Rate der fälschlich als Spam klassifizierter E-Mails (*false positive*), sowie die Rate der Spam-Nachrichten, die nicht als Solche erkannt werden (*false negative*).

3 Der Personenbezug in klassischen E-Mails

Offensichtlich ist der personenbezogene Charakter von E-Mails leicht zu erkennen. Meistens werden sie von einer natürlichen Person geschrieben (von automatischen Benachrichtigungsmails einmal abgesehen), und sind dafür gedacht, von einer anderen natürlichen Person gelesen zu werden. Folglich ergeben sich hier ganz klare Personenbezüge, die die rechtliche Wertung der automatisierten technischen Verarbeitung von E-Mails deutlich verkomplizieren.

Eine denkbare Möglichkeit zur Vermeidung dieser rechtlichen Problematik besteht in der Löschung bzw. hinreichenden Verschleierung des Personenbezuges von E-Mails. Angesichts der Fülle von personenbezogenen und personenbeziehbaren Informationen in einer typischen E-Mail stellt diese Option aber schwer erfüllbare Anforderungen an die technische Verarbeitung, um eine hinreichende Rechtskonformität zu erreichen. Nichtsdestotrotz lohnt es sich hier, sich die Details der Verarbeitung, und insbesondere die kritischen Personenbezüge in E-Mails genauer zu betrachten.

3.1 Sender und Empfänger

Zunächst existieren offensichtliche Personenbezüge in den E-Mail-Adressen von Sender und Empfängern einer E-Mail. Diese Daten weisen häufig direkt auf eine bestimmte natürliche Person oder eine zahlenmäßig begrenzte Gruppe von Personen hin, die erwartungsgemäß die versandte E-Mail lesen sollen, sowie auf typischerweise genau eine natürliche Person, die die E-Mail erstellt bzw. versandt hat[1]. Hier ist die Notwendigkeit zur Elimination des Personenbezuges offensichtlich.

3.2 Zustellende Mail-Server und weitere Zusatzinformationen

Eine E-Mail enthält üblicherweise eine Liste von Mail-Servern, die an der Zustellung der E-Mail beteiligt waren. Jeder Eintrag besteht dabei aus verschiedenen Informationsfeldern mit Hinweisen zur E-Mail-Weitergabe, beispielsweise der (DNS-)Name bzw. IP-Adresse des Mail-Servers, von dem eine E-Mail entgegengenommen wurde, sowie dem Zeitpunkt,

[1] Ersteller und Versender müssen dabei nicht zwangsweise identisch sein, und es können hier durchaus auch kollaborative Konstellationen mit mehreren Sendern existieren. Technisch abgebildet wird dies aber immer durch eine einzelne, explizit angegebene E-Mail-Adresse des Senders (`From:`-Header nach [Cro82, Res01, Res08]).

an dem dies erfolgte. Hier ist der Personenbezug nicht so trivial sichtbar. Zum Einen kann für manche Mail-Server ein direkter Personenbezug hergestellt werden (*"Der Mail-Server namens* `mail.laptop3.acme.com` *läuft auf dem Laptop des Kollegen X!"*). Zum Anderen können die weiteren Informationen aus diesen Einträgen mit den anderen personenbezogenen Daten aus der E-Mail verknüpft werden, um genauere Informationen zu erhalten. Beispielsweise geben die enthaltenen Zeitstempel relativ genau an, zu welchem Zeitpunkt eine E-Mail versandt wurde, was unter Umständen einen Personenbezug erlaubt (*"Um die Uhrzeit war aber nur noch Kollege Y im Büro!"*).

Darüber hinaus enthält eine E-Mail eine Reihe weiterer Datenfelder, die Informationen über den Sender bereitstellen. Beispielsweise können Angaben zum E-Mail-Programm enthalten sein, das für den Versand benutzt wurde ((`User-Agent`-Header), was ebenfalls ein Hinweis auf natürliche Personen sein kann (*"Nur Kollege Z schreibt Mails mit Emacs!"*).

Es ist aber festzuhalten, dass hier deutlich weniger Daten mit klarem Personenbezug vorliegen. Nichtsdestotrotz muss auch hier eine rechtliche Wertung und Abwägung hinsichtlich der Sensibilität solcher Daten erfolgen.

