

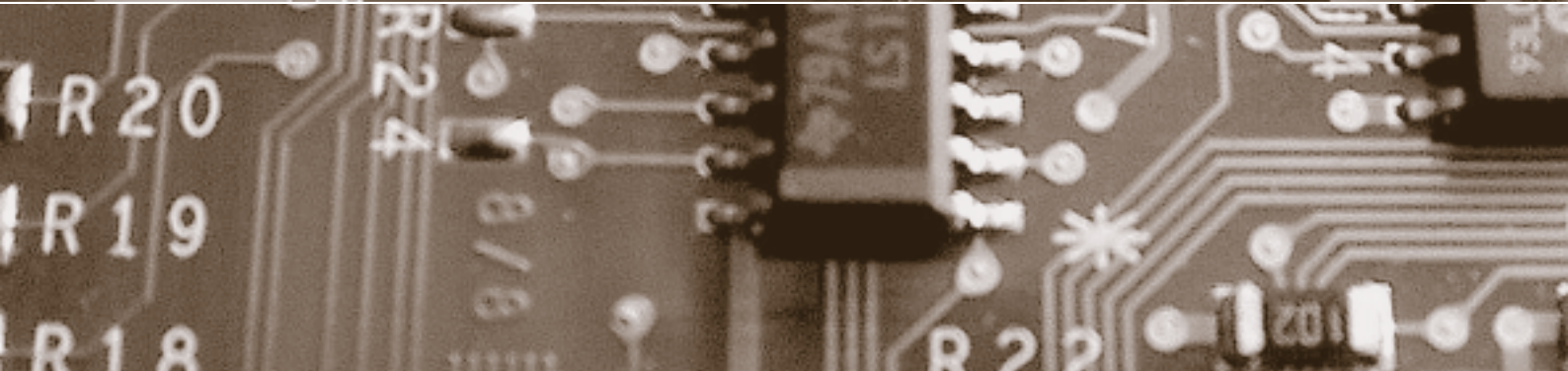
Schwerpunkt:

Anonymisierung

fokus: Das Recht auf Anonymität

fokus: Sind anonymisierte Daten anonym genug?

report: Drahtlose Sensornetze – eine Herausforderung



Herausgegeben von
Bruno Baeriswyl
Beat Rudin
Bernhard M. Hämmerli
Rainer J. Schweizer
Günter Karjoth

fokus



Schwerpunkt:

Anonymisierung

auftakt

Das Recht, in Ruhe gelassen zu werden
von Hans-Rudolf Merz

Seite 1

Der Schatten über der Anonymität
von Bruno Baeriswyl

Seite 4

Das Recht auf Anonymität
von Beat Rudin

Seite 6

zwischenakt

Der kleine Trick mit der Angst
von Urs Buess

Seite 13

Anonymisierung von genetischen Daten?
von Bruno Baeriswyl

Seite 14

Sind anonymisierte Daten anonym genug?
von Günter Karjoth

Seite 18

Anonymes E-Voting – eine Illusion?
von Rolf Oppliger

Seite 24

Folgerungskontrolle zum Schutz
von Information
von Joachim Biskup

Seite 28

Das Recht auf Anonymität ist ein Teil des Grundrechts auf informationelle Selbstbestimmung. In der Gesetzgebung finden wir etliche Gewährleistungen. Doch auch ausserhalb dieser Bereiche könnten mit Anonymisierungs- oder Pseudonymisierungslösungen in vielen Fällen die verfolgten Zwecke erreicht werden.

Das Recht auf Anonymität

Anonymisierung verhindert die Verletzung von Persönlichkeitsrechten. Ist das eine Lösung im Zusammenhang mit Biobanken? Jegliche Verwendung von Daten in einer Biobank setzt eine angemessene Aufklärung voraus.

Anonymisierung von genetischen Daten?

Wann reicht eine Anonymisierung aus, damit aus den anonymisierten Daten nicht doch wieder auf die betroffenen Personen zurückgeschlossen werden kann – und die Daten für den Forschungszweck trotzdem noch aussagekräftig genug sind?

Sind anonymisierte Daten anonym genug?

In der Theorie kann anonymes E-Voting mit Hilfe von blinden Signaturen relativ einfach realisiert werden. In der Praxis muss bei einer konkreten Realisierung eines E-Voting-Systems insbesondere darauf geachtet werden, dass nicht über verdeckte Kanäle Informationen über stimmberechtigte Personen z. B. in Tokens hineincodiert werden können.

