iubh INTERNATIONALE HOCHSCHULE

IT & TECHNIK

Grundbegriffe der Anonymisierung personenbezogener Daten

Ralf Kneuper

IUBH Internationale Hochschule

Main Campus: Erfurt

Juri-Gagarin-Ring 152

99084 Erfurt

Tel. 0421 – 1669 85 23

Kontakt/Contact: k.janson@iubh.de/

Autorenkontakt/Contact to the author(s):

Prof. Dr. Ralf Kneuper

IUBH Internationale Hochschule / IUBH University of Applied Sciences

Campus Bad Reichenhall

Kaiserplatz 1

83435 Bad Reichenhall

Telefon: +49- 173-3432005

Email: r.kneuper@iubh-fernstudium.de

IUBH Discussion Papers, Reihe: IT & Technik, Vol. 2, Issue 1 (1/2020)

ISSN-Nummer: **2512-319X**

Website: https://www.iubh-university.de/forschung/publikationen/

GRUNDBEGRIFFE DER ANONYMISIERUNG PERSONENBEZOGENER DATEN

Ralf Kneuper

Abstract:
Many organisations collect data about their processes, customers, the use of their products, and many other topics in order to analyse these data in the context of data mining, big data, machine learning, and similar approaches. However, in most cases such data refer to individual people, and the persons concerned rightly expect that their data are protected adequately and kept private. The resulting limitations on the use of the data often lead to conflicts with and limitations of the data analysis. A technique that helps to overcome these conflicts and limitations is the anonymisation of the data, modifying the data in such a way that they no longer refer to individuals but still allow certain forms of analysis. However, anonymising data turns out to be far more complex than just removing names and other identifiers, and there are many examples where apparently anonymised data were de-anonymised and could be assigned to the individuals concerned after all. Therefore, a number of systematic techniques for evaluating and achieving anonymity, such as k-anonymity and differential privacy, have been developed for this purpose. The current report therefore gives a first overview of the concept of anonymisation, the remaining threats to anonymity, and the main approaches used for anonymising data. The paper concludes with a summary of open research questions for further work.

Keywords:
anonymisation; anonymity; differential privacy; data protection; privacy

Einführung

Anonymisierung von Daten dient dazu, bei Daten, die sich auf bestimmte Personen beziehen, diesen Personenbezug wieder zu entfernen. Dies ist in erster Linie dann nötig, wenn Daten aus dem Produktivbetrieb für statistische Auswertungen und andere Datenanalysen genutzt werden sollen, beispielsweise bei der Veröffentlichung von Gesundheitsdaten für Forschungszwecke, oder der Auswertung von Kunden- und Bestelldaten. Aus Sicht des Datenschutzes ist in derartigen Fällen üblicherweise die Anonymisierung der Daten erforderlich, um die betroffenen Personen vor Missbrauch ihrer Daten zu schützen. Damit werden einerseits gesetzliche Anforderungen wie die Datenschutzgrundverordnung (DSGVO) erfüllt, andererseits auch im Eigeninteresse der Unternehmen Datenskandale vermieden.

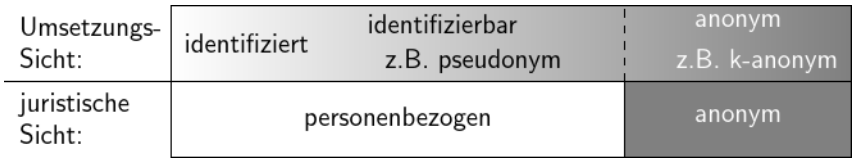
Allerdings ist die Anonymisierung von Daten wesentlich komplexer, als das auf den ersten Blick erkennbar ist. Dieser Beitrag soll daher eine Einführung in dieses Thema und die damit verbundenen Herausforderungen und Lösungsansätze geben, einschließlich Hinweisen auf offene Fragen, die weitere Forschung erfordern.

Identifizierbarkeit, Pseudonymisierung und Anonymisierung

Identifizierbarkeit und Anonymisierung

Die charakterisierende Eigenschaft von personenbezogenen Daten ist, dass sie sich auf eine identifizierte oder identifizierbare Person beziehen. Diese scheinbar eindeutige Definition ist bei genauerem Hinsehen allerdings deutlich komplexer, als dies zuerst den Anschein hat. Der Bezug auf eine *identifizierte* Person ist meist relativ problemlos zu erkennen, denn eine solche Identifikation geschieht durch den Namen oder einen ähnlichen Identifikator. Beim Bezug auf eine *identifizierbare* Person dagegen kann diese Identifizierung mit sehr unterschiedlichen Schwierigkeiten und Aufwand verbunden sein, so dass die Abgrenzung zwischen identifizierbaren und nicht identifizierbaren Daten in der praktischen Umsetzung oft nicht eindeutig möglich ist, auch wenn der Unterschied juristisch gesehen sehr eindeutig ist. Abbildung 1 stellt diese Situation grafisch dar. Eine sehr detaillierte Analyse dieser Unterscheidung zwischen identifizierenden und nicht identifizierenden Daten findet man beispielsweise in (Oswald, 2013).

Abbildung 1: Stufen der Identifizierbarkeit



Identifizierer

Identifizierer, die eine bestimmte Person identifizieren können, sind neben offensichtlichen Angaben wie dem Namen beispielsweise auch die Telefonnummer, die IP-Adresse, eine Steuer-oder Versicherungsnummer, sowie die Seriennummer IMEI und die Werbe-ID von Smartphones und ähnlichen Geräten. Dabei gibt es aber unterschiedliche Interpretationen in verschiedenen Ländern und Regionen. Insbesondere bei IP-Adressen war die Identifizierer-Eigenschaft in der EU bis zu einem entsprechenden EuGH-Urteil sowie der DSGVO umstritten, ist in der EU jetzt aber unzweifelhaft. In den USA dagegen gilt eine IP-Adresse nur mit Einschränkungen (abhängig von der relevanten gesetzlichen Grundlage etc.) als Identifizierer. Zumindest nach EU-Recht gilt aber normalerweise: Wenn man verschiedene Datensätze, die dieselbe Person betreffen, einander zuordnen kann, dann handelt es sich um Daten, die einer identifizierbaren Person zuordenbar sind, und nicht mehr um anonyme Daten. Das gilt beispielsweise auch bei Cookies. Ein Cookie, der zwischen Sitzungen erhalten bleibt, dient dazu, eine Person zu identifizieren und ist damit ein Identifizierer, auch wenn der Name der Person dabei unbekannt bleibt.

Anonyme und anonymisierte Informationen

Das Gegenteil von personenbezogenen Daten sind anonyme Informationen, definiert als Informationen, die sich *nicht* auf eine identifizierte oder identifizierbare Person beziehen. Dazu gehören neben Informationen über nicht identifizierbare Personen (z.B. „ein (nicht genauer identifizierter) Käufer des Produktes X ist sehr zufrieden damit“) oder über Gruppen von Personen (z.B. „20% der Käufer des Produktes X sind sehr zufrieden damit“) auch Informationen völlig ohne Personenbezug (z.B. „das Produkt X kostet €42).

Abhängig vom Schwierigkeitsgrad der Identifizierung spricht man auch von *absoluter* Anonymität, bei der eine Zuordnung zur betroffenen Person nicht möglich ist, und *relativer* Anonymität, bei der diese Zuordnung mehr

oder weniger schwierig, aber grundsätzlich möglich ist. Dabei ist allerdings fraglich, inwieweit eine absolute Anonymität tatsächlich erreichbar ist, denn bei entsprechendem Zusatzwissen, beispielsweise weil man die betroffene Person kennt, ist eine Re-Identifikation, fast immer möglich.

Ein spezieller Fall von anonymen Informationen sind anonymisierte Informationen, also Informationen, die ursprünglich personenbezogen waren, durch Informationsreduktion aber zu anonymen Informationen wurden.

