1. Workshop des German Record Linkage Centers
Record-Linkage Anwendungen: Ziele, Verfahren und Probleme, Nürnberg
04.-05. April 2012

Kryptographische Langzeitschlüssel

Rainer Schnell, Tobias Bachteler und Jörg Reiher

German Record Linkage Center und Universität Duisburg-Essen

05. April 2012

UNIVERSITÄT DUISBURG ESSEN

German RLC

Einleitung

- Längsschnittdaten bieten ein besonders hohes Analysepotential bezüglich vieler wichtiger Forschungsfragen, in den Sozialwissenschaften wie in den Nachbarfächern.
- Um Längsschnittdaten zu erhalten, muss ein *follow-up* der Untersuchungseinheiten über Zeit erfolgen.
- So könnte etwa in der Epidemiologie eine nationale Kohorte in den Datenbanken der Krankenkassen nachverfolgt werden. Im Bereich der Kriminologie müssten für die Berechnung von individuellen Rückfallrisiken Straftäter in den Straftatsverzeichnissen über Zeit nachverfolgt werden.
- Falls keine eindeutige nationale Identifizierungsnummer verfügbar ist, basiert das Linkage von Personendaten über Zeit üblicherweise auf Identifikatoren wie Namen und Geburtsdatum.
- Da die Nutzung solcher persönlich identifizierender Merkmale Datenschutzprobleme aufwirft, werden Methoden für ein datenschutzgerechtes Identitätsmanagement benötigt.

Pseudonyme oder kryptographische Schlüssel

- Ersetzen die „wahren“ Identitäten von Individuen.
- Im Gegensatz zur Anonymisierung ist ein Record-Linkage nach einer Pseudonymisierung immer noch möglich.
- Zu unterscheiden sind (Pommerening and Reng 2004):
  - Umkehrbare Pseudonyme, erlauben die Umkehrung des Pseudonymisierungsvorgangs
  - Einweg-Pseudonyme: Die wahren Identitäten können nicht zugeordnet werden.
- Hier werden nur Einweg-Pseudonyme betrachtet.
- In manchen Situationen werden Einweg-Pseudonyme benötigt, die nur aus einem Datenelement bestehen.
- Problem: Wie können langzeitstabile Pseudonyme konstruiert werden, die nicht umkehrbar sind und gleichzeitig gute Linkage-Ergebnisse erzielen?

Anforderungen

- Eine nahe liegende Lösung ist die Verschlüsselung der Identifikatoren mit einer Einweg-Hashfunktion, etwa SHA-1.
- Der resultierende Hashwert bildet das Pseudonym. Wenn zwei Pseudonyme übereinstimmen, repräsentieren sie vermutlich dasselbe Individuum.
- Allerdings wird dieses Verfahren abweichende Pseudonyme erzeugen, sowie kleinste Fehler oder Abweichungen in den Identifikatoren auftreten.
- Es existieren ziemlich einfache Verfahren zur Erzeugung von fehlertoleranten Pseudonymen, in die meist die langzeitstabilen Merkmale Vorname, Nachname, Geschlecht und Geburtsdatum eingehen.
- Problem: Deren Fehlertoleranz ist nicht besonders groß.
- Gesucht war daher eine Methode zur Erzeugung von Einweg-Pseudonymen mit gegenüber den bisher verfügbaren Methoden verbesserten Fehlertoleranz.

Basic Anonymous Linking Code

- Immer noch die üblicherweise verwendete Methode für das Linkage von de-identifizierten Daten (Herzog et al. 2007: 194).

