Landesbeauftragte
für Datenschutz und Informationsfreiheit
Nordrhein-Westfalen

■ April 2017 ■ Tel.: 0211-38 424 0 ■ poststelle@ldi.nrw.de

Anonymität in Zeiten von Big Data

Die zunehmende Flut an Informationen und die immer effizienteren Werkzeuge und Techniken zur Informationsverknüpfung und Informationsanalyse stellen eine ernst zu nehmende Gefahr für die Privatheit dar. Deshalb gilt es im Gegenzug wirksame Methoden zu etablieren, die eine unzulässige Gewinnung personenbeziehbarer Daten verhindern oder zumindest erheblich erschweren.

Wir sind im Informationszeitalter angekommen. Das Web ist überall präsent und nicht mehr wegzudenken. Durch die Nutzung der vielfältigen Möglichkeiten, die uns dadurch geboten werden, hinterlassen wir allerdings auch eine breite Datenspur. Die Liste der Daten, die über uns gesammelt werden können ist lang. Wir generieren Daten bei finanziellen Transaktionen, bei online-Einkäufen, in sozialen Netzwerken, als Standortdaten von Smartphones, in Form von Vitaldaten durch medizinische Apps, in Webprotokollen beim Surfen und vielen anderen Zusammenhängen. Die Gefahr, dass von uns differenzierte Persönlichkeitsprofile erstellt werden und sogar unsere Verhaltensweisen prognostizierbar werden, wächst mit den Daten, die wir hinterlassen. Big Data mit seinen Möglichkeiten zur Verknüpfung und Analyse riesiger Datenmengen könnte diesen Alptraum wahr werden lassen.

Das Bundesdatenschutzgesetz (BDSG) legt in § 4 fest, dass die Erhebung, Verarbeitung und Nutzung personenbezogener Daten nur zulässig ist, wenn das BDSG selbst oder eine andere Rechtsvorschrift dies erlaubt oder anordnet oder der Betroffene eingewilligt hat. Ferner unterliegt die Verarbeitung personenbezogener Daten einer strikten Zweckbindung. Eine Verarbeitung zu Zwecken, die nicht dem ursprünglichen Grund der Datenerhebung entsprechen oder eine Verarbeitung für eigene Geschäftszwecke ist nur in engen Grenzen erlaubt. Somit kann man davon ausgehen, dass eine Verknüpfung von personenbezogenen Daten, die aus unterschiedlichen Anwendungszusammenhängen stammen, sowie die Analyse personenbezogener Daten mit Analysemethoden des Big Data grundsätzlich unzulässig ist.

Einen Ausweg aus den durch die Datenschutzregelungen definierten engen Grenzen wird von vielen Big Data Protagonisten in der Anonymisierung gesehen. Nach dem Motto: Anonymisierte Daten sind keine personenbezogenen Daten mehr und unterliegen damit nicht den Datenschutzgesetzen. Das ist zwar grundsätzlich richtig, aber der Weg dorthin ist mit Fallstricken versehen.

Dieser Beitrag möchte das Thema Anonymität und Anonymisierung näher beleuchten. Aber nicht mit dem Ziel ein Kochrezept zur Anonymisierung von personenbezogenen Daten bereitzustellen, sondern Bewusstsein und Sensibilität für dieses komplexe Thema zu wecken. Es wird sich herausstellen, dass Anonymisierung in der heutigen Zeit und ihren Möglichkeiten zur Informationsverarbeitung und Informationsbeschaffung schwieriger ist, als gemeinhin angenommen.

Ein nicht ganz realitätsfernes Beispiel

Angenommen ein für die Pharmaindustrie tätiges Marktforschungsinstitut benötige Daten über die Krankheitsaufkommen in der Bevölkerung, um daraus die Entwicklung bestimmter Krankheiten bezogen auf Altersgruppen, Regionen und Geschlecht ableiten zu können, damit seine Kunden bereits vorhandene Medikamente gezielter vermarkten und darüber hinaus das Potential für die Entwicklung neuer lukrativer Medikamente abschätzen können. Aus diesem Grund wende sich das Marktforschungsinstitut an medizinische Einrichtungen (Krankenhäuser, Arztpraxen), um gegen Entgelt die erforderlichen Informationen zu erhalten.

