German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2013-02 | JUNE 6, 2013

Getting Big Data but avoiding Big Brother

Rainer Schnell

Getting Big Data but avoiding Big Brother.

Eine einführende Übersicht über technische Lösungen für die datenschutzgerechte Zusammenführung personenbezogener Daten in Deutschland*

Rainer Schnell

German Record Linkage Center

Universität Duisburg-Esssen

6.6.2013

Zusammenfassung

Für viele Forschungsprobleme liegen relevante Daten über die gleichen Personen in getrennten Datenbanken vor. Die Verknüpfung dieser Datenbanken würde die Untersuchung zahlreicher Fragestellungen erlauben, falls keine Datenschutzbedenken bestünden. Diese Einführung beschreibt in untechnischer Weise zwei neuere Verfahren zur Zusammenführung personenbezogener Datenbanken, die die Identität der Personen nicht erkennbar werden lassen.

Abstract

For many research problems relevant information on the same individuals is available in separate databases. Combining these databases would allow the study of many interesting problems, if no privacy concerns would exist. This untechnical introduction describes record linkage in general and two recently suggested methods for privacy preserving record linkage.

*Für hilfreiche Kommentare danke ich Manfred Antoni, Stefan Bender, Louise Dye, Marcel Noack und Sabrina Toregrozza.

1 Einleitung

In modernen Gesellschaften werden durch kommerzielle, bürokratische, organisatorische und nicht zuletzt medizinische Prozesse immer mehr personenbezogene Daten gespeichert. In kaum einem anderen Land sind die dabei entstehenden Datenbestände so vollständig voneinander getrennt und vor einem Missbrauch gesetzlich geschützt wie in Deutschland. Ein wesentliches Element dieser Verhinderung eines Missbrauchs personenbezogener Daten besteht in dem Verbot eines einheitlichen Personenkennzeichens durch das Bundesverfassungsgericht.[1] Im Gegensatz dazu besitzen alle skandinavischen Länder eine entsprechende Personenkennziffer, die in nahezu allen personenbezogenen Datenbeständen vorhanden ist. Eine solche einheitliche Personenkennziffer erlaubt die Zusammenführung unterschiedlichster Datenbestände und damit auch die Untersuchung vielfältiger Forschungsfragestellungen allein mit Hilfe vorhandener Daten. Dazu gehören zunächst einmal medizinische Fragestellungen, z.B. bezüglich des Langzeiterfolges von Therapien. Verfügt ein Land über kein einheitliches Personenkennzeichen, dann ist die Untersuchung vieler Fragestellungen schwierig oder faktisch unmöglich.

Im Bereich der Medizin ist in Deutschland durch das Fehlen eines zentralen Mortalitätsregisters schon die einfache Frage, wie lange ein Patient mit einer bestimmten Krebsart und einer bestimmten Therapie überlebt, faktisch kaum zu beantworten: Wechselt in Deutschland ein Krebspatient das Bundesland, dann gibt es keine tatsächlich nutzbare zentrale Möglichkeit, nach einem oder zwei Jahren zu klären, ob der Patient noch lebt oder nicht. Mit hohem manuellem Aufwand kann man einen Großteil der Patienten zwar finden, erhält aber keine Informationen über die Todesursache. Dies ist zwar auf einem anderen administrativen Weg prinzipiell auch möglich, bedarf aber eines solchen Aufwandes, dass dies in der Forschungspraxis in Deutschland bislang nur in Einzelfällen erfolgt. Damit ist der Behandlungserfolg mit vertretbarem Aufwand nicht über die Zeit hinweg beobachtbar. Dies ist die Folge der Einrichtung länderspezifischer Krebsregister unter Verzicht auf eine zentrale Patienten-Identifikation.

Vergleichbare Probleme ergeben sich für zahlreiche andere Forschungsfragestellungen. Ein Beispiel ist die Frage nach dem Effekt unterschiedlicher Maßnahmen nach einem Hirnschlag auf die Langzeiterfolge unterschiedlicher Rehabilitationsmaßnahmen. Ein anderes Beispiel ist die Frage, ob Kinder, die durch einen Kaiserschnitt geboren wurden, im Erwachsenenalter eine andere Mortalität besitzen. In den Sozialwissenschaften stellt sich z.B. die

[1] Siehe hierzu das Urteil zur Volkszählung vom 15. Dezember 1983 (BVerfGE 65, 1), vor allem Seite 61 der Begründung.