- Konstruktion (Herzog et al. 2007: 194):
  1. Die Identifikatoren werden in eine Zeichenkette konkateniert.
  2. Die Zeichenkette wird standardisiert (Löschen von Sonderzeichen, alle Buchstaben „to upper“, etc.)
  3. Eingabe in eine Einweg-Hashfunktion
- Beispiel:
  1. John O'Shea, 1.9.1967, male
  2. JOHNOSHEA01091967M
  3. 8017453af2064540453f02fab172f9aefaeb6310

Swiss Anonymous Linking Code

- Entwickelt von der Sektion Kryptologie des Eidgenössischen Militärdepartements im Auftrag des Schweizer Bundesamtes für Statistik (Office fédéral de la statistique 1997).
- Ziel: „. . . eine Methode zum Schutz der Vertraulichkeit der Personendaten in der Medizinischen Statistik zu entwickeln“.
- Konstruktion (auch Borst et al. 2001: 1240):
  1. Soundex(Nachname), Soundex(Vorname)
  2. Konkatenierung mit Geburtsdatum und Geschlecht
  3. Einweg-Hashfunktion (+ symmetrische Verschlüsselung)
- Beispiel:
  1. John O'Shea, 1.9.1967, male
  2. J500O20001091967M
  3. d000adaaa7f2b40a0ddf5f7b36f1bfde8f963e7f

(Encrypted) Statistical Linkage Key (SLK)

- Entwickelt am Australian Institute of Health and Welfare (AIHW)
- Zunächst als Klarschrift-Identifikator für das Home and Community Care program (Ryan et al. 1999)
- Falls verschlüsselt, bildet der SLK ein Pseudonym.
- Konstruktion (Karmel et al. 2010):
  - 2. und 3. Buchstabe des Vornamens
  - 2., 3. und 5. Buchstabe des Nachnamens
  - Konkatenierung mit Geburtsdatum und Geschlecht
  - Einweg-Hashfunktion
- Beispiel:
  1. John O'Shea, 1.9.1967, male
  2. OHSHA01091967M
  3. ab76990b084b82d3e06701c52d02485e8e2ba9fe

Eine neue Methode zur Konstruktion von kryptographischen Langzeitschlüsseln

- Wir schlagen vor, Einweg-Pseudonyme auf Basis von Bloom-Filtern zu verwenden.
- Ein Bloom-Filter ist eine Datenstruktur, die Bloom (1970) zur effizienten Prüfung von Mengenzugehörigkeiten vorgeschlagen hat.
- Bloom-Filter können auch genutzt werden um festzustellen, ob zwei Mengen approximativ übereinstimmen.
- Für unsere Lösung werden alle Identifikatoren in n-Gramme zerlegt und die resultierenden n-Gramm Mengen werden in einem einzigen Bloom-Filter gespeichert.
- Um zu prüfen, ob zwei Pseudonyme dieselbe Person repräsentieren, werden die beiden Bloom-Filter bitweise verglichen.

Bloom-Filter

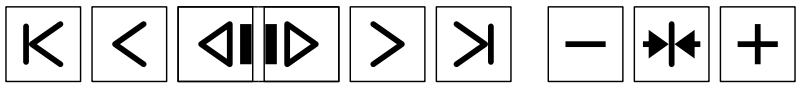
- Gegeben sei eine Menge $S = \{x_1, \ldots, x_n\}$.
- Ist z Element von S?
- Ein Bloom-Filter ist ein Bitarray der Länge l, in dem zunächst alle Bits auf 0 gestellt sind.
- Es werden k Hashfunktionen $h_1, \ldots, h_k$ verwendet, wobei jede auf den Bereich zwischen 0 und $l - 1$ abbildet.
- S wird im Bloom-Filter gespeichert, indem jedes Element $x_i \in S$ k mal abgebildet wird. Alle Bits mit den Indizes $h_j(x_i)$ werden auf 1 gestellt.
- Um die Mengenzugehörigkeit von z zu prüfen, wird z ebenfalls k mal abgebildet. Wenn alle k Indizes bereits auf 1 standen, ist z Element von S.
- Es gibt eine sehr kleine Wahrscheinlichkeit einer falsch positiv festgestellten Mengenzugehörigkeit.
- Als Hashfunktionen können ohne Weiteres Einweg-Hashfunktionen mit einem zusätzlichen geheimen Schlüssel (z. B. HMACs mit SHA-1 und MD5) verwendet werden.