Zu naiv wäre allerdings die Vorstellung, dass man durch einfaches Weglassen des Identifizierers bei einem Datensatz, also beispielsweise des Namens, grundsätzlich anonyme Informationen erhalte. Eine einfache Beispielrechnung zeigt, dass das oft nicht der Fall ist: Behält man bei einem Datensatz den Geburtstag (ohne Jahr) und die Postleitzahl, so gibt es in Deutschland etwa $30.000 * 365 \approx 11.000.000$ Kombinationen dieser Daten, bei rund 83 Mio. Einwohnern also im Durchschnitt eine Beschränkung auf nur etwa 7-8 Einwohner alleine durch diese beiden Angaben. Schon durch statistische Schwankungen ist daher damit zu rechnen, dass man in einer Reihe von Fällen die betroffene Person durch diese Angaben eindeutig identifizieren kann, und in den anderen Fällen benötigt man nur wenig zusätzliche Information zur Identifikation, beispielsweise das (ungefähre) Geburtsjahr. Eine in (Sweeney, 2002) beschriebene Studie in den USA zeigte, dass dort 87% der Bevölkerung alleine durch die Angabe von Postleitzahl, Geburtsdatum und Geschlecht eindeutig identifizierbar sind.

Beispiel: Anonymisierung einer Log-Datei
In einem Informationssystem werden Log-Dateien geschrieben, die die wesentlichen Aktivitäten und Änderungen im System protokollieren. Um diese Log-Dateien mit Techniken des Process Mining auszuwerten, war geplant, die in der Log-Datei protokollierten Namen der jeweiligen Benutzer durch ein zufällig generiertes Pseudonym zu ersetzen. Damit sollte erreicht werden, dass verschiedene Aktivitäten der gleichen Person einander zugeordnet werden können, ohne dass die Datenanalytiker die einzelnen Datensätze der Person zuordnen können.

Bei der Pilotierung des Vorgehens stellte sich aber schnell heraus, dass dies nicht ausreichte, da in den protokollierten Änderungen selbst teilweise Angaben über die jeweilige Person enthalten waren, wenn diese beispielsweise ihre E-Mail-Adresse geändert hatte.

Re-Identifikation
Neben leicht zu übersehenden identifizierenden Eigenschaften wie im Beispiel besteht die zweite große Gefahr für anonymisierte Daten darin, dass die Daten mit anderen verfügbaren Daten abgeglichen werden, die beispielsweise Name, Adresse und Geburtsdatum enthalten. Gemeinsam ermöglichen die beiden Datensammlungen dann eine Re-Identifikation der zu den sensiblen Daten gehörigen Personen. Dass dies nicht nur eine theoretische Möglichkeit ist, sondern ein reales Risiko, zeigen viele Beispiele, u.a. der ebenfalls in (Sweeney, 2002) beschriebene Fall, bei dem anonymisierte medizinische Daten (Diagnose, Medikation etc.) von rund 135.000 Staatsbediensteten in Massachusetts für Forschungszwecke veröffentlicht wurden. Durch Abgleich mit einem Wählerverzeichnis konnten die Daten in vielen Fällen anhand der gemeinsamen Attribute Postleitzahl, Geburtsdatum und Geschlecht zugeordnet und die Anonymität aufgehoben werden, darunter auch für den damaligen Gouverneur von Massachusetts. In einem anderen bekannten Fall veröffentlichte Netflix anonymisierte Daten ihrer Nutzer und deren Filmkonsums. Durch Abgleich gegen öffentlich verfügbare Daten in der Filmdatenbank IMDB, insbesondere dort veröffentlichte Filmbewertungen, konnte die Anonymisierung in vielen Fällen aufgehoben werden (Narayanan, Shmatikov 2008).

Pseudonymisierung
Ein Ansatz, um die Identifikation der betroffenen Personen für Unberechtigte zu erschweren, gleichzeitig für Berechtigte bei Bedarf aber weiterhin zu ermöglichen, ist die Pseudonymisierung der Daten. Darunter versteht man die Verarbeitung in einer Weise, dass die Identifizierung ohne zusätzliche, separat gespeicherte und gesicherte, Informationen (die Zuordnung von Pseudonym zu Identität) nicht möglich ist. Damit gelten die Daten immer noch als personenbezogen, aber die Pseudonymisierung ist ein Ansatz, um diese Daten wie gefordert zu schützen.

Die Herausforderungen an eine effektive Pseudonymisierung sind damit ähnlich wie die an eine effektive Anonymisierung, da die pseudonymen Daten für Dritte, die keinen Zugang zu den Zusatzinformationen haben, genauso wenig zuordenbar sein sollten wie anonyme Daten. Einen Überblick über relevante Anforderungen an eine effektive und datenschutzkonforme Pseudonymisierung findet man beispielsweise in (Schwartmann 2018).

Beispiele für Pseudonyme sind:

- IP-Adresse, evtl. in Kombination mit Datum und Uhrzeit
- Identifikator in einem Cookie
- kryptografische Identität, also ein öffentlicher Schlüssel (vgl (Boehme, Pesch 2017)
- kryptografischer Hash eines Identifizierers oder anderer Eigenschaften wie Wohnort und/oder Geburtsjahr. Dieser hilft beispielsweise, verschiedene Datensätze einander zuzuordnen, ohne den Betroffenen genau zu identifizieren. Auch dies kann allerdings zu Problemen führen, wenn die Anzahl der möglichen Identifizierer nicht sehr groß ist und damit eine Tabelle aller Werte erstellt werden kann, wie beispielsweise (Pandurangan 2014) für eine anonymisierte Aufstellung von Taxifahrten in New York zeigt. Man spricht hier auch von einem *Aufzählungsangriff*, da in diesem Fall die Pseudonymisierung durch eine Aufzählung aller Datensätze mit den zugeordneten Pseudonymen aufgehoben werden kann. Diese Form von Pseudonymen kann beispielsweise hilfreich sein bei der Zusammenführung von Daten, die an verschiedenen Quellen entstehen, beispielsweise Patientendaten verschiedener Krankenhäuser, die damit ohne zentrales Verzeichnis einheitliche Pseudonyme generieren können. Eine derartige Funktionalität ist beispielsweise mit dem 2019 veröffentlichten Entwurf des Digitale-Versorgung-Gesetzes[1] des deutschen Bundesgesundheitsministeriums vorgesehen, vorerst allerdings ohne technische Vorgaben zur Umsetzung. Der Ansatz ist allerdings nur dann sicher, wenn die Anzahl der möglichen verschiedenen Identifizierer hinreichend groß ist, so dass die Pseudonymisierung nicht über eine Tabelle aller Werte, also einen Aufzählungsangriff, aufgehoben werden kann. Selbst bei IPv4-Adressen ist eine solche Tabelle noch ohne großen Aufwand machbar.[2]
- ein beliebiges (zufälliges) Pseudonym mit Abbildungstabelle

Eine Variante der Pseudonymisierung ist die Nutzung von *Tokens*, wie sie in den Sicherheitsstandards der Zahlkartenindustrie (PCI-DSS) beschrieben ist, siehe (PCI Security Standards Council 2011) und (PCI Security Standards Council 2015).

Insgesamt wird deutlich, dass Anonymisierung und Pseudonymisierung von Daten komplexe Themen sind. Das folgende Kapitel behandelt daher Ansätze zur systematischen Anonymisierung und Pseudonymisierung von Daten ausführlicher.

Anonymisierung von personenbezogenen Daten

Je nach Aufgabenstellung ist der Personenbezug von Daten für eine Anwendung erforderlich oder auch nicht. Während beispielsweise für die Bearbeitung von Bestellungen natürlich die Identität der Kunden bekannt sein muss, ist das bei statistischen Auswertungen der Bestellungen meist nicht der Fall. Bei solchen statistischen Auswertungen ist daher eine Anonymisierung der Daten angemessen, insbesondere wenn die Daten nicht mehr für operative Zwecke benötigt werden, aber weiterhin Datenauswertungen gewünscht sind. Ähnliches gilt, wenn Daten veröffentlicht werden sollen, beispielsweise für Forschungszwecke oder zur Bedienung von Anfragen zur Informationsfreiheit.