3.3 Betreff und Mailtext

Schließlich enthält eine E-Mail üblicherweise eine Betreffszeile (`Subject`-Header), sowie einen Mailinhalt. Bei beiden Feldern handelt es sich klassischerweise um vom Verfasser frei befüllbare Textfelder, deren Personenbezug folglich sehr schwer zu taxieren ist. Offensichtlich ist der Personenbezug zum Verfasser, auch wenn er aus den textuellen Inhalten selbst nicht zwangsweise ersichtlich sein muss. Eine mit dem eigenen Namen oder Kürzel unterschriebene E-Mail ist heutzutage aber durchaus üblich, folglich kann die Offenlegung dieses Personenbezuges aus dem Freitext selbst nicht grundsätzlich ausgeschlossen werden. Darüber hinaus kann in beiden Textfeldern aber auch ein Bezug zum Empfänger bestehen, sowie zu beliebigen weiteren natürlichen Personen, wenn sie im Text genannt oder hinreichend charakterisiert werden.

3.4 Mail-Anhänge

Ebenso wie der Betreff und der Mailtext können Anhänge beliebige Bezüge zu verschiedenen Personen enthalten. Dies kann zum einen durch explizite Nennung von Namen oder anderer Identifikatoren in Textbestandteilen erfolgen. Aber auch viele andere Datenarten können Personenbezüge enthalten, z.B. ein Foto einer Person oder ein Röntgenbild als Teil einer Patientenakte.

4 Anonymisierung

4.1 Konzept

Das Ziel der Anonymisierung ist die Elimination des Personenbezuges aus Datensätzen, welche personenbezogene Daten enthalten. Dies hat den Zweck, dass diese Daten dann einfacher weiterverarbeitet werden können. Einfacher heißt hier: Verarbeitung auch ohne Einwilligung der betroffenen Person, über den zugestimmten Zweck hinaus, oder durch Personen, welche nicht zur Verarbeitung personenbezogener Daten geschult sind. In den meisten Fällen geht es um statistische Auswertungen, bei denen der konkrete Personenbezug nicht mehr relevant ist.

Anonymisierung kann auf verschiedene Arten erfolgen. Falls die Daten ohne weiteren Bezug zu anderen Daten verarbeitet werden, so können alle personenbezogenen Daten ersatzlos entfernt oder derart verfremdet werden, dass ein Personenbezug nicht mehr möglich ist. Bezogen auf das Beispiel von oben (Altersverifikation am Zigarettenautomaten) könnte die wie folgt erfolgen: zum einen könnte in dem anfallenden Datensatz (GEBURTSDATUM, ART DES ALTERSNACHWEISES) das Geburtsdatum gestrichen werden. Die verbleibenden Daten würden dann zwar noch Auswertungen erlauben, welche Karten (Führerschein, Bankkarte usw.) am häufigsten verwendet werden, ein Rückschluss auf die Person wäre aber nicht möglich. Zum anderen könnte statt des Geburtsdatums nur das Geburtsjahr gespeichert werden, zum Beispiel für statistische Auswertungen über das Alter von Rauchern. Auch hier ist ein Personenrückbezug nicht oder nur sehr schwer möglich.

In vielen Fällen sollen personenbezogene Daten aber später wieder mit Daten derselben Person in Beziehung gesetzt werden können, ohne auf die konkrete Person schließen zu können. Nehmen wir beispielsweise an, der Automatenaufsteller möchte beobachten, ob Kunden zwischen verschiedenen Zigarettenmarken wechseln. Der Automat muss also Daten der Art (KUNDEN-IDENTIFIKATOR, ZIGARETTENSORTE) speichern. Dabei soll der Kunden-Identifikator für jede Person immer der gleiche sein, gleichzeitig aber keinen Bezug zu dieser Person erlauben. Eine Möglichkeit wäre das Anlegen einer Tabelle, in der zu jedem Geburtsdatum (wir nehmen an dieser Stelle der Einfachheit halber an, das Geburtsdatum würde eine Person eindeutig identifizieren) ein generierter Identifikator gespeichert wird. Benutzt eine Person den Automaten ein weiteres Mal, wird der entsprechende Identifikator aus der Tabelle ausgelesen und zusammen mit der Zigarettensorte gespeichert. An dieser Stelle sieht man, dass Anonymisierung und Pseudonymisierung oft nicht scharf getrennt werden können: für den Statistiker, der später die Tabelle (KUNDEN-IDENTIFIKATOR, ZIGARETTENSORTE) auswertet, ist diese *anonymisiert*; für den Automaten, der die Zuordnung zum Geburtstag und damit zur Person besitzt, ist die Tabelle *pseudonymisiert*.