Der Cryptographic Long-term Key (CLK)

1. Ein Bloom-Filter der Länge 1 000 wird initialisiert.
2. Vorname wird in 2-Gramme zerlegt und mit 10 HMACs und dem geheimen Schlüssel K_1 im Filter gespeichert.
3. Nachname wird in 2-Gramme zerlegt und mit 10 HMACs und dem geheimen Schlüssel K_2 im Filter gespeichert.
4. Geburtstag wird in 1-Gramme zerlegt und mit 10 HMACs und dem geheimen Schlüssel K_3 im Filter gespeichert.
5. Geburtsmonat wird in 1-Gramme zerlegt und mit 10 HMACs und dem geheimen Schlüssel K_4 im Filter gespeichert.
6. Geburtsjahr wird in 1-Gramme zerlegt und mit 10 HMACs und dem geheimen Schlüssel K_5 im Filter gespeichert.
7. Geschlecht wird mit 10 HMACs und dem geheimen Schlüssel K_6 im Filter gespeichert.

Verdeutlichung

John O'Shea, 01.09.1967, male



Der Abgleich zweier CLKs

- Zwei identische Datenzeilen ergeben identische Bloom-Filter.
- Zwei sich sehr ähnliche Datenzeilen ergeben zwei sich sehr ähnliche Bloom-Filter.
- Mit Hilfe des Dice-Koeffizienten kann die Ähnlichkeit zweier Bloom-Filter als

