0829/14/DE
WP216

Stellungnahme 5/2014 zu Anonymisierungstechniken

Angenommen am 10. April 2014

DIE GRUPPE FÜR DEN SCHUTZ VON PERSONEN BEI DER VERARBEITUNG PERSONENBEZOGENER DATEN

eingesetzt durch die Richtlinie 95/46/EG des Europäischen Parlaments und des Rates vom 24. Oktober 1995,

gestützt auf Artikel 29 und auf Artikel 30 dieser Richtlinie,

gestützt auf ihre Geschäftsordnung,

HAT FOLGENDE STELLUNGNAHME ANGENOMMEN:

ZUSAMMENFASSUNG

In dieser Stellungnahme analysiert die Datenschutzgruppe die Wirksamkeit und die Grenzen der derzeit vorhandenen Anonymisierungstechniken vor dem Hintergrund des EU-Rechtsrahmens für den Datenschutz und spricht Empfehlungen für den Umgang mit diesen Techniken aus, wobei insbesondere das mit ihnen verbundene Restrisiko einer Identifizierung Berücksichtigung findet.

Die Datenschutzgruppe ist sich des potenziellen Werts der Anonymisierung bewusst und erkennt sie insbesondere als eine Strategie an, die Vorteile der „offenen Daten" für den Einzelnen und für die Gesellschaft insgesamt zu nutzen und zugleich die Risiken für die betroffenen Personen zu verringern. Allerdings haben Fallstudien und Forschungsarbeiten gezeigt, wie schwer es ist, einen tatsächlich anonymen Datenbestand zu generieren und dabei alle zugrunde liegenden Informationen zu erhalten, die für die zu bewältigende Aufgabe erforderlich sind.

Im Sinne der Richtlinie 95/46/EG und anderer einschlägiger Rechtsinstrumente der EU ist Anonymisierung das Ergebnis der Verarbeitung personenbezogener Daten mit dem Ziel, eine Identifizierung unwiderruflich unmöglich zu machen. Dabei sollten die für die Verarbeitung Verantwortlichen mehrere Faktoren in Betracht ziehen und alle Mittel berücksichtigen, die vernünftigerweise (entweder von dem für die Verarbeitung Verantwortlichen oder von einem Dritten) zur Identifizierung eingesetzt werden könnten.

Anonymisierung stellt eine Weiterverarbeitung personenbezogener Daten dar und muss als solche der Anforderung der Vereinbarkeit unter Berücksichtigung der Rechtsgrundlagen und Bedingungen der Weiterverarbeitung entsprechen. Darüber hinaus fallen anonymisierte Daten zwar nicht in den Anwendungsbereich der Datenschutzvorschriften, jedoch genießen die betroffenen Personen unter Umständen das Recht auf Schutz nach Maßgabe anderer Vorschriften (beispielsweise über den Schutz der Vertraulichkeit der Kommunikation).

In dieser Stellungnahme werden die wichtigsten Anonymisierungstechniken, d. h. Randomisierung und Generalisierung, beschrieben. Insbesondere werden Verfahren wie stochastische Überlagerung, Vertauschung, Differential Privacy, Aggregation, k-Anonymität, l-Diversität und t-Closeness erörtert. Es werden ihre Grundsätze, Stärken und Schwächen sowie die im Zusammenhang mit dem Einsatz der einzelnen Techniken häufig auftretenden Fehler und Mängel erläutert.

In dieser Stellungnahme wird die Robustheit der einzelnen Techniken auf der Grundlage von drei Kriterien untersucht:

(i) Ist es weiterhin möglich, eine Person herauszugreifen?

(ii) Ist es weiterhin möglich, eine Person betreffende Datensätze zu verknüpfen?

(iii) Können Informationen über eine Person durch Inferenz hergeleitet werden?

Die Kenntnis der wichtigsten Stärken und Schwächen der einzelnen Techniken ist hilfreich bei der Planung eines angemessenen Anonymisierungsverfahrens in einem bestimmten Kontext.

Das Verfahren der Pseudonymisierung wird ebenfalls behandelt, um einige Tücken und Missverständnisse aufzuzeigen: Pseudonymisierung ist keine Anonymisierungstechnik. Sie

verringert lediglich die Verknüpfbarkeit eines Datenbestands mit der wahren Identität einer betroffenen Person und stellt somit eine sinnvolle Sicherheitsmaßnahme dar.

Die Datenschutzgruppe kommt in dieser Stellungnahme zu dem Schluss, dass Anonymisierungstechniken geeignet sind, Garantien für den Schutz der Privatsphäre zu schaffen, und eingesetzt werden können, um wirksame Anonymisierungsverfahren zu entwickeln. Dies gilt allerdings nur, wenn ihre Anwendung ordnungsgemäß geplant wird. Das bedeutet, dass die Voraussetzungen (Kontext) und die Zielsetzung(en) des Anonymisierungsverfahrens klar festgelegt werden müssen, um die angestrebte Anonymisierung zu erreichen und zugleich zweckmäßige Daten hervorzubringen. Die Wahl der am besten geeigneten Lösung sollte auf der Grundlage einer Einzelfallbewertung erfolgen, nach Möglichkeit unter Heranziehung einer Kombination verschiedener Techniken und unter Berücksichtigung der in dieser Stellungnahme herausgearbeiteten praktischen Empfehlungen.

Schließlich sollten die für die Verarbeitung Verantwortlichen berücksichtigen, dass ein anonymisierter Datenbestand nach wie vor gewisse Restrisiken für die betroffenen Personen bergen kann. Denn zum einen sind Anonymisierung und Reidentifizierung aktive Forschungsbereiche, in denen regelmäßig neue Erkenntnisse veröffentlicht werden, zum anderen können selbst anonymisierte Daten (beispielsweise Statistiken) herangezogen werden, um vorhandene Personenprofile zu ergänzen, wodurch neue Datenschutzprobleme aufgeworfen werden. Anonymisierung sollte daher nicht als eine einmalige Aufgabe betrachtet werden, und die bestehenden Risiken sollten von den für die Verarbeitung Verantwortlichen regelmäßig neu bewertet werden.

1 Einleitung

Geräte, Sensoren und Netzwerke bringen riesige Datenmengen und immer neue Arten von Daten hervor. Angesichts dessen und der mittlerweile verschwindend geringen Kosten für die Speicherung von Daten bestehen ein zunehmendes öffentliches Interesse und ein wachsender Bedarf an der Wiederverwendung dieser Daten. „Offene Daten" bieten der Gesellschaft sowie dem Einzelnen und Organisationen klare Vorteile. Dies gilt allerdings nur, wenn das Recht eines jeden Bürgers auf den Schutz seiner personenbezogenen Daten und seiner Privatsphäre gewahrt bleibt.

Anonymisierung kann eine geeignete Strategie darstellen, um die Vorteile zu nutzen und zugleich die Risiken zu verringern. Wurde ein Datenbestand erfolgreich anonymisiert und die Identifizierung von Einzelpersonen zuverlässig ausgeschlossen, fallen die betreffenden Daten nicht mehr in den Anwendungsbereich der europäischen Datenschutzvorschriften. Allerdings belegen Fallstudien und Forschungsarbeiten, dass es sehr schwer ist, aus einem umfassenden Bestand personenbezogener Daten einen tatsächlich anonymen Datenbestand zu generieren und dabei alle zugrunde liegenden Informationen zu erhalten, die für die zu bewältigende Aufgabe erforderlich sind. Beispielsweise kann ein als anonym geltender Datenbestand unter Umständen mit einem anderen Datenbestand in einer Weise verknüpft werden, dass eine oder mehrere Personen identifiziert werden können.

In dieser Stellungnahme analysiert die Datenschutzgruppe die Wirksamkeit und die Grenzen der derzeit vorhandenen Anonymisierungstechniken vor dem Hintergrund des EU-Rechtsrahmens für den Datenschutz und spricht Empfehlungen für einen umsichtigen und verantwortungsvollen Einsatz dieser Techniken für die Planung eines Anonymisierungsverfahrens aus.

2 Definitionen und rechtliche Analyse

2.1. Definitionen im EU-Rechtsrahmen

In der Richtlinie 95/46/EG wird in Erwägungsgrund 26 im Hinblick auf die Anonymisierung festgestellt, dass anonymisierte Daten nicht in den Anwendungsbereich der Datenschutzvorschriften fallen:

> *„Die Schutzprinzipien müssen für alle Informationen über eine bestimmte oder bestimmbare Person gelten. Bei der Entscheidung, ob eine Person bestimmbar ist, sollten alle Mittel berücksichtigt werden, die vernünftigerweise entweder von dem Verantwortlichen für die Verarbeitung oder von einem Dritten eingesetzt werden könnten, um die betreffende Person zu bestimmen. Die Schutzprinzipien finden keine Anwendung auf Daten, die derart anonymisiert sind, dass die betroffene Person nicht mehr identifizierbar ist. Die Verhaltensregeln im Sinne des Artikels 27 können ein nützliches Instrument sein, mit dem angegeben wird, wie sich die Daten in einer Form*

anonymisieren und aufbewahren lassen, die die Identifizierung der betroffenen Person unmöglich macht."[1]

Dem Wortlaut von Erwägungsgrund 26 der Richtlinie ist eine Definition des Begriffs Anonymisierung zu entnehmen. Demnach müssen im Zuge der Anonymisierung von Daten hinreichend viele Elemente entfernt werden, sodass eine Identifizierung der betroffenen Person ausgeschlossen ist. Genauer gesagt müssen die Daten so verarbeitet werden, dass es selbst unter Verwendung „alle[r] Mittel [...], die vernünftigerweise" entweder von dem für die Verarbeitung Verantwortlichen oder von einem Dritten „eingesetzt werden könnten", nicht mehr möglich ist, eine natürliche Person zu bestimmen. Ein wichtiger Faktor ist, dass die Verarbeitung unumkehrbar sein muss. In der Richtlinie wird nicht klargestellt, wie ein solches Deidentifizierungsverfahren durchgeführt werden sollte oder könnte.[2] Das Hauptaugenmerk liegt dabei auf dem Ergebnis: Die Daten müssen so beschaffen sein, dass es mit „allen" Mitteln, die „vernünftigerweise" eingesetzt werden könnten, unmöglich ist, die betroffene Person zu bestimmen. Des Weiteren wird auf Verhaltensregeln als ein Instrument hingewiesen, um anzugeben, wie sich die Daten in einer Form anonymisieren und aufbewahren lassen, die die Identifizierung der betroffenen Person „unmöglich macht". Mit der Richtlinie wird also ganz klar ein sehr hoher Standard festgelegt.

In der Datenschutzrichtlinie für die elektronische Kommunikation (Richtlinie 2002/58/EG) wird ebenfalls in sehr ähnlichem Sinne auf „Anonymisierung" und „anonyme Daten" Bezug genommen. So heißt es in Erwägungsgrund 26:

„Verkehrsdaten, die für die Vermarktung von Kommunikationsdiensten oder für die Bereitstellung von Diensten mit Zusatznutzen verwendet wurden, sollten ferner nach der Bereitstellung des Dienstes gelöscht oder anonymisiert werden."

Dementsprechend schreibt Artikel 6 Absatz 1 vor:

„Verkehrsdaten, die sich auf Teilnehmer und Nutzer beziehen und vom Betreiber eines öffentlichen Kommunikationsnetzes oder eines öffentlich zugänglichen Kommunikationsdienstes verarbeitet und gespeichert werden, sind unbeschadet der Absätze 2, 3 und 5 des vorliegenden Artikels und des Artikels 15 Absatz 1 zu löschen oder zu anonymisieren, sobald sie für die Übertragung einer Nachricht nicht mehr benötigt werden."

In Artikel 9 Absatz 1 heißt es weiter:

„Können andere Standortdaten als Verkehrsdaten in Bezug auf die Nutzer oder Teilnehmer von öffentlichen Kommunikationsnetzen oder öffentlich zugänglichen Kommunikationsdiensten verarbeitet werden, so dürfen diese Daten nur im zur Bereitstellung von Diensten mit Zusatznutzen erforderlichen Maß und innerhalb des dafür erforderlichen Zeitraums verarbeitet werden, wenn sie anonymisiert wurden oder wenn die Nutzer oder Teilnehmer ihre Einwilligung gegeben haben."

Dem liegt der Gedanke zugrunde, dass das Ergebnis der Anonymisierung als ein auf personenbezogene Daten angewandtes technisches Verfahren nach dem aktuellen Stand der

[1] Zudem ist darauf hinzuweisen, dass dieser Ansatz auch in Erwägungsgrund 23 des Entwurfs der EU-Datenschutz-Grundverordnung herangezogen wird: „Um festzustellen, ob eine Person bestimmbar ist, sind alle Mittel zu berücksichtigen, die von dem für die Verarbeitung Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen aller Voraussicht nach zur Identifizierung der Person genutzt werden."

[2] Dieser Begriff wird auf S. 8 dieser Stellungnahme weiter ausgeführt.

Technik so dauerhaft sein sollte wie eine Löschung, d. h., es darf nicht möglich sein, die personenbezogenen Daten weiter zu verarbeiten.[3]

2.2. Rechtliche Analyse

Die Analyse des Wortlauts der in den wichtigsten EU-Datenschutzinstrumenten enthaltenen Bestimmungen über die Anonymisierung lässt vier zentrale Merkmale erkennen:

– Anonymisierung kann das Ergebnis einer Verarbeitung personenbezogener Daten sein, die mit dem Ziel vorgenommen wird, die Identifizierung der betroffenen Person unwiderruflich unmöglich zu machen.

– Es können verschiedene Anonymisierungstechniken ins Auge gefasst werden, da in den EU-Rechtsvorschriften diesbezüglich keine Vorgaben gemacht werden.

– Großes Augenmerk muss auf den kontextabhängigen Faktoren liegen: So sind „alle" Mittel zu berücksichtigen, die von dem für die Verarbeitung Verantwortlichen oder von einem Dritten „vernünftigerweise" zur Identifizierung genutzt werden könnten, wobei besonders darauf zu achten ist, welche Mittel dem aktuellen Stand der Technik entsprechend (angesichts zunehmender Rechenleistung und der steigenden Zahl verfügbarer Instrumente) „vernünftigerweise" eingesetzt werden könnten.

– Anonymisierung birgt einen Risikofaktor, der bei der Beurteilung der Validität jeglicher Anonymisierungstechniken Berücksichtigung finden muss – einschließlich der Möglichkeiten für eine Nutzung der mittels einer solchen Technik „anonymisierten" Daten – und dessen Schwere und Wahrscheinlichkeit bewertet werden sollten.

In dieser Stellungnahme wird der Begriff „Anonymisierungstechnik" verwendet und bewusst auf die Termini „Anonymität" oder „anonyme Daten" verzichtet, um zu betonen, dass mit jeder technisch-organisatorischen Maßnahme, die mit dem Ziel der „Anonymisierung" von Daten vorgenommen wird, ein Restrisiko verbunden ist.

2.2.1. Rechtmäßigkeit des Anonymisierungsverfahrens

Zunächst ist Anonymisierung eine Technik, die auf personenbezogene Daten angewendet wird, um eine unumkehrbare Deidentifizierung zu erreichen. Daher ist zunächst von der Annahme auszugehen, dass die personenbezogenen Daten im Einklang mit den geltenden

[3] An dieser Stelle ist daran zu erinnern, dass Anonymisierung auch in internationalen Normen wie beispielsweise der ISO 29100 definiert ist als der „Prozess, durch den personenbezogene Daten (*personally identifiable information* – PII) unumkehrbar in einer Weise verändert werden, dass die betroffene Person von dem für die Verarbeitung Verantwortlichen weder alleine noch in Zusammenarbeit mit Dritten unmittelbar oder mittelbar identifiziert werden kann" (ISO 29100:2011). Die Unmöglichkeit, die an den personenbezogenen Daten vorgenommen Änderungen mit dem Ziel einer unmittelbaren oder mittelbaren Identifizierung rückgängig zu machen, stellt demnach auch für die ISO einen zentralen Aspekt dar. Aus dieser Perspektive besteht eine beträchtliche Übereinstimmung mit den Grundsätzen und Begriffen, die der Richtlinie 95/46/EG zugrunde liegen. Gleiches gilt für die in einigen einzelstaatlichen Rechtsvorschriften (beispielsweise in Italien, Deutschland und Slowenien) vorgenommenen Definitionen, in denen das Hauptaugenmerk auf der Nichtidentifizierbarkeit liegt und auf einen „unverhältnismäßig hohen Aufwand" für die Reidentifizierung verwiesen wird (DE, SI). Das französische Datenschutzgesetz schreibt jedoch vor, dass die Daten selbst dann als personenbezogene Daten zu betrachten sind, wenn es äußerst schwierig und unwahrscheinlich ist, dass die betroffene Person reidentifiziert werden kann – es gibt hier also keinen Hinweis auf eine mögliche Prüfung hinsichtlich der „vernünftigerweise" einsetzbaren Mittel.

Rechtsvorschriften über die Speicherung von Daten in einem Format erhoben und verarbeitet wurden, das eine Identifizierung erlaubt.

In diesem Zusammenhang stellt das Anonymisierungsverfahren im Sinne der Verarbeitung solcher personenbezogenen Daten mit dem Ziel ihrer Anonymisierung eine Form der „Weiterverarbeitung" dar. Als solche muss diese Verarbeitung daraufhin geprüft werden, ob sie das Kriterium der Vereinbarkeit im Sinne der Leitlinien erfüllt, die von der Datenschutzgruppe in ihrer Stellungnahme 03/2013 zur Zweckbindung vorgelegt wurden.[4]

Demnach kann grundsätzlich jeder der in Artikel 7 der Richtlinie 95/46/EG genannten Gründe (einschließlich des berechtigten Interesses des für die Verarbeitung Verantwortlichen) als Rechtsgrundlage für eine Anonymisierung herangezogen werden, sofern auch die in Artikel 6 der Richtlinie festgelegten Anforderungen an die Qualität der Daten erfüllt sind und die spezifischen Umstände sowie alle in der Stellungnahme der Datenschutzgruppe zur Zweckbindung genannten Faktoren angemessen berücksichtigt werden.[5]

Darüber hinaus sind die in Artikel 6 Absatz 1 Buchstabe e der Richtlinie 95/46/EG (aber auch in Artikel 6 Absatz 1 und Artikel 9 Absatz 1 der Datenschutzrichtlinie für die elektronische Kommunikation) enthaltenen Bestimmungen zu berücksichtigen, da sie vorschreiben, dass personenbezogene Daten nicht länger, als es für die Realisierung der Zwecke, für die sie erhoben oder weiterverarbeitet werden, erforderlich ist, „in einer Form aufbewahrt werden, die die Identifizierung der betroffenen Personen ermöglicht".