Eine andere Möglichkeit zur Behandlung der Daten in diesem Anwendungsszenario ist die Verwendung einer Einwegfunktion. Diese wird auf das Geburtsdatum angewendet, und das Ergebnis wird als Kunden-Identifikator verwendet. Dies erlaubt keinen direkten Rückschluss auf das Geburtsdatum (und damit die Person), ist aber für bestimmte Angriffe anfällig (siehe unten).

4.2 Angriffe auf Anonymisierung

So einfach eine Anonymisierung wie oben beschrieben auch klingt, häufig lassen sich in anonymisierten Daten trotzdem Personenbezüge herstellen und damit der Zweck der Anonymisierung aushebeln (siehe hierzu auch [Jen13, BZ06]). Diese Methoden werden im folgenden beschrieben.

4.2.1 Zusätzliche Datenverknüpfungen

In den meisten Szenarien sind die anfallenden Daten deutlich komplexer als bei dem Zigarettenautomaten-Beispiel. Nehmen wir als Beispiel eine medizinische Akte einer Person. Hier ist schon deutlich schwerer zu bewerten, welche Datenfelder einen Personbezug herstellen können. Bei Daten wie Name, Adresse oder Versichertennummer ist dies leicht zu erkennen. Allerdings kann auch eine seltene Krankheit oder ein charakteristisches Röntgenbild einen Rückschluss auf eine konkrete Person zulassen. Hinzu kommt bei komplexen Daten die Möglichkeit der Kombination verschiedener Datenfelder zu eindeutig Personen-identifizierenden Informationen, obwohl jedes einzelne Datenfeld für sich genommen augenscheinlich relativ unverfängliche Daten enthält. Nimmt ein Patient beispielsweise eine Reihe von an sich weit verbreiteten Medikamenten, deren Kombination aber nur bei einer bestimmten (evtl. seltenen) Krankheit verschrieben wird, so ist auf diesem Weg wieder ein Personenbezug möglich.

Zusätzlich stehen die Daten, die ein System (in vermeintlich anonymisierter Form) speichert oder sogar veröffentlicht, nie alleine da. Zusammen mit anderen Daten lässt sich oft wieder eine Verbindung zu einer konkreten Person herstellen. Ist beispielsweise von einer medizinischer Akte der Hausarzt und der Ort, und zusätzlich von diesem Hausarzt der Terminplan mit Telefonnummer bekannt, so lässt sich über die Verbindung Telefonnummer $\leftrightarrow$ Vorwahlbereich $\leftrightarrow$ Ort ein Bezug von dieser Akte zu einer Telefonnummer und weiter über das Telefonbuch zu einer Person herstellen.

Hierbei ist das Prinzip der „kleinen Mengen“ zu beachten. Je kleiner der Personenkreis ist, auf dem ein anonymisierte Eigenschaft zutrifft (z.B. Hausarzt mit nur sehr wenigen Patienten), desto leichter ist der Bezug zu einer Person. Das Problem dabei ist, dass bei der Definition der Anonymisierung oft schwer abzuschätzen ist, wie groß der adressierte Personenkreis bei einzelnen konkreten Datensätzen später sein wird.

Umgekehrt lässt sich aber festhalten, dass sich jede Form der technischen Anonymisierung brechen lässt, sofern genügend große Mengen an Zusatzinformationen zur Verfügung stehen.

4.2.2 Angriffe auf Einwegfunktionen

Neben den oben erwähnten Methoden gibt es bei der Verwendung einer Einwegfunktion zur Generierung von Identifikatoren in anonymisierten Daten noch weitere Möglichkeiten zum Brechen der Anonymisierung. Sei h eine Einwegfunktion, $p \in P$ ein personenbezogenes Datum (z.B. Geburtsdatum) und $h(p)$ dann der Identifikator für diese Person in

einem anonymisierten Datensatz. Gemäß der Einwegeigenschaft von h lässt sich p nicht direkt aus $h(p)$ berechnen. Allerdings kann ein Angreifer natürlich Werte p' ausprobieren und dann den Vergleich $h(p') \stackrel{?}{=} h(p)$ durchführen. Hier kann man zwei verschiedene Angriffsarten unterscheiden:

Verifizierungs-Angriff Bei diesem Angriff hat der Angreifer bereits einen Verdacht welche Person sich hinten dem Identifikator $h(p)$ steckt und möchte dies verifizieren. In diesem Fall ist der Angriff sehr einfach: der Angreifer muss lediglich den Kandidaten p' gemäß der obigen Formel überprüfen.