Die medizinischen Einrichtungen wären aufgrund der bei ihnen vorhandenen Behandlungsdaten in der Lage, Tabellen der folgenden Art zu erstellen:

Name	Geburt	Geschlecht	PLZ	Diagnose
Meier , Franz	18.11.65	M	50670	Asthma
Schmitz, Heinz	03.05.65	M	50672	COPD
Becker, Karl	23.08.65	M	50678	Lungenkrebs
Wilms, Toni	14.01.76	M	51931	Depression
May, Josef	25.10.76	M	51939	Diabetes
Esser, Andrea	11.07.81	W	51105	Rheuma
Schulte, Willi	13.03.81	M	51145	Rheuma

Tabelle 1

Die Weitergabe solcher Informationen an das Marktforschungsinstitut wäre unzulässig, da diese Daten der ärztlichen Schweigepflicht unterliegen und außerdem auch nur zum Zwecke der Behandlung verarbeitet und genutzt werden dürfen.

Zur Umgehung von Gesetzesverstößen könnten die Einrichtungen nun auf die Idee kommen die entsprechenden Daten zu anonymisieren. Zu diesem Zweck eliminieren die Einrichtungen nun die Namen der Patientinnen und Patienten und geben Tabellen der folgenden Form an das Marktforschungsinstitut weiter:

Geburt	Geschlecht	PLZ	Diagnose
18.11.65	M	50670	Asthma
03.05.65	M	50672	COPD
23.08.65	M	50678	Lungenkrebs
14.01.76	M	51931	Depression
25.10.56	M	51939	Diabetes
11.07.70	W	51105	Rheuma
13.03.81	M	51145	Rheuma

Tabelle 2

An dieser Stelle stellt sich die Frage, sind die Daten der Tabelle 2 als anonym anzusehen?

Nun, was haben die medizinischen Einrichtungen gemacht? Sie haben das personenspezifische Attribut „Name" entfernt. Auf den ersten Blick ist damit der Personenbezug eliminiert. Was ist aber, wenn das Marktforschungsinstitut aus anderen Datenquellen beispielweise die folgenden Informationen hat oder sich besorgt (und das ist möglich)?

PLZ	Name
50670	Meier , Franz
50672	Schmitz, Heinz
50678	Becker, Karl
51931	Wilms, Toni
51939	May, Josef
51105	Esser, Andrea
51145	Schulte, Willi

Tabelle 3a

Name	Geburt
Meier , Franz	18.11.65
Schmitz, Heinz	03.05.65
Becker, Karl	23.08.65
Wilms, Toni	14.01.76
May, Josef	25.10.76
Esser, Andrea	11.07.81
Schulte, Willi	13.03.81

Tabelle 3b

Man erkennt leicht, dass sich mit diesen Zusatzinformationen die Ausgangsdaten rekonstruieren lassen, wenn man die Inhalte der Tabellen 3a und 3b mit der Tabelle 2 verknüpft. Das zeigt, ein Weglassen personenspezifischer Attribute gewährleistet noch keine Anonymität.

Einige wenige charakteristische Merkmale einer Person reichen oft schon aus, diese zu bestimmen. Eine Studie von L. Sweeney der Carnegie Mellon University hat gezeigt, dass 87 % der amerikanischen Bevölkerung allein durch Geburtsdatum, Geschlecht und ZIPCode (Postleitzahlen-Code des Postdienstleisters United States Postal Service) eindeutig identifizierbar sind. Das Erstaunliche ist, dass es sich um drei nicht personenspezifische Merkmale handelt, die aber zusammen einen eindeutigen Personenbezug ermöglichen. Man nennt solche Merkmale Quasi-Identifikatoren. Hierzu später mehr.