Anonymes E-Voting – eine Illusion?

impresum

digma: Zeitschrift für Datenrecht und Informationssicherheit, ISSN: 1424-9944, Website: www.digma.info

Herausgeber: Dr. iur. Bruno Baeriswyl, Dr. iur. Beat Rudin, Prof. Dr. Bernhard M. Hämmerli, Prof. Dr. iur. Rainer, J. Schweizer, Dr. Günter Karjoth

Redaktion: Dr. iur. Bruno Baeriswyl und Dr. iur. Beat Rudin

Rubrikenredaktor: Dr. iur. Amédéo Wermelinger

Zustelladresse: Redaktion digma, c/o Stiftung für Datenschutz und Informationssicherheit, Kirschgartenstrasse 7, CH-4010 Basel
Tel. +41 (0)61 270 17 70, redaktion@digma.info

Erscheinungsplan: jeweils im März, Juni, September und Dezember

Abonnementspreise: Jahresabo Schweiz: CHF 158.00, Jahresabo Ausland: Euro 112.00 (inkl. Versandkosten), Einzelheft: CHF 42.00

Anzeigenmarketing: Publimag AG, Europastrasse 30, Postfach, CH-8152 Glattbrugg
Tel. +41 (0)44 809 31 11, Fax +41 (0)44 809 32 22, www.publimag.ch, info@publimag.ch

Herstellung: Schulthess Druck AG, Arbenzstrasse 20, Postfach, CH-8034 Zürich

Verlag und Abonnementsverwaltung: Schulthess Juristische Medien AG, Zwingliplatz 2, Postfach, CH-8022 Zürich
Tel. +41 (0)44 200 29 99, Fax +41 (0)44 200 29 98, www.schulthess.com, zs.verlag@schulthess.com

**Die Crux der
Auskunft über
Verstorbene**

Die Verordnungsregelung zur Herausgabe von Daten an die Angehörigen von Verstorbenen ist anspruchsvoll, weil sie eine Interessenabwägung voraussetzt. Unter welchen Voraussetzungen ist ein Privatversicherer zur Auskunft an die Angehörigen berechtigt? Wann besteht eine Pflicht dazu?

**Datenschutz und
wirtschaftliche
Realität**

Unter welchen Voraussetzungen kann die Wirtschaft Datenschutz realistischerweise umsetzen? Der Diskussionsbeitrag aus dem Kreis des Vereins Unternehmens-Datenschutz fordert mehr Anreize (z. B. Steuererleichterungen) für erwiesenermaßen datenschutzkonform handelnde Unternehmen. Steuererleichterung für die Einhaltung von Gesetzen – eine aus Sicht der Redaktion etwas realitätsfremde Forderung.

**Drahtlose Sensor-
netze – eine
Herausforderung**

Drahtlose Sensornetze werden als die nächste Technologiewelle nach RFID gehandelt. Dabei offenbaren die im Beitrag erörterten Anwendungsfelder, dass es ratsam ist, datenschutzrechtliche, aber auch ethische Fragestellungen frühzeitig zu erörtern.

**Europarechtliche
Herausforde-
rungen**

Bund und Kantone stehen zurzeit im Evaluationsverfahren der EU im Hinblick auf die Assoziation der Schweiz an Schengen/Dublin. Passend dazu ist ein Buch erschienen, das umfassend die europarechtlichen Vorgaben darstellt, nach denen sich das schweizerische Datenschutzrecht künftig zu richten hat.

report



RECHT IN DER PRAXIS
Die Crux der Auskunft über Verstorbene
von Martin Hofer **Seite 34**

BETRUGSPRÄVENTION
Fraud Management: Kampf dem IT-Betrug
von Stefan Nöpflin **Seite 40**

RECHT UND PRAXIS
Datenschutz und wirtschaftliche Realität
von Esther Hefti
und Susanne Amrein-Fischer **Seite 42**

IT-SICHERHEIT
Unterwegs im World Wild Web
von Thomas Dübendorfer **Seite 46**

FORSCHUNG
Drahtlose Sensornetze – eine Herausforderung
von Dirk Westhoff
und Heinrich Stüttgen **Seite 48**

RECHTSPRECHUNG
Vertrauensarzt bis-repetitas
von Amédéo Wermelinger **Seite 50**

TRANSFER
Wie ist die Lage in der Informationssicherheit?
von Roland Portmann **Seite 52**

forum



BUCHBESPRECHUNG
Europarechtliche Herausforderungen
von Beat Rudin **Seite 54**

agenda **Seite 55**

schlussstakt
Wo sind die Liberalen in der Schweiz?
von Beat Rudin **Seite 56**

Cartoon
von Hanspeter Wyss

Sind anonymisierte Daten anonym genug?

Von den (begrenzten) technischen Möglichkeiten, persönliche Daten in eine perfekte anonyme Form zu wandeln



Dr. Günter Karjoth, IBM Forschungslabor Zürich, Rüschlikon
gka@zurich.ibm.com

Anonymisierte Daten verbergen die Identität ihrer Träger. Doch ihr Schutz ist nicht absolut – er hängt von der Umgebung ab, in der sie verwendet werden.

