IT-SICHERHEIT UND DATENSCHUTZ

KAPITEL 5 - ANONYMITÄTSMASSE

buchmann@hft-leipzig.de

LERNZIEL UND AUFBAU DIESES KAPITELS

- *Was ist ein Quasi-Identifier?*
- *Wie kann man Anonymität erreichen?*
- *Wie kann man Anonymität messen?*
  - k-Anonymity
  - l-Diversity
  - t-Closeness
  - Differential Privacy

- Lernziele
  - Sie können das Konzept des Quasi-Identifiers erklären und von Identifikatoren abgrenzen.
  - Ihnen sind die verschiedenen Anonymitätsmaße mit ihren Stärken und Schwächen vertraut.
  - Sie können Ansätze zur De-Anonymisierung erläutern

ZIELE DES DATENSCHUTZES

- *1 grundlegendes Gewährleistungsziel* **Datenminimierung**
  - *Umfang* der verarbeiteten Daten, *Zahl* der zugreifenden Stellen, *Ausmaß* der Verfügungsgewalt
- *6 elementare Gewährleistungsziele*
  - **Verfügbarkeit**: Daten für vorgesehene Zwecke verfügbar
  - **Integrität**: Daten halten Spezifikation aus Zweckbindung ein
  - **Vertraulichkeit**: Kein Zugriff durch Unbefugte
  - **Nichtverkettung**: Kein Kombinieren von Daten ohne Erlaubnis
  - **Transparenz**:Betroffene, Betreiber, Kontrollinstanzen können erkennen, wo welche Daten zu welchem Zweck vorliegen
  - **Intervenierbarkeit**: Betroffenenrechte (Benachrichtigung, Auskunft, Berichtigung, Sperrung, Löschung)

MOTIVATION

- Daten mit Personenbezug Dritten zugänglich machen
  - Dienstverbesserung, z.B., Optimierung häufig genutzter Funktionen
  - Statistik, z.B. im Gesundheitswesen
  - Erfüllung von Gesetzesanforderungen
  - Forschung
- Beispiel:
  - Erforschung von Nebenwirkungen von Medikamenten benötigt Diagnosen, Gesundheitszustand und Therapie-Informationen aller Patienten
- Wie Daten weitergeben, ohne...
  - die Privatheit der Betroffenen zu gefährden oder
  - den Nutzwert der Daten einzuschränken?

ANONYMISIERUNG

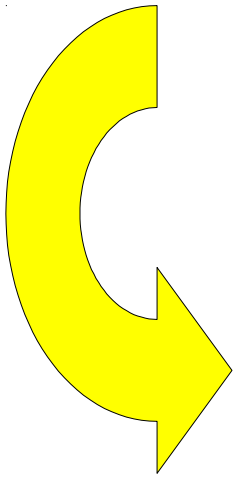
- DSGVO, Erwägungsgrund 26
  - Die Grundsätze des Datenschutzes sollten für alle Informationen gelten, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. (…) Die Grundsätze des Datenschutzes sollten daher nicht für **anonyme Informationen** gelten, d.h. für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene **Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann.** (…)

1. Wann ist ein Datensatz anonym?
2. Auf welche Weise kann man Anonymität erreichen?
3. Ist Anonymität als Schutzkonzept ausreichend?

Schlüssel

Sensibles Attribut

Name	Geb.	Geschl.	PLZ	Krankheit
Hans T.	19.04.75	M	76227	Impotenz
Peter T.	05.07.75	M	76228	Hodenkrebs
Klaus T.	17.01.75	M	76227	Sterilität
Jörg T.	23.04.81	M	76139	Schizophrenie
Uwe T.	30.12.81	M	76133	Diabetes
Melanie T.	05.07.83	W	76133	Magersucht
Inge T.	16.10.83	W	76131	Magersucht