Diese Bestimmung macht sehr deutlich, dass personenbezogene Daten zumindest „standardmäßig" anonymisiert werden sollten (im Einklang mit unterschiedlichen rechtlichen Anforderungen wie den in der Datenschutzrichtlinie für die elektronische Kommunikation enthaltenen Vorschriften über Verkehrsdaten). Möchte der für die Verarbeitung Verantwortliche solche personenbezogenen Daten aufbewahren, nachdem die ursprünglichen Verarbeitungszwecke oder die Zwecke der Weiterverarbeitung erreicht wurden, sollten Anonymisierungstechniken zur Anwendung kommen, um eine Identifizierung unwiderruflich unmöglich zu machen.

Dementsprechend ist die Datenschutzgruppe der Auffassung, dass die Anonymisierung als eine Form der Weiterverarbeitung personenbezogener Daten mit dem ursprünglichen Verarbeitungszweck vereinbar ist, allerdings nur unter der Voraussetzung, dass das Anonymisierungsverfahren geeignet ist, zuverlässig anonymisierte Informationen in dem in dieser Stellungnahme beschriebenen Sinne hervorzubringen.

Des Weiteren ist darauf hinzuweisen, dass die Anonymisierung den rechtlichen Einschränkungen entsprechen muss, die der Europäische Gerichtshof in seinem Urteil in der

---

[4] Stellungnahme 03/2013 der Artikel-29-Datenschutzgruppe, verfügbar unter http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

[5] Das bedeutet insbesondere, dass eine gründliche Beurteilung vorgenommen werden muss, bei der alle relevanten Umstände und vor allem die folgenden zentralen Faktoren zu berücksichtigen sind:
a) das Verhältnis zwischen den Zwecken, für die die personenbezogenen Daten erhoben wurden, und den Zwecken einer Weiterverarbeitung;
b) der Zusammenhang, in dem die personenbezogenen Daten erhoben wurden, sowie die berechtigten Erwartungen der betroffenen Personen hinsichtlich ihrer weiteren Verwendung;
c) die Art der personenbezogenen Daten und die Auswirkungen der Weiterverarbeitung auf die betroffenen Personen;
d) die vom für die Verarbeitung Verantwortlichen ergriffenen Sicherheits- und Schutzmaßnahmen zur Gewährleistung einer Verarbeitung nach Treu und Glauben sowie zur Verhütung unangemessener Auswirkungen auf die betroffenen Personen.

Rechtssache C-553/07 (*College van burgemeester en wethouders van Rotterdam/M.E.E. Rijkeboer*) bezüglich des Erfordernisses unterstrichen hat, die Daten in einem Format zu speichern, das eine Identifizierung erlaubt, um beispielsweise die Ausübung des Auskunftsrechts durch die betroffenen Personen zu ermöglichen. Der Gerichtshof urteilte: „*Nach Artikel 12 Buchstabe a der Richtlinie 95/46/EG [...] sind die Mitgliedstaaten verpflichtet, ein Recht auf Auskunft über die Empfänger oder Kategorien der Empfänger der Daten sowie den Inhalt der übermittelten Information vorzusehen, das nicht nur für die Gegenwart, sondern auch für die Vergangenheit gilt. Es ist Sache der Mitgliedstaaten, eine Frist für die Aufbewahrung dieser Information sowie einen darauf abgestimmten Zugang zu ihr festzulegen, die einen gerechten Ausgleich bilden zwischen dem Interesse der betroffenen Person am Schutz ihres Privatlebens, insbesondere mit Hilfe der in der Richtlinie 95/46 vorgesehenen Rechte und Rechtsbehelfe, auf der einen Seite und der Belastung, die die Pflicht zur Aufbewahrung der betreffenden Information für den für die Verarbeitung Verantwortlichen darstellt, auf der anderen Seite.*“

Dies ist insbesondere dann relevant, wenn sich der für die Verarbeitung Verantwortliche bei der Anonymisierung auf Artikel 7 Buchstabe f der Richtlinie 95/46/EG stützt: Es muss stets eine Abwägung des berechtigten Interesses des für die Verarbeitung Verantwortlichen und der Rechte und Grundfreiheiten der betroffenen Personen vorgenommen werden.

Beispielsweise ergab eine in den Jahren 2012 und 2013 durchgeführte Untersuchung der niederländischen Datenschutzbehörde über die Nutzung von Technologien für tief greifende Paketanalysen (*Deep Packet Inspection*) durch vier Mobilfunkbetreiber, dass für die Anonymisierung der Inhalte von Verkehrsdaten möglichst zeitnah nach der Erhebung dieser Daten ein rechtmäßiger Grund nach Artikel 7 Buchstabe f der Richtlinie 95/46/EG gegeben war. Denn nach Maßgabe von Artikel 6 der Datenschutzrichtlinie für elektronische Kommunikation sind Verkehrsdaten, die sich auf Teilnehmer und Nutzer beziehen und vom Betreiber eines öffentlichen Kommunikationsnetzes oder eines öffentlich zugänglichen Kommunikationsdienstes verarbeitet und gespeichert werden, so schnell wie möglich zu löschen oder zu anonymisieren. In diesem Falle ist eine entsprechende Rechtsgrundlage nach Artikel 7 der Datenschutzrichtlinie gegeben, da die Anonymisierung gemäß Artikel 6 der Datenschutzrichtlinie für elektronische Kommunikation zulässig ist. Dieser Sachverhalt lässt sich auch umgekehrt darstellen: Ist eine bestimmte Form der Datenverarbeitung gemäß Artikel 6 der Datenschutzrichtlinie für elektronische Kommunikation unzulässig, kann auch keine Rechtsgrundlage nach Artikel 7 der Datenschutzrichtlinie gegeben sein.

2.2.2. Mögliche Identifizierbarkeit betroffener Personen anhand anonymisierter Daten

In ihrer Stellungnahme 4/2007 zum Begriff „personenbezogene Daten“ hat sich die Datenschutzgruppe bei der Auseinandersetzung mit diesem Begriff auf die Hauptbausteine der in Artikel 2 Buchstabe a der Richtlinie 95/46/EG vorgenommenen Begriffsbestimmung konzentriert, einschließlich der in dieser Definition verwendeten Formulierung „eine bestimmte oder bestimmbare“ natürliche Person. In diesem Zusammenhang kam die Datenschutzgruppe unter anderem zu dem Schluss: „‚Anonymisierte Daten‘ sind somit anonyme Daten, die sich zuvor auf eine bestimmbare Person bezogen, die jedoch nicht mehr identifizierbar ist.“

Die Datenschutzgruppe hat daher bereits klargestellt, dass in der Richtlinie die Prüfung hinsichtlich der „Mittel [...], die vernünftigerweise [...] eingesetzt werden könnten“ als ein Kriterium empfohlen wird, mit dem beurteilt werden soll, ob das Anonymisierungsverfahren hinreichend robust ist, d. h. ob eine Identifizierung „vernünftigerweise“ unmöglich geworden ist. Der spezifische Kontext und die Umstände des Einzelfalls wirken sich unmittelbar auf die

Identifizierbarkeit aus. Im technischen Anhang zu dieser Stellungnahme wird analysiert, welche Auswirkungen die Wahl der am besten geeigneten Technik hat.

Wie bereits oben hervorgehoben, werden Forschung, Instrumente und Rechenleistung beständig weiterentwickelt. Daher ist es weder möglich noch sinnvoll, eine erschöpfende Auflistung der Umstände vorzunehmen, unter denen eine Identifizierung unmöglich ist. Dennoch sollen hier einige zentrale Faktoren herausgegriffen und erläutert werden.

Erstens ist darauf hinzuweisen, dass sich die für die Verarbeitung Verantwortlichen auf die konkreten Mittel konzentrieren sollten, die für eine Umkehrung der Anonymisierungstechnik erforderlich wären, wobei insbesondere die mit der Anwendung dieser Mittel verbunden Kosten und das dazu notwendige Fachwissen sowie die Wahrscheinlichkeit und die Schwere der Auswirkungen einer solchen Umkehrung zu berücksichtigen sind. Beispielsweise sollten die für die Verarbeitung Verantwortlichen den Aufwand und die Kosten der Anonymisierung (im Hinblick sowohl auf die zeitlichen als auch auf die sonstigen erforderlichen Ressourcen) gegen die zunehmende Verfügbarkeit kostengünstiger technischer Mittel zur Identifizierung von Personen in Datenbeständen, die zunehmende öffentliche Verfügbarkeit anderer Datenbestände (wie sie beispielsweise im Zusammenhang mit Strategien der „offenen Daten" verfügbar gemacht werden) und die zahlreichen Beispiele unvollständiger Anonymisierungsverfahren abwägen, die mit nachteiligen und zuweilen nicht wiedergutzumachenden Auswirkungen auf die betroffenen Personen verbunden sind.[6] Es ist festzuhalten, dass das Risiko einer Identifizierung im Zeitverlauf steigen kann und auch von den Entwicklungen in der Informations- und Kommunikationstechnologie abhängig ist. Etwaige Rechtsvorschriften müssen daher technisch neutral formuliert werden und idealerweise den im Wandel befindlichen Entwicklungspotenzialen der Informationstechnologie Rechnung tragen.[7]

Zweitens wird mit den „Mitteln", die „vernünftigerweise" eingesetzt werden könnten, um eine Person zu bestimmen, auf jene Mittel abgestellt, die „von dem für die Verarbeitung Verantwortlichen oder von einem Dritten" eingesetzt werden könnten. Entscheidend ist also die Erkenntnis, dass wenn der für die Verarbeitung Verantwortliche die ursprünglichen Daten (die eine Identifizierung zulassen) auf Ereignisebene nicht löscht und einen Teil dieses Datenbestands (beispielsweise nach der Entfernung oder Maskierung der Daten, die eine Identifizierung zulassen) weitergibt, beinhaltet der entstandene Datenbestand nach wie vor personenbezogene Daten. Nur wenn der für die Verarbeitung Verantwortliche die Daten auf einer Ebene aggregiert, auf der keine Einzelereignisse mehr identifizierbar sind, kann der entstandene Datenbestand als anonym bezeichnet werden. Ein Beispiel: Wenn eine Organisation Daten über individuelle Reisen erhebt, gelten die individuellen Reisemuster auf Ereignisebene für jede Partei nach wie vor als personenbezogene Daten, solange der für die Verarbeitung Verantwortliche (oder eine andere Person) weiterhin Zugang zu den ursprünglichen Rohdaten hat, selbst wenn direkte Identifikatoren aus dem an Dritte weitergegebenen Datenbestand entfernt wurden. Löscht der für die Verarbeitung Verantwortliche jedoch die Rohdaten und stellt Dritten lediglich statistische Daten auf einer hohen Aggregationsebene zur Verfügung – beispielsweise die Information „Montags

[6] Interessanterweise lautet der (am 21. Oktober 2013) vorgelegte Änderungsantrag des Europäischen Parlaments zum Erwägungsgrund 23 des Entwurfs der Datenschutz-Grundverordnung: „Bei der Prüfung der Frage, ob Mittel nach allgemeinem Ermessen aller Voraussicht nach zur Identifizierung der Person genutzt werden, sollten alle objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, herangezogen werden, wobei sowohl die zum Zeitpunkt der Verarbeitung verfügbare Technologie als auch die technologische Entwicklung zu berücksichtigen sind."

[7] Siehe Stellungnahme 4/2007 der Artikel-29-Datenschutzgruppe, S. 18.

verkehren auf der Route X 160 % mehr Passagiere als Dienstags“ – gelten diese als anonyme Daten.

Eine wirksame Anonymisierungslösung sorgt dafür, dass keine Partei in der Lage ist, eine Person aus einem Datenbestand herauszugreifen, eine Verbindung zwischen zwei Datensätzen eines Datenbestands (oder zwischen zwei unabhängigen Datenbeständen) herzustellen oder durch Inferenz Informationen aus einem solchen Datenbestand abzuleiten. Insgesamt ist also festzuhalten, dass die Entfernung der Elemente, die eine direkte Identifizierung erlauben, für sich genommen nicht ausreicht, um eine Identifizierung der betroffenen Person zuverlässig auszuschließen. Häufig werden weitere Maßnahmen erforderlich sein, um eine Identifizierung zu verhindern. Welche Maßnahmen das sind, ist wiederum abhängig von Kontext und Zwecken der Verarbeitung, der die anonymisierten Daten unterzogen werden sollen.

BEISPIEL:
Genetische Datenprofile sind ein Beispiel für personenbezogene Daten, bei denen das Risiko einer Identifizierung bestehen kann, wenn als einzige Anonymisierungstechnik die Identitäten der DNA-Geber entfernt werden, da jedes Profil einzigartig ist. In der Forschungsliteratur[8] wurde bereits belegt, dass durch die Kombination öffentlich verfügbarer Quellen für genetische Informationen (z. B. Genealogie-Datenbanken, Todesanzeigen, Ergebnisse von Suchmaschinenabfragen) und der Metadaten von DNA-Gebern (Zeitpunkt der Entnahme, Alter, Wohnort) die Identität einzelner Personen aufgedeckt werden kann, selbst wenn die DNA „anonym“ entnommen wurde.

Beide Kategorien von Anonymisierungstechniken – Randomisierung und Generalisierung von Daten –[9] haben Schwachstellen. Jedoch kann jede von ihnen unter gewissen Umständen und in bestimmten Kontexten geeignet sein, den angestrebten Zweck zu erreichen, ohne die Privatsphäre der betroffenen Personen zu gefährden. Es muss klar sein, dass „Identifizierung“ nicht nur auf die Möglichkeit abstellt, Namen und/oder Adressen betroffener Personen zu ermitteln, sondern auch auf eine potenzielle Identifizierbarkeit durch Herausgreifen, Verknüpfung und Inferenz. Des Weiteren ist es für die Anwendbarkeit der Datenschutzvorschriften unerheblich, welche Absichten der für die Verarbeitung Verantwortliche oder der Empfänger der Daten verfolgen. Solange die Daten eine Identifizierung erlauben, fallen sie unter die Datenschutzvorschriften.

Verarbeiteten Dritte einen Datenbestand, der einer Anonymisierungstechnik unterzogen wurde (der also von dem ursprünglichen für die Verarbeitung Verantwortlichen anonymisiert und freigegeben wurde), können sie dies rechtmäßig tun, ohne die Datenschutzbestimmungen berücksichtigen zu müssen, sofern sie nicht in der Lage sind, die betroffenen Personen im ursprünglichen Datenbestand (direkt oder indirekt) zu bestimmen. Jedoch müssen Dritte die oben genannten Faktoren berücksichtigen, die sich aus dem spezifischen Kontext und den besonderen Umständen ergeben (einschließlich der besonderen Merkmale der vom ursprünglichen für die Verarbeitung Verantwortlichen angewandten Anonymisierungstechniken), wenn sie über die Art der Verwendung und insbesondere der Kombination dieser anonymisierten Daten für ihre eigenen Zwecke entscheiden, da hieraus unterschiedliche Folgen für die Art ihrer Haftung entstehen können. Begründen diese Faktoren und Merkmale ein nicht hinnehmbares Risiko einer Identifizierung betroffener Personen, fällt die Verarbeitung erneut in den Anwendungsbereich der Datenschutzvorschriften.

[8] Siehe John Bohannon, „Genealogy Databases Enable Naming of Anonymous DNA Donors“, in: *Science*, Band 339, Nr. 6117 (18. Januar 2013), S. 262.

[9] Die wichtigsten Unterscheidungsmerkmale dieser beiden Anonymisierungstechniken werden in Abschnitt 3 („Technische Analyse“) unten beschrieben.

Die oben stehende Aufstellung soll keinesfalls erschöpfend sein, sondern vielmehr als allgemeine Orientierungshilfe für einen Ansatz zur Bewertung des Identifizierungspotenzials eines gegebenen Datenbestands dienen, der einer Anonymisierung mittels der verschiedenen verfügbaren Techniken unterzogen wird. Alle oben genannten Faktoren können gleichermaßen als Risikofaktoren gelten, die sowohl von den für die Verarbeitung Verantwortlichen bei der Anonymisierung von Datenbeständen als auch von Dritten bei der Nutzung dieser „anonymisierten" Datenbestände für ihre eigenen Zwecke abgewogen werden müssen.

2.2.3. Risiken der Nutzung anonymisierter Daten

Erwägt der für die Verarbeitung Verantwortliche die Nutzung von Anonymisierungstechniken, muss er die folgenden Risiken berücksichtigen:

– Ein häufiger Irrtum liegt in der Annahme, dass pseudonymisierte Daten mit anonymisierten Daten gleichzusetzen seien. Im Abschnitt „Technische Analyse" wird erläutert, dass pseudonymisierte Daten nicht mit anonymisierten Informationen gleichzusetzen sind, da sie nach wie vor das Herausgreifen einer einzelnen betroffenen Person sowie die Verknüpfung unterschiedlicher Datenbestände erlauben. Pseudonymisierte Daten sind geeignet, eine Identifizierung zu ermöglichen, und fallen daher weiterhin in den Anwendungsbereich der Rechtsvorschriften über den Datenschutz. Besonders relevant ist dies im Kontext wissenschaftlicher, statistischer oder historischer Forschungsarbeiten.[10]

BEISPIEL:
Ein typisches Beispiel für die Fehleinschätzungen bezüglich der Pseudonymisierung bietet ein Vorfall, der AOL (America On Line) betraf und allgemein bekannt wurde. Im Jahr 2006 wurde eine Datenbank mit 20 Mio. Suchwörtern veröffentlicht, die von etwa 650 000 Nutzern im Laufe von drei Monaten eingegeben worden waren. Als einzige Maßnahme für den Schutz der Privatsphäre der Nutzer wurde die Nutzer-ID durch einen numerischen Wert ersetzt. In der Folge konnten einige Nutzer öffentlich identifiziert und ihr Standort ausgemacht werden. Von Suchmaschinen verwendete Query Strings bergen auch nach ihrer Pseudonymisierung ein sehr großes Identifizierungspotenzial, vor allem wenn sie mit weiteren Merkmalen wie IP-Adressen oder anderen Konfigurationsparametern der Nutzer kombiniert werden.