Identifizierungs-Angriff Bei diesem Angriff geht es (ohne weiteres Vorwissen) um die Identifikation von p aus der Menge P. Hier muss also für zufällig gewählte $p' \in P$ der obige Vergleich solange durchgeführt werden, bis ein p' mit $h(p') = h(p)$ gefunden wurde. Die Komplexität des Angriffes hängt also von der Größe von P ab.

Im vorher erwähnten Beispiel (Generierung des Kundenidentifikators aus dem Geburtstag) ist P also die Menge aller Geburtstage. Damit ist $|P| < 40000$ (bei einem maximalen Alter von 109 Jahren). Dies ist natürlich viel zu klein, um einem Identifizierungs-Angriff zu widerstehen. Eine übliche Anforderung für ein Schutz vor diesem Angriff lautet: $|P| > 2^{80}$.

Die Verwendung von Einwegfunktionen zur Generierung von Identifikatoren ist also nur zu empfehlen, falls die Menge P groß genug ist und Verifizierungs-Angriffe ausgeschlossen werden können.

5 Anonymisierung zur E-Mail-Filterung

5.1 Ziel der Anonymisierung

Wie bei jeder Anonymisierung, ist das Ziel den Personenbezug aus den Daten, d.h. hier der E-Mail, zu entfernen. Dabei sollen aber E-Mail, welche zu der selben Spam-Welle gehören (wir bezeichnen diese Relation im folgenden mit $\simeq$), zueinander zugeordnet werden können. Formal wird also folgendes gesucht: eine Abbildung aus der Menge aller Nachrichten M in die Menge aller Nachrichtenfingerabdrücke F:

$$f : M \to F$$

dabei gilt für alle $m \in M$: $f(m)$ enthält keinen (bzw. keinen effizient berechenbaren) Personenbezug mehr und $f(m) = f(m')$ falls $m \simeq m'$, also für zwei Nachrichten aus der selben Spam-Welle. Gesucht ist also ein Identifikator für die E-Mail.

Auch wenn diese Anforderung schwächer klingt als bei den vorher diskutierten Szenarien findet man hier die gleichen Probleme wie zuvor. Insbesondere haben wir bisher nur Anonymisierung von strukturierten Daten betrachtet, bei denen bei jedem Datum der Typ

und die Art der Inhalts (z.B. Geburtstdatum) zugeordnet war. Dies erleichtert natürlich die Einschätzung des Schutzbedarfes des Datums und damit die notwendige Modifikation. Im Gegensatz dazu besteht eine E-Mail zum großen Teil aus unstrukturierten Daten. Insbesondere der Betreff (engl. *Subject*) und der Inhalt (engl. *Body*) sind Freitexte, welche beliebigen Inhalt annehmen können. Daher muss bei der Anonymisierung davon ausgegangen werden, dass diese schützenswerte Informationen enthalten, die nicht an den Spam-Detektor weitergegeben werden dürfen.

Ein weiteres Problem bei der gesuchten Anonymisierungs-Funktion ist, dass Spam-Versender sich der Existenz von derartiger Spam-Detektoren bewusst sind. Deshalb sind E-Mail, die zu der selben Spam-Welle gehören nicht gleich sondern werden auf geschickte Art und Weise variiert.

Bei der Vorstellung und Diskussion verschiedener Anonymisierungs-Funktion muss man sich also jeweils 2 Fragen stellen:

1. Ist es möglich (bzw. mit welchem Aufwand ist es möglich) aus dem Fingerabdruck der Nachricht, Teile der Original-Nachricht (um damit potentiell einen Personenbezug) zu gewinnen?
2. Wie präzise klassifiziert der Fingerabdruck die Nachrichten? Also:
   - Wie robust ist die Abbildungen gegen Veränderungen von Nachrichten innerhalb einer Spam-Welle?
   - Wie viele andere Nachrichten werden fälschlicherweise als zu der Spam-Welle gehörig eingestuft?