Allerdings reicht es für die Anonymisierung meist nicht aus, einfach die Identifizierer wie Name oder Kundennummer wegzulassen. Mit Hilfe einer Kombination anderer Datenfelder wie Postleitzahl, Geburtsdatum oder Geschlecht, der sogenannten Quasi-Identifizierer, kann die relevante Person oft wie oben erläutert trotzdem eindeutig identifiziert werden oder zumindest auf einen kleinen Kreis von Kandidaten eingeschränkt werden. Die folgende Beschreibung soll daher einen ersten Einblick in die Anonymisierung geben, damit bei Bedarf geeignete Verfahren und Algorithmen ausgewählt werden können. Ähnlich der Verschlüsselung von Daten ist auch hier stark davon abzuraten, eigene Verfahren zur Anonymisierung zu nutzen, sondern man sollte bewährte und geprüfte Verfahren einsetzen.

---

[1] Verfügbar unter http://dip21.bundestag.de/dip21/btd/19/134/1913438.pdf, abgerufen am 21.11.2019

[2] Der Autor hat selbst eine solche Tabelle für IPv4-Adressen auf einem gewöhnlichen, schon etwas älteren Laptop mit Hilfe eines einfachen PHP-Programms erstellt. Das Programm benötigte dafür etwa 5 Stunden.

Aus Datenschutzsicht ist dabei berücksichtigen, dass auch Anonymisierung eine Form der Verarbeitung personenbezogener Daten ist, d.h. auch hierfür muss es eine Rechtsgrundlage geben. Diese wird aber meist durch das „berechtigte Interesse" des Verantwortlichen gegeben sein, zumal die Anonymisierung ja meist auch im Interesse der Betroffenen liegt.

Grundbegriffe
Im Kontext der Bewertung der Anonymität von Daten wird zwischen verschiedenen Typen von Attributen von personenbezogenen Daten unterschieden:

- *Identifizierer* dienen dazu, einen Datensatz (in Datenbanken meist als Tupel bezeichnet) einer bestimmten Person zuordnen. Dazu gehört der Name als offensichtlicher Identifizierer, auch wenn er in der Praxis oft nicht eindeutig ist, sowie künstliche Identifizierer wie Kundennummer, Matrikelnummer, Steuernummer etc. Da bei der Anonymisierung die Identifikation ja gerade nicht gewünscht ist, werden hier die Identifizierer entfernt und daher im Folgenden nicht weiter betrachtet.
- Neben den geplanten Identifizierern gibt es in der Praxis auch viele andere Attribute, die zumindest in Kombination als Identifizierer dienen können und die daher als *Quasi-Identifizierer* bezeichnet werden.
- Die Identifizierung, also die Zuordnung eines Tupels zu einer Person, ist nur dann problematisch, wenn das Tupel auch vertrauliche oder *sensible Daten* enthält. Je nach Rahmenbedingungen können diese sich mit den Quasi-Identifizierern überschneiden. Die sensiblen Daten sind die zu schützenden Daten, deren Zuordnung zu einer Person durch die Anonymisierung verhindert werden soll.
- Schließlich gibt es möglicherweise auch *nicht sensible Daten*, deren Zuordnung zu einer Person unproblematisch wäre. Allerdings ist eine Zuordnung alleine der sensiblen Daten meist nicht einfach möglich, so dass die nicht sensiblen Daten meist zusammen mit den sensiblen Daten geschützt bzw. anonymisiert werden. Auch die nicht sensiblen Daten sind im Kontext der Anonymisierung allerdings nicht relevant und werden daher im Folgenden nicht explizit betrachtet.

Wenn Daten anonymisiert wurden, ergeben sich folgende Bedrohungen der Anonymität:

- Re-Identifikation (De-Anonymisierung) ist möglich, d.h. die anonymisierten Daten können, zumindest teilweise, wieder bestimmten Personen zugeordnet werden.
- Attribut-Ableitung: Es können Informationen über Einzelpersonen abgeleitet werden, beispielsweise weil ihre Identität auf eine Gruppe eingeschränkt werden kann, die alle die entsprechende Eigenschaft haben.
- Ableitung der Mitgliedschaft: Es kann abgeleitet werden, ob eine bestimmte Person zu einer betrachteten Datenmenge gehört. Auch diese Information kann problematisch sein, wenn es sich beispielsweise um die Besucher einer HIV-Beratungsstelle handelt. Neben der Mitgliedschaft kann auch die nicht-Mitgliedschaft eine sensible Information sein.

Dabei ist nicht nur die sichere Aufdeckung der Anonymität ein potentielles Problem, sondern auch bereits, wenn eine sensible Information mit hoher Wahrscheinlichkeit abgeleitet werden kann.

Zur Aufhebung einer bestehenden Anonymisierung gibt es eine Reihe verschiedener Techniken, beispielsweise die Verlinkung verschiedener Datensätze, die jeweils für sich unproblematisch sind, gemeinsam aber die Zuordnung sensibler Informationen erlauben, wie die oben beschriebenen Beispiele der Krankenversicherung in Massachusetts und von Netflix zeigen. Auch eine Veröffentlichung verschiedener Auswertungen über dem gleichen Datensatz kann durch Differenzbildung zur Aufhebung der Anonymität führen (Beispiel: „Anzahl der Prüflinge, die eine bestimmte Prüfung bestanden haben" vs. „Anzahl der Prüflinge außer Ralf Kneuper, die diese Prüfung bestanden haben"), ähnlich auch die gleiche Auswertung über verschiedene Stände eines Datensatzes („Ende Mai haben 120 von 131 Prüflingen die Prüfung bestanden" vs. „Ende Juni haben 120 von 132 Prüflingen die Prüfung bestanden". Wer jetzt – durch Zufall oder Recherche aufgrund dieser Angaben – weiß, dass ein bestimmter Studierender im Juni die Prüfung abgelegt hat, weiß damit auch, dass er diese nicht bestanden hat.).

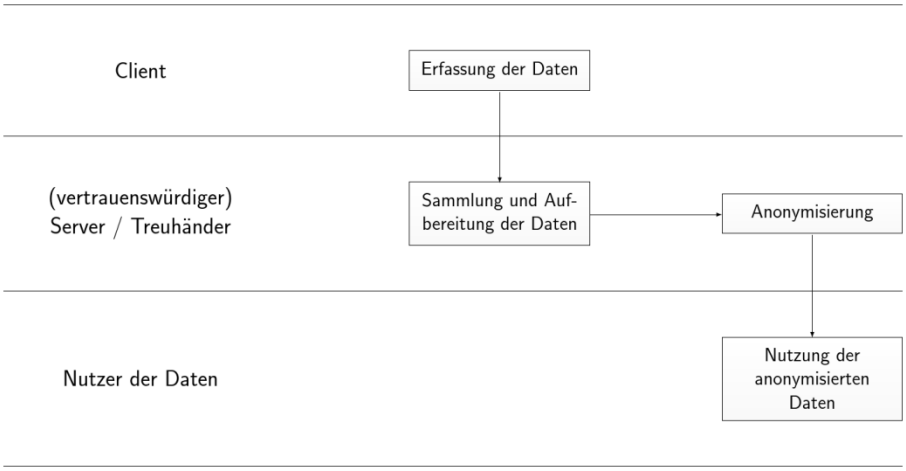
Als Werkzeuge zum Schutz vor diesen Bedrohungen gibt es u.a. die im Folgenden beschriebenen Methoden und Anonymitätsmodelle. Zuerst sollen aber die Szenarien, in denen Anonymisierung eingesetzt wird, kurz dargestellt werden.

Anonymisierungs-Szenarien

In der Literatur zur Anonymisierung werden verschiedene Szenarien zur Anonymisierung betrachtet. Das erste dieser Szenarien nutzt einen vertrauenswürdigen Treuhänder, der die Daten der Betroffenen sammelt (bzw. gesammelt hat), und diese dann anonymisiert und für Dritte zur Auswertung bereitstellt, siehe Abbildung 2.