– Ein zweiter Irrtum liegt in der Annahme, dass für die betroffenen Personen im Falle ordnungsgemäß anonymisierter Daten (die alle oben genannten Bedingungen und Kriterien erfüllen und definitionsgemäß nicht mehr in den Anwendungsbereich der Datenschutzrichtlinie fallen) keinerlei Garantien mehr gelten. Dies ist in erster Linie deswegen falsch, weil auf die Nutzung dieser Daten unter Umständen andere Rechtsvorschriften anwendbar sind. So verbietet beispielsweise Artikel 5 Absatz 3 der Datenschutzrichtlinie für elektronische Kommunikation die Speicherung von und den Zugang zu das Endgerät betreffenden „Informationen" aller Art (einschließlich nicht personenbezogener Daten) ohne die Zustimmung des Teilnehmers/Nutzers. Dieses Verbot ist Teil des übergreifenden Grundsatzes der Vertraulichkeit der Kommunikation.

– Ein dritter Fehler wäre bei der Nutzung ordnungsgemäß anonymisierter Daten unter bestimmten Umständen die Nichtberücksichtigung der Auswirkungen auf die betroffenen Personen. Dies gilt insbesondere im Falle von Verarbeitungen, die auf Profiling basieren. Der Schutz der Privatsphäre des Einzelnen ist in Artikel 8 der Europäischen Menschenrechtskonvention und Artikel 7 der Charta der Grundrechte der EU verankert. Selbst wenn die Datenschutzvorschriften unter Umständen keine Anwendung mehr auf derartige Daten finden, kann doch die Nutzung von Datenbeständen, die für den Einsatz durch Dritte anonymisiert und freigegeben wurden, eine Verletzung der Privatsphäre zur Folge

[10] Siehe Stellungnahme 4/2007 der Artikel-29-Datenschutzgruppe, S. 21 ff.

haben. Besondere Sorgfalt ist im Umgang mit anonymisierten Daten geboten, wann immer diese Informationen (häufig in Kombination mit anderen Daten) genutzt werden, um Entscheidungen zu treffen, die Auswirkungen (wenn auch indirekt) auf Personen haben. Wie die Datenschutzgruppe bereits in dieser Stellungnahme betont und insbesondere in ihrer Stellungnahme über den Begriff der Zweckbindung (Stellungnahme 03/2013)[11] präzisiert hat, sollten die berechtigten Erwartungen der betroffenen Personen hinsichtlich der Weiterverarbeitung ihrer Daten vor dem Hintergrund der relevanten kontextabhängigen Faktoren bewertet werden, wie beispielsweise der Art der Beziehung zwischen den betroffenen Personen und den für die Verarbeitung Verantwortlichen, der geltenden gesetzlichen Verpflichtungen und der Transparenz der Verarbeitungsvorgänge.

3 Technische Analyse, Robustheit der Technologien und typische Fehler

Die verschiedenen Anonymisierungsverfahren und -techniken weisen ein unterschiedliches Maß an Robustheit auf. In diesem Abschnitt werden die wichtigsten Elemente erläutert, die von den für die Verarbeitung Verantwortlichen bei der Anwendung dieser Verfahren und Techniken zu berücksichtigen sind, indem insbesondere darauf abgestellt wird, welches Schutzniveau mit einer bestimmten Technik erzielt werden kann, wobei der aktuelle Stand der Technik und drei Risiken in Betracht zu ziehen sind, die im Zusammenhang mit der Anonymisierung von wesentlicher Bedeutung sind:

- *Herausgreifen (singling out)*, d. h. die Möglichkeit, in einem Datenbestand einige oder alle Datensätze zu isolieren, welche die Identifizierung einer Person ermöglichen;
- *Verknüpfbarkeit*, d. h. die Fähigkeit, mindestens zwei Datensätze, welche dieselbe Person oder Personengruppe betreffen, zu verknüpfen (in derselben Datenbank oder in zwei verschiedenen Datenbanken). Ist ein Angreifer in der Lage (z. B. mittels Korrelationsanalyse) festzustellen, dass zwei Datensätze dieselbe Personengruppe betreffen, ohne jedoch einzelne Personen in dieser Gruppe herauszugreifen, bietet die betreffende Technik zwar einen Schutz vor dem „Herausgreifen", nicht aber vor der Verknüpfbarkeit;
- *Inferenz*, d. h. die Möglichkeit, den Wert eines Merkmals mit einer signifikanten Wahrscheinlichkeit von den Werten einer Reihe anderer Merkmale abzuleiten.

Eine Lösung, die Schutz vor diesen drei Risiken bietet, wäre somit robust und geeignet, eine Reidentifizierung mit den Mitteln, die vernünftigerweise entweder von dem für die Verarbeitung Verantwortlichen oder von einem Dritten eingesetzt werden könnten, zu verhindern. In diesem Zusammenhang weist die Datenschutzgruppe nachdrücklich darauf hin, dass die Techniken der Deidentifizierung und Anonymisierung Gegenstand laufender Forschungen sind, die bislang immer wieder gezeigt haben, dass keine Technik per se vor Mängeln gefeit ist. Grob gesagt lassen sich zwei Anonymisierungsansätze unterscheiden: Der erste basiert auf ***Randomisierung***, der zweite auf ***Generalisierung***. In dieser Stellungnahme werden auch andere Konzepte wie Pseudonymisierung, Differential Privacy, l-Diversität und t-Closeness erörtert.

[11] Verfügbar unter http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

In diesem Abschnitt werden die folgenden Fachtermini verwendet: Ein Datenbestand umfasst unterschiedliche Datensätze, die einzelne Personen (die betroffenen Personen) zum Gegenstand haben. Jeder Datensatz bezieht sich auf eine betroffene Person und umfasst eine Reihe von Werten (oder „Einträgen“, z. B.: 2013) für jedes Merkmal (z. B. Jahr). Ein Datenbestand ist eine Sammlung von Datensätzen, die entweder als Tabelle (oder Tabellensatz) oder als annotierter/gewichteter Graph dargestellt werden kann. Letzteres ist heute in zunehmendem Maße der Fall. Die in dieser Stellungnahme angeführten Beispiele beziehen sich auf Tabellen, sind aber auch auf andere grafische Darstellungen von Datensätzen anwendbar. Merkmalskombinationen, die sich auf eine betroffene Person oder eine Gruppe betroffener Personen beziehen, werden als Quasi-Identifikatoren bezeichnet. Mitunter kann ein Datenbestand auch mehrere Datensätze zu ein und derselben Person beinhalten. Ein „Angreifer“ ist ein Dritter (d. h. weder der für die Verarbeitung Verantwortliche noch der Auftragsverarbeiter), der zufällig oder absichtlich auf die Originaldatensätze zugreift.

3.1. Randomisierung

Als Randomisierung bezeichnet man eine Reihe von Techniken, welche die Daten in einer Weise verfälschen, dass die direkte Verbindung zwischen Daten und betroffener Person entfernt wird. Sind die Daten ausreichend unbestimmt, können sie nicht mehr einer bestimmten Person zugeordnet werden. Durch Randomisierung alleine wird die Einzigartigkeit der einzelnen Datensätze nicht eingeschränkt, da jeder Datensatz nach wie vor eine einzige betroffene Person zum Gegenstand hat. Sie kann jedoch vor Angriffen mittels Inferenztechniken sowie vor Inferenzrisiken schützen und mit Generalisierungstechniken kombiniert werden, um einen stärkeren Schutz der Privatsphäre zu erreichen. Unter Umständen sind weitere Techniken erforderlich, um zu gewährleisten, dass die Datensätze nicht die Identifizierung einer einzelnen Person erlauben.

3.1.1. Stochastische Überlagerung

Die Technik der stochastischen Überlagerung ist insbesondere dann sinnvoll, wenn Merkmale erhebliche negative Auswirkungen auf Personen haben können. Mit ihrer Hilfe werden Merkmale im Datenbestand so verändert, dass sie weniger genau sind, wobei jedoch die allgemeine Verteilung aufrechterhalten bleibt. Im Zuge der Verarbeitung eines Datenbestands entsteht der Eindruck präziser Werte, der jedoch in einem gewissen Maße täuscht. Wurde beispielsweise die Körpergröße einer Person ursprünglich auf den Zentimeter genau gemessen, könnte im anonymisierten Datenbestand die Körpergröße nur auf +-10 cm genau angegeben werden. Wird diese Technik wirksam angewandt, sind Dritte nicht in der Lage, eine Person zu identifizieren, die Daten wiederherzustellen oder zu ermitteln, wie die Daten verändert wurden.

Die stochastische Überlagerung muss in der Regel mit anderen Anonymisierungstechniken wie beispielsweise der Entfernung offensichtlicher Merkmale und Quasi-Identifikatoren kombiniert werden. Der Grad der Überlagerung sollte von der benötigten Informationsebene und den Auswirkungen einer Aufdeckung der geschützten Merkmale auf die Privatsphäre der betroffenen Personen abhängig gemacht werden.

3.1.1.1. Schutzniveau

- *Herausgreifen:* Es ist nach wie vor möglich, die Datensätze einer Einzelperson herauszugreifen (wobei diese unter Umständen keine Identifizierung zulassen), wenn auch die Datensätze weniger präzise sind.
- *Verknüpfbarkeit:* Es ist nach wie vor möglich, die Datensätze einer bestimmten Person zu verknüpfen, allerdings sind die Datensätze weniger präzise, wodurch es möglich wird, dass ein realer Datensatz mit einem im Nachhinein hinzugefügten Datensatz verknüpft wird (d. h. mit der „Überlagerung"). In einigen Fällen kann eine falsche Zuordnung für die betroffene Person ein erhebliches oder sogar ein größeres Risiko darstellen als eine korrekte.
- *Inferenz:* Angriffe mittels Inferenztechniken sind unter Umständen möglich, ihre Erfolgsrate wird jedoch geringer sein und es ist plausibel, dass sie einige falsch positive (und falsch negative) Ergebnisse hervorbringen.

3.1.1.2. Häufige Fehler

- Verwendung inkonsistenter Störgrößen: Ist die Überlagerung semantisch nicht glaubwürdig (d. h. ist sie „übertrieben" und missachtet die logischen Zusammenhänge zwischen den in einem Datenbestand erfassten Merkmalen), wird ein Angreifer, der auf die Datenbank zugreift, in der Lage sein, die Überlagerung herauszufiltern und mitunter die fehlenden Einträge zu rekonstruieren. Ist der Datenbestand ferner zu schwach besetzt[12], ist es unter Umständen weiterhin möglich, die überlagerten Dateneinträge mit einer externen Quelle zu verknüpfen.
- Annahme, dass die stochastische Überlagerung ausreichend ist: Die stochastische Überlagerung stellt eine ergänzende Maßnahme dar, die Angreifern den Zugriff auf personenbezogene Daten erschwert. Sofern nicht die Überlagerung die im Datenbestand enthaltenen Informationen überwiegt, kann nicht angenommen werden, dass die stochastische Überlagerung alleine als Anonymisierungstechnik ausreicht.

3.1.1.3. Schwachstellen der stochastischen Überlagerung

Ein sehr bekanntes Reidentifizierungs-Experiment wurde anhand der Kundendatenbank des Video-Anbieters Netflix durchgeführt. Wissenschaftler analysierten die geometrischen Eigenschaften dieser Datenbank, in der mehr als 100 Mio. Bewertungen auf einer Skala von 1 bis 5 erfasst waren, die von fast 500 000 Nutzern zu über 18 000 Filmen abgegeben worden waren. Diese Bewertungen wurden entsprechend internen Datenschutzleitlinien „anonymisiert", wobei alle Identitätsmerkmale der Kunden mit Ausnahme ihrer Bewertungen und deren Zeitpunkt entfernt wurden, und anschließend durch das Unternehmen veröffentlicht. Es wurde eine stochastische Überlagerung vorgenommen, indem die Bewertungen leicht verbessert oder verschlechtert wurden.

Nichtsdestotrotz wiesen die Wissenschaftler nach, dass 99 % der im Datenbestand erfassten Nutzer anhand von acht Bewertungen und den auf 14 Tage genau bekannten Zeitpunkten ihrer Abgabe und 68 % der Nutzer sogar anhand von nur zwei Bewertungen und den auf drei Tage genau bekannten Zeitpunkten ihrer Abgabe identifiziert werden konnten.[13]

[12] Dieser Begriff wird im Anhang (S. 30) weiter erörtert.

[13] Narayanan, A. und Shmatikov, V. (Mai 2008), „Robust de-anonymization of large sparse datasets", in: *Security and Privacy, 2008. SP 2008. IEEE Symposium on Security and Privacy* (S. 111 ff.), IEEE.

3.1.2. Vertauschung

Diese Technik fußt auf der Vertauschung der Merkmalswerte in einer Tabelle, sodass einige von ihnen künstlich mit anderen betroffenen Personen verknüpft werden. Ein solches Vorgehen ist sinnvoll, wenn es wichtig ist, dass die exakte Verteilung eines jeden Merkmals im Datenbestand erhalten bleibt.

Die Vertauschung kann als eine spezielle Form der stochastischen Überlagerung betrachtet werden. Bei der klassischen stochastischen Überlagerung werden Merkmale mithilfe von Zufallsgrößen verändert. Eine kohärente stochastische Überlagerung kann sich als schwierig erweisen. Zudem sorgt die leichte Modifizierung der Merkmalswerte unter Umständen nicht für einen hinreichenden Schutz der Privatsphäre. Alternativ dazu werden mithilfe von Vertauschungstechniken Merkmalswerte innerhalb des Datenbestands dadurch geändert, dass sie lediglich von einem Datensatz in einen anderen verschoben werden. Ein solches „Swapping“ sorgt dafür, dass Bandbreite und Verteilung der Merkmalswerte unverändert erhalten bleiben, nicht jedoch die Korrelationen zwischen den Merkmalswerten und den betroffenen Personen. Besteht zwischen zwei oder mehr Merkmalen ein logischer Zusammenhang oder eine statistische Korrelation und werden diese Merkmale unabhängig voneinander vertauscht, geht dieser Zusammenhang verloren. Es ist daher unter Umständen geboten, ganze Reihen miteinander in Zusammenhang stehender Merkmale zu vertauschen, um die logischen Beziehungen nicht zu zerstören, da andernfalls ein Angreifer die vertauschten Merkmale auffinden und die Vertauschung rückgängig machen kann.

Betrachtet man beispielsweise eine Untergruppe von Merkmalen in einem medizinischen Datenbestand, wie „Gründe für die Einweisung ins Krankenhaus/Symptome/zuständige Abteilung“, so besteht in den meisten Fällen ein starker logischer Zusammenhang zwischen den Werten, sodass die Vertauschung nur eines Wertes aufgespürt und sogar rückgängig gemacht werden könnte.

Ähnlich wie bei der stochastischen Überlagerung kann eine Vertauschung alleine keine Anonymisierung gewährleisten und sollte stets mit der Entfernung von offensichtlichen Merkmalen/Quasi-Identifikatoren kombiniert werden.

3.1.2.1. Schutzniveau

- *Herausgreifen:* Wie bei der stochastischen Überlagerung ist es nach wie vor möglich, die Datensätze einer Einzelperson herauszugreifen, wobei diese jedoch weniger präzise sind.
- *Verknüpfbarkeit:* Betrifft die Vertauschung Merkmale und Quasi-Identifikatoren, verhindert sie unter Umständen eine „korrekte“ Verknüpfung von Merkmalen innerhalb und außerhalb eines Datenbestands, erlaubt jedoch „fehlerhafte“ Verknüpfungen, da ein realer Eintrag einer anderen betroffenen Person zugeordnet werden kann.
- *Inferenz:* Es ist nach wie vor möglich, durch Inferenztechniken Informationen aus dem Datenbestand abzuleiten. Dies gilt insbesondere, wenn Merkmale korrelieren oder einen starken logischen Zusammenhang aufweisen. Weiß der Angreifer jedoch nicht, welche Merkmale vertauscht wurden, muss er in Betracht ziehen, dass seine Inferenztechnik auf einer falschen Hypothese basiert. Infolgedessen ist nur eine probabilistische Inferenz möglich.

3.1.2.2. Häufige Fehler

- Auswahl des falschen Merkmals: Die Vertauschung der Werte nicht sensitiver oder nicht risikobehafteter Merkmale führt zu keiner wesentlichen Verbesserung des Schutzes personenbezogener Daten. In diesem Fall bleiben die sensitiven/risikobehafteten Merkmalswerte mit dem ursprünglichen Merkmal verbunden, sodass ein Angreifer trotzdem in der Lage wäre, sensitive Informationen über einzelne Personen zu gewinnen.
- Zufällige Vertauschung von Merkmalswerten: Besteht ein starker Zusammenhang zwischen zwei Merkmalen, wird eine zufällige Vertauschung der Merkmalswerte nicht für hinreichende Garantien sorgen. Dieser häufige Fehler wird in Tabelle 1 veranschaulicht.
- Annahme, dass die Vertauschung ausreicht: Wie die stochastische Überlagerung kann auch die Vertauschung alleine nicht für Anonymität sorgen und sollte mit anderen Techniken, wie beispielsweise der Entfernung offensichtlicher Merkmale, kombiniert werden.

3.1.2.3. Schwachstellen der Vertauschung

Dieses Beispiel zeigt, dass eine zufällige Vertauschung von Merkmalswerten nur einen unzureichenden Schutz der Privatsphäre bewirken kann, wenn zwischen verschiedenen Merkmalen logische Zusammenhänge bestehen. Nach diesem Anonymisierungsversuch ist es ohne Weiteres möglich, das Einkommen jeder einzelnen Person in Abhängigkeit von ihrem Beruf (und dem Geburtsjahr) abzuleiten. Beispielsweise kann durch eine direkte Untersuchung der Daten gezeigt werden, dass der in der Tabelle aufgeführte CEO sehr wahrscheinlich im Jahr 1957 geboren wurde und das höchste Einkommen hat, während der Arbeitslose im Jahr 1964 geboren wurde und das geringste Einkommen hat.