Der amerikanische Filmverleiher Netflix veranstaltet gerade einen Wettbewerb zur Verbesserung seines Empfehlungssystems, um noch präzisere Resultate liefern zu können. Um dies testen zu können, veröffentlichte Netflix 10 Millionen Bewertungen von mehr als 480 000 ihrer Kunden über 18 000 Filmtiteln. Damit die Anonymität der Teilnehmer bewahrt bleibt, wurden aus den Daten persönliche Details entfernt und Namen durch Zufallszahlen ersetzt. Zwei Forscher aus Texas waren aber trotzdem in der Lage, beispielhaft einen Teil der Netflix-Daten zu deanonymisieren, indem sie Bewertungen und Zeitstempel mit öffentlichen Daten aus der Internet Movie Database (IMDB) verglichen. Sie zeigten damit, wie wenig zusätzliches Wissen notwendig ist, um Informationen aus den Netflix-Daten zu deanonymisieren. Kümmert es aber den durchschnittlichen Netflix-Kunden, dass seine Bewertungen aufgedeckt werden können? Die wichtigere Frage ist jedoch, ob es Netflix-Kunden gibt, deren Privatsphäre durch eine Analyse der Netflix-Daten aufgedeckt werden kann¹.

Datensammlungen

Ob im Gesundheitswesen, der Sozialfürsorge oder bei Kundenbindungsprogrammen – immer mehr personenbezogene Daten werden erfasst und verwaltet. Nach SWEENEY sind dabei die folgenden drei Trends im Sammeln von Daten über Personen unübersehbar²:

- *Sammele mehr.* Existierende personenspezifische Datensammlungen werden um weitere Felder ergänzt.
- *Sammele spezifischer.* Existierende statistische Datensammlungen werden durch personenspezifische Datensammlungen ersetzt.

- *Sammele, wenn immer du kannst.* Sobald eine Möglichkeit zum Sammeln von Daten gegeben ist, wird diese auch ausgenutzt.

Der «globale Speicherplatz pro Person» steigt unaufhaltsam. Wurden 1983 in den USA noch 15 Merkmale bei einer Geburt aufgezeichnet, so waren es 1996 schon 226 Merkmale. Nun ist es an sich nichts Schlechtes, wenn notwendige und aufschlussreiche persönliche Daten erhoben werden. Häufig ist dies schon von Gesetzes wegen notwendig, um eine zielgerichtete Verwendung von öffentlichen Mitteln zu erreichen. Statistische Ämter sammeln dafür alle gesundheitlichen, finanziellen oder andere statistisch relevante Daten über ihre Bevölkerung.

Aber nicht nur der Staat sammelt Daten. Immer mehr Händler führen eine Kundendatei mit detaillierten Kaufverhalten. Und wir selber geben persönliche Daten von uns preis, wenn wir einen Blog schreiben, unseren Freundeskreis offenlegen oder Urlaubs- und Partyerlebnisse teilen. Auch wenn diese Aussagen an verschiedenen Orten unter getrennten Identitäten gemacht worden sind, können diese Daten miteinander verknüpft werden, wenn sie genügend gemeinsame Merkmale beinhalten oder mit externen Quellen wie Telefonbücher, Wählerverzeichnisse etc. teilen. Ein Datenhändler mit mehreren Datenbanken ist in einer guten Lage, Datensätze in diesen Datenbanken zu deanonymisieren. Doch welchen Identifikationsschutz gibt es?

Anonymisierung

Einzelangaben der amtlichen Statistik unterliegen der strikten Geheimhaltung. Das deutsche Bundesstatistikgesetz (BStatG) beispielsweise ermöglicht aber eine Weitergabe von Einzeldaten zu Zwecken der Datenanalyse, wenn diese dem Befragten oder Betroffenen nicht zuzuordnen sind (absolute Anonymität) oder nur mit einem «unverhältnismässig grossen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können» (faktische Anonymität). Letztere Daten können aber nur für wissenschaftliche Zwecke verwendet und an entsprechende Forschungseinrichtungen abgegeben werden³, um der Wissen-

Tabelle 1: Medizinische Daten

Name	Geburtsdatum	Geschlecht	PLZ	Zivilstand	Tage	Diagnose
Hans Glück	11.3.59	Männlich	1072	Verheiratet	1	HIV
Robert Liebling	17.3.59	Männlich	1276	Verheiratet	7	Hepatitis
Emma Peel	1.7.60	Weiblich	1073	Ledig	2	Hepatitis
Isolde Isenthal	7.9.64	Weiblich	1077	Ledig	0	Brustschmerzen
John Steed	2.7.69	Männlich	1016	Geschieden	2	Tuberkulose
Lola Kornhaus	21.9.71	Weiblich	1267	Geschieden	4	Anämie
Molly Moon	24.12.78	Weiblich	1268	Geschieden	4	HIV

schaft die Nutzung anonymisierter Mikrodaten der amtlichen Statistik zu ermöglichen. Angeboten werden unter anderem Sozialstatistiken (z. B. Volkszählung, Einbürgerungen, Todesursachen), Wirtschaftsstatistiken (z. B. Erhebungen im Gewerbe, Einzelhandel und Tourismus), Finanz- und Steuerstatistiken, Rechtspflegestatistiken (z. B. Strafverfolgung und Strafvollzug) oder Agrar- und Umweltstatistiken.

Anonymisierte Daten sind derart verändert, dass sie nicht mehr einer Person zugeordnet werden können. Doch wie weit müssen persönliche Daten verändert werden, dass sie zwar immer noch die brauchbare Information beinhalten, ohne aber ihre Verbindung zu einer Person preisgeben?