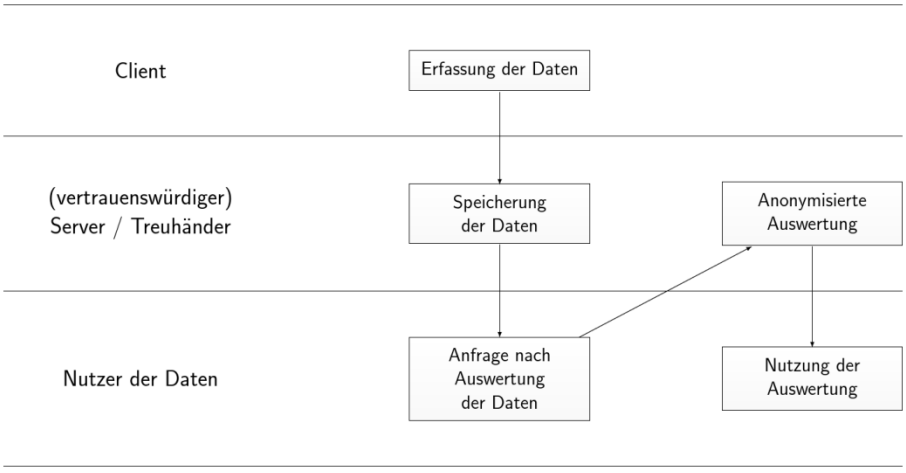
Das lässt dem Nutzer der Daten relativ viel Freiheit, die Daten so auszuwerten wie gewünscht. Umgekehrt erfordert dieses Szenario Vertrauen der Clients in den Treuhänder, und stellt hohe Anforderungen an die Qualität der Anonymisierung, da man nach Bereitstellung der Daten wenig oder keine Kontrolle darüber hat, wie diese anschließend ausgewertet oder mit anderen Daten kombiniert werden. Daneben sind die relevanten Daten in diesem Fall weiterhin personenbezogen und müssen entsprechend geschützt werden, was bei anonymisierten Daten nicht der Fall wäre.

Abbildung 2: Anonymisierungs-Szenario 1: mit vertrauenswürdigem Treuhänder, der anonymisierte Daten bereitstellt



Beim zweiten Szenario geht man ebenfalls von einem vertrauenswürdigen Treuhänder aus, der die personenbezogenen Daten speichert, in diesem Szenario aber nur anonymisierte Auswertungen (also keine Rohdaten, auch nicht in anonymisierter Form) herausgibt, siehe Abbildung 3. Das vereinfacht die Vorgehensweise zur Automatisierung und vor allem im Kontext der unten beschriebenen differentiellen Privatheit gibt es einige Algorithmen, die sich nur für diesen Fall eignen.

Abbildung 3: Anonymisierungs-Szenario 2: mit vertrauenswürdigem Treuhänder, der anonymisierte Auswertungen bereitstellt



Das dritte Szenario schließlich betrachtet den in Abbildung 4 gezeigten Fall, dass die Daten umgehend nach der Erfassung anonymisiert werden und dann nicht mehr als personenbezogene Daten vorliegen.

Abbildung 4: Anonymisierungsszenario 3: ohne Treuhänder, Anonymisierung beim Client

Vorgehensweise zur Anonymisierung}

Auch wenn es keine verbreitete Standard-Vorgehensweise zur Anonymisierung personenbezogener Daten gibt, sollten i.A. folgende Schritte berücksichtigt werden:

- Identifizierung der betroffenen Daten und deren geplanter Nutzung
- Initiale Risikobetrachtung
- Basis-Anonymisierung
- Erneute Risikobetrachtung
- Nach Bedarf weitergehende Anonymisierung auf Basis von Anonymitätsmodellen

Im Folgenden werden diese Schritte näher erläutert.

Identifizierung der betroffenen Daten und deren geplanter Nutzung

Zunächst einmal muss geklärt werden, um welche Daten es geht, wofür diese weiter benötigt werden und welchen Nutzen sie erbringen sollen. (Falls sie nicht mehr benötigt werden, sollten sie gelöscht werden und eine Anonymisierung ist überflüssig.) Dabei ist auch zu klären, dass die geplante Anonymisierung und insbesondere die anschließende Verwendung und ggf. Veröffentlichung der Daten auf einer geeigneten rechtlichen Grundlage basiert.

In diesem Zusammenhang sollte auch geklärt werden, welchen Typ die in den Daten enthaltenen Attribute haben, inwieweit es sich also um Identifizierer, Quasi-Identifizierer, sensible Daten oder andere Daten handelt, auch wenn diese Unterscheidung praktisch oft nicht eindeutig ist. Vor allem bei Quasi-Identifizierern ist dabei auch relevant, welche anderen Datensätze es gibt, über die eine De-Anonymisierung möglich sein könnte.

Initiale Risikobetrachtung

Im nächsten Schritt werden die mit den Daten verbundenen Risiken identifiziert, wobei sich das Risiko wie üblich aus der möglichen Schadenshöhe (anders betrachtet dem Schutzbedarf der Daten) in Kombination mit der Schadenswahrscheinlichkeit zusammensetzt:

- *Schadenshöhe/Schutzbedarf*: Wie hoch ist der mögliche Schaden für die Betroffenen, wenn die Anonymität der Daten mehr oder weniger umfassend aufgehoben wird, und welcher Schutz ist daher erforderlich? Soweit besondere Kategorien personenbezogener Daten betroffen sind, ist normalerweise von einem (mindestens) hohen Schutzbedarf auszugehen. Ansonsten ist hier jeweils der ungünstigste Fall zu betrachten, also der größte Schaden, der möglicherweise eintreten kann.
- *Schadenswahrscheinlichkeit*: Da die Schadenswahrscheinlichkeit stark von den eingesetzten Maßnahmen abhängt, berücksichtigt diese initiale Betrachtung nur die Verfügbarkeit der Daten, d.h. wer bzw. wie

viele Nutzer Zugriff auf die Daten bekommen. Sollen die Daten beispielsweise öffentlich verfügbar werden, so ist das Risiko eines Missbrauchs sehr viel größer als wenn die Daten auch nach der Anonymisierung nur für einen eng beschränkten Kreis von Nutzern verfügbar sind.

Basis-Anonymisierung durch Generalisierung und Unterdrückung von Daten
Bevor komplexere Ansätze zur Anonymisierung eingesetzt werden, beispielsweise die unten beschriebenen Vorgehensweisen auf Basis von Anonymitätsmodellen, sollten zuerst grundlegende Schritte durchgeführt werden, insbesondere natürlich die Löschung der Identifizierer. Vom *U.S. Department of Health & Human Services* gibt es unter dem Namen *Safe Harbor*[3] als Bestandteil der HIPAA-Regelungen zum Schutz von Gesundheitsinformationen eine definierte Liste von Informationen, die als Identifizierer (bzw. Quasi-Identifizierer, auch wenn dieser Begriff in Safe Harbor nicht verwendet wird) gelten und zur Anonymisierung entfernt werden müssen (Office for Civil Rights (OCR) 2012). Diese Liste kommt aus dem Gesundheitswesen, ist aber mit leichten Anpassungen auch für andere Bereiche anwendbar. Bei weniger kritischen Daten ist ggf. eine Lockerung dieser Regeln möglich, bei sehr sensiblen Daten auch eine Verschärfung erforderlich wie unten beschrieben angemessen.

Die folgenden Informationen müssen gemäß Safe Harbor zur Anonymisierung gelöscht werden:

- *Namen* von Individuen
- *Geographische Orte*, die sich auf Einheiten mit weniger als 20.000 Personen beziehen. In Deutschland, Österreich und der Schweiz bedeutet das beispielsweise, dass Postleitzahlen nicht vollständig verwendet werden dürfen, sondern je mindestens eine Stelle gestrichen werden muss (selbst ohne Berücksichtigung von Sonderfällen von Postleitzahlen mit besonders wenigen Einwohnern).[4]
- *Zeitdaten* genauer als das Jahr, die sich auf Einzelpersonen beziehen, beispielsweise Geburtsdatum oder Datum der Aufnahme in oder Entlassung aus einem Krankenhaus. Entsprechend sind auch Altersangaben genauer als ein Jahr zu löschen, außerdem Altersangaben über 89 Jahre, die nur aggregiert als eine Kategorie „90 Jahre und älter“ verwendet werden dürfen. Je nach Anwendungsbereich sollte diese Altersangabe auch niedriger gesetzt werden. Beispielsweise sollte bei einer Mitarbeiterliste die entsprechende Altersgrenze meist unterhalb des Rentenalters gesetzt werden, und es wird analog auch eine untere Altersgrenze erforderlich sein.
- Telephonnummern und Faxnummern
- E-Mail-Adressen
- Sozialversicherungsnummer
- Krankenaktennummer und Krankenversicherungsnummer
- Kontonummern
- Zertifikats- und Lizenznummern
- Fahrzeug-Kennzeichen und Fahrgestellnummern
- *Geräte-Identifizierer* und *-Seriennummern* wie beispielsweise die IMEI-Nummer eines Mobiltelefons oder die Werbe-ID von Android- oder iOS-Geräten
- Web Universal Resource Locators (URLs)

---

[3] Nicht zu verwechseln mit dem gleichnamigen Vertrag zwischen der EU und den USA zur Anerkennung eines angemessenen Datenschutzniveaus, der mittlerweile durch die Privacy Shield-Regelungen abgelöst wurde.