Jahr	Geschlecht	Beruf	Einkommen (vertauscht)
1957	M	Ingenieur	70 000
1957	M	CEO	5000
1957	M	Arbeitsloser	43 000
1964	M	Ingenieur	100 000
1964	M	Manager	45 000

Tabelle 1. Beispiel für eine unwirksame Anonymisierung durch Vertauschung der Werte korrelierender Merkmale

3.1.3. Differential Privacy

Das Konzept der Differential Privacy[14] zählt zur Kategorie der Randomisierungstechniken, verfolgt jedoch einen anderen Ansatz: Während die Überlagerung vorwiegend dann ins Spiel kommt, wenn ein Datenbestand freigegeben werden soll, kann das Konzept der Differential Privacy herangezogen werden, wenn der für die Verarbeitung Verantwortliche anonymisierte Ansichten eines Datenbestands generiert und zugleich eine Kopie der Originaldaten aufbewahrt. Solche anonymisierten Ansichten werden in der Regel mittels einer abgefragten Datenuntergruppe für einen Dritten erstellt. Die Untergruppe umfasst gewisse Überlagerungen, die im Nachhinein bewusst hinzugefügt wurden. Das Konzept der Differential Privacy verrät dem für die Verarbeitung Verantwortlichen, wie stark die

[14] Dwork, C. (2006), „Differential privacy“, in: *Automata, languages and programming* (S. 1 ff.), Springer Berlin Heidelberg.

Überlagerungen sein und in welcher Form sie hinzugefügt werden müssen, um den erforderlichen Schutz der Privatsphäre zu gewährleisten.[15] In diesem Zusammenhang ist eine kontinuierliche Überwachung (zumindest für jede neue Abfrage) von besonderer Bedeutung, um etwaige Möglichkeiten für eine Identifizierung von Personen in der abgefragten Datenuntergruppe auszumachen. Es muss jedoch klargestellt werden, dass durch die auf dem Konzept der Differential Privacy beruhenden Techniken die Originaldaten als solche nicht verändert werden. Berücksichtigt man alle Mittel, die von dem der für die Verarbeitung Verantwortlichen vernünftigerweise eingesetzt werden, ist dieser demnach in der Lage, Personen in den Abfrageergebnissen zu identifizieren, solange die Originaldaten erhalten bleiben. Derartige Abfrageergebnisse sind also als personenbezogene Daten zu betrachten.

Ein Vorteil von auf dem Begriff der Differential Privacy beruhenden Ansätzen ist die Tatsache, dass Datenuntergruppen autorisierten Dritten in Beantwortung einer konkreten Anfrage zur Verfügung gestellt werden und kein vollständiger Datenbestand freigegeben wird. Zu Prüfzwecken kann der für die Verarbeitung Verantwortliche eine Aufstellung aller Abfragen und Anfragen aufbewahren, um sicherzustellen, dass Dritte nicht auf Daten zugreifen, für die sie keine Berechtigung besitzen. Darüber hinaus kann das Abfrageergebnis Anonymisierungstechniken wie einer stochastischen Überlagerung oder Ersetzung unterzogen werden, um eine Verbesserung des Schutzes der Privatsphäre zu erreichen. In Forschungsarbeiten wird noch immer nach einem geeigneten interaktiven Abfrage-Antwort-Mechanismus gesucht, der zugleich in der Lage ist, Abfragen hinreichend präzise zu beantworten (d. h. mit möglichst geringen Überlagerungen) und zugleich den Schutz der Privatsphäre zu gewährleisten.

Um Angriffe durch Inferenztechniken und Verknüpfungen zu begrenzen, ist es erforderlich, die Abfragen eines Nutzers und die damit gewonnenen Informationen über die betroffenen Personen zu erfassen. Dementsprechend sollten auf dem Konzept der Differential Privacy beruhende Datenbanken nicht für Open-Source-Suchmaschinen bereitgestellt werden, bei denen nicht nachvollzogen werden kann, welcher Nutzer welche Abfragen gestellt hat.

3.1.3.1 Garantien

- *Herausgreifen:* Werden lediglich Statistiken bereitgestellt und werden auf die Datenuntergruppe geeignete Regelungen angewandt, sollte es unmöglich sein, anhand der Antworten eine einzelne Person herauszugreifen.
- *Verknüpfbarkeit:* Mithilfe von Mehrfachabfragen könnte es möglich sein, eine Verknüpfung von in zwei unterschiedlichen Antworten übermittelten Einträgen zu einer bestimmten Person vorzunehmen.
- *Inferenz:* Es ist möglich, durch die Anwendung von Inferenztechniken auf Mehrfachabfragen Informationen über Personen oder Gruppen abzuleiten.

3.1.3.2. Häufige Fehler

- Unzureichende Überlagerung: Um eine Verknüpfung mit Hintergrundwissen zu verhindern, dürfen nur möglichst wenige Hinweise darauf gegeben werden, ob eine bestimmte betroffene Person oder eine Gruppe betroffener Personen in der Datenuntergruppe erfasst ist. Aus datenschutztechnischer Sicht besteht die größte

[15] Siehe Ed Felten (2012), Protecting privacy by adding noise, verfügbar unter https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

Schwierigkeit darin, die Überlagerung der tatsächlichen Antworten so zu bemessen, dass zum einen der Schutz der Privatsphäre der betroffenen Personen und zum anderen die Zweckmäßigkeit der freigegebenen Antworten gewährleistet sind.

3.1.3.3 Schwachstellen der Differential Privacy

Unabhängige Bearbeitung jeder Abfrage: Die Kombination von Abfrageergebnissen kann unter Umständen zur Offenlegung von Informationen führen, die geheim bleiben sollten. Wird keine Abfragehistorie erstellt, kann ein Angreifer Mehrfachanfragen an eine auf dem Konzept der Differential Privacy beruhende Datenbank stellen und damit den Umfang der ausgegebenen Stichprobe schrittweise verringern, bis er ein bestimmtes Merkmal einer betroffenen Person oder einer Gruppe betroffener Personen eindeutig oder mit einer sehr hohen Wahrscheinlichkeit bestimmen kann. Des Weiteren darf man sich nicht darauf verlassen, dass die Daten für einen Dritten anonym sind, während der für die Verarbeitung Verantwortliche die betroffene Person nach wie vor in der Originaldatenbank mit den vernünftigerweise einsetzbaren Mitteln identifizieren kann.

3.2. Generalisierung

Generalisierung bildet die zweite Kategorie von Anonymisierungstechniken. Bei diesem Ansatz werden die Merkmale betroffener Personen durch die Veränderung der entsprechenden Größenskala oder -ordnung generalisiert, d. h. durch einen weniger spezifischen Wert ersetzt (d. h. durch die Angabe einer Region statt einer Stadt oder eines Monats statt einer Woche). Das Verfahren der Generalisierung kann zwar wirksam das Herausgreifen einzelner Personen verhindern, es ermöglicht jedoch nicht in jedem Falle eine effektive Anonymisierung. Insbesondere setzt es einen spezifischen und ausgefeilten quantitativen Ansatz voraus, um Verknüpfbarkeit und Inferenzen vorzubeugen.

3.2.1. Aggregation und k-Anonymität

Die Techniken der Aggregation und k-Anonymität zielen darauf ab, das Herausgreifen einer betroffenen Person zu verhindern, indem diese mit mindestens k anderen Personen zusammengefasst wird. Um dies zu erreichen, werden die Merkmalswerte in einem Maße generalisiert, dass alle k Personen denselben Merkmalswert aufweisen. Indem beispielsweise die Granularität der Standortdaten von der Stadt auf das Land vergröbert wird, trifft ein Merkmal auf eine größere Gruppe betroffener Personen zu. Geburtsdaten können durch die Bildung von Datumsintervallen oder die Gruppierung nach Monat oder Jahr generalisiert werden. Andere numerische Merkmale (z. B. Einkommen, Gewicht, Körpergröße oder Arzneimitteldosierungen) können durch Intervallwerte generalisiert werden. Diese Methoden können verwendet werden, wenn die Korrelation punktueller Merkmalswerte die Bildung von Quasi-Identifikatoren ermöglicht.

3.2.1.1. Schutzniveau

- *Herausgreifen:* Da nun k Personen dieselben Merkmalswerte aufweisen, sollte es nicht länger möglich sein, eine Einzelperson aus einer Gruppe von k Personen herauszugreifen.
- *Verknüpfbarkeit:* Die Verknüpfbarkeit wird zwar eingeschränkt, es ist jedoch nach wie vor möglich, Datensätze nach Gruppen von k Personen zu verknüpfen. Innerhalb dieser Gruppe beträgt dann die Wahrscheinlichkeit, dass zwei Datensätze denselben Pseudo-Identifikatoren entsprechen, 1/k (und ist damit womöglich signifikant höher als die Wahrscheinlichkeit, dass diese Einträge nicht verknüpfbar sind).

- *Inferenz:* Die größte Schwachstelle des Modells der k-Anonymität liegt darin, dass es keinen Schutz vor irgendeiner Form von Angriffen durch Inferenztechniken bietet. Denn wenn alle k Personen ein und derselben Gruppe angehören und bekannt ist, zu welcher Gruppe eine Person gehört, ist es sehr einfach, den Wert einer Eigenschaft zu ermitteln.

3.2.1.2. Häufige Fehler

- Fehlen einiger Quasi-Identifikatoren: Ein maßgeblicher Parameter für die k-Anonymität ist der Schwellenwert von k. Je höher der Wert von k, desto stärker ist der Schutz der Privatsphäre. Ein häufiger Fehler besteht darin, den Wert k durch die Reduzierung der berücksichtigten Gruppe von Quasi-Identifikatoren künstlich anzuheben. Die Reduzierung der Quasi-Identifikatoren macht es einfacher, Cluster aus k Personen zu bilden, da die anderen Merkmale geeignet sind, eine Identifizierung zu ermöglichen (insbesondere wenn einige von ihnen sensitive Merkmale sind oder eine sehr hohe Entropie aufweisen, wie dies bei sehr seltenen Merkmalen der Fall ist). Ein entscheidender Fehler ist es, bei der Auswahl des zu generalisierenden Merkmals nicht alle Quasi-Identifikatoren zu berücksichtigen. Können einige Merkmale genutzt werden, um eine Einzelperson aus einem Cluster aus k Personen herauszugreifen, ist die Generalisierung nicht geeignet, einige Personen zu schützen (siehe das Beispiel in Tabelle 2).
- Geringer k-Wert: Wird ein geringer k-Wert angestrebt, ist dies ähnlich problematisch. Ist k zu klein, ist das Gewicht der einzelnen Personen in einem Cluster zu signifikant und Angriffe mittels Inferenztechniken haben eine höhere Erfolgsquote. Ist beispielsweise k gleich 2, ist die Wahrscheinlichkeit, dass die beiden Personen dieselbe Eigenschaft aufweisen, höher, als bei einem k-Wert von über 10.
- Keine Gruppierung von Personen mit demselben Gewicht: Die Gruppierung von Personen mit einer ungleichmäßigen Verteilung von Merkmalen kann ebenfalls problematisch sein. Das Gewicht der Datensätze einzelner Personen innerhalb des Datenbestands ist dann unterschiedlich groß: Während einige einen signifikanten Teil der Einträge ausmachen, sind die anderen nahezu unbedeutend. Daher ist es wichtig, dass k groß genug ist, um zu verhindern, dass Personen einen zu großen Anteil der Einträge innerhalb eines Clusters ausmachen.

3.2.1.3. Schwachstellen der k-Anonymität

Das größte Problem der k-Anonymität liegt darin, dass dieses Konzept keinen Schutz vor Angriffen durch Inferenztechniken bietet. Weiß der Angreifer im folgenden Beispiel, dass eine bestimmte Person im Datenbestand erfasst ist und 1964 geboren wurde, weiß er auch, dass diese Person einen Herzinfarkt hatte. Ist ferner bekannt, dass dieser Datenbestand von einer französischen Organisation bereitgestellt wurde, ist klar, dass alle Personen in Paris leben, da die Pariser Postleitzahlen mit der Ziffernfolge 750* beginnen.

Jahr	Geschlecht	PLZ	Diagnose
1957	M	750*	Herzinfarkt
1957	M	750*	Cholesterin
1957	M	750*	Cholesterin
1964	M	750*	Herzinfarkt
1964	M	750*	Herzinfarkt

Tabelle 2. Beispiel für eine schlecht geplante k-Anonymisierung

3.2.2. L-Diversität und t-Closeness

Das Konzept der l-Diversität erweitert die k-Anonymität, um sicherzustellen, dass keine deterministischen Angriffe mittels Inferenztechniken mehr möglich sind, indem dafür gesorgt wird, dass die einzelnen Merkmale in jeder Äquivalenzklasse mindestens l verschiedene Werte aufweisen.

Eine grundlegende Zielsetzung ist die Begrenzung des Auftretens von Äquivalenzklassen mit einer geringen Variabilität der Merkmalswerte, sodass für Angreifer mit Hintergrundwissen über eine bestimmte betroffene Person stets eine signifikante Unsicherheit gewährleistet ist.

Das Konzept der l-Diversität ist geeignet, Daten vor Angriffen mittels Inferenztechniken zu schützen, wenn die Merkmalswerte gut verteilt sind. Es ist jedoch darauf hinzuweisen, dass diese Technik nicht das Durchsickern von Informationen verhindern kann, wenn die Merkmalswerte innerhalb einer Partition uneinheitlich verteilt sind, eine geringe Bandbreite aufweisen oder semantisch ähnlich sind. Letztendlich bietet das Konzept der l-Diversität Raum für Angriffe mittels probabilistischer Inferenz.

Der Ansatz der t-Closeness verfeinert das Konzept der l-Diversität, indem versucht wird, Äquivalenzklassen zu bilden, die der ursprünglichen Verteilung der Merkmalswerte in der Tabelle ähneln. Diese Technik ist zweckdienlich, wenn es wichtig ist, die ursprünglichen Daten möglichst wenig zu verändern. Hierfür wird eine weitere Bedingung für die Äquivalenzklasse eingeführt: Es müssen nicht nur mindestens *l* verschiedene Werte in jeder Äquivalenzklasse vertreten sein, es ist auch erforderlich, dass jeder Wert so oft vertreten ist, dass die ursprüngliche Verteilung für jedes einzelne Merkmal abgebildet wird.

3.2.2.1. Schutzniveau

- *Herausgreifen:* Wie k-Anonymität können auch l-Diversität und t-Closeness gewährleisten, dass es unmöglich ist, Datensätze zu einer Person aus einer Datenbank herauszugreifen.
- *Verknüpfbarkeit:* L-Diversität und t-Closeness stellen im Hinblick auf die Einschränkung der Verknüpfbarkeit keine Verbesserung dar. Es stellt sich dieselbe Problematik wie bei jedem Cluster: Die Wahrscheinlichkeit, dass dieselben Einträge zu derselben betroffenen Person gehören, ist größer als 1/N (wobei N die Zahl der betroffenen Personen in der Datenbank bezeichnet).
- *Inferenz:* Die wichtigste Verbesserung, die l-Diversität und t-Closeness gegenüber der k-Anonymität mit sich bringen, liegt darin, dass Angriffe mittels Inferenztechniken gegen Datenbanken, die l-Diversität oder t-Closeness aufweisen, niemals zu 100 % sicher sein können.

3.2.2.2. Häufige Fehler

- Schutz sensitiver Merkmale durch ihre Vermischung mit anderen sensitiven Merkmalen: Es genügt nicht, zwei Werte eines Merkmals in einem Cluster zu haben, um einen Schutz der Privatsphäre zu bewirken. Tatsächlich sollte die Verteilung der Werte sensitiver Merkmale in jedem Cluster der Verteilung dieser Werte in der gesamten Population ähneln oder zumindest innerhalb des Clusters einheitlich sein.

3.2.2.3. Schwachstellen der l-Diversität

In der unten stehenden Tabelle ist l-Diversität bezüglich des Merkmals „Diagnose“ gewährleistet. Weiß ein Angreifer jedoch, dass eine in der Tabelle erfasste Person 1964 geboren ist, ist er nach wie vor in der Lage, mit sehr hoher Wahrscheinlichkeit anzunehmen, dass diese Person einen Herzinfarkt hatte.

Jahr	Geschlecht	PLZ	Diagnose
1957	M	750*	**Herzinfarkt**
1957	M	750*	**Cholesterin**
1957	M	750*	**Cholesterin**
1957	M	750*	**Cholesterin**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Cholesterin**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Herzinfarkt**
1964	M	750*	**Herzinfarkt**

Tabelle 3. Tabelle mit l-Diversität, in der die Werte des Merkmals „Diagnose“ nicht einheitlich verteilt sind

Name	Geburtsdatum	Geschlecht
Smith	1964	M
Rossi	1964	M
Dupont	1964	M
Jansen	1964	M
Garcia	1964	M

Tabelle 4. Weiß ein Angreifer, dass diese Personen in Tabelle 3 erfasst sind, könnte er durch Inferenz ableiten, dass sie einen Herzinfarkt hatten.

4. Pseudonymisierung

Im Zuge der Pseudonymisierung wird ein Merkmal (in der Regel ein einzigartiges Merkmal) in einem Datensatz durch ein anderes ersetzt. Die natürliche Person kann daher nach wie vor mit großer Wahrscheinlichkeit indirekt identifiziert werden. Dementsprechend kann Pseudonymisierung alleine niemals einen anonymen Datenbestand hervorbringen. Sie wird in dieser Stellungnahme dennoch erörtert, da mit ihrem Einsatz zahlreiche Missverständnisse und Fehler verbunden sind.

Pseudonymisierung verringert die Verknüpfbarkeit eines Datenbestands mit der wahren Identität einer betroffenen Person und stellt somit eine sinnvolle Sicherheitsmaßnahme, aber kein Anonymisierungsverfahren dar.

Das Ergebnis der Pseudonymisierung kann unabhängig vom ursprünglichen Wert sein (wie im Falle einer von dem für die Verarbeitung Verantwortlichen generierten Zufallszahl oder eines von der betroffenen Person gewählten Beinamens) oder von den Originalwerten eines Merkmals oder einer Reihe von Merkmalen abgeleitet sein, beispielsweise durch Hashfunktionen oder Verschlüsselungsverfahren.