Was versteht der Gesetzgeber unter Anonymisieren?

In § 3 Abs. 6 des Bundesdatenschutzgesetzes heißt es:

„Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können."

Anonymität ist also dann gegeben, wenn ein Personenbezug gänzlich ausgeschlossen ist oder wenn der Aufwand, der zur Herstellung eines Personenbezugs erforderlich ist, unverhältnismäßig hoch ist. Die zweite Alternative der Legaldefinition bezeichnet man auch als faktische Anonymität. Also eine Anonymität, die unter praktischen Erwägungen gewährleistet ist, weil ein Angreifer zur De-Anonymisierung einen so hohen Aufwand betreiben müsste, dass dieser nicht mehr Verhältnis zum potentiellen Nutzen stehen würde und damit wahrscheinlich von einem Angriffsversuch absieht.

Wir gehen im Weiteren von einer faktischen Anonymisierung aus, auch wenn diese noch mit gewissen Restrisiken behaftet ist. Es ist nämlich im Einzelfall schwer abschätzbar, wann man den Aufwand für eine De-Anonymisierung als zu groß betrachten kann, weil etliche Faktoren in diese Bewertung einfließen können. Dennoch ist die faktische Anonymisierung in einer Informationsgesellschaft der einzig gangbare Weg, denn die Analyse von Informationen kann auch einen gesellschaftlichen Nutzen haben. Hingegen führt eine Anonymisierung, die eine Re-Identifikation sicher ausschließt, zu einem so großen Informationsverlust, dass aus den Daten keine nutzbaren Schlüsse mehr gezogen werden können.

Methoden zur Anonymisierung und Anonymitätsmaße

Eine notwendige Voraussetzung einer jeden Anonymisierung ist die Eliminierung der expliziten bzw. direkten Identifikationsmerkmale, wie Namen, Anschriften, Personenkennzeichen (etwa Steuernummer, Krankenversichertennummer, Sozialversicherungsnummer), Bankverbindungen oder auch Telefonnummern.

Dies reicht aber im Allgemeinen nicht aus, da es genügend Merkmale gibt, die eine Person indirekt identifizieren können. Damit zurück zu den bereits erwähnten Quasi-Identifikatoren.

Quasi-Identifikator

Wir gehen von einer personenbezogenen Tabelle T aus, in der bereits alle Attribute eliminiert wurden, die direkt identifizierende Merkmale beschreiben. Unter einem Quasi-Identifikator versteht man nun eine Teilmenge der verbleibenden Attribute der Tabelle T, die durch Verknüpfung mit korrelierendem Wissen zur Personenidentifikation führt. Das korrelierende Wissen zeichnet sich dabei dadurch aus, dass es Informationen zu den Attributen enthält, die den Quasi-

Identifikator ausmachen und darüber hinaus mindestens ein Attribut eines direkten Identifikationsmerkmals.

Bezogen auf unser Beispiel bildet die Attributmenge QI={Geburt, Geschlecht, PLZ} den Quasi-Identifikator der Tabelle 2.

Aus den Tabellen 3a und 3b lässt sich unter Ableitung des Geschlechts aus dem Namen die folgende Tabelle erzeugen:

Name	**Geburt**	**Geschlecht**	**PLZ**
Meier , Franz	18.11.65	M	50670
Schmitz, Heinz	03.05.65	M	50672
Becker, Karl	23.08.65	M	50678
Wilms, Toni	14.01.76	M	51931
May, Josef	25.10.76	M	51939
Esser, Andrea	11.07.81	W	51105
Schulte, Willi	13.03.81	M	51145

Tabelle 4

Die Tabelle 4 enthält den Quasi-Identifikator QI der Tabelle 2 und das direkt identifizierende Attribut „Name". Damit enthält Tabelle 4 korrelierendes Wissen zu Tabelle 2 und durch Verknüpfung dieser beiden Tabellen kann die ursprüngliche Tabelle 1 wieder hergestellt werden. Die Verknüpfung mit korrelierendem Wissen hat für das sensible zu schützende Merkmal „Diagnose" den Personenbezug wieder hergestellt und damit die jeweilige Krankheit der Patienten und Patientinnen offengelegt.