[4] In Deutschland gibt es rund 83 Mio. Einwohner und rund 30.000 verschiedene verwendete Postleitzahlen, also im Durchschnitt rund 2.800 Einwohner pro Postleitzahl.

In Österreich gibt es rund 9 Mio. Einwohner und rund 2.600 verschiedene verwendete Postleitzahlen, also im Durchschnitt rund 3.500 Einwohner pro Postleitzahl.

In der Schweiz gibt es rund 8,5 Mio. Einwohner und rund 4.100 verschiedene verwendete Postleitzahlen, also im Schnitt rund 2.100 Einwohner pro Postleitzahl.

- IP-Adressen sollen gemäß Safe Harbor ebenfalls gelöscht werden. Bei der Analyse der Nutzung von Webseiten (Web Analytics}) werden die IP-Adressen der Besucher allerdings weiterhin benötigt. Safe Harbor sieht für diesen Fall keine Ausnahme vor, da er im Kontext von Gesundheitsdaten normalerweise nicht relevant ist.
  Analog zum Umgang mit geographischen Orten sollte in solchen Fällen zumindest eine entsprechende Verkürzung der IP-Adresse genutzt werden, um Einheiten mit mindestens etwa 20.000 Personen zu erreichen. Etwas vereinfacht wird im Folgenden davon ausgegangen, dass eine IP-Adresse jeweils eindeutig einer Person zugeordnet werden kann und daher Einheiten von mindestens 20.000 IP-Adressen benötigt werden.
  Um ausgehend von einer aktuellen Weltbevölkerung von knapp 8 Milliarden Menschen Einheiten von mindestens 20.000 Menschen zu bilden, dürfen maximal 400.000 ($\approx 2^{18}$) solcher Einheiten unterschieden werden, wobei wir der Einfachheit halber und nicht ganz korrekt von einer annähernd gleichmäßigen Verteilung der Menschen auf die Einheiten ausgehen. Anders formuliert darf die Identifikation der Einheit, in diesem Fall also die IP-Adresse, auf maximal 18 Bit genau angegeben werden, d.h. IP-Adressen müssen zur Anonymisierung auf maximal 18 Bit gekürzt werden, unabhängig davon, ob es sich um eine IPv4-Adresse mit 32 Bit oder eine IPv6-Adresse mit 128 Bit handelt.
  Allerdings ist die Umsetzung dieser Regelung eher schwierig, da das am weitesten verbreiteten System für Web Analytics, nämlich Google Analytics, auch bei einer „Anonymisierung" nur ein Oktett einer IPv4-Adresse löscht, also die Adresse auf 24 Bits kürzt, entsprechend Einheiten von knapp 500 Personen. Bei Matomo dagegen ist die Anzahl der löschenden Oktette als Parameter einstellbar und sollte auf 2 gesetzt werden. Da es bei Web Analytics aber normalerweise nicht – wie bei Safe Harbor angenommen – um Gesundheitsdaten geht, scheint auch eine Verkürzung auf 24 Bits vertretbar, solange es nicht um Webseiten mit besonders sensiblen Daten geht und auch keine anderen Parameter erfasst werden, die in Kombination mit der (verkürzten) IP-Adresse zur Identifizierung verwendet werden können.
- *Biometrische Identifizierer*, einschließlich Fingerabdruck und akustischem Fingerabdruck
- Portrait-Fotos
- Andere *eindeutige Identifikationsnummern*, Charakteristika oder Codes

Die beschriebenen Anpassungen werden allgemeiner auch als *Generalisierung* (von Daten) und *Unterdrückung* (von Daten) bezeichnet:

- Generalisierung: Wenn es viele verschiedene mögliche Werte eines Quasi-Identifizierers gibt, dann ist dessen Informationsgehalt relativ hoch und jeder Wert schränkt die Anzahl der möglichen Personen stark ein. Um das zu reduzieren, kann man durch Generalisierung die Anzahl der Werte reduzieren, indem man sie also verallgemeinert oder zusammenfasst. So wird gemäß der beschriebenen Regeln beispielsweise statt dem genauen Geburtsdatum nur das Geburtsjahr angegeben, Postleitzahlen oder IP-Adressen werden verkürzt gespeichert.
- Unterdrückung bezeichnet die Löschung bestimmter Daten, in diesem Fall beispielsweise von kompletten Spalten / Attributen wie Name oder Sozialversicherungsnummer. Alternativ können auch einzelne Felder unterdrückt werden, beispielsweise immer dann, wenn ein bestimmter Wert nur selten vorkommt (Ausreißer). Beispielsweise ist in vielen Fällen die Angabe des Berufs unproblematisch. Wenn aber in einer Mitarbeiterdatenbank der Beruf beispielsweise Vorstandsvorsitzender ist, dann gilt das nicht mehr und das Feld mit diesem Wert muss unterdrückt werden.[5] Auch die Verkürzung einer IP-Adresse wie oben beschrieben stellt eine Form der Unterdrückung von Daten dar.

[5] *Es reicht natürlich nicht aus, alleine dieses eine Feld eines Datensatzes zu unterdrücken, sondern es muss darüber hinaus eine angemessene Anzahl weiterer Felder unterdrückt werden, da sonst die Unterdrückung selbst wieder eine eindeutige Markierung wäre.*

Erneute Risikobetrachtung
Ziel einer erneuten Risikobetrachtung ist in erster Linie die Beantwortung der Frage, wie hoch die Schadenswahrscheinlichkeit und damit das verbleibende Risiko nach Durchführung der Basis-Anonymisierung ist. Hierzu gehört insbesondere eine Abschätzung, wie eindeutig ein Datensatz eine Person eingrenzt, und welche anderen Daten verfügbar sind, mit denen eine Verlinkung zur Aufhebung der Anonymität möglich wäre.

Dabei sind vor allem evtl. nicht durchgeführte Schritte der Basis-Anonymisierung zu berücksichtigen, wenn beispielsweise bestimmte Quasi-Identifizierer für die relevante Anwendung benötigt werden und daher nicht gelöscht oder zumindest generalisiert werden sollen. Das Risiko kann sich auch mit der Zeit ändern, beispielsweise durch neue Methoden zur Datenanalyse. Daher sollte bei wiederholter Veröffentlichung von anonymen Daten auch die Risikobetrachtung regelmäßig aktualisiert werden.

Eventuell muss zu diesem Zeitpunkt auch der Schutzbedarf neu bewertet werden, um neue Erkenntnisse zu berücksichtigen.

Weitergehende Anonymisierung
Abhängig vom verbleibenden Restrisiko kann eine weitergehende Anonymisierung auf Basis der unten eingeführten Anonymitätsmodelle angemessen sein. Diese Modelle dienen in erster Linie dazu, den Grad der erreichten und der geforderten Anonymität zu quantifizieren, um über weitere Anonymisierungsschritte zu entscheiden.

Die wichtigsten dabei verwendeten Maßnahmen, über die oben eingeführte Generalisierung und Unterdrückung hinaus, sind die Vertauschung von Daten sowie das Hinzufügen von Rauschen.