Im folgenden sollen nun verschiedene Anonymisierungsverfahren vorgestellt und entsprechend analysiert werden.

5.2 Anonymisierungsfunktionen

5.2.1 Kryptographische Hash-Funktionen über E-Mail-Bestandteile

Eine der einfachsten Arten die Anonymisierung durchzuführen, ist die Verwendung einer kryptographischen Hash-Funktion wie SHA-1 [EJ01] oder MD5 [Riv92].

Die erste Variante wäre hier die gesamte E-Mail-Nachricht (also E-Mail-Header und -Body) zu hashen und das Ergebnis als Fingerabdruck zu verwenden. Da aber der Versandweg zu unterschiedlichen Empfängern und Zeitstempel selbst für identische Kommunikationspartner verschieden sind, ist dieser Gesamt-Hash für zwei E-Mails nie gleich. Dieser Identifikator eignet sich also offensichtlich nicht zur Klassifizierung von E-Mails.

Die zweite Variante ist die Unterteilung der E-Mail in ihre semantischen Bestandteile: E-Mail-Body, Sender, Empfänger, Betreff sowie sonstige E-Mail-Header und die Hash-Bildung für jeden einzelnen dieser Teile. Der Fingerabdruck würde dann also aus einen Hash-Tupel bestehen.

Für die E-Mail-Header, welche den Versandweg inkl. Zeitstempel enthalten, gilt das gleiche wie oben: diese sind für zwei E-Mails nie gleich, der Hash darüber also ungeeignet für die Klassifizierung.

Der Empfänger (genauer: der Inhalt des `To`-Feldes im E-Mail-Header) kann ein Hinweis auf eine Spam-Kampagne sein, falls hier ein generischer Eintrag (z.B. „Alle Sparkassen-Kunden“) enthalten ist. In den meisten Fällen ist aber auch bei Spam-Nachrichten hier der tatsächliche Empfänger enthalten, weswegen es für die Klassifizierung ungeeignet ist.

Der Absender (genauer: der Inhalt des `From`-Feldes im E-Mail-Header) kann ein Indikator für eine Spam-Welle sein. Typischerweise wird dieser innerhalb einer Spam-Welle nicht geändert. Allerdings wird bei Spam-Mails auch häufig ein falscher, legitimer Absender verwendet. Die Klassifizierung alleine aufgrund des Hashes des Absenders kann also zu vielen *false positive* Einstufungen führen. Das Hauptproblem aus Datenschutzsicht ist der Personenbezug dieses Feldes, relevant natürlich primär bei Nicht-Spam-Mails. Da E-Mail-Adressen teilweise sehr kurz sind, ist hier ein Identifizierungs-Angriff auf den Hashwert des Absenders möglich. Zusätzlich ist natürlich immer ein Verifizierungs-Angriff möglich. Hierfür muss der Spam-Detektor bereits die Kenntnis einer konkreten E-Mail-Adresse habe und gewinnt das Wissen, dass von dieser Adresse eine Nachricht verschickt wurde.

Der Betreff ist grundsätzlich ein guter Indikator für eine Spam-Welle. Durch die meist große Länge ist hier die Gefahr eines Identifizierungs-Angriff auf den Hash des Betreffs auch sehr gering. Auch ein Verifizierungs-Angriff ist hier typischerweise nicht möglich. Allerdings ist der Hash-Wert natürlich gegen keinerlei Veränderung des Textes resistent. Da aber Spammer den Betreff oftmals innerhalb einer Spam-Welle variieren (z.B. Betreff enthält den Namen des Empfängers oder eine eindeutige ID), ist hier die Gefahr eines *false negative* sehr groß. Umgekehrt kann es hier aber auch zu *false positive* Einstufungen kommen, falls der Betreff nur aus einen kurzen Standardtext, wie „Information“ oder „Wichtig“, besteht.

Für den E-Mail-Body gilt fast dasselbe wie für den Betreff: Angriffe auf dem Hash sind praktisch nicht möglich, allerdings sind Variationen innerhalb einer Spam-Welle sehr häufig und führen zu unterschiedlichen Hash-Werten. Typische Variationen, die von Spammer vorgenommen werden, sind: Umstellung von Absätzen, Einfügen des Namens des Empfängers, zufällige Menge an unsichtbaren Zeichen (z.B. Leerzeichen am Ende) oder Hyperlinks, die eine ID enthalten, um den Empfänger wiederzuerkennen.