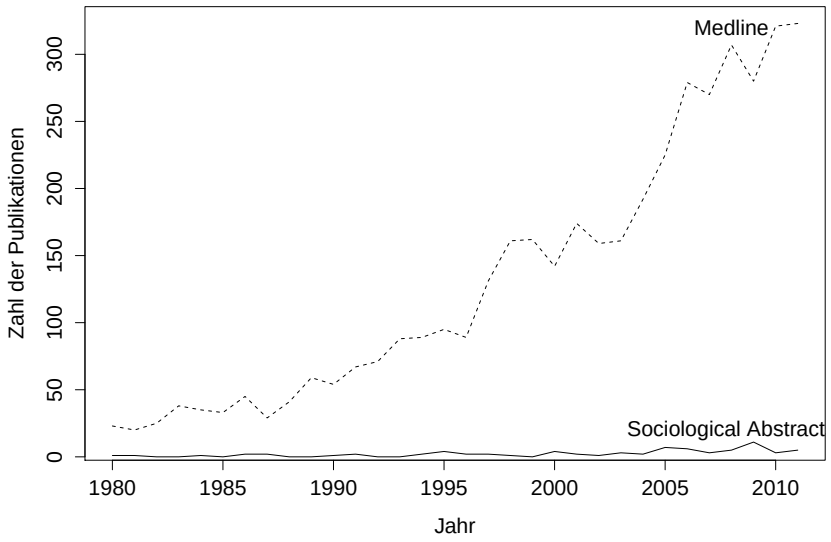

Abbildung 1: Anzahl der Record-Linkage Anwendungen in der Medizin und in der Soziologie

Frage nach der Rückfallwahrscheinlichkeit eines Straftäters, wenn die Daten der Strafverfolgungsbehörden aus Datenschutzgründen z.B. nach Ablauf von 10 oder 15 Jahren gelöscht wurden. In Deutschland lassen sich solche Fragen unter anderem aufgrund der Trennung verschiedener Datenbanken, z.B. zwischen den Bundesländern oder zwischen gesetzlichen und privaten Krankenversicherern und dem Fehlen einheitlicher Personenkennzeichen in der Regel kaum mit den verfügbaren Daten beantworten.

2 Verknüpfung personenbezogener Datenbanken

Die Möglichkeiten der Verknüpfung unterschiedlicher Datenbanken zur Untersuchung neuer Fragestellungen wird in der statistischen Literatur „Record-Linkage“ genannt. Record-Linkage wird in der Forschungsliteratur in zunehmenden Maße genutzt. Die Anwendungen reichen dabei von Fragen der Geschichtswissenschaft (z.B. die Veränderung der Haushaltszusammensetzung im viktorianischen England), über medizinische Probleme (z.B. Unfallhäufigkeiten von hyperaktiven Kindern) bis hin zur Ersetzung klassischer Volkszählungen mit Befragungen durch reine Registerzensen. Die Entwicklung verläuft

in den einzelnen Wissenschaftsgebieten mit unterschiedlicher Geschwindigkeit: Die Abbildung 1 zeigt die Entwicklung der Anzahl der Record-Linkage Anwendungen in der medizinischen Literatur einerseits und in der sozialwissenschaftlichen Literatur andererseits.[2]

Ein Anstieg in der sozialwissenschaftlichen Literatur ist bislang kaum erkennbar, während in der Medizin nunmehr pro Jahr mehr als 300 Arbeiten unter Verwendung von Record-Linkage veröffentlicht werden. Angesichts der Vorteile des Record-Linkage (Größe der Datenbestände, geringe Kosten, relativ hohe Datenqualität bei vielen – aber nicht allen – Datenbeständen) wird die Nutzung vorhandener Datenbestände durch Record-Linkage für Forschungszwecke auch in den Sozialwissenschaften sicherlich zunehmen.

Prinzipiell werden die Möglichkeiten des Einsatzes von Record-Linkage mit der wachsenden Zahl personenbezogener Datenbanken immer größer. Die Datenbanken sind natürlich um so interessanter für Forschungsfragen, je mehr relevante Informationen sie zu einer Forschungsfrage enthalten und je größer die Abdeckung der interessierenden Population in der Datenbank ist. Viele administrative Datenbanken enthalten nahezu die gesamte Population (z.B. Einwohnermelderegister) oder zumindest für viele Fragen relevante Subgruppen (z.B. sozialversicherungspflichtig Beschäftigte). Je größer die Abdeckung der Population wird und je mehr Merkmale in den Datenbanken enthalten sind, desto größer werden in der Regel die Befürchtungen, dass die Daten nicht nur für Forschungszwecke genutzt werden. Entsprechend müssen die Merkmale, die einen Personenbezug erlauben, vor einer Einsicht geschützt werden.

3 Verknüpfung personenbezogener Datenbanken in Deutschland

In der Bundesrepublik verwenden verschiedene bürokratische Organisationen verschiedene Personenkennziffern (Matrikel-, Personalausweis-, Wehrpass-, Sozialversicherungs-, Rentenversicherungsnummer etc.). Dateien mit mehreren verschiedenen Kennziffern für dieselbe Person in unterschiedlichen Organisationen existieren selten. Unter solchen Umständen bleiben als mögliche Identifikationsvariablen in der Regel nur der Name, das Geburtsdatum und die Adresse der Person.