Mikrodaten

Mikrodaten sind die Originaldaten statistischer Erhebungen, die einen hohen Detaillierungsgrad besitzen und aus Datenschutzgründen nicht öffentlich zugänglich sind. Durch Verdichtung und Plausibilisierung⁴ werden aus den Mikrodaten veröffentlichbare Makrodaten gewonnen, die keine Rückschlüsse auf die Originaldaten zulassen. Makrodaten liegen häufig in unterschiedlichen Verdichtungsstufen vor, regional-systematisch z. B. pro Gemeinde, Kreis und Bundesland. Mikrodaten haben aber den grossen Vorteil, dass der Empfänger auf diesen Daten seine eigenen Analysen durchführen kann.

Um die Problematik des Schutzes persönlicher Daten zu illustrieren, betrachten wir im Folgenden eine Sammlung (Datei, Datenbank) von Mikrodaten als eine Tabelle, welche aus n Tupeln (Zeilen in der Tabelle oder Datensätze einer Datei) mit jeweils m Attributen (Merkmalen) besteht. Jedem Attribut ist ein Wertebereich zugeordnet (den verschiedenen Ausprägungen eines Merkmals). Ausserdem nehmen wir an, dass jedes Tupel zu genau einer Person gehört und keine Person mehrmals in der Tabelle vorkommt.

Tabelle 1 zeigt eine Datensammlung, welche über den Gesundheitszustand von Patienten ei-

nes Krankenhauses erhoben worden ist. Sie besteht aus sieben Tupeln mit den Attributen <Name>, <Geburtsdatum>, <Geschlecht>, Postleitzahl des Patientenwohnorts (<PLZ>), <Zivilstand>, Verweildauer in Tagen (<Tage>) und <Diagnose>.

Datenverknüpfungsproblem

Ein grundsätzliches Problem ist die Frage, mit welchen anderen Datensätzen diese (anonymisierte) Datei abgeglichen (verknüpft) werden kann. In einer häufig zitierten Studie⁵ hat SWEENEY, basierend auf den statistischen Daten des Jahres 1990, gezeigt, dass 87% der amerikanischen Bevölkerung von immerhin 248 Millionen Menschen allein durch drei einfache demografische Merkmale, nämlich Geschlecht, 5-stellige Postleitzahl und Geburtsdatum (Jahr, Monat

87% der US-Bevölkerung sind allein durch drei einfache demografische Merkmale (Geschlecht, Postleitzahl und Geburtsdatum) eindeutig gekennzeichnet.

und Tag), eindeutig gekennzeichnet sind. Auch wenn eine weitere Studie⁶ «nur» auf eine Trefferquote von 61% der Bevölkerung kommt, welche durch diese drei Merkmale eindeutig charakterisiert werden, zeigt es sich, dass es nur «einige wenige Merkmale braucht, um eine Person ein-

Kurz & bündig

Waren früher veröffentlichte Informationen meistens in statistischer Form, besteht heute ein steigender Bedarf an Mikrodaten, die dem Empfänger eine eigene Analyse ermöglichen. Eine Anonymisierung durch Löschen oder Verschlüsseln der expliziten Identifikationsmerkmale ist aber nicht ausreichend, weil die so veränderten Daten immer noch ohne grossen Aufwand personenbezogene Informationen preisgeben können, wenn sie mit anderen Informationen verknüpft werden. Das einfache Konzept der k -Anonymität definiert den Grad des Schutzes der sensitiven Daten einer Person durch die Grösse der Gruppe, in der sie sich in den anonymisierten Daten befindet. Aber dieses Mass ist nicht perfekt.

deutig zu identifizieren». Wie kann man verhindern, dass private Informationen über eine Person bekannt werden, in dem öffentliche oder leicht erreichbare Daten mit den anonymisierten Daten, sei es durch Beobachtungen oder social engineering, verglichen (verknüpft) werden? Als weitere

Das Konzept der k -Anonymität definiert den Grad des Schutzes sensibler Daten einer Person durch die Grösse der Gruppe, in der sie sich in den anonymisierten Daten befindet.

Herausforderung gilt es dabei eine Balance zwischen dem Schutz der Identität und der Datenbrauchbarkeit zu erreichen.

Identifikatoren

Um personenbezogene Daten zu anonymisieren, müssen sie derart verändert werden, dass sie nicht mehr einer Person zuordbar sind. Dazu unterscheidet man die Attribute der Mikrodaten wie folgt.

- Ein *Identifikator* ist ein Attribut, welches eine Person eindeutig identifiziert. Beispiele sind Ausweisnummern, Konto- und Sozialversicherungsnummern (staatliche Kennnummern) und Matrikelnummern.
- Ein *Quasi-Identifikator* ist eine Menge von Attributen, welche sich auch in anderen Tabellen befinden und damit Verknüpfungen ermöglichen, die eine Person bis zu einem gewissen Genauigkeitsgrad identifizieren kann. Beispiele sind Geschlecht, Alter und Telefonnummer⁷.
- *Sensitive Attribute* sind die Attribute, welche Informationen über eine Person darstellen, die nicht mit ihr verbunden werden sollen. Beispiele sind Gehalt, Religion, politische Haltung und Gesundheitszustand.