Zusammenfassend lässt sich sagen, dass die Hash-Wert-Tupel auf E-Mail-Bestandteilen keinen guten Fingerabdruck ergibt. Die Hash-Werte sind entweder gegen Personenbezugs-Angriffe anfällig oder beziehen sich auf Daten, die (zumindest häufig) innerhalb einer Spam-Welle nicht identisch sind und damit keine Identifizierung ermöglichen.

5.2.2 Eigenschaften des Textes

Wie im vorherigen Abschnitt gesehen, ist insbesondere der E-Mail-Betreff und der E-Mail-Body für die Identifizierung einer E-Mail geeignet. Andere Daten sind entweder in jeder E-Mail unterschiedlich (Versandwegprotokoll) oder erlauben Identifizierungs-Angriffe (Sender oder Empfänger).

Die simple Hash-Bildung über Betreff oder Body hat das Problem, dass selbst kleinste Änderungen an den Texten, den Hash-Wert verändern und damit eine Klassifizierung unmöglich machen.

Eine anderer Ansatz zur Berechnung des Fingerabdrucks besteht darin, Eigenschaften der Nachricht zu verwenden, welche sich innerhalb einer Spam-Welle nicht oder zu geringfügig ändern. Beispiele für solche Eigenschaften sind:

- Anzahl der Wörter (bei Betreff, Body)
- Anzahl der Sätze (bei Body)
- Anzahl der Absätze (bei Body)
- Häufigkeitsverteilung der Buchstaben (bei Body)

Bei diesen Eigenschaften ist ein Angriff zur Wiederherstellung des Personenbezuges fast ausgeschlossen. Weiterhin sind diese Metriken sehr robust gegen Veränderungen innerhalb einer Spam-Welle (wie im Abschnitt zuvor diskutiert). Da hier keine Hash-Werte verwendet werden, sind auch unscharfe Vergleiche möglich. Allerdings sind diese Metriken auch nicht sehr präzise: die Wahrscheinlichkeit, dass eine E-Mail außerhalb der Spam-Welle den gleichen Fingerabdruck hat, ist sehr groß.

5.2.3 Anonymisierungsverfahren höherer Ordnung

Eine Kombination der Ideen der beiden vorherigen Kapitel lautet wie folgt: unterteile den E-Mail-Body (evtl. auch den E-Mail-Betreff) in kleinere „Häppchen“ und bilde für jeden Teil einen Hash-Wert. Damit erhält man eine Folge von Hash-Werten als Fingerabdruck der E-Mail. Mögliche Unterteilungen sind dabei:

- MIME-Teile (getrennt durch *Multipart Boundary*)
- Absätze (getrennt durch Leerzeile)
- Sätze (getrennt durch .)
- Wörter (getrennt durch Leerzeichen)

Dabei stellen MIME-Teile typischerweise eine zu grobe Unterteilung (ähnlich wie gesamter E-Mail-Body) dar. Wörter wiederum erlauben aufgrund ihrer geringen Länge einen Identifizerungs-Angriff. Außerdem ist hier der Gesamtfingerabdruck sehr groß (Anzahl der Wörter $\times$ Hash-Wert-Länge).

Die Hash-Werte der Sätze oder Absätze hingegeben sind resistent gegen Identifizierungs-Angriffe, sofern man sehr kurze (kürzer als ca. 15. Zeichen) Sätze/Absätze bei der Hash-Wertbildung ignoriert. Weiterhin ist die Gefahr eines *false positive* sehr gering: zwei E-Mails, die nichts miteinander zu tun haben, werden nur sehr geringer Wahrscheinlichkeit den gleich Satz/Absatz enthalten.

Bei der Verwendung von Hash-Werten aller Sätze erreicht man offensichtlich eine höhere Präzision als bei Absätzen, welche aber durch geringere Effizienz erkauft wird.

Gegen welche Veränderung in der E-Mail eine Menge von Hash-Werten von Sätzen/Absätzen robust ist und wie eine Klassifizierung von E-Mails erfolgt, hängt auch von der Organisation der Menge der Hash-Werte ab. Dies soll im folgenden diskutiert werden.