[2]Die Abbildung basiert auf Auszählungen des Schlüsselbegriffs „record linkage“ in den Datenbanken „Pubmed“ und „Sociological Abstracts“.

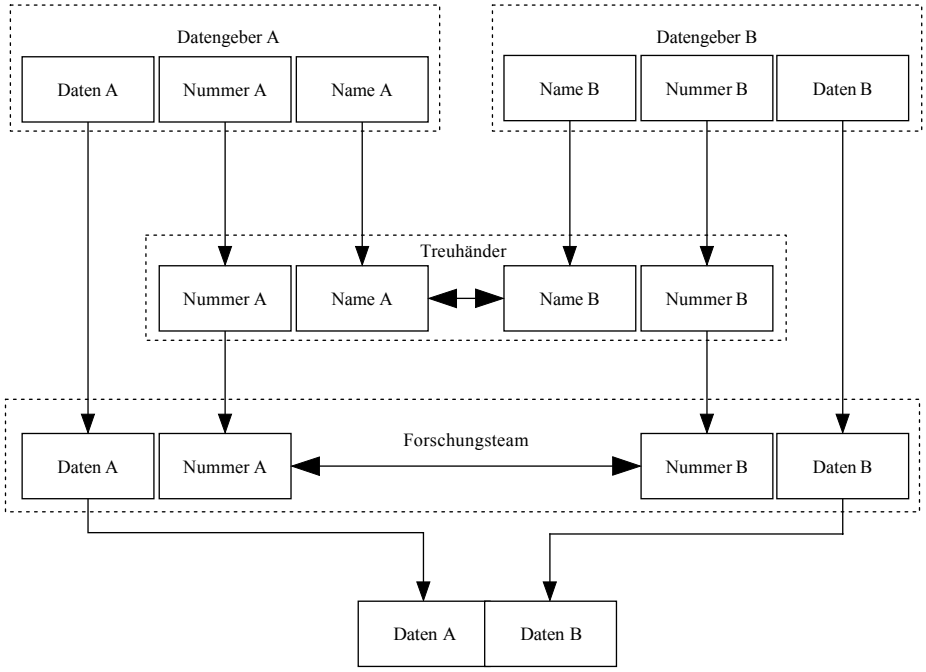

Abbildung 2: Record-Linkage mit einem Datentreuhänder

Die Weitergabe solcher Identifikationsvariablen ist fast immer heikel. Eine einfache Möglichkeit zur Vermeidung des Problems ist die Verwendung einer Institution als „Datentreuhänder“ (vgl. Abbildung 2). Im Falle zweier Datengeber mit den Datenbeständen A und B werden die Datenbestände unabhängig voneinander mit beliebigen, aber eineindeutigen neuen Kennziffern versehen. Die neuen Kennziffern werden jeweils den Datenbeständen zugefügt. Das Forschungsteam erhält die beiden Datensätze A und B mit ihrer jeweiligen neuen Identifikationsnummer, aber ohne die ursprünglichen Identifikationsvariablen. Die beiden Datengeber erstellen jeweils eine neue Datei, die nur die ursprünglichen Identifikationsvariablen (Name etc.) und die jeweilige neue Kennziffer enthalten. [3] Diese beiden Dateien werden einem Datentreuhänder übergeben, der nun zwar die ursprünglichen Identifikationsvariablen und die neuen Identifikationsnummern kennt, aber nicht die Daten aus A und B. Der Datentreuhänder verknüpft die beiden Auszüge

[3] In der Regel kann in Deutschland ein solches Linkage nur für Personen durchgeführt werden, die einer Zusammenführung explizit zugestimmt haben. Nach der natürlich immer erforderlichen Absprache mit den Datenschutzbehörden gibt es aber auch in Deutschland für einige Datenbestände die Möglichkeit, über ein solches Treuhändermodell auch vollständige Datenbestände zusammenzuführen. Dies setzt immer langwierige Verhandlungen mit den Datengebern und den Datenschutzbeauftragten voraus. Erfahrungsgemäß scheint auch international eine Dauer dieser Verhandlungsphase von zwei Jahren üblich zu sein.

über die ursprünglichen Identifikationsvariablen, führt also die Zuordnung der Records aufgrund gleicher oder ähnlicher Identifikationsvariablen durch. Danach löscht der Treuhänder die ursprünglichen Identifikationsvariablen und erstellt eine neue Datei, die nur die Paare aus beiden neuen Kennziffern enthält. Diese Datei wird dem Forschungsteam übermittelt. Mit dieser Datei können die Datenbestände A und B zusammengeführt werden, ohne dass das Forschungsteam die Identität der Personen kennt.

