Anonymisierungsverfahren

Dr. M Sariyar / Dr. J Drepper
TMF

Outline

- Relevante Grundbegriffe
- Privacy-Kriterien und Risikomodellierung
- Anonymisierungsverfahren

Relevante Grundbegriffe

Anonymisierungsbegriff

- **ISO 29100:2011:** “Anonymization is the **process** by which personally **identifiable** information (PII) is **irreversibly** altered in such a way that a PII principal can no longer be **identified** directly or indirectly, either by the PII controller alone or in collaboration with any other party.”

- **Wiederholung**: Das **Re-Identifizierungsrisiko** in den Daten soll reduziert und dennoch die **Nützlichkeit** der Daten für vielfältige Analysen erhalten werden

Gegen welche Risiken wird gesichert?

- Reidentification
  - Singling out: einen Datensatz, der zu einem Inviduum gehört, isoliert
  - Record Linkage: Datensätze als zu einem Individuum gehörig klassifiziert
- Attribute disclosure
  - Attribute Linkage: Sensitive Werte für Individuen geschlussfolgert
  - Probabilistische Inferenz: Erhöhung der Wahrscheinlichkeit für die Schlussfolgerung über sensitive Werte
- Membership disclosure
  - Table Linkage: Schlussfolgere die Präsenz eines Individuums

Arten von Attributen

1. Global-eindeutige (z.B. Sozialversicherungs-Nr.) und direkte Identifikatoren (z.B. Name)
   => Unbedingt Löschen zum Erreichen von Anonymität

2. **Quasi-Identifikatoren** (z.B. PLZ, Alter, Geschlecht) => QIDs

3. Sensitive Attribute (z.B. Krankheitsstatus)

4. Non-Sensitive Attribute

Quasi-Identifier

OECD-Definition für Quasi-Identifier:

“Variable values or combinations of variable values within a dataset that are not structural uniques but might be empirically unique and therefore in principle uniquely identify a population unit.”

Arten von Attributen

Irrelevant		QIDs		SensAttr
ID	Geschlecht	Geburtsdatum	PLZ	ICD-10 Code
6	M	1980	10117	Q90.1
8	F	1966	10117	F31.1
1	M	1979	10118	F31.0
9	M	1988	11067	F31.9
11	F	1965	11910	G50.1
4	F	1983	11934	F34.8
10	M	1973	12002	F34.8
3	F	1967	12033	F31.9
2	M	1989	12200	F31.1
5	F	1959	12200	G50.1
12	M	1976	13011	Q90.1
7	M	1975	13135	Q90.0

Privacy-Kriterien und Risikomodellierung

Häufig genannte syntaktische Privacy-Kriterien für personen-beziehbare strukturierte Daten

- ***k*-Anonymity**: Datensätze mit gleichen Werten für die QIDs tauchen mindestens k mal auf (Äquivalenzklasse) => Re-Identifikationsrisiko wird auf maximal 1/k festgelegt!

- **Distinctive *l*-Diversity**: Es gibt mindestens l verschiedene Ausprägungen des sensitiven Attributs in einer Äquivalenzklasse

- **Alternativen zu syntaktischen Kriterien**
  - Risk-based models (häufig in der Literatur zu Statistical Disclosure Control)
  - Semantic privacy models (e.g., differential privacy)

k-anonymity und l-diversity

PLZ	Alter	Krankheit
476**	2*	COPD
476**	2*	COPD
476**	2*	COPD
4790*	≥40	AIDS
4790*	≥40	COPD
4790*	≥40	Krebs
476**	3*	COPD
476**	3*	Krebs
476**	3*	Krebs

keine l-Diversität

mind. 2-Diversität

Anonymisierungsverfahren

Einige Anonymisierungsverfahren für Tabellen

- **Generalisierung und Suppression** (Details in QIDs verstecken)
  - Ersetze Werte in der höheren Ebene einer Generelisierungshierarchie
  - Full-domain oder lokale (subtree, cell) Generalisierung
  - Suppression

- **Perturbation**: z.B.
  - Additive Noise (z.B. Randomization),
  - Data swapping
  - Microaggregation: teile Datensatz in homogene Cluster der Länge k und ersetze alle Attributwerte im Cluster durch einen Wert (Mittelwert, Modalwert, etc.)

Rahmen für Anonymisierung

- Komplexe Aufgabe, da
  - Viele Verfahren existieren, die miteinander kombiniert werden können
  - Methoden oft geeignet zu parametrisieren sind

- Kontext ist zu berücksichtigen:
  - Anwendung (Nutzer, Typen & Prozess. von Daten, gewünschte Analysen, Release-Mechanismus, etc.)
  - Risiken sind zu modellieren (unterschiedliche Möglichkeiten)
  - Nutzen ist zu bewerten (unterschiedliche Möglichkeiten)
    - General purpose metric: z.B. information loss
    - Special purpose metric: z.B. für logistische Regression

Aktivitäten der TMF: Ressourcen und Unterstützung

- 2013: Weiterentwicklung des OpenAnonymizer zum TMF-ANON-Tool
  - unterstützt k-Anonymisierung u. l-Diversifizierung

- 2015: TMF-Workshop zu Anonymisierungstools:
  - ANON: a flexible tool for achieving k-anonymous and l-diverse tables
  - ARX: Comprehensive Tool for Anonymizing Biomedical Data
  - MuArgus: Software to produce safe microdata
  - sdcMicro and sdcMicroGUI: R-packages for SDC

  (Nachbericht und Folien s. www.tmf-ev.de/news/1706)

- 2016: TMF erarbeitet aktuell ein Schulungskonzept
  - Erste Evaluationsschulung am 18.5.2016 durchgeführt
  - Folgetermin voraussichtlich am 7.7.2016 (bereits ausgebucht)