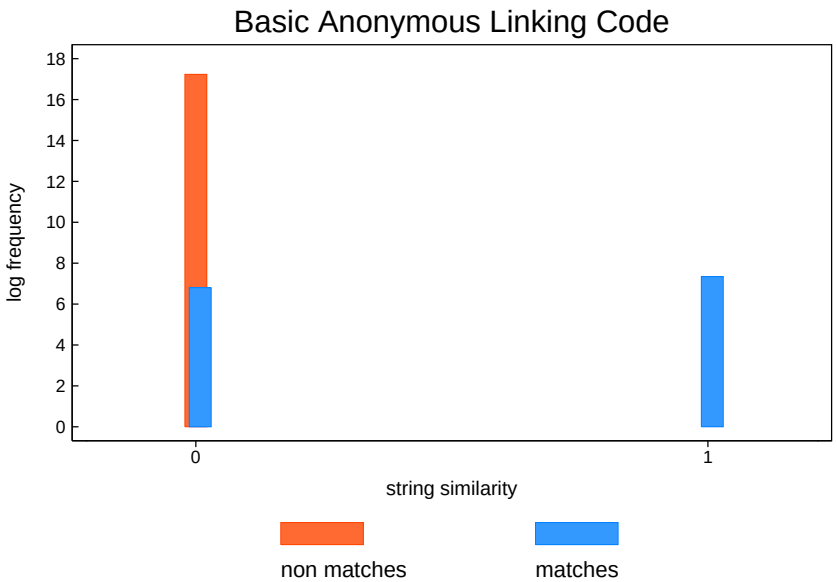
$$D_{A,B} = \frac{2h}{(a+b)} \tag{1}$$

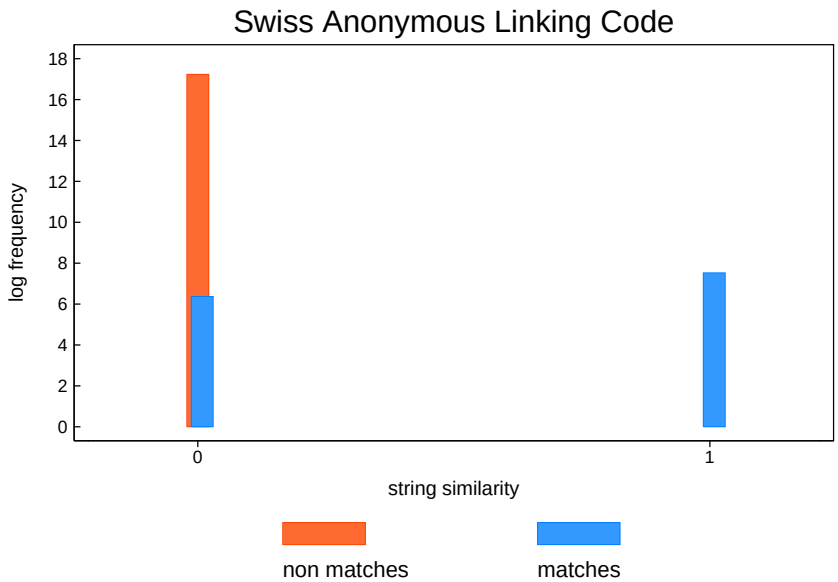
bestimmt werden, wobei h die Zahl der gemeinsam in beiden Filtern, a die Zahl der in Filter A und b die Zahl der in Filter B auf 1 gestellten Bitpositionen ist.

- Das heißt, die Ähnlicheit zwischen den Datenzeilen kann durch die Ähnlichkeit der CLKs approximiert werden.

Daten

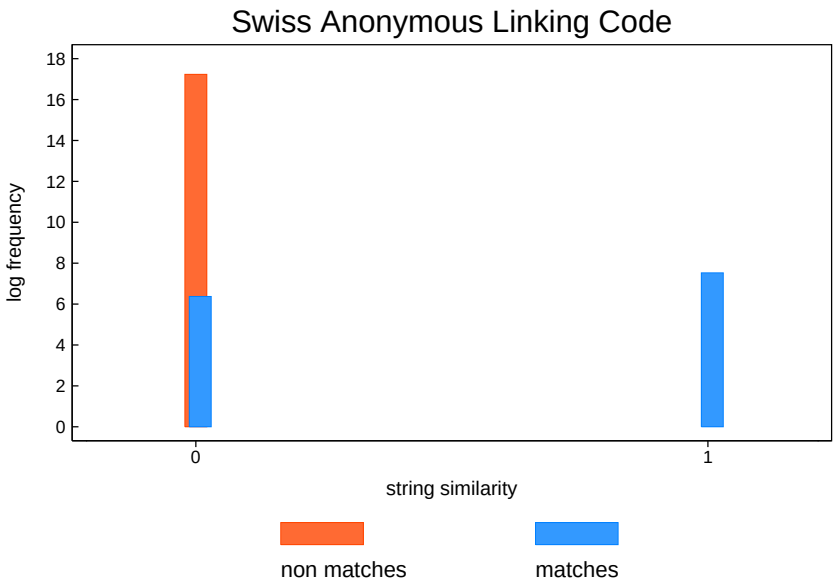
- *A*-File: $n = 2,500$, *B*-File: $n = 10,000$
- $25,000,000$ Vergleiche: $2,000$ *matches*, also $24,998,000$ *non matches*.
- Simulierte Daten basierend auf Telefonbuch-Einträgen, simulierte Fehler im *B*-File.
- CLK und der Klarschriftabgleich (Damerau-Levenshtein): Schwellenwert der Klassifikation am optimalen F-score.
- CLK: Schwelle .84, Damerau-Levenshtein: Schwelle .83
- Identifikatoren:
  1. Vorname
  2. Nachname
  3. Geburtstag
  4. Geburtsmonat
  5. Geburtsjahr
  6. Geschlecht