- *Vertauschung* ist eine sinnvolle Maßnahme, wenn in der späteren Nutzung der Daten der Bezug zwischen der betrachteten Spalte und anderen Spalten nicht wichtig ist, wohl aber Auswertungen wie Mittelwert und Streuung über die Spalte berechnet werden sollen. In diesem Fall kann man Werte der gleichen Spalte zwischen verschiedenen Datensätzen (typischerweise Personen) vertauschen.
- Auch beim *Hinzufügen von Rauschen* werden Daten verändert, indem systematisch kleine Fehler eingebaut werden, die idealerweise die Ergebnisse der benötigten Auswertungen nicht oder zumindest nur geringfügig beeinträchtigen, gleichzeitig aber Aussagen über Einzelpersonen so verfälschen, dass sie nicht mehr verwendbar sind. Werden beispielsweise gemäß einer bekannten Wahrscheinlichkeitsverteilung Zufallszahlen zu den Einzelwerten addiert, so sind die Einzelwerte kaum noch zu verwenden, aber dank der bekannten Verteilung lässt sich dieses „Rauschen" etwa bei der Berechnung des Mittelwertes wieder herausrechnen.

Anonymisierung auf Basis von Anonymitätsmodellen

Eine nicht ausreichend durchdachte Anonymisierung kann wieder aufgehoben werden, wie die Beispiele oben gezeigt haben, beispielsweise durch Kombination mit anderen, öffentlich verfügbaren Daten. Um das zu verhindern, wurden diverse Anonymitätsmodelle entwickelt, die den Grad der Anonymität bewerten, also die Schwierigkeit, anonymisierte Daten zu re-identifizieren. Aufbauend darauf gibt es geeignete Algorithmen, die helfen, Anonymität nach diesen Modellen zu erreichen. Die bekanntesten Anonymitätsmodelle sind *k*-Anonymität (mit diversen Ergänzungen) sowie differentielle Privatheit („differential privacy"), die unten genauer erläutert werden. Einen Überblick über die Vielzahl dieser Verfahren und ihre Stärken und Schwächen gibt beispielsweise (Gkoulalas-Divanis et al. 2014).

Für alle diese Modelle gilt, dass sie sich auf Auswertungen größerer Datenmengen beziehen, bei denen die Identität der Einzelpersonen nicht wichtig ist, sondern bei denen Aussagen über eine größere Menge von Personen gemacht werden sollen. Gleichzeitig sollen meist aber verschiedene Datenpunkte, die sich auf die gleiche Person beziehe, einander zugeordnet werden können. Ein Beispiel dafür sind Auswertungen über den Erfolg von Werbeanzeigen, also zur Frage, welcher Anteil von Personen, denen eine bestimmte Anzeige gezeigt wurde, das beworbene Produkt gekauft hat. Noch wesentlich kritischer sind Auswertungen von Gesundheitsdaten, beispielsweise zum Erfolg bestimmter Therapien oder bestimmter Krankenhäuser.

Wichtig ist in diesem Kontext zu verstehen, dass das Ziel derartiger Anonymitätsmodelle eine *Bewertung der erreichten Anonymität* ist. Daraus folgt natürlich die Frage, *wie* ein bestimmter Grad von Anonymität erreicht

werden kann, aber dies ist nicht ein Teil des Modells selbst. Im Folgenden werden daher *k*-Anonymität und differentielle Privatheit als die bekanntesten Anonymitätsmodelle vorgestellt, allerdings nur mit kurzen Hinweisen auf die zugehörigen Algorithmen und Verfahren.

k-Anonymität

Das Konzept der *k*-Anonymität wurde 2002 von L. Sweeney entwickelt, um den Grad der Anonymität einer Menge von Daten zu bewerten (Sweeney 2002), üblicherweise dargestellt als (Datenbank-)Tabelle, bei der die Identifizierer (Schlüssel) bereits entfernt wurden. Dieser Ansatz unterstützt in erster Linie das erste, in Abbildung 2 dargestellte, der oben beschriebenen Anonymisierungs-Szenarien.

Anonymitätsmaß

Eine Menge von Daten wird als *k*-anonym bezeichnet, wenn es zu den Informationen über eine bestimmte Person jeweils mindestens *k-1* weitere Personen gibt, die auf Grundlage der Quasi-Identifizierer nicht von der Person unterschieden werden können. Man bezeichnet eine solche Gruppe von *k* (oder mehr) Mitgliedern dann auch als Äquivalenzklasse.

Definition: Eine Tabelle heißt *k*-anonym (für $k \in \mathbb{N}$), wenn jede Kombination von Zuordnungen von Werten für die Quasi-Identifizierer in der Tabelle mindestens *k* mal vorkommt (oder alternativ überhaupt nicht).

Je größer der Parameter *k* gewählt ist, desto größer ist der Grad der Anonymität, wobei ein hoher Grad der Anonymität allerdings oft auf Kosten der Genauigkeit der möglichen Auswertungen geht.

Umsetzung

Die wichtigsten Maßnahmen, um *k*-Anonymität bei einer vorhandenen Datenmenge zu erreichen, sind Generalisierung und Unterdrückung von Daten wie oben beschrieben, siehe (Samarati und Sweeney, 1998). Die von Safe Harbor beschriebenen Maßnahmen helfen daher, eine höhere *k*-Anonymität zu erreichen, auch wenn damit kein bestimmter Wert $k > 1$ garantiert werden kann.

Um diese Ansätze auch bei größeren Datenmenge umzusetzen, ist entsprechende Software-Unterstützung notwendig, wobei die Nutzung entsprechender Bibliotheken zu empfehlen ist. Abgesehen davon, dass dies Entwicklungsaufwand spart, hilft die Nutzung solcher Bibliotheken insbesondere auch, die Erfahrung von Spezialisten auf dem Gebiet der Anonymisierung zu nutzen und Fehler zu vermeiden, die sonst möglicherweise die erwünschte Anonymisierung wieder aufheben. Ein Beispiel für eine als Open Source verfügbare Bibliothek zur Umsetzung von *k*-Anonymität und verwandten Ansätzen ist ARX, siehe (ARX o.J.), (Prasser et al. 2014).

Bewertung

Nutzung von *k*-Anonymität (mit hinreichend hohem Wert *k*) schützt gut gegen eine Re-Identifizierung von Daten, aber nur schlecht gegen die Ableitung von Attributen oder Gruppen-Mitgliedschaft. Beispielsweise kann insbesondere bei kleinem *k* eine Ableitung sensibler Attribute noch möglich sein, wenn alle Mitglieder der Äquivalenzklasse den gleichen oder zumindest ähnliche Werte für dieses Attribut haben. Auch Hintergrundwissen kann dazu beitragen, bestimmte Mitglieder einer Äquivalenzgruppe auszuschließen. Weiß man beispielsweise, dass die gesuchte Person zu einer bestimmten Äquivalenzgruppe gehört und erst sieben Jahre alt ist, kann es sich nicht um eines der Mitglieder mit Schwangerschaftsbeschwerden handeln.

Da die Definition von *k*-Anonymität auf der Identifikation der Quasi-Identifizierer aufbaut, kann auch eine fehlerhafte Liste dieser Quasi-Identifizierer dazu führen, dass *k*-Anonymität einer Datensammlung nicht den erwarteten Schutz gibt.

Um die genannten Probleme zu reduzieren, gibt es eine Reihe von Erweiterungen der *k*-Anonymität, insbesondere *i*-Diversität und *t*-Closeness wie in (Gkoulalas-Divanis et al. 2014) und (Article 29 Data Protection Working Party 2014) beschrieben.

Differentielle Privatheit (Differential Privacy)

Differentielle Privatheit versucht, die für k-Anonymität geltenden Einschränkungen durch einen völlig anderen Ansatz zu überwinden, indem nämlich der Wissenszuwachs, der durch einen einzelnen Datensatz erreichbar ist, beschränkt wird. Die Ergebnisse von Auswertungen werden also durch die Aufnahme einer weiteren Peron nur geringfügig verändert, so dass man auch nur sehr wenige Informationen über diese Person schlussfolgern kann. Differentielle Privatheit wird meist im Kontext des Anonymisierungsszenarios 2 betrachtet (siehe Abbildung 3),

also mit Nutzung eines vertrauenswürdigen Servers oder Treuhänders. Der Ansatz ist aber auch für Szenario 3 (Abbildung 4) geeignet und wird dann auch als lokale differentielle Privatheit bezeichnet (Nguyen et al. 2016).