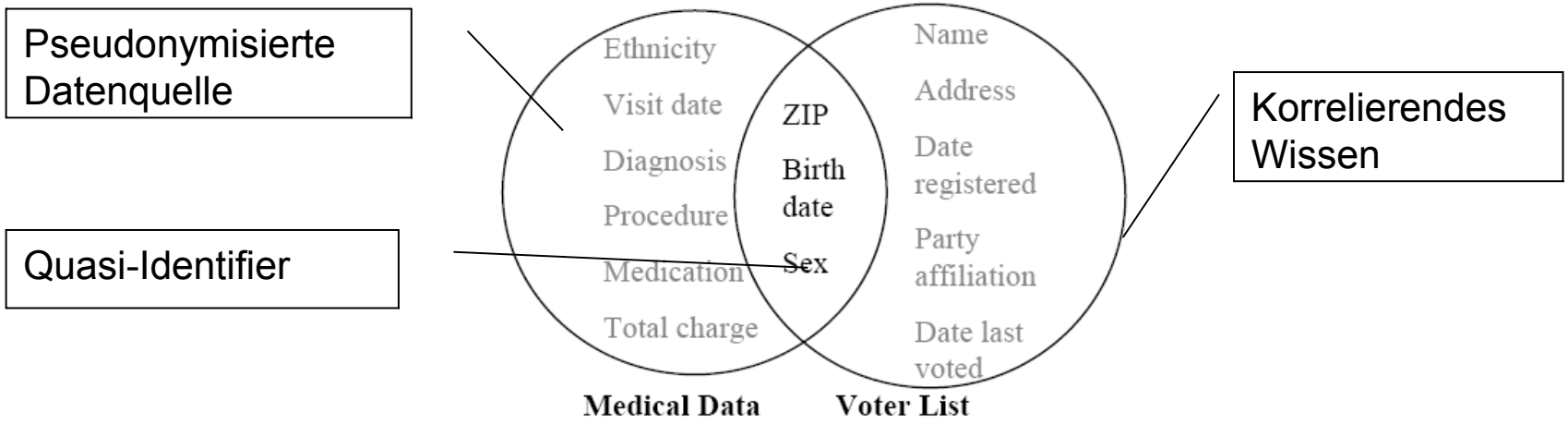
Attribut „Name“ gelöscht

Geb.	Geschl.	PLZ	Krankheit
19.04.75	M	76227	Impotenz
05.07.75	M	76228	Hodenkrebs
17.01.75	M	76227	Sterilität
23.04.81	M	76139	Schizophrenie
30.12.81	M	76133	Diabetes
05.07.83	W	76133	Magersucht
16.10.83	W	76131	Magersucht

Ziel erreicht?

QUASI-IDENTIFIER

- Zwischen 63% und 87% der amerikanischen Bevölkerung eindeutig anhand der Attribute {*Geburtsdatum*, *PLZ*, *Geschlecht*} identifizierbar



- *Re-Identifikation* durch *Verknüpfung* (linking) von *korrelierendem Wissen*

  William Weld (ehem. Gov) lebt in Cambridge und ist Wähler,
  6 Personen haben seinen Geburtstag, 3 sind männlich, 1 in seiner PLZ

DEFINITION QUASI-IDENTIFIER

- Gegeben sei
  - einen Population aus Individuen U
  - eine personenspezifische Tabelle $T(A_1...A_n)$ mit Attributen A_1 bis A_n
  - außerdem $f_c: U \rightarrow T$ und $f_g: T \rightarrow U'$ mit $U \subseteq U'$
- Q_T (ein Quasi-Identifier von T) besteht aus einem Set von Attributen $(A_i...A_j) \subseteq (A_1...A_n)$ für das gilt: $\exists p \in U: f_g (f_c (p)[Q_T]) = p$

WIE KANN MAN ANONYMITÄT ERREICHEN?

- Was kann man mit den Daten ***tun,*** damit sie anonym werden?

WAS ANONYMISIEREN?

- Es geht um die Zuordnbarkeit von Informationen

  → wir müssen uns nur den Quasi-Identifier anschauen

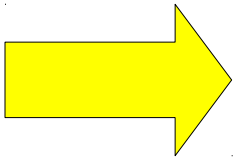
Name	Geb.	Geschl.	PLZ	Krankheit
Hans T.	19.04.75	M	76227	Impotenz
Peter T.	05.07.75	M	76228	Hodenkrebs
Klaus T.	17.01.75	M	76227	Sterilität
Jörg T.	23.04.81	M	76139	Schizophrenie
Uwe T.	30.12.81	M	76133	Diabetes
Melanie T.	05.07.83	W	76133	Magersucht
Inge T.	16.10.83	W	76131	Magersucht

- Wann welches Anonymisierungsverfahren einsetzen?