Weiter kann es noch Attribute geben, welche in keine der oberen Kategorien fallen. In Tabelle 1 stellt das Attribut «Name» einen Identifikator dar. Einen Quasi-Identifikator bildet die Gesamtheit der Attribute «Geburtsdatum», «Geschlecht», «Postleitzahl» und «Zivilstand». Attribut «Diagnose» ist ein sensibles Attribut und Attribut «Verweildauer» fällt in keine dieser Kategorien.

Um eine Tabelle zu anonymisieren, entfernt man als erstes die Identifikatoren: in unserem Beispiel die Namen der Patienten, indem sie entweder unterdrückt oder verschlüsselt werden. Reicht dies nicht aus? Nimmt man nun an, dass es nur eine Frau gibt, welche am 24.12.1978 geboren ist und in 1268 wohnt, so kann man leicht aus einer so anonymisierten Tabelle sie und insbesondere ihre Krankheit herauslesen. Wie die amerikanischen Studien gezeigt haben, ist diese Annahme recht wahrscheinlich.

k -Anonymität

Das Konzept der k -Anonymität⁸ definiert den Grad des Schutzes der sensiblen Daten einer Person durch die Grösse der Gruppe, in der sie sich in den anonymisierten Daten befindet. Es verlangt, dass Daten nur freigegeben werden, wenn jede Kombination von Werten des Quasi-Identifikators (mit gleicher Wahrscheinlichkeit) auf mindestens k -Individuen abgebildet wird. Oder anders ausgedrückt: für jeden Tupel in der k -anonymisierten Tabelle gibt es noch mindestens $k-1$ weitere Tupel, deren Quasi-Identifikator die gleichen Werte hat.

Die Durchsetzung der k -Anonymität erfordert die Vorbestimmung des Quasi-Identifikators. Dieser hängt aber vom Wissen des Empfängers ab, welches die Möglichkeiten zur Verknüpfung bestimmt. Da nicht jeder Empfänger Zugriff auf alle öffentlichen Daten hat, kann es für eine Tabelle mehrere Quasi-Identifikatoren geben. Damit der Datenbesitzer nicht wissen muss, welche Attribute sich auch in externen Tabellen befinden und damit möglicherweise dem Empfänger zur Verfügung stehen, kann man einfachheitshalber nur einen einzigen Quasi-Identifikator definieren, welcher aus allen Attributen besteht. Die Bestimmung des richtigen Quasi-Identifikators ist im Allgemeinen recht schwierig. In Bezug auf unsere Beispielstabelle müssen wir zu Recht fragen, ob nicht auch das Merkmal «Tage» Teil des Quasi-Identifikators sein sollte.

Schutztechniken

Da eine Anonymisierung immer mit einem Informationsverlust verbunden ist, gilt es jene Attribute abzuschwächen, welche den geringsten Einfluss auf die spätere Analyse haben. Dabei gibt es zwei Methoden, dieses Ziel zu erreichen. Bei der *Generalisierung* werden Attribute durch eine Gruppierung ihrer Ausprägungen vergrößert. So kann das Alter der Patienten vergrößert werden, in dem nur noch Monat und Jahr der Geburt betrachtet wird. Reicht diese Generalisierung nicht aus, können z. B. 5-Jahre-Altersgruppen gebildet werden. Hier ist jedoch zu beachten, dass Altersgruppen nicht gleichmässig gefüllt sein werden, da es z. B. sicher viel mehr Patienten in der Altersgruppe der 60- bis 65-Jährigen gibt als in der Gruppe der 90- bis 95-Jährigen. Bei der *Unterdrückung* wird ein Attributwert vollständig entfernt und durch ein Platzhaltersymbol ersetzt. Es kann auch eine ganze Zeile gelöscht werden. Diese Methode ist von Vorteil, wenn einzelne Werte «Ausreisser» darstellen, welche wesentlich zu einem hohen Aufdeckungsrisiko beitragen. Unterdrückung kann als extreme Form der Generalisierung betrachtet werden. Neben der Generalisierung und Unterdrückung gibt es

Tabelle 2: Anonymisierte Medizinische Daten (Release 1)

Name	Geburtsdatum	Geschlecht	PLZ	Zivilstand	Tage	Diagnose
–	[50–59]	Männlich	1***	Verheiratet	1	HIV
–	[50–59]	Männlich	1***	Verheiratet	7	Hepatitis
–	[60–69]	Person	10**	Single	2	Hepatitis
–	[60–69]	Person	10**	Single	0	Brustschmerzen
–	[60–69]	Person	10**	Single	2	Tuberkulose
–	[70–79]	Weiblich	126*	Geschieden	4	Anämie
–	[70–79]	Weiblich	126*	Geschieden	4	HIV

noch weitere Methoden Mikrodaten zu verändern⁹. Das Vertauschen der Werte oder das Hinzufügen von Rauschen verändert aber die Werte, so dass nur noch statistische Analysen wie Mittelwertberechnung möglich sind.