Abwandlungen dieses Modells werden für verschiedenste Anwendungen verwendet. Administrativ besonders aufwändige Varianten liegen des Treuhändermodell liegen z.B. den Krebsregistern mit ihrer Trennung von Register- und Vertrauensstellen zu Grunde (Richter/Krieg 2008).

4 Anonyme Verknüpfung von Datenbanken

Möchte man verhindern, dass der Treuhänder die Identifikatoren kennt, dann kann man die Identifikatoren durch die Datengeber verschlüsseln lassen. Hierzu eignet sich prinzipiell jedes moderne Verschlüsselungsverfahren.

In der Kryptografie wurden viele verschiedene solcher Verfahren entwickelt. Die gesamte Klasse dieser Verfahren wird als Einweg-Hash-Funktionen (im Englischen: „keyed HMACs“: keyed Hashed Message Authentification Codes) bezeichnet. Das gemeinsame Merkmal all dieser Verfahren ist die Unumkehrbarkeit einer Verschlüsselung: Aus dem Ergebnis der Verschlüsselung kann nicht mehr auf die Eingabe geschlossen werden. Beispiele für diese Verfahren sind MD-5 oder SHA-1. MD-5 ist vielen Anwendern von CD-Brennern bekannt: Mit MD-5 werden Prüfsummen über ganze CDs berechnet. Ändert sich auch nur ein Bit oder eine Zahl bei einer Kopie, liefert MD-5 eine völlig andere Prüfsumme. Aus der Ähnlichkeit der Ergebnisse einer solchen Einwegfunktion kann nicht mehr auf die Ähnlichkeit der Eingaben geschlossen werden. Bei einem „keyed“ HMAC wird zusätzlich ein Passwort verwendet, so dass das Ergebnis der Verschlüsselung vom zu verschlüsselnden Text und einem Passwort abhängt.

Verschlüsselt man die Identifikatoren mit solchen HMACs führt dies zum Verlust aller Fälle, bei denen die Identifikatoren durch kleine Datenfehler (z.B. vertauschte Buchstaben im Namen) nicht vollständig übereinstimmen. Im Regelfall sind die Fälle mit nicht vollständig identischen Identifikatoren inhaltlich anders als Fälle ohne Veränderungen in den Identifikatoren. Daher kann man sich nicht auf die fehlerfreien Fälle beschränken, wenn man Aussagen über alle Fälle machen möchte. Aus diesem Grund werden z.B. in Krebsregistern phonetische Codes aus den Namen der Personen gebildet und dann die phonetischen Codes verschlüsselt (Borst/Allaert/Quantin 2001).

Phonetische Codes basieren auf Gruppen ähnlich klingender Wortbestandteile, wobei alle Bestandteile derselben Gruppe denselben Code erhalten. Verschiedene ähnlich klingende Schreibweisen eines Namens erzeugen daher den gleichen phonetischen Code. Der international am weitesten verbreitete phonetische Code ist „Soundex“ (vgl. z.B. Christen 2012a:74). Soundex führt z.B. für die Namen Engel, Engall, Engehl, Ehngel, Ehngehl, Enngeel, Enkel, Ehnkiehl und Ehenekiehl zum selben Code: E524. Dieses Beispiel illustriert das Problem phonetischer Codes allgemein: Sehr verschiedene Namen haben den gleichen Code und bedingen daher möglicherweise falsche Identifizierungen. Phonetische Codes erlauben auch keine Ähnlichkeitsberechnungen, da alle Varianten eines Namens zum gleichen Code führen. Daher führen solche Codes auch zu erheblichen Verlusten bei relativ geringen Abweichungen (z.B. Bengel hat den Code B524). Daher stellen verschlüsselte phonetische Codes keine ideale Lösung dar.

5 Anonyme und fehlertolerante Datenverknüpfungsmethoden

Die Probleme phonetischer Codes bei der Verknüpfung personenbezogener Daten haben zur Ausbildung eines eigenen Forschungsgebietes in der Informatik geführt: „Privacy Preserving Record Linkage“ oder kurz PPRL. In den letzten 10 Jahren sind innerhalb dieses Forschungsgebietes zahlreiche Methoden entwickelt worden, die fehlertolerante und anonyme Datenverknüpfungen erlauben (Vatsalan/Christen/Verykios 2013). Die meisten dieser Verfahren benötigen aber entweder in der Praxis untragbare Rechenzeiten oder zahlreiche Interaktionen zwischen den Datenhaltern. Nur sehr wenige dieser PPRL-Verfahren wurden bislang in inhaltlichen Forschungsprojekten eingesetzt. Eines dieser Verfahren wurde von der Arbeitsgruppe des Verfassers entwickelt und soll daher etwas ausführlicher dargestellt werden.