RAUSCHEN HINZUFÜGEN

- Wert_neu := Wert + Zufallszahl (Wertebereich +/-X)

Geb.	Geschl.	PLZ
19.04.75	M	76227
05.07.75	M	76228
17.01.75	M	76227
23.04.81	M	76139
30.12.81	M	76133
05.07.83	W	76133
16.10.83	W	76131

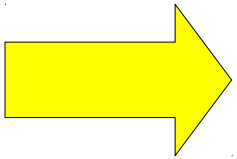


Geb.	Geschl.	PLZ
11.04.75	M	76225
25.06.75	W	76231
10.01.75	M	76226
03.05.81	M	76139
23.12.81	M	76136
15.07.83	M	76134
21.10.83	W	76129

DUMMY-DATENSÄTZE HINZUFÜGEN

- plausible, echt aussehende Daten künstlich generieren

Geb.	Geschl.	PLZ
19.04.75	M	76227
05.07.75	M	76228
17.01.75	M	76227
23.04.81	M	76139
30.12.81	M	76133
05.07.83	W	76133
16.10.83	W	76131

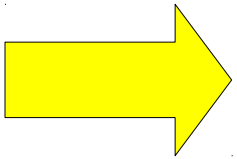


Geb.	Geschl.	PLZ
19.04.75	M	76227
05.07.75	M	76228
17.01.75	M	76227
23.04.81	M	76139
30.12.81	M	76133
05.07.83	W	76133
16.10.83	W	76131
27.05.76	W	76331
23.01.85	M	76222

INFORMATIONEN UNTERDRÜCKEN

- Einzigartige Tupel oder Attribute löschen

Geb.	Geschl.	PLZ
19.04.75	M	76227
05.07.75	M	76228
17.01.75	M	76227
23.04.81	M	76139
30.12.81	M	76133
05.07.83	W	76133
16.10.83	W	76131

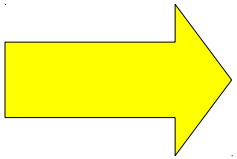


Geschl.	PLZ
M	76227
M	76227
	76133
	76133

DATEN VERTAUSCHEN

- Zufälliges Verändern der Reihenfolge der Werte

Geb.	Geschl.	PLZ
19.04.75	M	76227
05.07.75	M	76228
17.01.75	M	76227
23.04.81	M	76139
30.12.81	M	76133
05.07.83	W	76133
16.10.83	W	76131

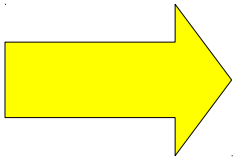


Geb.	Geschl.	PLZ
30.12.81	M	76131
16.10.83	W	76133
17.01.75	M	76139
05.07.83	M	76227
19.04.75	W	76228
23.04.81	M	76133
05.07.75	M	76227

DATEN GENERALISIEREN

- Werte auf Intervalle oder Obermengen abbilden

Geb.	Geschl.	PLZ
19.04.75	M	76227
05.07.75	M	76228
17.01.75	M	76227
23.04.81	M	76139
30.12.81	M	76133
05.07.83	W	76133
16.10.83	W	76131



Geb.	Geschl.	PLZ
20. Jh	{M,W}	76***
20. Jh	{M,W}	76***
20. Jh	{M,W}	76***
20. Jh	{M,W}	76***
20. Jh	{M,W}	76***
20. Jh	{M,W}	76***
20. Jh	{M,W}	76***

K-ANONYMITÄT

- Wie kann man Anonymität ***messen?***

IDEE DER K-ANONYMITÄT

- Daten werde in einer Form preisgegeben, dass keine Rückschlüsse auf ein **einzelnes** Individuum gezogen werden können.