Werden die Hash-Werte als **Folge** von Werten übertragen, so muss der Empfänger beim Vergleich mit einer Referenz-Mail (d.h. Mail, die zu einer Spam-Welle gehört) bei dieser ebenfalls eine Folge von Hash-Werten der Sätze/Absätze berechnen und dann diese Werte untereinander vergleichen. Dabei lassen sich folgende Eigenschaften feststellen:

- Satz/Absatz identisch und an der gleichen Position
- Satz/Absatz identisch aber an anderer Position
- Satz/Absatz unterschiedlich

Man kann davon ausgehen, dass die Veränderungen innerhalb einer Spam-Welle zum einen im Permutieren der Reihenfolge der Sätze/Absätze bestehen und „echte“ Veränderungen nur einen sehr kleinen Teil der Sätze/Absätze betrifft. Damit lassen sich hier mittels eines Schwellwertvergleich mit sehr hoher Präzision E-Mail klassifizieren. Der offensichtliche Nachteil dieses Systems ist die Laufzeit: für jede E-Mail sind $O(n^2)$ Vergleiche (mit n Anzahl der Sätze/Absätze) notwendig.

Eine Modifikation dieses Verfahrens überträgt die Hash-Werte als **geordnete Folge**. Dabei kann nur noch die Identität/Ungleichheit von Sätzen/Absätzen festgestellt werden, was für unsere Klassifizierung ausreichend ist. Die Laufzeit beim Spam-Erkenner ist hier auf $O(n)$ Vergleiche reduziert, allerdings benötigt der E-Mail-Provider pro E-Mail einen zusätzlichen Sortierschritt (bei guten Sortieralgorithmen $O(n \log(n))$).

Eine weitere Möglichkeit die Menge der Hash-Werte anzuordnen ist schließlich ein **Hash-Baum**. Hier könnte beispielsweise die Wurzel den Hash-Wert der gesamten Nachricht enthalten, die Ebene darunter die Hash-Werte der Absätze und die Ebene darunter die Hash-Werte der Sätze. Diese Struktur hat den Vorteil, dass der Spam-Erkenner adaptiv arbeiten kann. Bei hoher Auslastung führt er nur einen (unpräzisen) Vergleich auf der obersten Ebene durch; bei geringerer Auslastung kann er auf tieferen Ebenen durch größere Anzahl von Vergleichen einer höhere Präzision erreichen.

6 Verwandte Arbeiten

Analog zur hier vorgestellten Architektur zur effektiven Anonymisierung von E-Mail-Daten existiert eine Reihe von Vorarbeiten, die das Problem der Anonymisierung oder hinreichend abstrakter Pseudonymisierung ebefalls beleuchtet haben. Zur Begriffsdefinition maßgeblich sind die Arbeiten von Pfitzmann und Hansen zu nennen, deren Werk zur Terminologie von Anonymität und Pseudonymität einen wichtigen Beitrag zur Formalisierung der jeweiligen Konzepte geleistet hat (vgl. [PH10]). Auch die Arbeiten von

Kerschbaum haben aufgezeigt, wie sich hinreichend anonymisierte Daten dennoch effektiv und effizient zur Weiterverarbeitung durch Dritte eignen. Sein Schema zur Pseudonymisierung von Ereignissen auf einer Zeitachse illustriert dieses Konzept vorbildlich (vgl. [Ker07]). Schließlich finden sich für den Bereich der effektiven Anonymisierung von IP-Adressen in Netzwerk-Verkehrsdaten zahlreiche Arbeiten. Als gute Einführung sei hier auf Xu et al. verwiesen (vgl. [XFAM01]).

7 Zusammenfassung und Ausblick

Das in dieser Arbeit vorgestellte Konzept zur Anonymisierung von E-Mails mit dem Ziel der effektiven Weiterverarbeitung durch Dritte unter Eliminierung des Personenbezuges zeigt auf, wie eine mögliche Ausgestaltung einer datenschutzfreundlicheren Informationsverarbeitung aussehen könnte. Durch die beschriebenen Techniken lässt sich der Personenbezug unter Umständen soweit aus einer anonymisierten E-Mail herauslösen, dass eine Rückführung dieser Mail zu den originären Akteuren nahezu unmöglich wird. Gleichzeitig bleibt das anonymisierte Ergebnis zur Weiterverarbeitung verfügbar, etwa um – wie im beschriebenen Szenario veranschaulicht – eine verteilte Architektur zur Spam-Erkennung zu realisieren.