Zumeist werden die folgenden Pseudonymisierungstechniken eingesetzt:

- Verschlüsselung mit einem Geheimschlüssel: In diesem Fall kann der Inhaber des Schlüssels jede betroffene Person ohne Weiteres durch die Entschlüsselung des Datenbestands reidentifizieren, da die personenbezogenen Daten nach wie vor im Datenbestand vorhanden sind, wenn auch in verschlüsselter Form. Bei dem neuesten Stand der Technik entsprechenden Verschlüsselungsverfahren ist eine Entschlüsselung ausschließlich mithilfe des Schlüssels möglich.
- Hashfunktion (Streuwertfunktion): Dabei handelt es sich um eine Funktion, die eine beliebig große Eingabemenge auf eine bestimmte Zielmenge abbildet (die Eingabemenge kann entweder ein einzelnes Merkmal oder eine Reihe von Merkmalen umfassen) und nicht umkehrbar ist; das bedeutet, dass das im Falle der Verschlüsselung aufgezeigte Risiko der Umkehrung nicht mehr besteht. Ist jedoch der Bereich der Eingabewerte der Hashfunktion bekannt, kann der korrekte Wert eines bestimmten Datensatzes erschlossen werden, in dem die Eingabewerte durch eine Hashfunktion geleitet werden. Wurde beispielsweise ein Datenbestand durch Hashen der Landes-ID pseudonymisiert, können die Werte ohne Weiteres abgeleitet werden, indem der Angreifer alle möglichen Eingabewerte hasht und das Ergebnis mit den Werten im Datenbestand vergleicht. Hashfunktionen werden in der Regel so angelegt, dass sie relativ schnell zu verarbeiten sind. Gehashte Datenbestände sind anfällig für Angriffe mittels der Exhaustionsmethode (Brute-Force-Methode).[16] Zudem ist es möglich, Rainbow-Tables zu erzeugen, welche die Rekonstruktion einer umfassenden Reihe von Hashwerten ermöglicht.

  Die Verwendung gesalzener Hashes (wobei die Merkmalswerte mit einem Zufallswert – dem „Salz“ (Salt) – versehen werden, bevor sie gehasht werden), kann die Wahrscheinlichkeit einer möglichen Ableitung der Eingabewerte verringern. Auch hier ist es jedoch unter Umständen nach wie vor möglich, den hinter einem solchen

[16] Im Zuge solcher Angriffe werden alle plausiblen Eingabewerte durchprobiert, um Korrespondenztabellen zu generieren.

gesalzenen Hash verborgenen Ursprungswert eines Merkmals mit den vernünftigerweise einsetzbaren Mitteln zu berechnen.[17]

- Schlüsselabhängige kryptologische Hashfunktionen: Dies ist eine besondere Form der Hashfunktion, bei der ein Geheimschlüssel als zusätzlicher Eingabewert verwendet wird (der Unterschied zu gesalzenen Hashes liegt darin, dass das Salz in der Regel nicht geheim ist). Der für die Verarbeitung Verantwortliche kann die Funktion unter Verwendung des Geheimschlüssels erneut auf das Merkmal anwenden, für den Angreifer ist es jedoch bedeutend schwieriger, alle möglichen Eingabewerte durch die Funktion zu leiten, ohne den Schlüssel zu kennen, da die Zahl der zu probierenden Möglichkeiten hinreichend groß ist, um zu erreichen, dass hierfür ein unverhältnismäßiger Aufwand erforderlich wäre.

- Deterministische Verschlüsselung oder schlüssellose kryptologische Hashfunktion: Diese Technik entspricht der Auswahl einer Zufallszahl als Pseudonym für jedes Merkmal in der Datenbank mit anschließendem Löschen der Korrespondenztabelle. Diese Vorgehensweise erlaubt[18] es, das Risiko der Verknüpfbarkeit zwischen den personenbezogenen Daten im Datenbestand und den in einem anderen Datenbestand, in dem ein anderes Pseudonym verwendet wird, enthaltenen personenbezogenen Daten derselben Person zu verringern. Angesichts der modernen Algorithmen muss ein Angreifer eine enorme Rechenleistung aufbringen, um die Funktion zu entschlüsseln oder für alle möglichen Eingabewerte zu wiederholen, da kein Schlüssel verfügbar ist und jeder mögliche Schlüssel ausprobiert werden müsste.

- Tokenisierung: Diese Technik kommt in der Regel (wenn auch nicht ausschließlich) im Finanzsektor zur Anwendung, um Karten-IDs durch Werte zu ersetzen, die für einen Angreifer weniger zweckmäßig sind. Sie wurde auf der Grundlage der vorstehend erläuterten Techniken entwickelt und basiert auf der Anwendung eines Einweg-Verschlüsselungsmechanismus oder der Zuweisung von fortlaufenden Nummern oder nicht mathematisch aus den Originaldaten abgeleiteten Zufallszahlen im Wege einer Indexfunktion.

4.1. Schutzniveau

- *Herausgreifen:* Es ist nach wie vor möglich, Datensätze einzelner Personen herauszugreifen, da die Person weiterhin anhand eines einzigartigen Merkmals identifiziert wird, das im Zuge der Pseudonymisierungsfunktion erzeugt wurde (das pseudonymisierte Merkmal).

- *Verknüpfbarkeit:* Datensätze, in denen für eine bestimmte Person dasselbe pseudonymisierte Merkmal verwendet wird, können ganz einfach verknüpft werden. Selbst wenn für ein und dieselbe betroffene Person unterschiedliche pseudonymisierte Merkmale verwendet werden, kann eine Verknüpfung mittels anderer Merkmale unter Umständen nach wie vor möglich sein. Nur wenn kein anderes Merkmal im Datenbestand verwendet werden kann, um eine betroffene Person zu identifizieren, und wenn jede Verbindung zwischen dem Originalmerkmal und dem pseudonymisierten Merkmal ausgeschlossen wurde (unter anderem durch die Löschung der Originaldaten), bestehen zwischen zwei Datenbeständen, in denen

[17] Dies gilt insbesondere dann, wenn die Art des Merkmals bekannt ist (Name, Sozialversicherungsnummer, Geburtsdatum usw.). Um die erforderliche Rechenlast zu erhöhen, könnte auch eine Hashfunktion zur Schlüsselableitung durchgeführt werden. Hierbei wird der berechnete Wert unter Verwendung eines kurzen Salzwertes mehrmals gehasht.

[18] In Abhängigkeit von den anderen Merkmalen im Datenbestand und der Löschung der Originaldaten.

unterschiedliche pseudonymisierte Merkmale verwendet werden, keine offensichtlichen Kreuzverweise mehr.

- *Inferenz:* Angriffe mittels Inferenztechniken zwecks Ermittlung der wahren Identität einer betroffenen Person sind innerhalb eines Datenbestands sowie über mehrere unterschiedliche Datenbanken, die für eine Person dasselbe pseudonymisierte Merkmal verwenden, hinweg möglich. Gleiches gilt, wenn die Pseudonyme selbsterklärend sind und die wahre Identität der betroffenen Person nicht ordnungsgemäß maskieren.

4.2. Häufige Fehler

- Annahme, dass ein pseudonymisierter Datenbestand anonymisiert ist: Häufig gehen für die Verarbeitung Verantwortliche davon aus, dass die Entfernung oder Ersetzung von einem oder mehreren Merkmalen ausreicht, um einen Datenbestand zu anonymisieren. Zahlreiche Beispiele belegen, dass dies nicht der Fall ist. Die einfache Änderung der ID verhindert nicht, dass Angreifer eine betroffene Person identifizieren, wenn der Datenbestand nach wie vor Quasi-Identifikatoren enthält oder die Werte anderer Merkmale noch immer geeignet sind, eine Person zu identifizieren. In vielen Fällen kann eine Person in einem pseudonymisierten Datenbestand ebenso einfach identifiziert werden wie im ursprünglichen Datenbestand. Es sollten zusätzliche Schritte unternommen werden, bevor ein Datenbestand als anonymisiert betrachtet wird, wie beispielsweise die Entfernung und Generalisierung von Merkmalen, die Löschung der Originaldaten oder zumindest eine starke Aggregation der Daten.
- Häufige Fehler beim Einsatz der Pseudonymisierung als Technik zur Verringerung der Verknüpfbarkeit:
    - Verwendung desselben Schlüssels in unterschiedlichen Datenbanken: Der Ausschluss der Verknüpfbarkeit unterschiedlicher Datenbestände ist in hohem Maße von der Verwendung eines kryptologischen Algorithmus und davon abhängig, dass eine Einzelperson in unterschiedlichen Kontexten verschiedenen pseudonymisierten Merkmalen entspricht. Es ist daher wichtig, niemals denselben Schlüssel für unterschiedliche Datenbanken zu verwenden, um die Verknüpfbarkeit einzuschränken.
    - Verwendung unterschiedlicher Schlüssel („rotierender Schlüssel“) für unterschiedliche Personen: Es mag verführerisch sein, unterschiedliche Schlüssel für verschiedene Personengruppen zu verwenden und den Schlüssel nach einer bestimmten Anzahl pseudonymisierter Datensätze zu ändern (indem beispielsweise derselbe Schlüssel für die Erfassung von zehn Einträgen zu derselben Person zu verwendet wird). Wird diese Vorgehensweise jedoch nicht ordnungsgemäß geplant, kann sie zur Entstehung von Mustern führen und damit die angestrebten Vorteile teilweise zunichtemachen. Wird beispielsweise der Schlüssel mittels spezifischer Regeln für bestimmte Personen geändert, erleichtert dies die Verknüpfbarkeit der eine bestimmte Person betreffenden Einträge. Verschwinden wiederkehrende pseudonymisierte Daten aus der Datenbank und erscheinen zugleich neue Daten, kann dies ebenfalls ein Hinweis darauf sein, dass beide Datensätze dieselbe natürliche Person betreffen.

- Aufbewahrung des Schlüssels: Wird ein Geheimschlüssel gemeinsam mit den pseudonymisierten Daten aufbewahrt, ist der Angreifer im Falle einer Kompromittierung der Daten unter Umständen in der Lage, die pseudonymisierten Daten ohne Weiteres mit den entsprechenden Originalmerkmalen zu verknüpfen. Gleiches gilt, wenn der Schlüssel zwar unabhängig von den Daten, jedoch ungesichert gespeichert wird.

4.3. Schwachstellen der Pseudonymisierung

- *Gesundheitswesen*

1. Name, Adresse, Geburtsdatum	2. Zeitraum des Bezugs bestimmter Sozialleistungen	3. Body-Mass-Index	6. Referenznr. der Untersuchungskohorte
[redacted]	< 2 Jahre	15	QA5FRD4
	> 5 Jahre	14	2B48HFG
	< 2 Jahre	16	RC3URPQ
	> 5 Jahre	18	SD289K9
	< 2 Jahre	20	5E1FL7Q

Tabelle 5. Beispiel für eine Pseudonymisierung durch Hashen von Name, Adresse und Geburtsdatum, die ohne Weiteres rückgängig gemacht werden kann

Es wurde ein Datenbestand generiert, um den Zusammenhang zwischen dem Gewicht einer Person und dem Bezug bestimmter Sozialleistungen zu untersuchen. Der Original-Datenbestand enthielt Name, Adresse und Geburtsdatum der betroffenen Personen. Diese Merkmale wurden jedoch gelöscht. Die Referenznummern der Untersuchungskohorte wurden anhand der gelöschten Daten unter Anwendung einer Hashfunktion generiert. Obwohl Name, Adresse und Geburtsdatum aus der Tabelle gelöscht wurden, können die Referenznummern der Untersuchungskohorte einfach berechnet werden, wenn Name, Adresse und Geburtsdatum einer betroffenen Person sowie die verwendete Hashfunktion bekannt sind.

- *Soziale Netzwerke*

Es wurde gezeigt[19], dass sensitive Informationen über betroffene Personen aus Graphen sozialer Netzwerke extrahiert werden können, obwohl die entsprechenden Daten „Pseudonymisierungs"-Techniken unterzogen wurden. Der Betreiber eines sozialen Netzwerks nahm fälschlicherweise an, eine Pseudonymisierung sei hinreichend robust, um nach dem Verkauf von Daten an andere Unternehmen zu Marketing- und Werbezwecken eine Identifizierung zu verhindern. Der Betreiber ersetzte die Klarnamen der Nutzer durch Nicknamen, was jedoch ganz eindeutig nicht ausreichte, um die Nutzerprofile zu anonymisieren, da die Beziehungen zwischen den einzelnen Personen einzigartig sind und als Identifikator herangezogen werden können.

- *Standorte*

Wissenschaftler des MIT[20] analysierten in jüngster Zeit einen pseudonymisierten Datenbestand, in dem die in einem Zeitraum von 15 Monaten erhobenen räumlich-

[19] Narayanan, A. und Shmatikov, V., „De-anonymizing social networks", in: *30th IEEE Symposium on Security and Privacy*, 2009.

[20] De Montjoye, Y.-A., Hidalgo, C., Verleysen, M. und Blondel, V. (2013), „Unique in the Crowd: The privacy bounds of human mobility", in: *Nature*, Nr. 1376.

zeitlichen Bewegungskoordinaten von 1,5 Millionen Menschen in einem Gebiet mit einem Radius von 100 km erfasst waren. Sie zeigten, dass 95 % der Population anhand von vier Standorten und über 50 % der betroffenen Personen anhand von nur zwei Standorten herausgegriffen werden können (wobei einer dieser Orte bekannt ist, sehr wahrscheinlich das „Zuhause“ oder der „Arbeitsplatz“). Somit blieb sehr wenig Raum für den Schutz der Privatsphäre, obwohl die Identitäten der Personen mittels der Ersetzung ihrer tatsächlichen Merkmale durch andere Kennungen pseudonymisiert worden waren.

5. Schlussfolgerungen und Empfehlungen

5.1. Schlussfolgerungen

Die Techniken der Deidentifizierung und Anonymisierung sind Gegenstand intensiver Forschungen. In dieser Stellungnahme wurde umfassend gezeigt, dass jede Technik ihre eigenen Vor- und Nachteile aufweist. In den meisten Fällen ist es nicht möglich, Mindestempfehlungen für die zu verwendenden Parameter abzugeben, da jeder Datenbestand auf Einzelfallbasis bewertet werden muss.

Häufig kann ein anonymisierter Datenbestand nach wie vor gewisse Restrisiken für die betroffenen Personen bergen. Denn selbst wenn es nicht mehr möglich ist, den Datensatz einer Person präzise auszumachen, ist es unter Umständen nach wie vor möglich, mithilfe anderer (nicht zwangsläufig öffentlich) verfügbarer Quellen Informationen über diese Person zu gewinnen. Es ist darauf hinzuweisen, dass neben den unmittelbaren Auswirkungen der Folgen einer unzureichenden Anonymisierung auf die betroffenen Personen (Belästigung, Zeitaufwand und das Gefühl des Kontrollverlusts aufgrund der Aufnahme in einen bestimmten Cluster, sei es unwissentlich oder ohne vorherige Zustimmung) weitere, mittelbare Nebenwirkungen einer mangelhaften Anonymisierung auftreten können, wann immer eine betroffene Person aufgrund der Verarbeitung anonymisierter Daten von einem Angreifer irrtümlich ins Visier genommen wird. Letzteres gilt insbesondere, wenn der Angreifer in arglistiger Absicht handelt. Daher weist die Datenschutzgruppe nachdrücklich darauf hin, dass Anonymisierungstechniken geeignet sind, Garantien für den Schutz der Privatsphäre zu schaffen, allerdings nur, wenn sie ordnungsgemäß geplant werden. Das bedeutet, dass die Voraussetzungen (Kontext) und die Zielsetzung(en) des Anonymisierungsverfahrens klar festgelegt werden müssen, um die angestrebte Anonymisierung zu erreichen.

5.2. Empfehlungen

- Einige Anonymisierungstechniken weisen gewisse Beschränkungen auf. Diese Beschränkungen müssen sorgfältig geprüft werden, bevor eine bestimmte Technik von den für die Verarbeitung Verantwortlichen zur Planung eines Anonymisierungsverfahrens eingesetzt wird. Diese müssen dem mit der Anonymisierung verfolgten Zweck Rechnung tragen, wie beispielsweise dem Schutz der Privatsphäre von Personen bei der Veröffentlichung eines Datenbestands oder der Möglichkeit, einem Datenbestand eine bestimmte Information zu entnehmen.
- Keine der in dieser Stellungnahme beschriebenen Techniken erfüllt zuverlässig die Kriterien einer wirksamen Anonymisierung (d. h. Unmöglichkeit des Herausgreifens einer bestimmten Person, keine Verknüpfbarkeit verschiedener Datensätze zu einer Person und Ausschluss von Inferenzen bezüglich einer Person). Da es jedoch durchaus möglich ist, einige dieser Risiken mit einer bestimmten Technik vollständig oder teilweise auszuschließen, ist eine sorgfältige Planung erforderlich, indem die Anwendung einer

bestimmten Technik entsprechend der gegebenen Situation ausgestaltet wird und auch Kombinationen dieser Techniken herangezogen werden, um die Robustheit des Ergebnisses zu verbessern.

Die unten stehende Tabelle bietet einen Überblick über die Stärken und Schwächen der oben erörterten Techniken im Hinblick auf die drei grundlegenden Kriterien:

	Besteht das Risiko, dass Einzelpersonen ausgewählt werden können?	**Besteht das Risiko der Verknüpfbarkeit?**	**Besteht das Risiko der Inferenz?**
Pseudonymisierung	Ja	Ja	Ja
Stochastische Überlagerung	Ja	Unter Umständen nein	Unter Umständen nein
Ersetzung	Ja	Ja	Unter Umständen nein
Aggregation oder k-Anonymität	Nein	Ja	Ja
L-Diversität	Nein	Ja	Unter Umständen nein
Differential Privacy	Unter Umständen nein	Unter Umständen nein	Unter Umständen nein
Hashen/Tokenisierung	Ja	Ja	Unter Umständen nein

Tabelle 6. Stärken und Schwächen der untersuchten Techniken

- Die Wahl der am besten geeigneten Lösung sollte auf der Grundlage einer Einzelfallbewertung erfolgen. Eine Lösung (d. h. ein vollständiges Anonymisierungsverfahren), die diese drei Kriterien erfüllt, wäre robust und geeignet, eine Identifizierung mit Mitteln, die vernünftigerweise entweder von dem für die Verarbeitung Verantwortlichen oder von einem Dritten eingesetzt werden könnten, zu verhindern.
- Erfüllt ein Vorschlag eines der Kriterien nicht, sollte eine gründliche Evaluierung der hinsichtlich einer Identifizierung bestehenden Risiken vorgenommen werden. Diese Evaluierung sollte der zuständigen Behörde vorgelegt werden, wenn die einzelstaatlichen Rechtsvorschriften vorgeben, dass die Behörde das Anonymisierungsverfahren bewerten oder genehmigen muss.