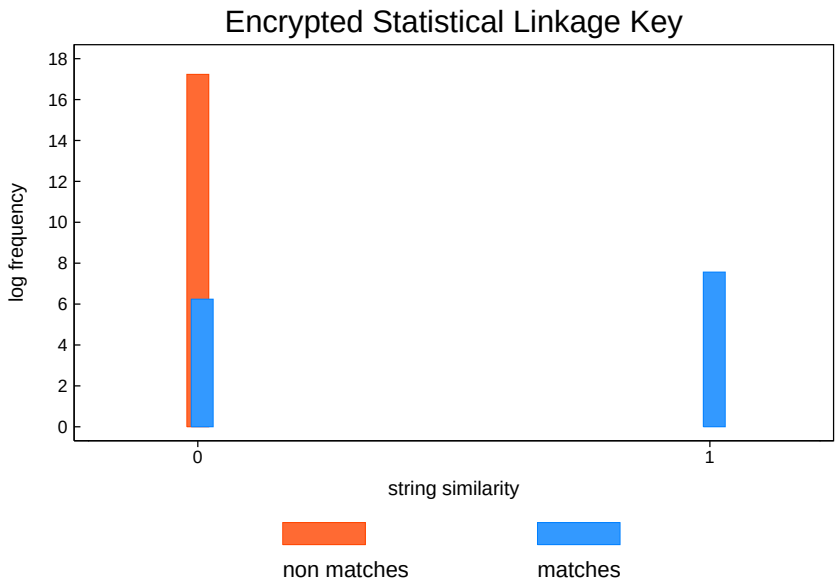




TP	FP
1265	0
FN	**TN**
735	24998000

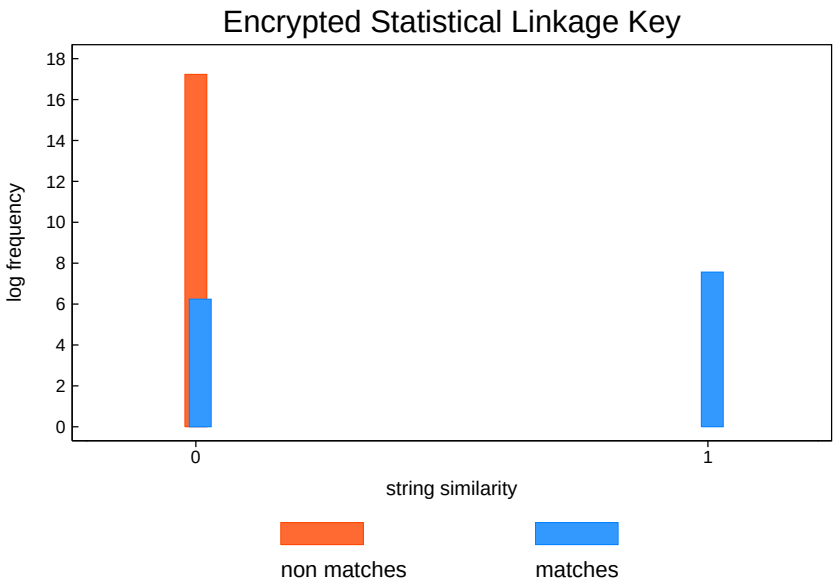
TP	FP
1521	0
FN	**TN**
479	24998000

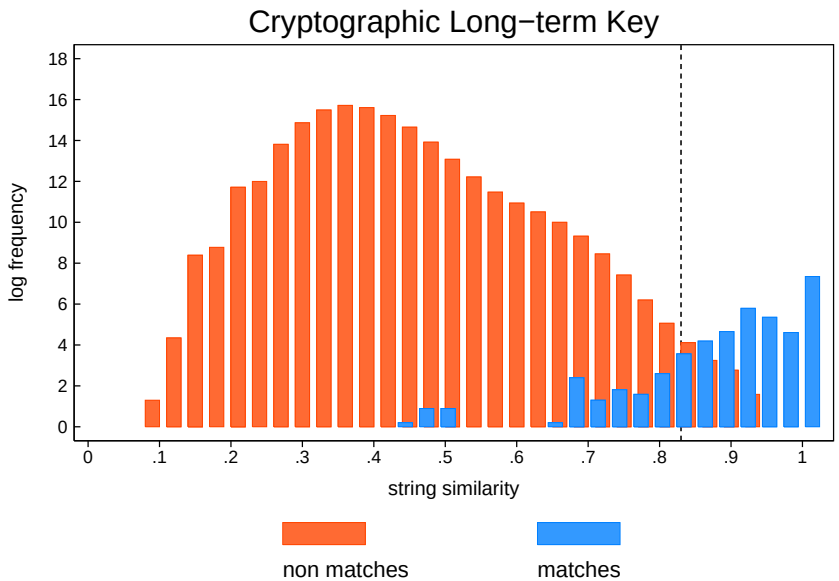




TP	FP
1521	0
FN	**TN**
479	24998000

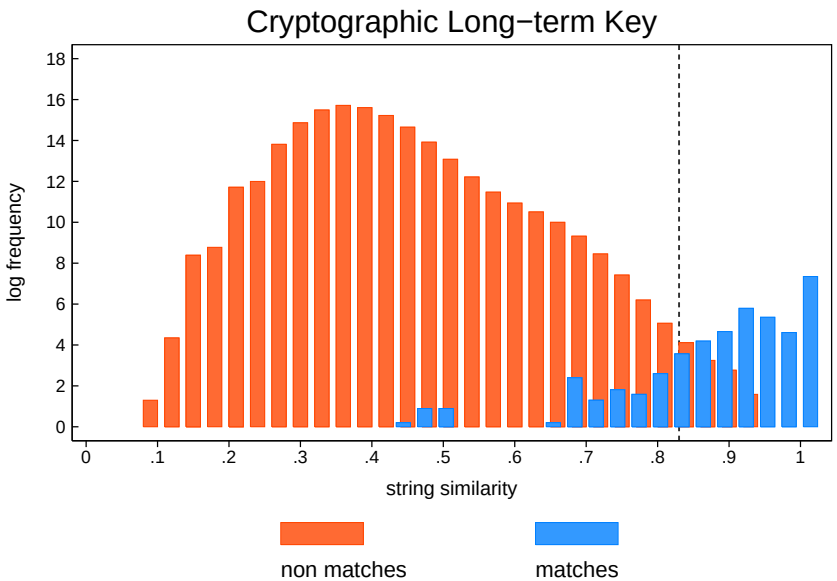
TP	FP
1580	0
FN	**TN**
420	24998000

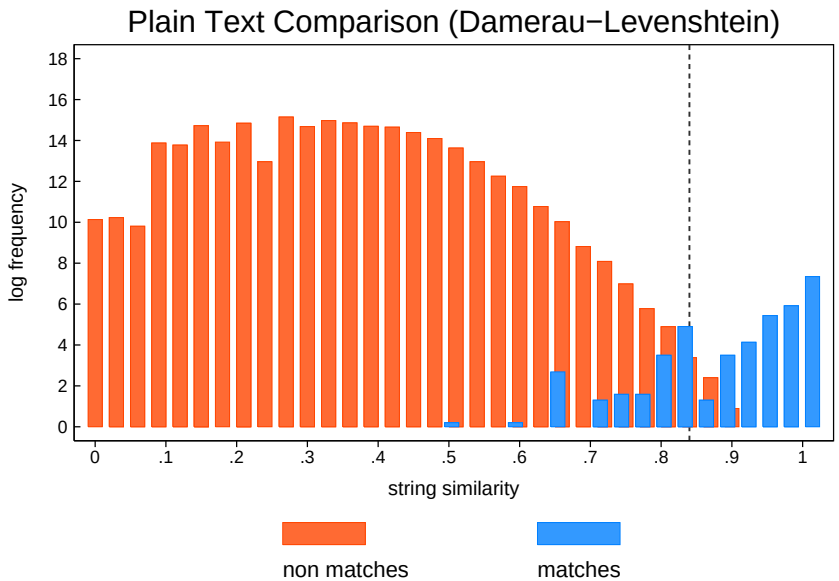




TP	FP
1580	0
FN	**TN**
420	24998000

TP	FP
1953	50
FN	**TN**
47	24997950





TP	FP
1953	50
FN	**TN**
47	24997950

TP	FP
1945	22
FN	**TN**
55	24997978

Laufzeiten

- Abgleich von $2,500 \times 10,000 = 25,000,000$ Paaren
- 2.80GHz Pentium D Prozessor mit Windows XP und 2 GB RAM. Alle Routinen wurden in Java implementiert.

Tabelle 1: *Laufzeiten in Minuten*

Methode	***Laufzeit***
Basic Anonymous Linkage Code	53:36
Swiss Anonymous Linkage Code	53:10
Encrypted Statistical Linkage Key	53:34
Cryptographic Long-term Key	72:08
Klarschrift Damerau-Levenshtein	111:32

Schlussfolgerungen

- Vorgestellt wurde eine neue, fehlertolerante und unumkehrbare Methode zur Erzeugung von Einweg-Pseudonymen.
- Ein CLK besteht aus einem einzelnen Bloom-Filter.