Schnell/Bachteler/Reiher (2009) schlugen eine neue Klasse von Verfahren zum PPRL vor. Neu ist hierbei die Anwendung sogenannter Bloom-Filter für die Berechnung von Namensähnlichkeiten. Bloom-Filter basieren auf der mehrfachen Abbildung von Objekten in einen binären Vektor.

Beim vorgeschlagenen Verfahren werden (wie bei vielen anderen Verfahren auch) die Namen nicht direkt miteinander verglichen, sondern es wird geprüft, wieviele Bestandteile zwei Namen miteinander gemeinsam haben. Die Namen werden zunächst in Gruppen von je zwei Buchstaben eingeteilt, daher werden diese Gruppen Bigramme genannt. Zum Beispiel besteht der Name „SMITH“ aus den Bigrammen „_S“, „SM“,MI“, „IT“, „TH“ und „H_ “.

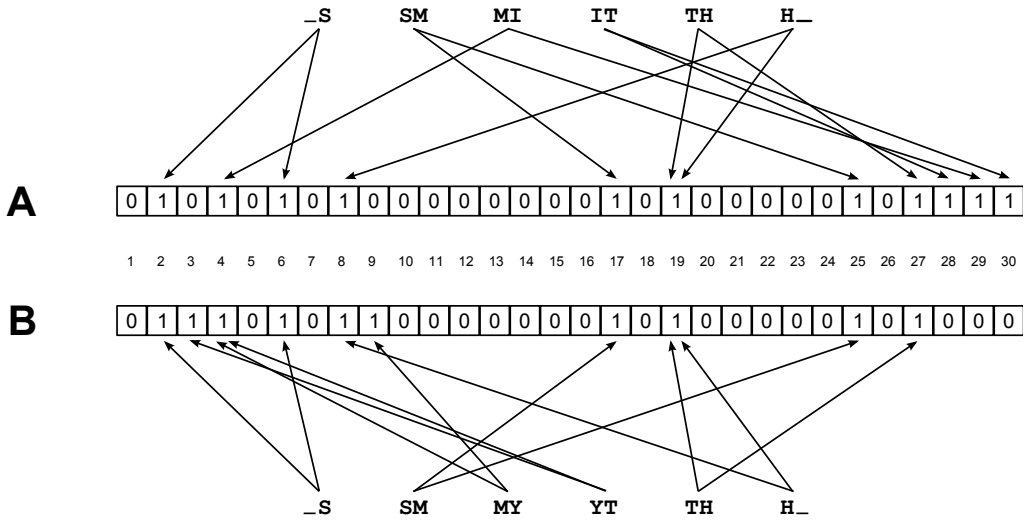

Abbildung 3: Beispiel für die Verschlüsselung der Namen „SMITH“ und „SMYTH“ in zwei binäre Vektoren

„SMYTH“ besteht aus den Bigrammen „_S“ „SM“, „MY“, „YT“, „TH“ und „H_“. Die beiden Namen haben 4 Bigramme (_S,SM,TH,H_) gemeinsam, insgesamt besitzen die beiden Namen zusammen 12 Bigramme. Ein Maß für die Ähnlichkeit zweier Bigramm-Mengen besteht in dem doppelten Anteil der gemeinsamen Bigramme an der Zahl insgesamt vorhandener Bigramme. Im Beispiel ist dieses Maß $2*4/12 \approx 0.667$. Man bezeichnet dieses Maß als Dice-Koeffizient. Namen mit höherem Dice-Koeffizient sind einander ähnlicher als Namen mit kleinem Dice-Koeffizient. Die Grundidee des neuen Verfahrens besteht darin, die Bigramme mit mehreren, unabhängigen HMACs in einen binären Vektoren abzubilden. Eine solche Abbildung wird in der Informatik allgemein als Bloom-Filter bezeichnet. Die Idee, die Ähnlichkeit der resultierenden binären Vektoren als Approximation der Bigramm-Ähnlichkeit zu verwenden, wurde von der Arbeitsgruppe erstmalig vorgeschlagen und als Safelink-Verfahren bezeichnet. Für Safelink werden alle Identifikatoren einzeln in einen eigenen Bloom-Filter gespeichert.