Um die Risiken hinsichtlich einer Identifizierung einzudämmen, sollten die folgenden bewährten Verfahren in Erwägung gezogen werden:

<u>Bewährte Anonymisierungsverfahren</u>

Grundsätzlich:

- Der für die Verarbeitung Verantwortliche darf niemals nach dem Prinzip „Freigeben und Vergessen“ handeln. Aufgrund des Restrisikos der Identifizierung sollte er:
    - 1. regelmäßig neue Risiken ermitteln und die Restrisiken erneut evaluieren,
    - 2. prüfen, ob die Kontrollmechanismen für die ermittelten Risiken ausreichen, und diese gegebenenfalls entsprechend anpassen, UND

- 3. die Risiken überwachen und steuern.

- Im Rahmen der Bewertung dieser Restrisiken ist (gegebenenfalls) auch das Identifizierungspotenzial des nicht anonymisierten Teils des Datenbestands zu berücksichtigen, wobei insbesondere die Möglichkeit seiner Kombination mit dem anonymisierten Teil des Datenbestands sowie potenzielle Korrelationen zwischen Merkmalen (z. B. zwischen Daten über Standort und Wohlstandsniveau) in Betracht zu ziehen sind.

Kontextabhängige Faktoren:

- Es sollte klar beschrieben werden, welche Zwecke mit dem anonymisierten Datenbestand verfolgt werden, da diese eine zentrale Rolle bei der Bestimmung des Risikos der Identifizierung spielen.

- Damit sind unweigerlich auch Überlegungen bezüglich aller relevanten kontextabhängigen Faktoren verbunden, wie beispielsweise der Art der Originaldaten, der vorhandenen Kontrollmechanismen (einschließlich Sicherungsmaßnahmen zur Einschränkung des Zugangs zu den Datenbeständen), der Stichprobengröße (quantitative Faktoren), der Verfügbarkeit öffentlicher Informationsquellen (auf welche sich die Empfänger stützen können) und der Art der beabsichtigten Datenfreigabe gegenüber Dritten (beschränkt, uneingeschränkt, z. B. über das Internet, usw.).

- Zudem sind die möglichen Angreifer zu bedenken, wobei die Attraktivität der Daten für gezielte Angriffe zu berücksichtigen ist (auch diesbezüglich stellen die Sensitivität der Informationen und die Art der Daten Schlüsselfaktoren dar).

Technische Faktoren:

- Der für die Verarbeitung Verantwortliche sollte die angewandte(n) Anonymisierungstechnik(en) offenlegen, insbesondere wenn er beabsichtigt, den anonymisierten Datenbestand freizugeben.

- Offensichtliche (z. B. seltene) Merkmale/Quasi-Identifikatoren sollten aus dem Datenbestand entfernt werden.

- Werden Techniken der stochastischen Überlagerung (zur Randomisierung) eingesetzt, sollte der Grad der Überlagerung der Datensätze anhand des Wertes eines Merkmals (d. h., es sollte keine übertriebene Überlagerung vorgenommen werden), der Auswirkungen einer Aufdeckung der geschützten Merkmale auf die betroffenen Personen und/oder gegebenenfalls der schwachen Besetzung des Datenbestands bestimmt werden.

- Wird (bei der Randomisierung) auf Differential Privacy gesetzt, sollte die Notwendigkeit einer Nachverfolgung der Abfragen berücksichtigt werden, um Abfragen zu ermitteln, welche die Privatsphäre gefährden können, da das Gefährdungspotenzial mit jeder neuen Abfrage steigt.

- Wird auf Generalisierungstechniken zurückgegriffen, ist es von grundlegender Bedeutung, dass sich der für die Verarbeitung Verantwortliche auch im Hinblick auf ein und dasselbe Merkmal nicht auf ein Generalisierungskriterium beschränkt. Das bedeutet, dass unterschiedliche Granularitäten der Standortdaten oder unterschiedliche Zeitintervalle gewählt werden sollten. Die Wahl des anzuwendenden Kriteriums muss sich nach der Verteilung der Merkmalswerte in der gegebenen Population richten. Nicht alle Verteilungen sind für eine Generalisierung geeignet – d. h., bei der Generalisierung kann kein allgemeingültiger Ansatz verfolgt werden. Die Variabilität innerhalb der Äquivalenzklassen ist zu gewährleisten. Beispielsweise sollte in Abhängigkeit von den oben genannten „kontextabhängigen Faktoren“ (Stichprobengröße usw.) eine spezifische

Schwelle festgelegt werden, unterhalb derer eine spezifische Stichprobe verworfen (oder ein anderes Generalisierungskriterium herangezogen) werden sollte.

ANHANG

Leitfaden zu Anonymisierungstechniken

A.1. Einleitung

Anonymität wird in den Mitgliedstaaten der EU unterschiedlich ausgelegt. In einigen Ländern wird sie mit rechenleistungsbedingter Anonymität (d. h., selbst für den für die Verarbeitung Verantwortlichen in Zusammenarbeit mit Dritten sollte es aufgrund der hierfür erforderlichen Rechenleistung schwierig sein, eine der betroffenen Personen direkt oder indirekt zu identifizieren), in anderen mit perfekter Anonymität gleichgesetzt (d. h., selbst für den für die Verarbeitung Verantwortlichen in Zusammenarbeit mit Dritten sollte es unmöglich sein, eine der betroffenen Personen direkt oder indirekt zu identifizieren). In beiden Fällen bezeichnet jedoch die „Anonymisierung" das Verfahren, in dem Daten anonymisiert werden. Der Unterschied liegt darin, wie hoch das Risiko einer Reidentifizierung sein darf.

Für anonymisierte Daten sind die verschiedensten Nutzungsformen vorstellbar, von sozialwissenschaftlichen Erhebungen über statistische Analysen bis hin zur Entwicklung neuer Dienstleistungen oder Produkte. Zuweilen können selbst Aktivitäten, die einem solchen allgemeinen Zweck dienen, Auswirkungen auf bestimmte betroffene Personen haben und die vermeintliche Anonymität der verarbeiten Daten zunichtemachen. Hierfür gibt es zahlreiche Beispiele, von der Auflegung gezielter Marketinginitiativen bis hin zur Durchführung öffentlicher Maßnahmen auf der Grundlage von Nutzerprofilen oder Verhaltens- und Bewegungsmustern[21].

Bedauerlicherweise gibt es neben allgemeinen Aussagen keine ausgereifte Metrik für die Vorabbewertung des für die Reidentifizierung nach der Verarbeitung erforderlichen zeitlichen oder sonstigen Aufwands oder die Auswahl des am besten geeigneten Verfahrens, um die Wahrscheinlichkeit zu verringern, dass sich eine freigegebene Datenbank auf eine identifizierte Gruppe betroffener Personen bezieht.

Die „Kunst der Anonymisierung", wie diese Verfahren in der wissenschaftlichen Literatur[22] zuweilen genannt werden, bildet einen neuen Forschungszweig, der noch immer in den Kinderschuhen steckt. Es gibt viele Verfahren, um das Identifizierungspotenzial von Datenbeständen zu senken, allerdings ist darauf hinzuweisen, dass die meisten dieser Verfahren eine Verknüpfung der verarbeiteten Daten mit betroffenen Personen nicht verhindern. Es hat sich herausgestellt, dass unter bestimmten Umständen eine Identifizierung von Personen anhand von als anonym geltenden Datenbeständen gelingen kann, während in anderen Situationen falsch positive Ergebnisse erzielt wurden.

Grob gesagt lassen sich zwei Anonymisierungsansätze unterscheiden: Der erste basiert auf Generalisierung, der zweite auf Randomisierung. Die Erörterung der Einzelheiten und Feinheiten dieser Verfahren erlaubt neue Einblicke in das Identifizierungspotenzial von Daten und neue Erkenntnisse über den Begriff der personenbezogenen Daten selbst.

[21] Siehe hierzu beispielsweise den TomTom betreffenden Fall in den Niederlanden (siehe das in Abschnitt 2.2.3 erläuterte Beispiel).

[22] Jun Gu,Yuexian Chen, Junning Fu, HuanchunPeng, Xiaojun Ye (2010), „Synthesizing: Art of Anonymization, Database and Expert Systems Applications Lecture Notes", in: *Computer Science*, Springer, Band 6261, 2010, S. 385ff.

A.2. „Anonymisierung“ durch Randomisierung

Eine Option der Anonymisierung ist die Veränderung der tatsächlichen Werte, um die Herstellung einer Verbindung zwischen anonymisierten Daten und Originaldaten zu verhindern. Dieses Ziel kann durch eine ganze Palette von Verfahren – von der Überlagerung bis hin zum Daten-Swapping (Vertauschung) – erreicht werden. Es ist darauf hinzuweisen, dass die Entfernung eines Merkmals einer extremen Form der Randomisierung dieses Merkmals gleichkommt (indem das Merkmal vollständig überlagert wird).

Zuweilen liegt der Zweck der Verarbeitung insgesamt weniger in der Freigabe eines randomisierten Datenbestands, sondern vielmehr darin, den Zugang zu Daten mittels Abfragen zu ermöglichen. Das Risiko für die betroffene Person entsteht in diesem Fall aus der Wahrscheinlichkeit, mit der ein Angreifer in der Lage ist, ohne Wissen des für die Verarbeitung Verantwortlichen Informationen aus einer Reihe unterschiedlicher Abfragen abzuleiten. Um die Anonymität der in den Datenbestand aufgenommenen Personen zu wahren, sollten keine Rückschlüsse darauf möglich sein, dass eine betroffene Person im Datenbestand erfasst ist. Hierzu ist die Verknüpfung zu jeglicher Form von Hintergrundwissen, über das ein Angreifer unter Umständen verfügt, aufzulösen.

Werden die Abfrageergebnisse angemessen überlagert, kann dies das Risiko einer Reidentifizierung weiter verringern. Dieses Konzept, in der Literatur bekannt als Differential Privacy[23], unterscheidet sich von den vorstehend beschriebenen Verfahren dahin gehend, dass es denjenigen, die Daten zur Verfügung stellen, eine stärkere Kontrolle über die Datenzugriffe erlaubt als im Falle einer öffentlichen Freigabe. Mit der Überlagerung werden zwei Hauptziele verfolgt: zum einen der Schutz der Privatsphäre der im Datenbestand erfassten betroffenen Personen, zum anderen die Gewährleistung der Zweckmäßigkeit der freigegebenen Informationen. Insbesondere muss der Grad der Überlagerung in einem angemessenen Verhältnis zur Abfrageintensität stehen (eine zu genaue Beantwortung zu vieler Abfragen zu einzelnen Personen erhöht die Wahrscheinlichkeit einer Identifizierung). Heute muss die erfolgreiche Anwendung der Randomisierung auf Einzelfallbasis bewertet werden, da keine Technik eine absolut verlässliche Methodik bietet. So gibt es Beispiele für Informationslecks bezüglich der Merkmale betroffener Personen (ob diese im Datenbestand erfasst sind oder nicht) selbst in Fällen, in denen der Datenbestand von dem für die Verarbeitung Verantwortlichen als randomisiert erachtet wurde.

Es könnte hilfreich sein, konkrete Beispiele zu erläutern, um die potenziellen Schwachstellen der Randomisierung als Instrument der Anonymisierung aufzuzeigen. Beispielsweise können im Rahmen des interaktiven Zugriffs auf Datenbanken vermeintlich datenschutzfreundliche Abfragen ein Risiko für die betroffenen Personen darstellen. Weiß der Angreifer, dass ein Datenbestand, der Informationen über die Häufigkeit des Merkmals A in einer Population P enthält, eine Untergruppe S von Personen umfasst, so kann er mithilfe der beiden Abfragen „Wie viele Personen in Population P weisen das Merkmal A auf?“ und „Wie viele Personen in Population P, mit Ausnahme der Personen, die der Untergruppe S angehören, weisen das Merkmal A auf?“ ohne Weiteres (durch Subtraktion) die Zahl der Personen in der Untergruppe S bestimmen, die tatsächlich das Merkmal A aufweisen, und zwar entweder eindeutig oder durch Likelihood-Inferenz. In jedem Fall könnte ein erhebliches Risiko für die Privatsphäre der Personen in der Untergruppe S bestehen, insbesondere in Abhängigkeit von der Art des Merkmals A.

[23] Cynthia Dwork, „Differential Privacy“, International Colloquium on Automata, Languages and Programming (ICALP) 2006, S. 1 ff.

Zudem ist davon auszugehen, dass wenn eine betroffene Person nicht im Datenbestand erfasst ist, ihre Beziehung zu Daten innerhalb des Datenbestands jedoch bekannt ist, die Freigabe des Datenbestands ein Risiko für ihre Privatsphäre darstellen kann. Ist beispielsweise bekannt, dass „der Merkmalswert A der Zielperson um die Größe X vom Durchschnittswert der Population abweicht“, kann der Angreifer personenbezogene Daten einer bestimmten betroffenen Person exakt ableiten, indem er einfach den Datenbankverwalter um die datenschutzfreundliche Operation der Extraktion des Durchschnittswerts von Merkmal A bittet.

Die Veränderung der tatsächlichen Werte in einer Datenbank durch gewisse relative Unschärfen ist ein Verfahren, das sehr sorgfältig geplant werden muss. Die Überlagerung muss hinreichend sein, um die Privatsphäre zu schützen, jedoch zugleich so geringfügig bleiben, dass die Zweckmäßigkeit der Daten gewährleistet ist. Ist beispielsweise die Zahl der betroffenen Personen mit einem bestimmten Merkmalswert sehr klein oder die Sensitivität des Merkmals hoch, kann es sinnvoll sein, nicht die tatsächliche Zahl anzugeben, sondern ein Intervall zu nennen oder den Sachverhalt zu beschreiben, beispielsweise durch „eine geringe Anzahl von Fällen, die möglicherweise gegen null geht“. Auf diese Weise bleibt der Schutz der Privatsphäre der betroffenen Personen gewahrt, da selbst wenn bekannt ist, dass die freigegebenen Daten überlagert wurden, in jedem Fall ein gewisses Maß an Unsicherheit bestehen bleibt. Was die Zweckmäßigkeit der Daten betrifft, so können die Ergebnisse – sofern die Veränderung der Daten durch Unschärfen ordnungsgemäß geplant wurde – nach wie vor für statistische Zwecke oder als Grundlage für Entscheidungen herangezogen werden.

Die Randomisierung einer Datenbank und der Zugriff nach dem Konzept der Differential Privacy erfordern weiterführende Überlegungen. Erstens kann der korrekte Überlagerungsgrad je nach Kontext (Art der Abfrage, Umfang der in der Datenbank erfassten Population, Art des Merkmals und sein Identifizierungspotenzial) erheblich schwanken, sodass keine allgemeingültige Lösung ins Auge gefasst werden kann. Darüber hinaus kann sich der Kontext im Zeitverlauf ändern, sodass der interaktive Mechanismus entsprechend angepasst werden muss. Die Kalibrierung der Überlagerung erfordert die Nachverfolgung der kumulativen Risiken für den Schutz der Privatsphäre, die jeder interaktive Mechanismus für die betroffenen Personen bedeutet. Der Datenzugriffsmechanismus sollte dementsprechend mit Warnhinweisen versehen werden, die angezeigt werden, wenn ein gewisses Maß an Gefährdung der Privatsphäre erreicht ist und die betroffenen Personen einem konkreten Risiko ausgesetzt sind, wenn eine neue Abfrage durchgeführt wird. Diese Warnungen unterstützen den für die Verarbeitung Verantwortlichen bei der Entscheidung darüber, in welchem Maße die tatsächlichen personenbezogenen Daten im Einzelfall überlagert werden müssen.

Andererseits gibt es auch die Möglichkeit, Merkmalswerte zu löschen (oder zu verändern). Eine häufig herangezogene Lösung für den Umgang mit gewissen atypischen Merkmalswerten ist die Löschung entweder der Daten, welche die atypischen Personen betreffen, oder der atypischen Werte. Im letzteren Fall ist es wichtig zu gewährleisten, dass das Fehlen des Wertes an sich nicht für die Identifizierung einer betroffenen Person genutzt werden kann.

Im Folgenden wird die Randomisierung durch die Ersetzung von Merkmalen untersucht. Ein großes Missverständnis beim Umgang mit Anonymisierungsverfahren ist deren Gleichsetzung mit Verschlüsselungs- oder Codierungsverfahren. Dieses Missverständnis fußt auf zwei Annahmen: a) Sobald einige Merkmale eines Datensatzes in einer Datenbank (z. B. Name, Adresse, Geburtsdatum) verschlüsselt oder im Wege eines Codierungsverfahrens wie einer kryptologischen Hashfunktion durch eine scheinbar zufällige Zeichenfolge ersetzt wurden, ist

dieser Datensatz „anonymisiert“, und b) die Anonymisierung ist effektiver, wenn die Länge des Schlüssels angemessen ist und der Verschlüsselungsalgorithmus dem Stand der Technik entspricht. Da dieses Missverständnis unter den für die Verarbeitung Verantwortlichen sehr verbreitet ist, besteht hier Klärungsbedarf. Gleiches gilt für die Pseudonymisierung und die damit vermeintlich verbundene Verringerung der Risiken.

Zunächst einmal werden mit diesen Techniken vollkommen unterschiedliche Zielsetzungen verfolgt: Verschlüsselung als eine Sicherungsmaßnahme soll die Vertraulichkeit eines Kommunikationskanals zwischen bestimmten Parteien (Menschen, Geräten, Softwareanwendungen oder Hardwarekomponenten) gewährleisten, um ein Abhören oder eine unbeabsichtigte Offenlegung zu verhindern. Die Codierung entspricht einer semantischen Übersetzung der Daten anhand eines Geheimschlüssels. Auf der anderen Seite besteht das Ziel der Anonymisierung darin, die Identifizierung von Personen zu verhindern, indem eine verdeckte Verknüpfung von Merkmalen mit einer betroffenen Person ausgeschlossen wird.