Das Beispiel zeigt also, dass die in den meisten Fällen noch übliche Praxis der Elimination von direkten Identifikationsmerkmalen nicht ausreicht. Es ist zwar eine notwendige Maßnahme aber noch keine hinreichende. Im Folgenden werden Methoden und Konzepte erläutert, die eine Personenidentifikation durch Verknüpfung mit korrelierendem Wissen ausschließt bzw. deren Wahrscheinlichkeit stark verringern kann.

k-Anonymität

Der k-Anonymität liegt die Idee zu Grunde, die zu den Quasi-Identifikatoren gehörenden Daten zu Gruppen mit gleichem Informationsgehalt zusammenzufassen, so dass die hinter den Daten stehenden Individuen nicht mehr unterscheidbar sind und damit eine Verknüpfung mit korrelierendem Wissen nicht mehr eindeutig möglich ist. Anonymität wird damit durch die Gruppe gewährleistet. Je größer die Gruppe, je größer ist das Maß an Anonymität bzw. je kleiner ist die Wahrscheinlichkeit als Angehöriger einer Gruppe mit bestimmten Merkmalen identifiziert zu werden.

An dieser Stelle erkennt man bereits, Anonymität ist eine mit Wahrscheinlichkeiten behaftete Größe. Der Parameter k definiert bei der k-Anonymität die Mindestgröße der Gruppen. Er ist damit gleichzeitig das Maß der Anonymität. In einer Gruppe von k Individuen liegt die Wahrscheinlichkeit bei 1/k ein einzelnes Individuum korrekt zu identifizieren. Also beispielsweise liegt die Trefferquote in einer Gruppe mit 100 Individuen bei 1%.

Es gibt verschiedene Ansätze, um k-Anonymität zu erreichen:

1. Dummy-Datensätze hinzufügen
2. Unterdrücken von Informationen durch Löschung
3. Vertauschen von Daten
4. Verallgemeinerung von Daten

Wenden wir die k-Anonymität auf Tabelle 2 durch Verallgemeinerung an, erhält man beispielweise die folgende Tabelle:

Geburt	**Geschlecht**	**PLZ**	**Diagnose**
65	M	5067*	Asthma
65	M	5067*	COPD
65	M	5067*	Lungenkrebs
76	M	5193*	Depression
76	M	5193*	Diabetes
81	*	511**	Rheuma
81	*	511**	Rheuma

Tabelle 5

Hierbei wurden die Geburtsdaten auf das Geburtsjahr reduziert und die Postleitzahlen um die letzte bzw. die letzten beiden Stellen gekürzt. Darüber hinaus wurde das Geschlecht in den beiden letzten Datensätzen eliminiert. Dadurch entstehen drei Gruppen mit jeweils identischen Daten im Quasi-Identifikator. Da die kleinste Gruppe aus zwei Datensätzen besteht, haben wir es hier mit einer 2-Anonymität zu tun, also k=2. Die Tabelle 5 gewährleistet insgesamt also nur eine 50%ige Anonymität, was natürlich bei weitem zu wenig ist.

Wenn wir jetzt eine Verknüpfung mit dem korrelierenden Wissen der Tabelle 4 versuchen, stellt sich heraus, dass beispielweise der Datensatz von Franz Meier mit den Datensätzen 1, 2 und 3 der Tabelle 5 verknüpfbar wäre. Somit ist er nicht eindeutig einem Individuum der Gruppe zuzuordnen und damit kann man nur sagen, dass er jeweils mit einer Wahrscheinlichkeit von 33% an Asthma COPD oder Lungenkrebs leidet.