- Die Linkage Ergebnisse der CLKs sind deutlich besser als die der bisherigen Verfahren und nahe an einem Klarschriftabgleich.

Literatur 1

Bloom BH 1970. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM* 13 (7) 422–426.

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

Herzog T, Scheuren F, Winkler WE 2007. *Data Quality and Record Linkage Techniques*. New York: Springer.

Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y 2010. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: The experience of the PIAC cohort study. *BMC Health Services Research* 10 (41).

Kuzu M, Kantarcioglu M, Durham E, Malin B 2011. A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. Fischer-Hübner S, Hopper N (Hg.) *Privacy Enhancing Technologies: Proceedings of the 11th Privacy Enhancing Technologies Symposium: 27–29 July 2011; Waterloo, Canada*. Berlin: Springer, S. 226–245

Literatur 2

Office fédéral de la statistique (Hg.) 1997. *La protection des données dans la statistique médicale*. Neuchâtel.

Pommerening K, Reng M 2004. *Secondary use of the EHR via pseudonymisation*. Bos L, Laxminarayan S, Marsh A (Hg.) Medical Care and Compunetics 1. Amsterdam: IOS Press, S. 441–446.

Ryan T, Holmes B, Gibson D 1999. A national minimum data set for Home and Community Care. AIHW cat. no. AGE 13, Canberra: AIHW.

Schnell R, Bachteler T, Reiher J 2009. Privacy-preserving record linkage using Bloom filter. *BMC Medical Informatics and Decision Making* 9 (41).

Schnell R, Bachteler T, Reiher J 2011. A novel error-tolerant anonymous linking code. German RLC Working Paper No. wp-grlc-2011-02.

Kontakt

- rainer.schnell@uni-due.de
- German Record Linkage Center
  - www.record-linkage.de
  - recordlinkage@iab.de

Annex 1: Standardization of Files before Matching

1. Resolve Umlauts and ß
2. Remove Special Characters (only ASCII)
3. Delete Blanks
4. to upper

Annex 2: Simulation Details

- Data Simulator based on "generate2.py" (Pudjijono and Christen 2009)
- Variety of error types, including typos, phonetic errors, ocr-errors
- Error Probabilities
  - First name: $p = .2$
  - Last name: $p = .15$
  - Full birth date: $p = .05$
  - Sex: $p = .05$

  - Independent probabilities
  - Based on real "gold standard" data (PASS)

Christen P, Pudjijono A 2009. Accurate synthetic generation of realistic personal information. *P*roceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp. 507–514.