Weder Verschlüsselung noch Codierung sind für sich genommen für die Zielsetzung geeignet, die Identifizierung einer betroffenen Person auszuschließen, da die Originaldaten – mindestens für den für die Verarbeitung Verantwortlichen – weiterhin verfügbar oder ableitbar sind. Die alleinige Durchführung einer semantischen Übersetzung personenbezogener Daten, wie es beim Codieren geschieht, schließt nicht die Möglichkeit aus, die Daten in ihre ursprüngliche Struktur zurückzuführen, sei es durch die umgekehrte Anwendung des Algorithmus, durch Angriffe mittels der Exhaustionsmethode oder infolge einer Datenpanne. Eine dem Stand der Technik entsprechende Verschlüsselung kann für einen verbesserten Datenschutz sorgen, indem sie bewirkt, dass die Daten für Dritte, die den Schlüssel nicht kennen, unverständlich sind, sie führt jedoch nicht zwangsläufig zu einer Anonymisierung. Solange der Schlüssel oder die Originaldaten verfügbar sind (wenn auch nur für eine vertrauenswürdige dritte Partei, mit der die sichere Hinterlegung von Schlüsseln vertraglich vereinbart wurde), ist die Möglichkeit der Identifizierung einer betroffenen Person nicht zuverlässig ausgeschlossen.

Es ist irreführend, als Maßstab für den Grad der „Anonymisierung“ eines Datenbestands ausschließlich auf die Robustheit des Verschlüsselungsmechanismus abzustellen, da viele andere technische und organisatorische Faktoren die Gesamtsicherheit eines Verschlüsselungsmechanismus oder einer Hashfunktion ebenfalls beeinflussen. In der Literatur wird über zahlreiche erfolgreiche Angriffe berichtet, bei denen der Algorithmus vollständig umgangen wurde, indem entweder Schwächen bei der Aufbewahrung der Schlüssel (wenn es z. B. einen weniger sicheren Standardmodus gibt) oder andere menschliche Faktoren (z. B. schwache Passwörter für die Schlüsselwiederherstellung) ausgenutzt wurden. Schließlich sei noch darauf hingewiesen, dass eine Verschlüsselung mit einer bestimmten Schlüsselgröße die Vertraulichkeit für einen bestimmten Zeitraum sicherstellen soll (die meisten der aktuellen Schlüssel müssen um das Jahr 2020 vergrößert werden), während für das Ergebnis eines Anonymisierungsverfahrens keine zeitliche Begrenzung gelten sollte.

Im Folgenden sollen nun die Grenzen der Randomisierung (oder Ersetzung oder Entfernung) von Merkmalen erläutert werden. Dabei werden verschiedene Beispiele für eine mangelhafte Anonymisierung durch Randomisierung aus den letzten Jahren sowie die Gründe für derartige Fehlschläge angeführt.

Ein Fall, der es zu einiger Berühmtheit gebracht hat, war die Freigabe eines unzureichend anonymisierten Datenbestands im Zusammenhang mit dem Netflix-Preis.[24] Untersucht man einen generischen Datensatz in einer Datenbank, in der einige Merkmale einer betroffenen Person randomisiert wurden, kann jeder Datensatz in zwei Subdatensätze {randomisierte Merkmale, ursprüngliche Merkmale} aufgeteilt werden, wobei die ursprünglichen Merkmale alle möglichen Kombinationen aus vermeintlich nicht personenbezogenen Daten darstellen können. Eine spezifische Beobachtung, die anhand des im Zuge der Ausschreibung des Netflix-Preises freigegebenen Datenbestands gemacht werden kann, beruht auf der Überlegung, dass jeder Datensatz durch einen Punkt in einem mehrdimensionalen Raum dargestellt werden kann, wobei jedes klare Merkmal eine Koordinate darstellt. Bei Anwendung dieser Technik kann jeder Datenbestand als eine Konstellation aus Punkten in einem solchen mehrdimensionalen Raum betrachtet werden, die eine sehr schwache Besetzung aufweisen kann, d. h., dass die Punkte sehr weit auseinanderliegen können. Tatsächlich können sie soweit auseinanderliegen, dass nach der Aufteilung des Raums in große Regionen jede dieser Regionen nur einen Datensatz enthält. Selbst mittels Überlagerung gelingt es nicht, die Datensätze so nahe zusammenzurücken, dass eine Region mehrere Datensätze enthält. Im Netflix-Experiment beispielsweise waren die Datensätze hinreichend einzigartig, wenn nur acht Filmbewertungen berücksichtigt wurden, die zu einem auf 14 Tage genau bekannten Zeitpunkt abgegeben worden waren. Nach der Überlagerung sowohl der Bewertungen als auch der Zeitpunkte ihrer Abgabe waren keine Überlappungen zwischen Regionen mehr festzustellen. Mit anderen Worten: Alleine die Auswahl von nur acht bewerteten Filmen stellte einen einzigartigen Fingerabdruck der abgegebenen Bewertungen dar, der keinen zwei betroffenen Personen in der Datenbank gemeinsam war. Auf der Grundlage dieser geometrischen Beobachtung matchten die Wissenschaftler den vermeintlich anonymen Netzflix-Datenbestand mit einer anderen öffentlichen Datenbank mit Filmbewertungen (IMDB) und spürten damit Nutzer auf, die innerhalb derselben Zeitintervalle Bewertungen für dieselben Filme abgegeben hatten. Da die meisten Nutzer eine Eins-zu-Eins-Korrespondenz aufwiesen, konnten die der IMDB-Datenbank entnommenen Hilfsinformationen in den freigegebenen Netflix-Datenbestand importiert werden, sodass alle vermeintlich anonymisierten Datensätze durch Identitäten ergänzt werden konnten.

Es ist wichtig, darauf hinzuweisen, dass es sich hier um eine allgemeine Eigenschaft handelt: Der unveränderte Teil jeder „randomisierten" Datenbank birgt noch immer ein sehr großes Identifizierungspotenzial, das von der Seltenheit der Kombination der unveränderten Merkmale abhängig ist. Diesen Vorbehalt sollten für die Verarbeitung Verantwortliche stets im Auge behalten, wenn sie sich für eine Randomisierung als Verfahren für eine beabsichtigte Anonymisierung entscheiden.

Zahlreichen derartigen Reidentifizierungs-Experimenten lag ein ähnlicher Ansatz zugrunde, bei dem zwei Datenbanken auf dieselbe Teilregion projiziert wurden. Es handelt sich hier um eine sehr leistungsfähige Reidentifizierungsmethode, die in letzter Zeit in unterschiedlichen Bereichen zur Anwendung gekommen ist. So wurde beispielsweise ein Identifizierungs-Experiment am Beispiel eines sozialen Netzwerks[25] durchgeführt, bei dem der mithilfe von Kennungen anonymisierte soziale Graph der Nutzer ausgewertet wurde. Als Merkmale für die Identifizierung wurden in diesem Falle die Kontaktlisten aller Nutzer herangezogen, da gezeigt wurde, dass zwei Personen nur mit äußerst geringer Wahrscheinlichkeit dieselbe

[24] Arvind Narayanan und Vitaly Shmatikov: „Robust de-anonymization of large sparse datasets", in: *IEEE Symposium on Security and Privacy 2008*, S. 111 ff.

[25] Backstrom, L., Dwork, C. und Kleinberg, J. M. (2007), „Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography", in: *Proceedings of the 16th International Conference on World Wide Web WWW'07*, S. 181 ff.

Kontaktliste haben. Basierend auf dieser intuitiven Annahme stellte man fest, dass der Teilgraph der internen Verbindungen zwischen einer sehr geringen Zahl von Knoten bereits einen topologischen Fingerabdruck darstellt, der innerhalb des Netzwerks verborgen ist und ermittelt werden kann, und dass ein großer Teil des gesamten sozialen Netzwerks identifiziert werden kann, sobald dieses Teilnetzwerk identifiziert wird. Um die Möglichkeiten eines ähnlichen Angriffs anhand einiger Werte aufzuzeigen, wurde nachgewiesen, dass anhand von weniger als zehn Knoten (die Millionen unterschiedliche Konfigurationen innerhalb des Teilnetzes zulassen, von denen jede potenziell einen topologischen Fingerabdruck darstellt) ein erfolgreicher Reidentifizierungsangriff gegen ein soziales Netzwerk aus mehr als 4 Mio. pseudonymisierten Knoten und 70 Mio. Verknüpfungen durchgeführt werden und die Privatsphäre zahlreicher Verbindungen gefährdet sein kann. Es ist zu betonen, dass dieser Reidentifizierungsansatz nicht speziell auf soziale Netzwerke zugeschnitten ist, sondern hinreichend allgemein ist, um potenziell an andere Datenbanken angepasst zu werden, in denen die Beziehungen zwischen Nutzern erfasst werden (z. B. Telefonkontakte, E-Mail-Korrespondenz, Dating-Seiten usw.).

Ein anderer Weg, um vermeintlich anonymisierte Datensätze zu identifizieren, basiert auf der Analyse des Schreibstils (Stilometrie).[26] Es wurde bereits eine Reihe von Algorithmen entwickelt, um aus geparsten Texten Metriken zu gewinnen, wobei unter anderem die Häufigkeit der Verwendung bestimmter Wörter, das Vorkommen spezifischer grammatischer Muster und die Art der Zeichensetzung herangezogen wurden. Alle diese Eigenschaften können verwendet werden, um vermeintlich anonyme Texte mit dem Schreibstil eines identifizierten Autors zu verknüpfen. Wissenschaftler haben die Schreibstile aus mehr als 100 000 Blogs ausgewertet und sind heute in der Lage, den Autor eines Postings mit einer Trefferquote von nahezu 80 % automatisch zu identifizieren. Es ist davon auszugehen, dass diese Technik weiter an Genauigkeit gewinnt, wenn auch andere Signale wie der Standort oder andere im Text enthaltene Metadaten berücksichtigt werden.

Das Identifizierungspotenzial der Verwendung der Semantik eines Datensatzes (d. h. des verbleibenden, nicht randomisierten Teils eines Datensatzes) ist ein Thema, das sowohl von der Forschungsgemeinde als auch von der Wirtschaft weiter untersucht werden sollte. Die jüngste Wiederherstellung der Identitäten von DNA-Gebern (2013)[27] zeigt, dass seit dem allseits bekannten AOL-Skandal (2006) nur sehr geringe Fortschritte gemacht wurden. Damals wurde eine Datenbank mit 20 Mio. Suchwörtern veröffentlicht, die von etwa 650 000 Nutzern im Laufe von drei Monaten eingegeben worden waren. Dies führte zur Identifizierung und zur Bestimmung der Standorte einer Reihe von AOL-Nutzern.

Standortdaten bilden eine weitere Datenfamilie, die selten alleine durch die Entfernung der Identitäten der betroffenen Personen oder die Verschlüsselung einiger Merkmale anonymisiert werden kann. Die Bewegungsmuster von Menschen sind unter Umständen hinreichend einzigartig, dass der semantische Teil der Standortdaten (die Orte, an denen sich die betroffene Person zu einem bestimmten Zeitpunkt aufgehalten hat) selbst ohne weitere

[26] http://33bits.org/2012/02/20/is-writing-style-sufficient-to-deanonymize-material-posted-online/.

[27] Genetische Daten sind ein bedeutsames Beispiel für sensitive Daten, bei denen das Risiko einer Reidentifizierung bestehen kann, wenn als einzige Technik der „Anonymisierung“ die Identitäten der DNA-Geber entfernt werden. Siehe das in Abschnitt 2.2.2 oben erläuterte Beispiel. Siehe auch John Bohannon, „Genealogy Databases Enable Naming of Anonymous DNA Donors“, in: *Science*, Bd. 339, Nr. 6117 (18. Januar 2013), S. 262.

Merkmale geeignet ist, zahlreiche Eigenschaften einer betroffenen Person zu offenbaren.[28] Dies wurde in repräsentativen wissenschaftlichen Studien vielfach nachgewiesen.[29]

In diesem Zusammenhang ist vor der Verwendung von Pseudonymen als einem Verfahren zur Gewährleistung eines angemessenen Schutzes der betroffenen Personen vor der Offenlegung ihrer Identität oder Merkmale zu warnen. Basiert die Pseudonymisierung auf der Ersetzung einer Identität durch einen anderen einzigartigen Code, wäre es blauäugig, dieses Verfahren als eine robuste Deidentifizierung zu betrachten, da damit die Komplexität der Identifizierungsmethoden und die mannigfaltigen Kontexte, in denen sie zur Anwendung kommen können, unberücksichtigt blieben.

A.3. „Anonymisierung" durch Generalisierung

Ein einfaches Beispiel soll helfen, den auf der Generalisierung von Merkmalen basierenden Ansatz zu erläutern.

Gegeben sei ein Fall, in dem ein für die Verarbeitung Verantwortlicher beschließt, eine einfache Tabelle mit drei Arten von Informationen oder Merkmalen freizugeben – einer für jeden Datensatz einzigartigen Identifikationskennung, einer Standortkennung, welche die betroffene Person mit ihrem Wohnort verknüpft, und einer Eigenschaftskennung, die eine Eigenschaft der betreffenden Person zeigt. Zugleich sei angenommen, dass diese Eigenschaft zwei unterschiedliche Werte annehmen kann, die generisch mit {P1, P2} angegeben werden:

Serien-ID	**Standort-ID**	**Eigenschaft**
Nr. 1	Rom	P1
Nr. 2	Madrid	P1
Nr. 3	London	P2
Nr. 4	Paris	P1
Nr. 5	Barcelona	P1
Nr. 6	Mailand	P2
Nr. 7	New York	P2
Nr. 8	Berlin	P1

Tabelle A1. Stichprobe betroffener Personen, nach Standort und den Eigenschaften P1 und P2

Weiß jemand, hier als der Angreifer bezeichnet, dass eine bestimmte betroffene Person (die Zielperson), die in Mailand wohnt, in der Tabelle erfasst ist, so kann er aus der Untersuchung der Tabelle den Schluss ziehen, dass Nr. 6 als einzige betroffene Person mit dieser Standortkennung die Eigenschaft P2 aufweist.

Dieses sehr einfache Beispiel zeigt die wichtigsten Faktoren jedes Identifizierungsverfahrens auf, das auf einen Datenbestand angewendet wird, der einem vermeintlichen Anonymisierungsverfahren unterzogen wurde. Das heißt, es gibt einen Angreifer, der (zufällig

[28] Dieses Problem wurde in einigen einzelstaatlichen Rechtsvorschriften in Angriff genommen. Beispielsweise werden in Frankreich veröffentlichte Standortstatistiken durch Generalisierung und Vertauschung anonymisiert. So veröffentlicht das INSEE Statistiken, die durch die Aggregation aller Daten auf ein Gebiet von 40 000 Quadratmetern generalisiert wurden. Die Granularität des Datenbestands reicht aus, um die Zweckmäßigkeit der Daten zu gewährleisten, während Vertauschungen Deanonymisierungsangriffe in schwach besetzten Gebieten verhindern. Insgesamt bietet bei dieser Datenfamilie eine Aggregation und Vertauschung einen starken Schutz vor Angriffen mittels Inferenz- und Deanonymisierungstechniken (http://www.insee.fr/en/).

[29] De Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. und Blondel, V.D. (2013), „Unique in the Crowd: The privacy bounds of human mobility", in: *Nature* 3, Artikel Nr. 1376.

oder absichtlich) Hintergrundwissen über einige oder alle im Datenbestand erfassten Personen hat. Das Ziel des Angreifers ist es, dieses Hintergrundwissen mit den Daten des freigegebenen Datenbestands zu verknüpfen, um ein klareres Bild von den Merkmalen dieser betroffenen Personen zu gewinnen.

Um zu erreichen, dass eine Verknüpfung der Daten mit jeglicher Art von Hintergrundwissen weniger effektiv oder aufwendiger wäre, könnte sich der für die Verarbeitung Verantwortliche auf die Standort-ID konzentrieren und die Stadt, in der die betroffenen Personen wohnen, durch ein größeres Gebiet, wie beispielsweise das Land, ersetzen. Bei dieser Vorgehensweise sähe die Tabelle aus wie folgt:

Serien-ID	Standort-ID	Eigenschaft
Nr. 1	Italien	P1
Nr. 2	Spanien	P1
Nr. 3	Vereinigtes Königreich	P2
Nr. 4	Frankreich	P1
Nr. 5	Spanien	P1
Nr. 6	Italien	P2
Nr. 7	USA	P2
Nr. 8	Deutschland	P1

Tabelle A2. Generalisierung von Tabelle A1 nach Land

Mittels dieser neuen Datenaggregation wird erreicht, dass das Hintergrundwissen des Angreifers über eine bestimmte betroffene Person (d. h. „die Zielperson lebt in Rom und ist in der Tabelle erfasst“) keine eindeutigen Schlussfolgerungen über ihre Eigenschaft zulässt, da die in der Tabelle erfassten in Italien lebenden Personen unterschiedliche Eigenschaften aufweisen, namentlich P1 und P2. Für den Angreifer verbleibt eine 50%-ige Unsicherheit bezüglich der Eigenschaft der Zielperson. Dieses sehr einfache Beispiel zeigt die Auswirkung der Generalisierung auf das Anonymisierungsverfahren. Tatsächlich könnte dieser Kniff zwar eingesetzt werden, um die Wahrscheinlichkeit der Identifizierung einer in Italien lebenden Zielperson wirksam zu halbieren, im Hinblick auf in anderen Ländern (z. B. in den USA) ansässige Zielpersonen ist er jedoch nicht effektiv.

Des Weiteren kann ein Angreifer nach wie vor Informationen über eine in Spanien lebende Zielperson gewinnen. Verfügt er über Hintergrundwissen von der Art „die Zielperson lebt in Madrid und ist in der Tabelle erfasst“ oder „die Zielperson lebt in Barcelona und ist in der Tabelle erfasst“, kann der Angreifer mit einer 100%-igen Sicherheit ableiten, dass die Zielperson die Eigenschaft P1 aufweist. Daher ist festzustellen, dass Generalisierung nicht für die gesamte im Datenbestand erfasste Population denselben Schutz vor Verletzungen der Privatsphäre oder Angriffen mittels Inferenztechniken bietet.