Tabelle 2 zeigt eine mögliche Anonymisierung von Tabelle 1, welche alleine durch Generalisierung den Grad der 2-Anonymität erfüllt. Dazu wurden u. a. die Werte <Ledig> und <Geschieden> des Attributs <Zivilstand> in eine neue Gruppe <Single> zusammengefasst, welches noch eine gemeinsame semantische Eigenschaft charakterisiert. Für Merkmal <Geschlecht> führen wir die Generalisierung <Person> ein. Da <Person> den allgemeinsten Wert darstellt, ist ihre Verwendung gleichbedeutend einer Unterdrückung dieses Attributes. Der Datenbestand in Tabelle 2 besteht jetzt aus drei Blöcken, wobei jeder Block mindestens zwei Tupel enthält und damit das Schutzmass erfüllt.

Die Daten in Tabelle 3 wurden durch Unterdrückung einer Zeile (John Steed) gewonnen.

Dadurch mussten die Werte der anderen Zeilen nicht mehr so stark generalisiert werden, um den gleichen Anonymitätsgrad zu erreichen. Wird ein Tupel vollständig unterdrückt, muss nicht not-

Daten, die mittels Generalisierung und Unterdrückung anonymisiert sind, erlauben eine korrekte Interpretation.

wendigerweise der ganze Tupel aus der Tabelle entfernt werden. Hat es k oder mehr Tupel, deren Quasi-Identifikator aus den Null-Werten besteht, können deren sensitiven Merkmale in der Tabelle aufgeführt werden, da auch sie die Bedingung der k -Anonymität erfüllen.

Generalisierungshierarchien

Da bei der Generalisierung der Wert eines gegebenen Attributs durch einen weniger spezifischen Wert ersetzt wird, beruht diese Methode auf der Definition einer Generalisierungshierarchie.

Literatur, weiterführende Links

- V. CIRIANI/S. DE CAPITANI DI VIMERCATI/S. FORESTI/P. SAMARATI, «K-Anonymity» in: Security in Decentralized Data Management, Advances in Information Security 33, Springer (2007).
- P. GOLLE, Revisiting the uniqueness of simple demographics in the US population, in: Proc. 5th ACM Workshop on Privacy in Electronic Society (2006).
- K. LEFEVRE/D.J. DEWITT/R. RAMAKRISHNAN, Incognito: efficient full-domain k -anonymity, in: ACM SIGMOD International Conference on Management of Data (2005).
- A. MACHANAVAJHALA/D. KIFER/J. GEHRKE/M. VENKITASUBRAMANIAM, L -diversity: Privacy beyond k -anonymity. ACM Trans. Knowl. Discov. Data, 1(1):1556-4681 (2007).
- J. MERZ/D. VORGRIMLER/M. ZWICK, De facto anonymised microdata file on income tax statistics 1998, MPRA Paper 5740, Universitätsbibliothek München (2006), <<http://mpra.ub.uni-muenchen.de/5740/>>.
- A. NARAYANAN/V. SHMATIKOV, How To Break Anonymity of the Netflix Prize Dataset, <<http://arxiv.org/abs/cs/0610105>> (2007).
- P. SAMARATI: Protecting Respondents' Identities in Microdata Release, IEEE Trans. on Knowledge and Data Engineering, 13(6): 1010–1027 (2001),
- L. SWEENEY, Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, (2000).
- L. SWEENEY. «Information Explosion» in: Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. Urban Institute, Washington, D.C. (2001).
- L. SWEENEY. Achieving k -anonymity privacy protection using generalization and suppression. International J. of Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):571–588 (2002a).
- L. SWEENEY. k -anonymity: A model for protecting privacy. International J. of Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570 (2002b). (alle Links letztmals kontrolliert am 15.1.2008)



Tabelle 3: Anonymisierte Medizinische Daten (Release 2)

Name	Geburtsdatum	Geschlecht	PLZ	Zivilstand	Tage	Diagnose
–	[55–59]	Männlich	1***	Verheiratet	1	HIV
–	[55–59]	Männlich	1***	Verheiratet	7	Hepatitis
–	[60–64]	Weiblich	107*	Ledig	2	Hepatitis
–	[60–64]	Weiblich	107*	Ledig	0	Brustschmerzen
–	–	–	–	–	–	–
–	[70–79]	Weiblich	126*	Geschieden	4	Anämie
–	[70–79]	Weiblich	126*	Geschieden	4	HIV

chie, in der der allgemeinste Wert an der Wurzel steht und die Blätter für die spezifischsten Werte stehen. Bei einer Generalisierung werden daher die Werte, welche durch die Blätter dargestellt sind, durch einen Wert in den Knoten auf dem Weg zur Wurzel ersetzt. Diese Generalisierung

optimale Generalisierung zu finden, d.h., deren Veränderung der Tabelle zum geringsten Informationsverlust führt, oder, anders ausgedrückt, die «nützlichsten» Daten erhält. Um ein Maß für die «Brauchbarkeit» einer anonymisierten Tabelle zu haben, kann der Abstand der Werte innerhalb der Generalisierungshierarchie verwendet werden. Für das Merkmal «PLZ» gilt z.B. $10^{**} < 1073$ und damit Abstand 2. Bezüglich einer gegebenen Generalisierungshierarchie hat eine Tabelle eine *k*-minimale Generalisierung¹¹, wenn es keine andere Generalisierung gibt, welche ebenfalls *k*-anonym ist und alle Werte im Quasi-Identifikator gleich oder kleiner sind. Eine minimale *k*-anonyme Tabelle generalisiert nur, was nötig ist.