Annex 3: Damerau-Levenshtein Distance

- Minimum number of insertions, deletions, substitutions, and transpositions to convert string *a* into string *b*
- In our experiments the best performing similarity function (outperforms Bi-, Trigrams, and Jaro-Variants)

Hall PAV, Dowling GR 1980. Approximate string matching. *A*CM Computing Surveys 12 (4) 381–402.

Annex 4: Similarity Thresholds for CLK and Damerau-Levenshtein

- We calculated Recall, Precision, and F-Score for different similarity thresholds

$$recall = \frac{\sum TP}{\sum TP + \sum FN} \tag{2}$$

$$precision = \frac{\sum TP}{\sum TP + \sum FP} \tag{3}$$

$$F - Score = \frac{2}{\left(\frac{1}{precision} + \frac{1}{recall}\right)} \tag{4}$$

- We took the threshold for which F-Score is at maximum
- CLK: max F-Score .986, threshold .87
- Damerau-Levenshtein: max F-Score .986, threshold .81

Annex 5: Quantin & Colleagues

1. Identifiers ***separately*** transformed according to phonetic rules.
2. Codes encrypted with a one-way hash function.
3. A third party performs probabilistic record linkage on the resulting hash values.

Bouzelat H, Quantin C, Dusserre L 1996. Extraction and anonymity protocol of medical file. Cimino JJ (ed.) *Proceedings of the 1996 AMIA Annual Fall Symposium: 26-30 October 1996; Washington, DC*. Philadelphia: Hanley & Belfus, pp. 323–327.

Annex 6: PID-Generator (Pommerening)

- PID = "Pseudonyme Patientenidentifikatoren"
- Reversible Pseudonyms
- Trusted central organization (Data Base + Pseudonymisation Service) generates content free pseudonyms
- Holds a list of associations: Identification Data (IDAT) – Pseudonyms
- New records: Comparison with list based on IDAT
- Record linkage to combine data files is based solely on content free pseudonyms

---

Pommerening K, Reng M 2004. *Secondary use of the EHR via pseudonymisation*. Bos L, Laxminarayan S, Marsh A (Eds.) Medical Care and Compunetics 1. Amsterdam: IOS Press, pp. 441–446.

Annex 7: Soundex-Codierung

1. Der erste Buchstabe der Zeichenkette bleibt erhalten.
2. In allen anderen Positionen werden die Buchstaben a, e, h, i, o, u, w und y gelöscht.
3. Den verbliebenen Buchstaben werden gemäß folgender Regeln Zahlen zugeordnet:
   - b, f, p, v $\rightarrow$ 1
   - c, g, j, k, q, s, x, z $\rightarrow$ 2
   - d, t $\rightarrow$ 3
   - l $\rightarrow$ 4
   - m, n $\rightarrow$ 5
   - r $\rightarrow$ 6
4. Der resultierende Code wird auf 4 Stellen reduziert

- Beispiele: Hilbert $\rightarrow$ H416; Mayer $\rightarrow$ M600; Mayr $\rightarrow$ M600

Annex 8: Hash-based Message Authentication Codes (HMACs)

- A ***hash function*** converts an input string of variable length into an output string of fixed length: $H = h(I)$. H is called *hash value* or *hash sum*.
- ***One-way hash function***: Determining I from H is infeasible. Examples: SHA-1, MD5, SHA-512
- A (public) MAC-Algorithm takes a message I and a secret key k and produces a ***Message Authentication Code (MAC)***. The MAC is send along with the message. If the receiver knows the secret key k, he can generate MAC' from the message. If MAC == MAC', the message is authenticated.
- Using a one-way function along with a secret key is a way to generate MACs.
- e. g. ***HMAC***: $H = h((k \oplus pad1) + h((k \oplus pad2) + I))$.
  - h is a one-way hash function.
  - $pad1$ and $pad2$ are (public) constants.
  - $\oplus$ is the bitwise XOR operation.