Dieser Argumentation folgend, könnte man versucht sein, den Schluss zu ziehen, dass eine stärkere Generalisierung – beispielsweise nach Kontinent – hilfreich sein könnte, um jegliche Verknüpfung zu verhindern. Bei dieser Vorgehensweise sähe die Tabelle aus wie folgt:

Serien-ID	Standort-ID	Eigenschaft
Nr. 1	Europa	P1
Nr. 2	Europa	P1
Nr. 3	Europa	P2
Nr. 4	Europa	P1
Nr. 5	Europa	P1
Nr. 6	Europa	P2
Nr. 7	Nordamerika	P2
Nr. 8	Europa	P1

Tabelle A3. Generalisierung von Tabelle A1 nach Kontinent

Bei dieser Form der Aggregation wären mit Ausnahme der in den USA lebenden Person alle in der Tabelle erfassten betroffenen Personen vor Angriffen durch Verknüpfung und Identifizierung geschützt, da jegliche Hintergrundinformation von der Art „die Zielperson lebt in Madrid und ist in der Tabelle erfasst“ oder „die Zielperson lebt in Mailand und ist in der Tabelle erfasst“ mit einer gewissen Wahrscheinlichkeit auf die Eigenschaft einer bestimmten betroffenen Person schließen lassen (P1 mit einer Wahrscheinlichkeit von 71,4 % und P2 mit einer Wahrscheinlichkeit von 28,6 %), aber keine direkte Verknüpfung ermöglichen würde. Der Preis für diese weitere Generalisierung wäre ein offensichtlicher und drastischer Informationsverlust: Die Tabelle ermöglicht nicht die Aufdeckung möglicher Korrelationen zwischen Eigenschaften und Wohnort, es lässt sich also nicht feststellen, ob ein bestimmter Wohnort mit einer höheren Wahrscheinlichkeit auf eine der beiden Eigenschaften schließen lässt. Hingegen lässt die Tabelle lediglich sogenannte „marginale“ Verteilungen erkennen, namentlich die absolute Wahrscheinlichkeit des Auftretens der Eigenschaften P1 und P2 in der gesamten Population (in unserem Beispiel 62,5 % bzw. 37,5 %) und auf den beiden Kontinenten (71,4 % bzw. 28,6 % in Europa sowie 100 % und 0 % in Nordamerika).

Das Beispiel zeigt ferner, dass das Verfahren der Generalisierung die praktische Zweckmäßigkeit der Daten beeinflusst. Gegenwärtig sind einige Planungsinstrumente verfügbar, um im Vorhinein (d. h. vor der Freigabe des Datenbestands) den am besten geeigneten Grad der Generalisierung eines Merkmals zu bestimmen und so die Risiken einer Identifizierung der in einer Tabelle erfassten betroffenen Personen zu verringern, ohne die Zweckmäßigkeit der freigegebenen Daten übermäßig zu beeinträchtigen.

K-Anonymität

Ein auf der Generalisierung von Merkmalen basierendes Verfahren, mit dem versucht wird, Angriffe durch Verknüpfung zu verhindern, ist die sogenannte k-Anonymität. Dieses Verfahren geht auf ein Ende der 1990er Jahre durchgeführtes Reidentifizierungs-Experiment anhand eines vermeintlich anonymisierten Datenbestands zurück, der von einem im Gesundheitswesen tätigen privaten US-amerikanischen Unternehmen freigegeben wurde. Die Anonymisierung bestand in diesem Falle in der Entfernung der Namen der betroffenen Personen, wobei der Datenbestand allerdings noch immer Gesundheitsdaten und andere Merkmale enthielt, wie beispielsweise die Postleitzahl (die Standort-ID für den Wohnort), das Geschlecht und das vollständige Geburtsdatum. Dieselbe Dreierkombination aus Merkmalen {Postleitzahl, Geschlecht, vollständiges Geburtsdatum} war auch in anderen öffentlich verfügbaren Registern (z. B. im Wählerverzeichnis) enthalten und konnte somit von einer Wissenschaftlerin herangezogen werden, um die Identitäten bestimmter betroffener Personen

mit den Merkmalen im freigegebenen Datenbestand zu verknüpfen. Das Hintergrundwissen des Angreifers (der Wissenschaftlerin) könnte so ausgesehen haben: „Ich weiß, dass die betroffene Person im Wählerverzeichnis als einzige Person eine bestimmte Dreierkombination aus Merkmalen {Postleitzahl, Geschlecht, vollständiges Geburtsdatum} aufweist. Der freigegebene Datenbestand enthält einen Datensatz mit dieser Dreierkombination." Es wurde empirisch nachgewiesen[30], dass die große Mehrheit (mehr als 80 %) der betroffenen Personen in dem für diese Forschungsarbeit herangezogenen öffentlichen Register eindeutig einer bestimmten Dreierkombination zugeordnet werden und somit identifiziert werden konnte. Dementsprechend ist festzustellen, dass die Daten in diesem Falle nicht ordnungsgemäß anonymisiert worden waren.

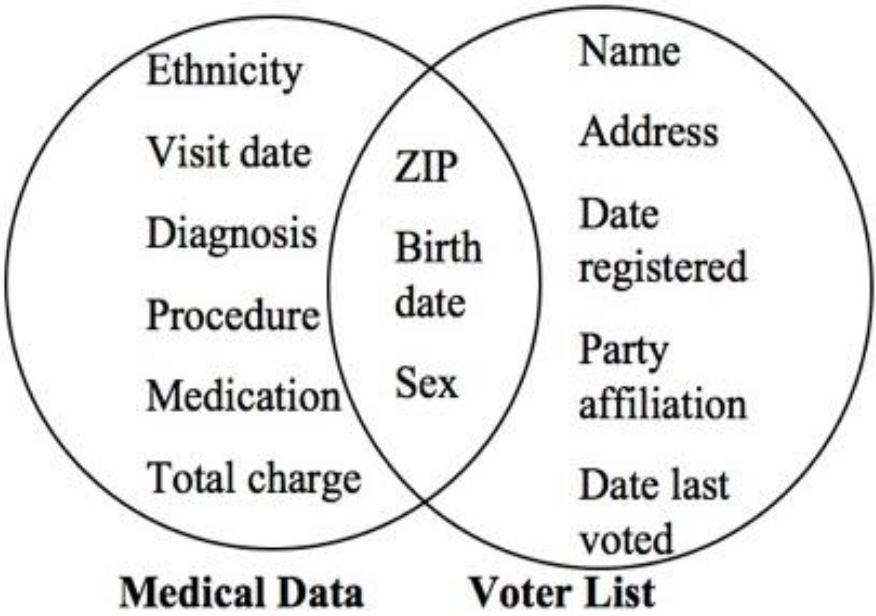

Abbildung A1. Reidentifizierung durch Datenverknüpfung

Legende zur Abbildung A1	
Ethnicity	Ethnische Zugehörigkeit
Visit date	Datum des Arztbesuchs
Diagnosis	Diagnose
Procedure	Behandlung
Medication	Medikation
Total charge	Gesamtkosten
ZIP	PLZ
Birth date	Geburtsdatum
Sex	Geschlecht
Name	Name
Address	Adresse
Date registered	Datum der Registrierung
Party affiliation	Parteimitgliedschaft
Date last voted	Datum der letzten Wahlteilnahme
Medical Data	Medizinische Daten
Voter List	Wählerverzeichnis

[30] Sweeney, L. (1997), „Weaving Technology and Policy Together to Maintain Confidentiality", in: *Journal of Law, Medicine &Ethics*, 25, Nr. 2 und 3, S. 98 ff.

Um die Effektivität ähnlicher Angriffe durch Verknüpfung zu verringern, wurde vorgeschlagen, die für die Verarbeitung Verantwortlichen sollten zunächst den Datenbestand prüfen und jene Merkmale, die vernünftigerweise von einem Angreifer für die Verknüpfung mit einer anderen Hilfsquelle genutzt werden könnten, gruppieren. Jede dieser Gruppen sollte mindestens *k* identische Kombinationen aus generalisierten Merkmalen umfassen (d. h., jede Gruppe sollte eine Äquivalenzklasse von Merkmalen darstellen). Datenbestände sollten demnach erst nach der Aufteilung in solche homogenen Gruppen freigegeben werden. Die für die Generalisierung ausgewählten Merkmale werden in der Literatur als Quasi-Identifikatoren bezeichnet, da sie, wenn ihre realen Werte bekannt wären, eine unmittelbare Identifizierung betroffener Personen zulassen würden.

Die Schwächen einer unzureichenden Planung der k-Anonymisierung von Tabellen wurden in zahlreichen Identifizierungs-Experimenten aufgezeigt. Solche Schwächen sind beispielsweise gegeben, wenn die übrigen Merkmale in einer Äquivalenzklasse identisch (wie dies bei der Äquivalenzklasse der in Spanien lebenden betroffenen Personen in dem in Tabelle A2 veranschaulichten Beispiel der Fall ist) oder sehr ungleich verteilt sind, wobei ein bestimmtes Merkmal eine besonders hohe Prävalenz aufweist, oder auch wenn eine Äquivalenzklasse nur sehr wenige Datensätze umfasst. In diesen beiden Fällen ist eine Likelihood-Inferenz möglich. Eine weitere Schwäche ergibt sich, wenn kein signifikanter „semantischer" Unterschied zwischen den realen Merkmalen der Äquivalenzklassen besteht (d. h. wenn zwar die quantitativen Werte dieser Merkmale tatsächlich unterschiedlich sind, ihre numerischen Werte jedoch sehr nah beieinanderliegen, oder wenn sie Teil einer Reihe semantisch ähnlicher Merkmale sind, z. B. derselben Skala für das Kreditrisiko oder derselben Krankheitsfamilie angehören), sodass es nach wie vor möglich ist, dem Datenbestand mittels Angriffen durch Verknüpfung eine große Menge von Informationen über betroffene Personen zu entnehmen.[31] In diesem Zusammenhang ist unbedingt zu beachten, dass wann immer ein Merkmal schwach besetzt ist (wenn z. B. eine bestimmte Eigenschaft in einem geografischen Gebiet kaum vertreten ist) und es mit einer ersten Aggregation nicht gelingt, die Daten so zu gruppieren, dass die unterschiedlichen Eigenschaften hinreichend häufig auftreten (wenn z. B. in einem geografischen Gebiet weiterhin eine geringe Zahl einiger weniger Eigenschaften auszumachen ist), ist eine weitere Aggregation von Merkmalen erforderlich, um die angestrebte Anonymisierung zu erreichen.

L-Diversität

Aufbauend auf diesen Beobachtungen wurden im Laufe der Jahre verschiedene Varianten der k-Anonymität vorgeschlagen und einige Planungskriterien für die Verbesserung des Verfahrens der Anonymisierung durch Generalisierung entwickelt, um die Risiken von Angriffen durch Verknüpfung zu verringern. Sie basieren auf den probabilistischen Eigenschaften von Datenbeständen. Im Einzelnen wird eine weitere Anforderung hinzugefügt, namentlich dass jedes Merkmal in einer Äquivalenzklasse mindestens *l* Mal erscheint, sodass ein Angreifer nur mit einer signifikanten Unsicherheit Aussagen über ein Merkmal treffen kann, selbst wenn er über Hintergrundwissen über eine bestimmte betroffene Person verfügt. Dies kommt der Anforderung gleich, dass eine ausgewählte Eigenschaft in einem Datenbestand (oder Partition) mindestens mit einer bestimmten Häufigkeit auftreten muss:

[31] Es ist darauf hinzuweisen, dass Korrelationen auch dann hergestellt werden können, wenn die Datensätze bereits nach Merkmalen gruppiert wurden. Weiß der für die Verarbeitung Verantwortliche, welche Arten von Korrelationen er prüfen möchte, kann er die relevantesten Merkmale auswählen. Beispielsweise sind die Ergebnisse von PEW-Umfragen nicht anfällig für Angriffe mittels feingranularer Inferenztechniken und nach wie vor geeignet, um Korrelationen zwischen demografischen Merkmalen und Interessen auszumachen (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx).

Mit diesem Kniff könnte das Risiko einer Reidentifizierung verringert werden. Dies ist die Zielsetzung des auf L-Diversität basierenden Anonymisierungsverfahrens. Ein Beispiel für dieses Verfahren wird in den Tabellen A4 (Originaldaten) und A5 (Ergebnis des Verfahrens) dargestellt. Das Beispiel zeigt, dass mittels einer ordnungsgemäßen Planung die Generalisierung der Standort-ID und der Altersangaben der in Tabelle A4 erfassten Personen die Unsicherheit hinsichtlich der tatsächlichen Merkmale der einzelnen betroffenen Personen in der Erhebung erheblich erhöht wird. Weiß der Angreifer beispielsweise, dass eine betroffene Person zu der ersten Äquivalenzklasse gehört, kann er nicht zuverlässig feststellen, ob eine Person die Eigenschaften X, Y oder Z aufweist, da in dieser (und jeder anderen) Äquivalenzklasse mindestens ein Datensatz enthalten ist, der diese Eigenschaften ausweist.

Seriennummer	Standort-ID	Alter	Eigenschaft
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Tabelle A4. Tabelle mit nach Standort, Alter und den drei Eigenschaften X, Y und Z gruppierten Personen

Seriennummer	Standort-ID	Alter	Eigenschaft
1	11*	<50	X
4	11*	<50	Y
9	11*	<50	Z
10	11*	<50	Z
5	23*	>50	Z
6	23*	>50	X
7	23*	>50	Y
8	23*	>50	Y
2	12*	<50	X
3	12*	<50	Y
11	12*	<50	Z
12	12*	<50	Z

Tabelle A5. Beispiel für eine Version von Tabelle A4 mit l-Diversität

T-Closeness:

In dem spezifischen Fall, in dem die Merkmale innerhalb einer Partition uneinheitlich verteilt sind, eine geringe Bandbreite von Werten aufweisen oder semantisch ähnlich sind, kommt ein als t-Closeness bekannter Ansatz zur Anwendung. Dieser stellt eine weitere Verbesserung der Anonymisierung durch Generalisierung dar und besteht in einem Verfahren, in dem die Daten so in Äquivalenzklassen aufgeteilt werden, dass die ursprüngliche Verteilung der Merkmale im Originaldatenbestand weitestmöglich abgebildet wird. Zu diesem Zweck wird im

Wesentlichen das folgende zweistufige Verfahren durchgeführt. Tabelle A6 stellt den Originaldatenbestand dar und enthält die realen Merkmalswerte der betroffenen Personen, gruppiert nach Standort, Alter, Einkommen und zwei Kategorien semantisch ähnlicher Eigenschaften, namentlich (X1, X2, X3) und (Y1, Y2, Y3) (z. B. ähnliche Kreditrisikoklassen, ähnliche Krankheiten). In der ersten Tabelle wurde für *l-Diversität* gesorgt, wobei l=1 (Tabelle A7), indem die Datensätze in semantisch ähnliche Äquivalenzklassen gruppiert wurden, wodurch nur eine unzureichende Anonymisierung erzielt wurde. Anschließend wurde der Datenbestand weiter verarbeitet, um t-Closeness und eine höhere Variabilität in jeder Partition zu erreichen (Tabelle A8). Nach dem zweiten Schritt beinhaltet tatsächlich jede Äquivalenzklasse Datensätze mit Merkmalen beider Kategorien. Es ist darauf hinzuweisen, dass Standort-ID und Alter in den verschiedenen Verfahrensschritten unterschiedliche Granularitäten aufweisen: Das bedeutet, dass für jedes Merkmal unterschiedliche Generalisierungskriterien erforderlich sein könnten, um die angestrebte Anonymisierung zu erreichen, was wiederum eine spezifische Planung und eine angemessene Rechenleistung seitens der für die Verarbeitung Verantwortlichen verlangt.

Seriennummer	Standort-ID	Alter	Einkommen	Eigenschaft
1	1127	29	30 000	X1
2	1112	22	32 000	X2
3	1128	27	35 000	X3
4	1215	43	50 000	X2
5	1219	52	120 000	Y1
6	1216	47	60 000	Y2
7	1115	30	55 000	Y2
8	1123	36	100 000	Y3
9	1117	32	110 000	X3

Tabelle A6. Tabelle mit nach Standort, Alter, Einkommen und zwei Merkmalskategorien gruppierten Personen

Seriennummer	Standort-ID	Alter	Einkommen	Eigenschaft
1	11**	2*	30 000	X1
2	11**	2*	32 000	X2
3	11**	2*	35 000	X3
4	121*	>40	50 000	X2
5	121*	>40	120 000	Y1
6	121*	>40	60 000	Y2
7	11**	3*	55 000	Y2
8	11**	3*	100 000	Y3
9	11**	3*	110 000	X3

Tabelle A7. Version von Tabelle A6 mit *l*-Diversität

Seriennummer	Standort-ID	Alter	Einkommen	Eigenschaft
1	112*	<40	30 000	X1
3	112*	<40	35 000	X3
8	112*	<40	100 000	Y3
4	121*	>40	50 000	X2
5	121*	>40	120 000	Y1
6	121*	>40	60 000	Y2
2	111*	<40	32 000	X2
7	111*	<40	55 000	Y2
9	111*	<40	110 000	X3

Tabelle A8. Version von Tabelle A6 mit *t*-Closeness

Es ist nachdrücklich darauf hinzuweisen, dass das Ziel einer Generalisierung der Merkmale betroffener Personen mittels solcher fundierter Verfahren zuweilen nur im Hinblick auf einige wenige und nicht für alle Datensätze erreicht werden kann. Bewährte Verfahren sollten sicherstellen, dass alle Äquivalenzklassen Personen mit zahlreichen unterschiedlichen Merkmalen umfassen, sodass keine Angriffe mittels Inferenztechniken mehr möglich sind. In jedem Fall verlangt dieser Ansatz eine gründliche Bewertung der verfügbaren Daten durch die für die Verarbeitung Verantwortlichen sowie eine kombinatorische Evaluierung verschiedener Alternativen (beispielsweise verschieden große Intervalle von Merkmalswerten, unterschiedliche Granularitäten der Standort- oder Altersangaben usw.). Mit anderen Worten, eine Anonymisierung durch Generalisierung kann nicht das Ergebnis eines ersten groben Versuchs der für die Verarbeitung Verantwortlichen sein, analytische Merkmalswerte in einem Datensatz durch Intervalle zu ersetzen. Vielmehr sind spezifischere quantitative Ansätze vonnöten, wie beispielsweise die Evaluierung der Entropie der Merkmale innerhalb der einzelnen Partitionen oder die Messung der Abstände zwischen den ursprünglichen Verteilungen der Merkmale und der Verteilung in den einzelnen Äquivalenzklassen.