Die Abbildung 3 illustriert das Verfahren für die beiden Namen SMITH und SMYTH mit Bigrammen und einem (unrealistisch kurzen) Bloom-Filter mit 30 Bits und zwei Hash-Funktionen. Die Namen werden in Bigramme zerlegt und jedes Bigramm der beiden Namen in den Bloom-Filtern A und B gespeichert. So ergibt z.B. das gemeinsame Bigramm „_S“ den Hash-Funktionswert 1 für die erste Hash-Funktion und den Wert 5 für die zweite Hash-Funktion: Entsprechend werden die Bits an den Positionen 1 und 5 in beiden Bloom-Filtern auf den Wert 1 gesetzt. Im Gegensatz dazu sind die Bigramme „YT“ (Hash-Werte 2 und 3) und „IT“ (Hash-Werte 27 und 29) nur in

einem Namen vorhanden, daher werden in den beiden Bloom-Filtern unterschiedliche Bits auf den Wert 1 gesetzt. Nach der Speicherung aller Bigramme in die Bloom-Filter sind 8 identische Bit-Positionen in beiden Bloom-Filtern auf 1 gesetzt. Insgesamt sind 11 Bits im Filter A und 10 Bits in Filter B gleich 1 gesetzt. Der Dice-Koeffizient ergibt sich als $2 * 8/21 \approx 0.762$. Die Bigrammähnlichkeit der beiden Namen (0.667) wird durch die Ähnlichkeit der Bloom-Filter (0.762) approximiert.[4]

Da aber für die Speicherung kryptographische Einwegfunktionen genutzt wurden, können aus den Bloom-Filtern die Eingabenamen nicht mehr rekonstruiert werden. Dadurch wird eine fehlertolerante Verknüpfung zweier Datenbanken durch eine Forschungsgruppe oder eine Drittpartei bei vollständiger Anonymität der Identifikatoren möglich.

Für den Einsatz des Verfahrens sind eine Reihe von Entscheidungen notwendig. Weitgehend unkritisch ist die Wahl, ob Bi- oder Trigramme eingesetzt werden: Unsere Simulationen zeigen kaum Unterschiede zwischen den Ergebnissen. Ebenso ist die Wahl der Länge der Bloom-Filter innerhalb eines Intervalls von 500-1000 Bits eher unproblematisch. Die Wahl der eigentlichen Hash-Funktion erscheint ebenfalls kaum eine Rolle zu spielen. Kritisch ist hingegen die Zahl der Hash-Funktionen im Verhältnis zur Länge der Bloom-Filter: Je höher die Zahl der Hash-Funktionen, desto sicherer wird der Bloom-Filter gegenüber einem Angriff. Bei einem Bloom-Filter der Länge 500 sollte die Zahl der Hash-Funktionen mindestens bei 15 liegen. Liegt die Zahl der Funktionen deutlich über 30, verringert sich die Güte der Approximation der Bigrammähnlichkeit durch die Bloom-Filter. Bei der Wahl von 15 Funktionen wird in der Regel ein akzeptabler Kompromiss zwischen Sicherheit und Güte erreicht.

Trotzdem sind mit großem Aufwand solche Verschlüsselungen unter – für den Angreifer – optimalen Bedingungen (z.B. Verfügbarkeit der Häufigkeiten der Identifikatoren) prinzipiell angreifbar, d.h. einzelnen Bloom-Filtern lassen sich zumindest die häufigsten zugrundeliegenden Namen zuordnen (Kuzu et al. 2011). Da dies auf jeden Fall verhindert werden muss, schlug die Arbeitsgruppe des Verfassers die Verwendung eines gemeinsamen Bloom-Filters für alle Identifikatoren vor (Schnell/Bachteler/Reiher 2011). Alle Identifikatoren werden hierbei mit unterschiedlicher Anzahl von HMACs und unterschiedlichen Passwörtern in den gleichen Bloom-Filter abgebildet.

Für die Verknüpfung medizinischer Qualitätssicherungsdaten zu Geburten kann für ein Neugeborenes diese Menge von Identifikatoren aus den Merkmalen Geburtsjahr, Geburtsmonat, Geburtstag, Geschlecht, Vorname

[4]Die Approximation wird mit realistischen Werten für die Länge der Bloom-Filter und die Zahl der Hash-Funktionen besser als in diesem Anschauungsbeispiel.

(bei Geburt), Nachname (bei Geburt), Geburtsort und Geburtsland bestehen (Heller 2013). Diese Identifikatoren werden verwendet, weil sie leicht zu erheben sind und auch für große Datenbanken in der Regel zu einer einzigartigen Kombination für jede Person führen. Jedes dieser Merkmale wird anschließend in die Menge direkt aufeinanderfolgender Buchstaben (Bigramme) zerlegt (auch die numerischen Werte, z.B. 01 in „ 0“, „01“ und „1 “). Jedes Bigramm eines Identifikators wird dann mit einem für den jeweiligen Identifikator spezifischen Passwort und einer spezifischen Anzahl von Hash-Funktionen in den gleichen Bloom-Filter abgebildet. Die Zahl der Hash-Funktionen pro Identifikator richtet sich dabei nach der Zahl der möglichen Ausprägungen und der Verteilung der Ausprägungen der Merkmale in der Datenbank.