Die k-Anonymität hat leider einige Schwächen, die durch Angriffe ausgenutzt werden können. Zwei solcher Angriffe werden kurz erläutert:

1. **Homogenitätsangriff**

   Durch die Gruppenbildung kann es vorkommen, dass nicht nur die Quasi-Identifikatoren innerhalb einer Gruppe die gleichen Werte haben, sondern auch die zu schützenden sensiblen Attribute.

   In Tabelle 5 ist in der letzten Gruppe die Diagnose für alle Gruppenmitglieder gleich. Versucht man nun eine Verknüpfung mit der Tabelle 4, kann man die infrage kommenden Personen Andrea Esser und Willi Schulte nicht eindeutig genau einer Zeile in Tabelle 5 zuordnen. Das braucht man aber auch nicht, da das sensible Attribut Diagnose in beiden Fällen den gleichen Wert hat. Man weiß also, dass beide Personen an Rheuma leiden.

2. **Angriff durch spezifisches Hintergrundwissen**

   Spezifisches Hintergrundwissen kann durch Ausschlussverfahren die eindeutige Identifizierung einer Person ermöglichen. Verknüpft man beispielweise die Daten von Josef May aus Tabelle 4 mit den Daten aus Tabelle 5, so wäre eine Verknüpfung sowohl mit dem 4. als auch mit dem 5. Datensatz möglich. Weiß man allerdings, dass Josef May nicht an einer Depression leidet, so kann es folglich nur noch Diabetes sein.

l-Diversität

Die k-Anonymität bietet keinen ausreichenden Schutz vor der Zuordnung von sensiblen Attributen zu realen Personen. Homogenitätsangriffe werden durch die l-Diversität verhindert. Ein Schutz vor Offenbarung durch spezifisches Hintergrundwissen kann es zwar letztlich nicht geben, aber die Wahrscheinlichkeit kann durch l-Diversität reduziert werden.

Wir gehen von einer k-anonymisierten Tabelle T aus. Dann besteht diese Tabelle aus Datengruppen G, die sich jeweils in den Werten der Quasi-Identifikatoren nicht unterscheiden. Eine Gruppe G wird als l-divers bezeichnet, wenn sie mindestens l „gut repräsentierte Werte" für die sensiblen Attribute besitzt. Die Tabelle T ist l-divers, wenn jede Datengruppe G der Tabelle l-divers ist.

Was heißt nun „gut repräsentiert"? Hier gibt es fünf verschiedene Varianten, die eine gute Repräsentation der Werte des sensiblen Attributs zum Ziel haben. An dieser Stelle sei nur auf die sogenannte Entropie-l-Diversität eingegangen. Bei dieser Variante kommt es darauf an, dass jede Datengruppe G einer k-Anonymen Tabelle T, l unterschiedliche Werte für das sensible Attribut aufweist. Hierbei wird das sensible Attribut einer Person dadurch geschützt, dass es unter l-1 anderen Attributwerten versteckt wird. Folglich würde ein Angreifer l-1 spezifische Hintergrundwissen benötigen, um durch Ausschlusskriterien den richtigen Wert des sensiblen Attributs zu ermitteln. Für l größer als 1 ist trivialerweise eine Homogenitätsattacke bereits ausgeschlossen.

Die Tabelle 5 unseres Beispiels besitzt nur Entropie-1-Diversität, da die Gruppe, die aus den letzten beiden Datensätzen gebildet wird, nur einen Wert für das sensitive Attribut Diagnose enthält, nämlich Rheuma. Damit wird eine Homogenitätsattacke möglich und folglich auch ein Angriff mit 0 spezifischem Hintergrundwissen (wegen l=1).