Die Umformung einer Tabelle in eine *k*-anonyme Form kann sehr rechenintensiv sein. Dies gilt insbesondere, wenn Änderungen nicht nur ganze Spalten oder Zeilen betreffen müssen, sondern auch auf einzelne Zellen angewendet werden können. Im allgemeinsten Fall ist die Aufgabe eine optimale Lösung zu finden NP vollständig¹². Dennoch wurde in den letzten Jahren eine Vielzahl von Methoden zur *k*-Anonymisierung personenbezogener Daten entwickelt. Zu den bekanntesten implementierten Verfahren gehören DataFly, MinGen und Incognito¹³. Während DataFly sehr effizient arbeitet, liefert es im Gegensatz zu den beiden anderen Verfahren keine optimale Lösung (*k*-minimale Tabelle). Da jedoch bei der Generalisierung die Laufzeit exponentiell

Die Erstellung einer Generalisierungshierarchie erfordert für die Bestimmung der richtigen Quasi-Identifikatoren ein Verständnis der Daten.

kann für kontinuierliche (Geburtsdatum, Einkommen, etc.) wie auch kategorisierte Werte (Geschlecht, Zivilstand, etc.) angewandt werden. Es können daher verschiedene generalisierte Tabellen erzeugt werden, je nach dem welche Generalisierungsschritte angewendet werden.

Werden Daten mit diesen beiden Techniken anonymisiert, kann dem Empfänger gesagt werden, wie diese Umformungen zustande gekommen sind, was eine korrekte Interpretation der Daten erlaubt. Ferner ist die Information, welche über eine betroffene Person gemacht wird, «wahrheitsgetreu», was nachfolgende Bewertungen für Betrugserkennung oder Gesundheitsfürsorge ermöglicht, welche personenspezifische Muster enthalten¹⁰.

Minimale *k*-Anonymisierung

Wie gezeigt kann es mehrere Anonymisierungen einer Tabelle geben. Daher ist es wichtig, eine

Fussnoten

- 1 NARAYANAN/SHMATIKOV (2006).
- 2 SWEENEY (2001).
- 3 Vgl. <<http://www.forschungsdatenzentrum.de/datenangebot.asp>>.
- 4 Beispielsweise werden automatisch generierte Korrekturfaktoren eingerechnet und fehlende oder unplausible Daten durch Schätz- oder Prognosewerte aufgefüllt.
- 5 SWEENEY (2000).
- 6 GOLLE (2006).
- 7 Die Nummer eines Mobiltelefons ist aber meistens ein eindeutiger Identifikator.
- 8 SAMARATI (2001), SWEENEY (2002b).
- 9 CIRIANI/DE CAPITANI DI VIMERCATI/FORESTI/SAMARATI (2007).
- 10 SWEENEY (2002a).
- 11 SAMARATI (2001).
- 12 Dies bedeutet, dass weder ein Algorithmus bekannt ist, der effizienter arbeitet als das bloße Durchprobieren aller Möglichkeiten, noch jemand beweisen konnte, dass es solch einen Algorithmus nicht gibt.
- 13 SWEENEY (2002a).
- 14 MACHANAVAJHALA/KIFER/GEHRKE/VENKITASUBRAMANIAM (2007).
- 15 MERZ/VORGRIMMLER/ZWICK (2005).

mit der Anzahl der Tupel in der Tabelle zunimmt, wird in der Praxis wohl Geschwindigkeit vor Qualität gelten und damit das einfachere Verfahren den Vorzug finden.

Angriffe

Aber auch ein Qualitätsmass wie k -Anonymität hat seine Grenzen. Versuche, eine partielle oder totale Re-Identifikation einer anonymisierten Tabelle zu erreichen, können in zwei Arten unterschieden werden. Einem Massenangriff ähnlich versucht man mit Hilfe einer externen Tabelle zusätzliches Wissen zu bekommen, mit dem man die Identität von so vielen Individuen wie möglich aufdecken möchte. Andererseits versucht man in einem gezielten Angriff herauszufinden, ob eine spezielle Person in der Datenbank (nicht) aufgeführt ist.