Der resultierende Bloom-Filter bildet einen kryptografischen Code, der nicht umkehrbar ist und zusammen mit den Passwörtern aus den individuellen Identifikatoren einer Person gebildet werden kann. Daher wird dieser Code als „kryptografischer Langzeitschlüssel“ („cryptographic longterm key“: CLK) bezeichnet. Bei sinnvoller Wahl der Parameter des Verfahrens funktioniert dieses Verfahren auch für große Datenbestände (Kuzu et al. 2013). Ein kryptografischer Angriff auf einen CLK ist erheblich schwieriger als ein Angriff auf einen einzelnen Bloom-Filter. Bislang wurde kein einziger erfolgreicher Angriff auf einen CLK berichtet. Die Arbeitsgruppe widmet sich derzeit der kryptografischen Analyse und der Implementierung des Verfahrens für sehr große Datenbestände, wie z.B. Krebsregistern, Mortalitätsregistern und Zensen.

6 Anwendungen und offene Probleme der Forschung zum Record-Linkage

Die Arbeitsgruppe des Autors hat Bloom-Filter-Verfahren bisher zweimal für die Deduplizierung von Krebsregistern (jeweils mehr als 300.000 Fälle) verwendet. Schon drei Jahre nach der Erstveröffentlichung gab es Anwendungen des Verfahrens für sehr große Datensätze, wie z.B. in den USA durch Weber et al. (2012), in der Schweiz durch Kuehni et al. (2012) und in Brasilien durch Santos et al. (2011). Modifikationen des Verfahrens sind Gegenstand laufender Bemühungen mehrerer Arbeitsgruppen.

International setzt die Forschung derzeit den Schwerpunkt auf die Probleme, die bei der tatsächlichen Umsetzung der Linkage-Verfahren in der Praxis entstehen. Dies sind vor allem Probleme durch große Fallzahlen und fehlende Identifikatoren. Das in der Praxis bedeutsamste Problem ist dabei das Problem sehr großer Fallzahlen. Alle Verfahren zur fehlertoleranten anony-

men Datenverknüpfung funktionieren in kleinen Datenbeständen mit wenigen Tausend Fällen. Bei größeren Fallzahlen versagen fast alle Verfahren. Bei zwei Dateien mit jeweils einer Million Fälle müssen fast $5 * 10^{11}$ Paare verglichen werden, was selbst bei 10.000 Vergleichen pro Sekunde mehr als 1.5 Jahre dauern würde. Daher versucht man Verfahren zu entwickeln, die mit Sicherheit die ähnlichsten Paare auch dann finden, wenn man nur die ähnlichsten Paare vergleicht. Dazu müssen Möglichkeiten gefunden werden, für jeden Fall nur die ähnlichsten anderen Fälle zu finden. Generell werden solche Verfahren als „Blocking"- oder „Indexing"-Verfahren bezeichnet (Christen 2012b). Mit den besten solcher Verfahren lässt sich das Problem mit $1.000.000 * 1.000.000$ Fällen innerhalb von wenigen Stunden auf einem Standardrechner lösen. Beschränkt man den Vergleich potentieller Paare auf Teilgruppen („blocks"), z.B. nur auf Personen gleicher Geburtsorte oder gleicher Geburtsjahrgänge (wobei diese dann getrennt und auf andere Art verschlüsselt werden), dann lassen sich auch sehr große Datensätze, wie sie in der Forschung mit Routinedaten im Gesundheitswesen oder bei Zensen anfallen, anonym, sicher und fehlertolerant verknüpfen. Derzeit suchen mehrere Arbeitsgruppen weltweit nach Verfahren, um die Blocks zu vergrößern, ohne dass die Rechenzeit deutlich ansteigt. Da dieses Problem sehr ähnlich zu den zentralen Problemen von Suchmaschinen im Allgemeinen ist, handelt es sich um ein sehr aktives Forschungsgebiet in der Informatik.

7 Infrastruktureinrichtungen für das Record-Linkage

Um die Zahl und die Qualität von Record-Linkage-Anwendungen in den Fachwissenschaften dauerhaft zu steigern und so neue Datenquellen für die Forschung unter Einhaltung des Datenschutzes zu erschließen, wurde im August 2011 das Deutsche Zentrum für Record Linkage (German RLC) gegründet. Angesiedelt an zwei Standorten, im Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (IAB) in Nürnberg und der Arbeitsgruppe des Verfassers am Institut für Soziologie der Universität Duisburg-Essen, steht es allen Fachrichtungen offen. Das German RLC wird seit 2011 von der Deutschen Forschungsgemeinschaft (DFG) gefördert.