Wir müssen also die Tabelle weiter verändern. Dies tun wir diesmal, indem wir zwei Dummy-Datensätze hinzufügen:

Geburt	Geschlecht	PLZ	Diagnose
65	M	5067*	Asthma
65	M	5067*	COPD
65	M	5067*	Lungenkrebs
76	M	5193*	Depression
76	M	5193*	Diabetes
81	*	511**	Rheuma
81	*	511**	Rheuma
81	*	511**	Bluthochdruck
81	*	511**	Leukämie

Tabelle 6

Tabelle 6 besitzt damit 2-Anonymität und Entropie-2-Diversität. Das ist natürlich für die Praxis viel zu wenig, soll aber als Beispiel ausreichen.

Die l-Diversität gleicht zwar einige Schwächen der k-Anonymität aus, aber auch sie selbst ist mit Schwachstellen behaftet und empfindlich gegenüber bestimmten Angriffen. An dieser Stelle sei nur der Ähnlichkeitsangriff erwähnt. Dieser ist möglich, wenn die Werte des sensiblen Attributs innerhalb einer Datengruppe Gemeinsamkeiten aufweisen. In Tabelle 6 kann man in der ersten Gruppe keinen Datensatz eindeutig einer Person zuordnen, da jeweils sowohl Franz Meier, Heinz Schmitz als auch Karl Becker infrage kommen. Die spezifischen Krankheiten der Datengruppe fallen allerdings alle unter die Kategorie der Atemwegserkrankungen. Damit kann man folgern, dass alle drei Personen eine Erkrankung haben, die etwas mit den Atemwegen zu tun hat.

t-Closeness

Die Methode der t-Closeness bietet gegenüber der l-Diversität eine Verbesserung, indem sie die sensiblen Attributwerte nicht nur gruppenbezogen betrachtet, sondern jeweils ins Verhältnis zu allen sensiblen Attributwerten der gesamten Tabelle setzt. Hierbei wird eine Distanz zwischen den Attributwerten einer Gruppe zur gesamten Tabelle, die nicht größer als ein Wert t sein darf, definiert. Dadurch erreicht man, dass sich eine Gruppe von jeder anderen Gruppe einer Tabelle in ihren sensiblen Attributwerten kaum unterscheidet. Die Distanzmes-

sung ist allerdings ein schwieriges Problem und eine Erläuterung würde den Rahmen dieses Beitrags sprengen.

Praktische Umsetzung

Anonymisieren ist keine einfache Sache. Alle vorgestellten Methoden und Konzepte haben eines gemeinsam: Die Ausgangsdaten müssen in allen Fällen verändert werden, sei es durch Verallgemeinerung, Löschung, Verfälschung oder Hinzufügung. Dadurch entstehen zwangsläufig Informationsverluste. Die Kunst besteht nun darin die Daten in einer Weise zu verändern, dass das erforderliche Maß an (statistischer) Aussagekraft noch erhalten bleibt und gleichzeitig ein hinreichendes Maß an Anonymität gewährleistet ist. Allerdings sind die Wege zur Erreichung dieser Ziele nicht deterministisch. Es gibt immer verschiedene Lösungsmöglichkeiten. Für die IT-Fachleute unter der Leserschaft an dieser Stelle der Hinweis, dass Anonymisierung zu den NP-harten Problemen gehört. Zudem dürften in der Zukunft weitere Methoden und Konzepte zur Anonymisierung entwickelt werden und damit einhergehend auch neue Angriffsvarianten. Was heute noch als hinreichend anonym angesehen werden kann, könnte sich morgen schon als Trugschluss erweisen.

Fazit

In Zeiten von Big Data ist Anonymisierung entscheidend für die Wahrung der Privatheit. Dabei ist Anonymisieren eine komplexe Aufgabe und erfordert Spezialkenntnisse. Jeder sei davor gewarnt, ohne diese Kenntnisse, Daten in die Hände Dritter zu geben, in dem naiven Glauben sie anonymisiert zu haben.