Werden von ein und derselben Tabelle mehrere k -Anonymisierungen veröffentlicht, so können Tupel anhand ihrer Position in der Tabelle miteinander verknüpft werden, falls nicht deren Reihenfolge vertauscht wird. Bei komplementären Veröffentlichungen ist darauf zu achten, dass es keinen Quasi-Identifikator einer Person gibt, der unterschiedliche Tupel in zwei k -anonymisierten Tabellen identifiziert und beide Tupelmengen nur einen gemeinsamen Tupel haben, da sonst dieser Tupel der Repräsentant der Person ist. Es ist auch von Vorteil, wenn Bereiche verschoben werden. Ist ein Attribut in der ersten Anonymisierung in Altersgruppen von 50–59, 60–69 etc. eingeteilt, könnte man bei einer anderen Anonymisierung die Bereiche 45–54, 55–64 etc. verwenden.

Weiter kann die Homogenität von Merkmalen zur Aufdeckung der Identität einer Person führen, wenn die gesamte Gruppe, in der sich diese Person verbirgt, die gleichen Merkmale aufweist. Ferner ist es möglich, durch spezifisches Hintergrundwissen auch bei unterschiedlichen Merkmalen durch Ausschluss von sensitiven Attributwerten auf die richtigen zu schliessen. Attribute sind dann problematisch, wenn nur wenige Merkmalsträger dieser Beschreibung entsprechen (z.B. der einzige Apotheker eines Ortes). Ist in einem Datensatz für das Jahr 2007 als Vermögen 56 Milliarden Dollar angegeben, ist es nicht schwer, auf den Merkmalsträger zu schliessen.

Ausblick

Da k -Anonymität keinen ausreichenden Schutz bietet, wenn sensitive Attributwerte durch Hintergrundwissen vorhergesagt oder ausgeschlossen werden können, wurde das Konzept der l -Diversität entwickelt¹⁴. Analog zur k -Anonymität wird der sensitive Attributwert in einer Menge von $l-1$ anderen Attributwerten

versteckt. Ein q -Block ist eine Menge von Tupeln, deren Quasi-Identifikator die gleichen Werte haben. Ein q -Block ist l -divers, wenn das sensitive Attribut in mindestens l Ausprägungen in dem q -Block vorkommt, und eine Tabelle ist l -divers, wenn ihre q -Blöcke alle l -divers sind. Bei einem Angriff mit Hintergrundwissen

Es ist sehr schwer zu verhindern, dass mit geeignetem Hintergrundwissen auch aus anonymisierten Daten Beziehungen zu Personen hergestellt werden können.

müssen damit mindestens $l-1$ andere Werte ausgeschlossen werden. Aber auch dieses Konzept hat Schwächen, die zur Definition weiterer Schutzkonzepte führten.

Der Erfolg einer k -Anonymisierung hängt im Wesentlichen von der Generalisierungshierarchie ab. Deren Erstellung ist aber nicht trivial und erfordert ein Verständnis der Daten für die Bestimmung der richtigen Quasi-Identifikatoren. In einer Einkommensteuertabelle sind das Alter des Steuerzahlers und die Anzahl seiner Kinder kritische Merkmale, da diese leicht zu erhalten und meistens von hoher Qualität sind. Obwohl Informationen über den Wohnort und das Geschlecht ebenfalls leicht zu erhalten sind, sind sie für einen Angriff weniger geeignet, um Individuen zu unterscheiden¹⁵.

Derzeitige Forschungsarbeiten betreffen nicht nur die Frage der geeigneten Algorithmen, sondern auch der Auswahl der Quasi-Identifikatoren und der Grad der Anonymität. Ist es möglich zu quantifizieren, um wie viel besser ein Anonymisierungsgrad k von 20 im Vergleich zu 10 schützt? Weiter wird auch an einem besseren Schutz für wiederholte Anonymisierungen wie auch für die verteilte Erzeugung von anonymisierten Tabellen gearbeitet.

Ob persönliche Daten einer Person zugeordnet werden können oder nicht, hängt von vielen Faktoren ab. Wie wir gesehen haben, ist es sehr schwer zu verhindern, dass mit geeignetem Hintergrundwissen auch aus anonymisierten Daten Beziehungen zu Personen hergestellt werden können. Dies sollten wir bedenken, wenn wir «freiwillig» unsere Vorlieben im Internet publizieren. Es kommt nicht darauf an, wie sensitiv die Daten für einen sind, sondern wie charakteristisch. Denn dies bestimmt den Aufwand, der nötig ist, sie mit anderen Daten zu verknüpfen, um die Aufdeckung der Identität zu ermöglichen. ■

Meine Bestellung

- 1 Jahresabonnement digma (4 Hefte des laufenden Jahrgangs)
à **CHF 158.00** bzw. bei Zustellung ins Ausland **EUR 123.00** (inkl. Versandkosten)

Name _____ Vorname _____

Firma _____

Strasse _____

PLZ _____ Ort _____ Land _____

Datum _____ Unterschrift _____

Bitte senden Sie Ihre Bestellung an:

Schulthess Juristische Medien AG, Zwingliplatz 2, CH-8022 Zürich

Telefon +41 44 200 29 19

Telefax +41 44 200 29 18

E-Mail: zs.verlag@schulthess.com

Homepage: www.schulthess.com

Schulthess 