Die folgende Beschreibung gibt eine erste kurze Einführung in differentielle Privatheit, eine umfassende Einführung einschließlich der zugehörigen Algorithmen ist beispielsweise in (Dwork, Roth 2013) oder (Wood et al. 2018) zu finden.

Die zentrale Idee der differentiellen Privatheit besteht darin, dass den Basisdaten ein statistisches Rauschen mit einer bekannten Verteilung hinzugefügt wird. Für jeden einzelnen Datensatz ist dann nicht mehr erkennbar, inwieweit die Daten durch dieses Rauschen verfälscht wurden. Für die Gesamtheit der Daten lässt sich aber (bei hinreichend großer Datenmenge) das Rauschen dank der bekannten Verteilung wieder herausrechnen, und statistische Auswertungen sind damit weiterhin möglich. Die einfachste Form, derartiges Rauschen hinzuzufügen, ist die sogenannte *randomisierte Antwort* (engl. „randomised response"), daneben ist auch die Nutzung der Laplace-Verteilung verbreitet.

Randomisierte Antwort

Das Prinzip der randomisierten Antwort kommt aus den Sozialwissenschaften und wird dort u.a. bei Umfragen für Fragen eingesetzt, bei denen die korrekte Antwort den Befragten möglicherweise unangenehm ist und daher das Risiko einer falschen Antwort hoch ist, beispielsweise „Haben Sie im letzten Monat etwas gestohlen?". Der Algorithmus in Abbildung 5 beschreibt eine mögliche Umsetzung dieses Prinzips für den Fall einfacher Ja-Nein-Fragen.

Abbildung 5: Randomisierte Antwort

Algorithmus 1 Beispiel für randomisierte Antwort (Randomised Response)

```
Wirf Münze (unbeobachtet)
if Zahl then
    Gebe korrekte Antwort                          ▷ Wahrscheinlichkeit 50%
else
    Wirf zweite Münze (ebenfalls unbeobachtet)
    if Zahl then
        Gebe Antwort „nein"                        ▷ Wahrscheinlichkeit 25%
    else
        Gebe Antwort „ja"                          ▷ Wahrscheinlichkeit 25%
    end if
end if
```

Durch Abzug des eingefügten statistischen Rauschens vom Umfrageergebnis, im Beispiel-Algorithmus also der durch den Münzwurf verursachten 25% Ja-Antworten und 25% Nein-Antworten, erhält man das korrekte Ergebnis, wenn auch mit deutlich erhöhter Ungenauigkeit. Der wesentliche Vorteil dieses Ansatzes ist, dass bei jedem einzelnen Befragten nicht mehr erkennbar ist, ob die gegebene Antwort das Ergebnis des Münzwurfs oder die korrekte Antwort auf die Frage wiedergibt. Man spricht daher auch von der „plausiblen Abstreitbarkeit" (engl. plausible deniability) der Antworten. Dies funktioniert natürlich nur, solange die gleichen Personen nicht wiederholt befragt werden, da sonst auch für eine einzelne Person das statistische Rauschen herausgerechnet werden kann.

Das Hinzufügen von statistischem Rauschen wird durch eine „randomisierte Funktion" modelliert, also eine Funktion, deren Ergebnis von einem oder mehreren integrierten Zufallsereignissen abhängt, im Beispiel von Algorithmus 1 den beiden Münzwürfen. Um nun den durch einen Datensatz erreichbaren Wissenszuwachs zu messen, wird das Ergebnis dieser randomisierten Funktion für verschiedene Datensammlungen betrachtet, die sich nur in einem Datensatz unterscheiden, und der Unterschied wie folgt durch einen Parameter $\varepsilon \in R^+$ beschrieben:

Definition: Eine randomisierte Funktion *f* ermöglicht *ε-differentielle Privatheit*, wenn für alle Datensammlungen D1 und D2, die sich in maximal einem Datensatz unterscheiden, und alle S ⊆ Bild(f), gilt

$$P[f(D_1) \in S] \leq e^{\varepsilon} \times P[f(D_2) \in S]$$

Hierbei beschreibt P[X] die Wahrscheinlichkeit eines Zufallsereignisses X gemäß der in *f* enthaltenen Verteilung.

Der Parameter ε beschreibt also den Grad der erreichten Anonymität, wobei im Gegensatz zum Parameter *k* bei der *k*-Anonymität hier ein kleiner Wert von ε einen hohen Grad der Anonymität beschreibt.

Im obigen Beispiel der randomisierten Antwort ist *f* der beschriebene Algorithmus, Bild(f) die Menge der möglichen Antworten {ja, nein}, und *f* ermöglicht differentielle Privatheit, wenn ε = ln 3 ≈ 1, 1 (oder größer) gewählt wird, siehe Claim 3.5 in (Dwork, Roth 2013).

Lokale differentielle Privatheit
Wie oben erwähnt geht differentielle Privatheit normalerweise davon aus, dass es einen vertrauenswürdigen Server oder Treuhänder gibt, der die differentielle Privatheit der Auswertungen sicherstellt. Für viele Anwendungen reicht das aber nicht aus, wie beispielsweise für Auswertungen von einzelnen Client-Anwendungen wie beispielsweise Web-Browsern. Für diesen Fall gibt es die sogenannte *lokale* differentielle Privatheit, bei der das statistische Rauschen nicht der Gesamtheit der auszuwertenden Daten hinzugefügt wird, sondern den einzelnen Clients. Das führt allerdings zu der Schwierigkeit, dass das Rauschen über alle Clients hinweg sehr groß werden kann und die Auswertung der Ergebnisse beeinträchtigt.

Auch für (lokale) differentielle Privatheit gibt es Open-Source-Bibliotheken. Diese wurden von Google veröffentlicht[6], wobei die Bibliothek zur lokalen differentiellen Privatheit das ebenfalls von Google entwickelte und in (Erlingsson , Pihur, Korolova 2014) beschriebene Verfahren RAPPOR umsetzt.

Bewertung
Differentielle Privatheit bietet (bei korrekter Umsetzung) eine mathematisch nachweisbare Sicherheit der Anonymisierung, bei der auf semantischer Ebene die fließende Informationsmenge beschränkt wird und die dadurch unabhängig von sonstigem Vorwissen eines Angreifers ist. Dies ist eine wesentliche Verbesserung gegenüber *k*-Anonymität und verwandten Modellen, bei denen auf syntaktischer Ebene bestimmte Informationsflüsse beschränkt werden, wodurch ein Angreifer mit Vorwissen oder mit einer anderen Angriffsform als erwartet die Anonymisierung möglicherweise aufbrechen kann.

Daher gilt differentielle Privatheit derzeit als der beste Ansatz zur Anonymisierung und wird, zumindest nach eigener Aussage der betroffenen Unternehmen, u.a. von Apple, Google und Uber eingesetzt. Es sind aber nur wenige Details zur tatsächlichen Anwendung von differentieller Privatheit bei den genannten Unternehmen öffentlich bekannt.

Die Kehrseite der hohen Sicherheit von differentieller Privatheit ist, dass die Anwendung relativ komplex ist und eine tiefgehende Beschäftigung mit dem Thema erfordert. Außerdem erfordert differentielle Privatheit je nach Rahmenbedingungen in manchen Fällen, dass den Ausgangsdaten relativ viel Rauschen hinzugefügt werden muss, so dass der Nutzen der anonymisierten Daten ggf. deutlich reduziert ist.

Auch bei differentieller Privatheit sind allerdings noch folgende Einschränkungen der erreichbaren Anonymität zu berücksichtigen:

- Die Daten der einzelnen Person haben immer noch Einfluss auf das Gesamtergebnis, denn sonst könnte man sie (wie auch die Daten aller anderen Personen) weglassen. Differentielle Privatheit mit einem hinreichend kleinen Wert für ε stellt sicher, dass dieser Einfluss gering ist und Rückschlüsse auf die Person kaum noch möglich sind.
- Ergebnisse können weiterhin Aussagen über Personen machen, aber nicht mehr spezifisch über die einzelne Person, sondern als Teil einer Gruppe.