Das Zentrum bietet Dienstleistungen zum Record-Linkage an. Die Serviceleistungen des German RLC umfassen die Beratung bei der Planung und Realisierung von Datenverknüpfungsprojekten, die Ausführung von Datenverknüpfungen als Auftragsarbeiten sowie die Bereitstellung entsprechender

Software zum Record-Linkage (MergeToolBox, siehe Schnell/Bachteler/Reiher 2005). Für mehrere große epidemiologische und sozialwissenschaftliche Projekte dient das German RLC als Clearing-Stelle für die Verknüpfung von sensiblen Datenbeständen. So werden beispielsweise für die Qualitätssicherungsstelle Hessen (GQH) drei Datenbanken über unterschiedliche Behandlungsstadien bei Schlaganfallpatienten verlinkt. Auf der epidemiologischen Seite werden im Projekt AeKo (Arbeitsmedizinische Forschung in epidemiologischen Kohortenstudien) im Rahmen einer Studie des Universitätsklinikums Essen Angaben der Befragten einer arbeitsmedizinischen Erhebung mit den Beschäftigungsdaten der Bundesagentur für Arbeit zusammengeführt. Auf der sozialwissenschaftlichen Seite wurden unter anderem Dienstleistungen für die meisten der derzeit laufenden Großprojekte (SAVE, SOEP, PASS, PAIRFAM) erbracht.

Weitere Informationen sind auf der Homepage des German RLC unter www.record-linkage.de zu finden.

Literatur

Borst, F., Allaert, F.-A., and Quantin, C. (2001). The Swiss solution for anonymous chaining patient files. In Patel, V., Rogers, R., and Haux, R., editors, *Proceedings of the 10th World Congress on Medical Informatics: 2–5 September 2001; London*, pages 1239–1241, Amsterdam. IOS Press.

Christen, P. (2012a). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer, Berlin.

Christen, P. (2012b). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1537–1555.

Heller, G. (2013). Pseudonymisierte Verknüpfung von Daten der gesetzlichen Qualitätssicherung. Technical Report 15. Mai, Aqua-Institut, Göttingen.

Kuehni, C. E., Rueegg, C. S., Michel, G., Rebholz, C. E., Strippoli, M.-P. F., Niggli, F. K., Egger, M., and von der Weid, N. X. (2012). Cohort profile: the Swiss Childhood Cancer Survivor Study. *International Journal of Epidemiology*, 41(6):1553–1564.

Kuzu, M., Kantarcioglu, M., Durham, E., and Malin, B. (2011). A constraint satisfaction cryptanalysis of bloom filters in private record linkage. In

Fischer-Hübner, S. and Hopper, N., editors, *The 11th Privacy Enhancing Technologies Symposium*, pages 226–245, Berlin. Springer.

Kuzu, M., Kantarcioglu, M., Durham, E. A., Toth, C., and Malin, B. (2013). A practical approach to achieve private medical record linkage in light of public resources. *Journal of the American Medical Informatics Association*, 20(2):285–292.

Richter, A. and Krieg, V. (2008). Pseudonymisierungslösungen in den Krebsregistern Schleswig-Holstein und Nordrhein-Westfalen. Technical report, TMF Workshop ID-Management, Berlin.

Santos, L. M. P., Guanais, F., Porto, D. L., de Morais Neto, O. L., Stevens, A., Escalante, J. J. C., de Oliveira, L. B., and Modesto, L. (2011). Peso ao nascer entre crianças de famílias de baixa renda beneficiárias e não beneficiárias do programa bolsa família da região nordeste. In Ministério da Saúde, editor, *Saúde Brasil 2010*, pages 271–293. Brasilia.

Schnell, R., Bachteler, T., and Reiher, J. (2005). MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. *ZA-Information*, 56:93–103.

Schnell, R., Bachteler, T., and Reiher, J. (2009). Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making*, 9(41):1–11.

Schnell, R., Bachteler, T., and Reiher, J. (2011). A novel Error-Tolerant anonymous linking code. Working Paper WP-GRLC-2011-02, German Record Linkage Center, Nuremberg.

Vatsalan, D., Christen, P., and Verykios, V. S. (2013). A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969.

Weber, S. C., Lowe, H., Das, A., and Ferris, T. (2012). A simple heuristic for blindfolded record linkage. *Journal of the American Medical Informatics Association*, 19(1e):e157–e161.

8 Kontakt

Prof. Dr. Rainer Schnell
Universität Duisburg-Essen
Lotharstr. 65
47057 Duisburg

www.uni-due.de/methods

oder

German Record Linkage Center
Regensburger Str. 104
90478 Nürnberg

email: recordlinkage@iab.de
homepage: www.record-linkage.de

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 104
D-90478 Nuremberg

Template layout

Christine Weidmann