Es zeichnet sich ab, dass die gesetzliche Umgebung der einschlägigen datenschutzrechtlichen Vorschriften in naher Zukunft deutlich stärkere Auflagen an die verarbeitenden Organisationen personenbezogener Daten stellen werden. Dies erfordert geeignete Lösungen für verschiedenste Szenarien und Anwendungsgebiete, stets mit ähnlichen Anforderungen wie im Beispielszenario beschrieben. Entsprechend steht zu erwarten, dass der Bedarf an geeigneten Anonymisierungs- und Pseudonymisierungsverfahren analog der in dieser Arbeit vorgestellten Techniken deutlich steigen wird.

Entsprechend besteht die naheliegende Weiterführung dieser Arbeit in der Implementierung und Praxiserprobung des vorgestellten Anonymisierungssystems. Insbesondere ist eine hinreichende Erprobung der Anonymisierungs- und Re-Identifizierungsmöglichkeiten von E-Mail-Nachrichten anhand geeignet großer und realistischer Datensätze erforderlich.

Für zukünftige Forschungstätigkeiten empfiehlt sich ferner die umfassende Analyse äquivalenter Anonymisierungsverfahren, unter Berücksichtigung der Stärke ihrer Anonymisierungswirkung, ihrer Rückauflösbarkeit, ihrer Verkettbarkeit mit anderen Datensätzen und ihrer Eignung für weiterführende Verarbeitungsschritte. Es bedarf somit der ausführlichen Erfassung, Katalogisierung, Formalisierung, Implementierung und Umsetzung möglichst vieler entsprechender Anonymisierungstechniken. Die Durchführung und entsprechend die Förderung von Forschungsarbeiten in diesem Kontext ist folglich dringend erforderlich.

8 Danksagung

Die in dieses Dokument eingebrachten Arbeiten von M. Jensen wurden vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Projektes “Monitoring durch Informationsfusion und Klassifikation zur Anomalieerkennung (MonIKA)” gefördert.

Literatur

[Bun78] Bundesrepublik Deutschland. Bundesdatenschutzgesetz (BDSG). 1978.

[BZ06] Michael Barbaro und Tom Zeller. A Face Is Exposed for AOL Searcher No. 4417749. *http://select.nytimes.com/gst/abstract.html?res=F10612FC345B0C7A8CDDA10894DE404482*, 2006.

[Cro82] D. Crocker. *STANDARD FOR THE FORMAT OF ARPA INTERNET TEXT MESSAGES*. Internet Engineering Task Force, August 1982. Obsoleted by RFC 2822, updated by RFCs 1123, 2156, 1327, 1138, 1148.

[EJ01] D. Eastlake 3rd und P. Jones. *US Secure Hash Algorithm 1 (SHA1)*. Internet Engineering Task Force, September 2001. Updated by RFCs 4634, 6234.

[Jen13] Meiko Jensen. Challenges of Privacy Protection in Big Data Analytics. In *Proceedings of the IEEE 2nd International Congress on Big Data (BigData 2013)*, Seiten 235–238, 2013.

[Ker07] Florian Kerschbaum. Distance-preserving pseudonymization for timestamps and spatial data. In *WPES*, Seiten 68–71, 2007.

[PH10] Andreas Pfitzmann und Marit Hansen. A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf, August 2010. v0.34.

[Res01] P. Resnick. *Internet Message Format*. Internet Engineering Task Force, April 2001. Obsoleted by RFC 5322, updated by RFCs 5335, 5336.

[Res08] P. Resnick. *Internet Message Format*. Internet Engineering Task Force, Oktober 2008.

[Riv92] R. Rivest. *The MD5 Message-Digest Algorithm*. Internet Engineering Task Force, April 1992. Updated by RFC 6151.

[uE] Europäisches Parlament und Europäischer Rat. Richtlinie 95/46/EG zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr. *COM/95/375 COM/92/422 COM/90/314-2*.

[XFAM01] Jun Xu, Jinliang Fan, Mostafa Ammar und Sue B. Moon. On the design and performance of prefix-preserving IP traffic trace anonymization. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, IMW ’01, Seiten 263–266, New York, NY, USA, 2001. ACM.