Offene Forschungsthemen

Anonymisierung als Forschungsthema kann aus verschiedenen Sichten betrachtet werden: Zuerst einmal stellt sich hier die grundlegende Frage, wann Anonymisierung aus ethischer und aus rechtlicher Sicht eingesetzt werden sollte. Dabei ist zu berücksichtigen, dass selbst eine Anonymisierung von personenbezogenen Daten ein

[6] https://github.com/google/differential-privacy, https://github.com/google/rappor

Risiko darstellt und es immer wieder dazu kommt, dass eine solche Anonymisierung ganz oder teilweise aufgehoben wird.

Wünschenswert wäre daher die Definition einer Vorgehensweise zur Entscheidung über eine Anonymisierung. Diese sollte auf Kriterien basieren wie den ethischen und rechtlichen Rahmenbedingungen, der Art der Daten und der damit verbundenen Gefährdung der Betroffenen, und dem mit der Nutzung der Daten erreichbaren Nutzen. Mögliche Empfehlungen könnten dann beispielsweise sein, dass Daten nicht oder zumindest nur von Beginn an anonym erfasst werden dürfen, oder dass die Daten anonymisiert werden sollten, wobei hier auch das zu verwendende Anonymisierungsmodell, die entsprechenden Parameter (k bei k-Anonymität, ε bei differentieller Privatheit, etc.) und der weitere Schutz der anonymisierten Daten festzulegen sind.

Aus technischer Sicht können die offenen Forschungsthemen zur Anonymisierung personenbezogener Daten in zwei große Gruppen gegliedert werden:

- Weiterentwicklung der theoretischen Konzepte, beispielsweise geeigneter Algorithmen, mit denen k-Anonymität oder differentielle Privatheit in relevanten Anwendungsszenarien unterstützt werden kann.
- Unterstützung der praktischen Anwendung der vorhandenen theoretischen Konzepte. Die existierenden Beschreibungen sind meist sehr mathematisch ausgerichtet, so dass sie für nicht-Spezialisten kaum anwendbar sind. Es gibt zwar verschiedene Open-Source-Bibliotheken zur Unterstützung, aber auch hierzu existiert wenig Dokumentation, die deren Einsatz und Parametrisierung beschreibt.

In der Praxis führt der zweite Punkt dazu, dass Anonymisierung meist nur ad hoc durchgeführt wird, weitgehend beschränkt auf die Löschung von offensichtlichen Identifizierern sowie die Generalisierung der wichtigsten Quasi-Identifizierer. Insbesondere die in der Forschung als „Gold-Standard" der Anonymisierung bezeichnete differentielle Privatheit ist in der praktischen Anwendung nur in wenigen, großen Anwendungsfällen mit Teams von Spezialisten angekommen, so bei der kommenden US-Volkszählung[7] oder für ausgewählte Anwendungsfälle bei den bereits genannten Unternehmen Google, Apple und Uber.

Hilfreich für eine breitere Anwendung der vorhandenen Verfahren zur Anonymisierung wären daher detaillierte Fallstudien, sowie Anleitungen und Beschreibungen, die diese Verfahren auch für nicht-Spezialisten nutzbar machen könnten.

Solche Anleitungen und Beschreibungen sollten helfen, die theoretischen Konzepte für praktische Anwendungsfälle zu bewerten, die geeigneten Konzepte auszuwählen und sie in der Praxis umzusetzen, ggf. mit Hilfe vorhandener SW-Bibliotheken. Idealerweise sollte die Anwendung damit so einfach werden, dass zumindest einfache Anwendungsfälle, wie sie im Kontext des Datenschutzes häufig vorkommen, auch von „normalen" Business-Analytikern, Softwareentwicklern und Datenschutzbeauftragten ohne tiefgehende Spezialisierung umgesetzt werden können.

[7] https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census_bureau_adopts.html

Literaturverzeichnis:

Article 29 Data Protection Working Party (2014)	Opinion 05/2014 on Anonymisation Techniques. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf , abgerufen am 24.11.2019.
ARX (o.J.)	*ARX – Data Anonymization Tool. A comprehensive software for privacy-preserving microdata publishing*. URL: https://arx.deidentifier.org/, abgerufen am 05.01.2020.
Böhme, Rainer und Pesch, Paulina (2017)	Technische Grundlagen und datenschutzrechtliche Fragen der Blockchain-Technologie. DuD Datenschutz und Datensicherheit, Vol. 41, Heft 8, S. 473-481.
Dwork, Cynthia und Roth, Aaron (2013)	The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, Vol. 9, Nr. 3-4, S. 211–407.
Erlingsson, Úlfar, Pihur, Vasyl und Korolova, Aleksandra (2014)	Rappor: Randomized aggregatable privacy-preserving ordinal response. In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS '14*, S. 1054–1067, New York, USA, ACM.
Gkoulalas-Divanis, Aris, Loukides, Grigorios und Sun, J (2014)	Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of biomedical informatics, 50.
Narayanan, Arvind und Shmatikov, Vitaly (2008)	*Robust de-anonymization of large sparse datasets*. In: 2008 IEEE Symposium on Security and Privacy, S. 111-125.
Nguyen, Thong T., Xiao, Xiaokui, Yang, Yin, Hui, Siu Cheung, Shin, Hyejin und Shin, Junbum (2016)	*Collecting and analyzing data from smart device users with local differential privacy.* URL: https://arxiv.org/pdf/1606.05053, abgerufen am 04.01.2020.
Office for Civil Rights (OCR) (2012)	Guidance Regarding Methods for De-identificationof Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html, abgerufen am 22.11.2019.
Oswald, Malcolm (2013)	Anonymisation standard for publishing health and social care data specification. Supporting guidance: Drawing the line between identifying and non-identifying data. NHS The Information Centre for Health and Social Care (UK). URL: https://digital.nhs.uk/binaries/content/assets/legacy/pdf/b/e/1523202010guid.pdf, abgerufen am 26.10.2019.
Pandurangan, Vijay (2014)	*On taxis and rainbows*. URL: https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1 , abgerufen am 21.11.2019.
PCI Security Standards Council (2011)	*PCI DSS Tokenization Guidelines.* URL: https://www.pcisecuritystandards.org/documents/Tokenization_Guidelines_Info_Supplement.pdf abgerufen am 20.11.2019
PCI Security Standards Council (2015)	*Tokenization Product Security Guidelines.* URL: https://www.pcisecuritystandards.org/documents/Tokenization_Product_Security_Guidelines.pdf abgerufen am 20.11.2019
Prasser, Fabian, Kohlmayer, Florian, Lautenschläger, Ronald und Kuhn, Klaus A. (2014)	ARX – A Comprehensive Tool for anonymizing biomedical data. In: *AMIA Annual Symposium Proceedings*, S. 984-993, Washington (DC), USA.

Samarati, Pierangela und Sweeney, Latanya (1998)	*Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* In: Proceedings of the IEEE Symposium on Research in Security and Privacy (S&P), Oakland, CA, 1998. URL: https://dataprivacylab.org/dataprivacy/projects/kanonymity/index3.html abgerufen am 23.11.2019
Schwartmann, Rolf und Weiß, Steffen (2018)	*Anforderungen an den datenschutzkonformen Einsatz von Pseudonymisierungslösungen.* Bundesministerium des Innern, für Bau und Heimat, 2018. https://www.gdd.de/downloads/anforderungen-an-datenschutzkonforme-pseudonymisierung, abgerufen am 21.11.2019.
Sweeney, Latanya (2002)	*k-Anonymity: a model for protecting privacy*. In: International Journal on Uncertainty, Fuzzyness and Knowledge-based Systems, Vol. 5/10, S. 557-570.
Wood, Alexandra, Altman, Micah, Bembenek, Aaron, Bun, Mark, Gaboardi, Marco, Honaker, James, Nissim, Kobbi, O'Brien, David, Steinke, Thomas und Vadhan, Salil (2018)	*Differential Privacy: A Primer for a Non-Technical Audience*. Vanderbilt Journal of Entertainment & Technology Law, Vol. 21, Nr. 1, S. 209–276.