- k Datensätze formen eine Äquivalenzklasse
  - k-Anonymität schützt mit einer Konfidenz von 1/k vor einer ‚korrekten' Verknüpfung einer Person mit ihren sensitiven Attributen

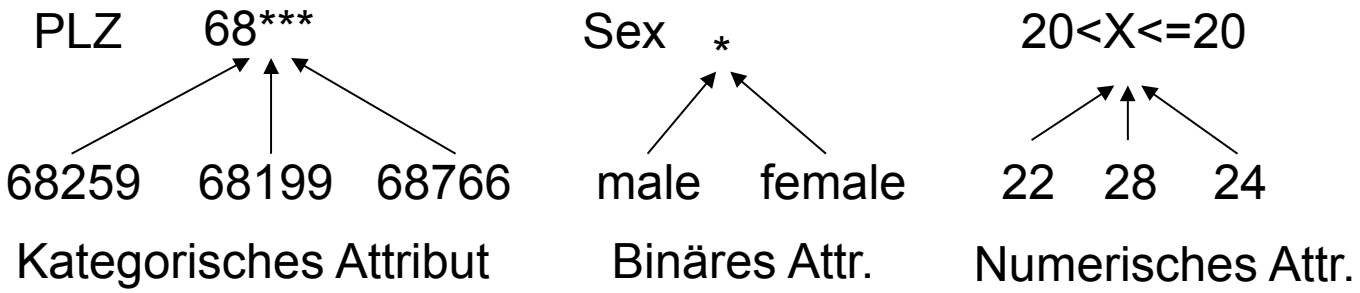
DEFINITION K-ANONYMITÄT

- Gegeben sei
  - eine personenspezifische Tabelle eine personenspezifische Tabelle $T(A_1...A_n)$ mit Attributen A_1 bis A_n
  - der dazugehörige Quasi-Identifier Q_T

- Tabelle T ist k-anonym genau dann, wenn jede Sequenz von Werten aus $T[Q_T]$ mindestens k mal in $T[Q_T]$ vorkommt.
  - Jedes Tupel ist von k-1 anderen Tupeln (bis auf die sensiblen Attribute) nicht unterscheidbar.

    → Es werden nur die Quasi-Identifier betrachtet!

K-ANONYMITÄT DURCH GENERALISIERUNG



Beispiel einer genera-lisierten Tabelle für k=2

Geb.	Sex	PLZ	Krankheit
..75	M	7622*	Impotenz
..75	M	7622*	Hodenkrebs
..75	M	7622*	Sterilität
..81	M	7613*	Schizophrenie
..81	M	7613*	Diabetes
..83	W	7613*	Magersucht
..83	W	7613*	Magersucht

Anonym?

PROBLEM: HOMOGENITÄTSANGRIFF

- Identifizierende Attribute sind generalisiert, es entstehen jedoch Gruppen mit identischen sensiblen Attributen [Mac06].

Geb.	Sex	PLZ	Krankheit
..75	M	7622*	Impotenz
..75	M	7622*	Hodenkrebs
..75	M	7622*	Sterilität
..81	M	7613*	Schizophrenie
..81	M	7613*	Diabetes
..83	W	7613*	Magersucht
..83	W	7613*	Magersucht

Beispiel einer generalisierten Tabelle für k=2

PROBLEM: KORRELIERENDES WISSEN

- Korrelierendes Wissen (Background Knowledge Attack)
  - Zusatzwissen erlaubt beispielsweise durch Ausschlussverfahren die eindeutige Zuordnung zu einer Person.

Geb.	Sex	PLZ	Krankheit
..75	M	7622*	Impotenz
..75	M	7622*	Hodenkrebs
..75	M	7622*	Sterilität
..81	M	7613*	Schizophrenie
..81	M	7613*	Diabetes
..83	W	7613*	Magersucht
..83	W	7613*	Magersucht

Beispiel einer generalisierten Tabelle für k=2

WEITERE HERAUSFORDERUNGEN (1/2)

- Sortierungsbasierte Angriffe (Unsorted Matching Attack)
  - Werden anonymisierte Tabellen GT1 und GT2 in gleicher Sortierung veröffentlicht, kann originaler Datenbestand (PT) rekonstruiert werden

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

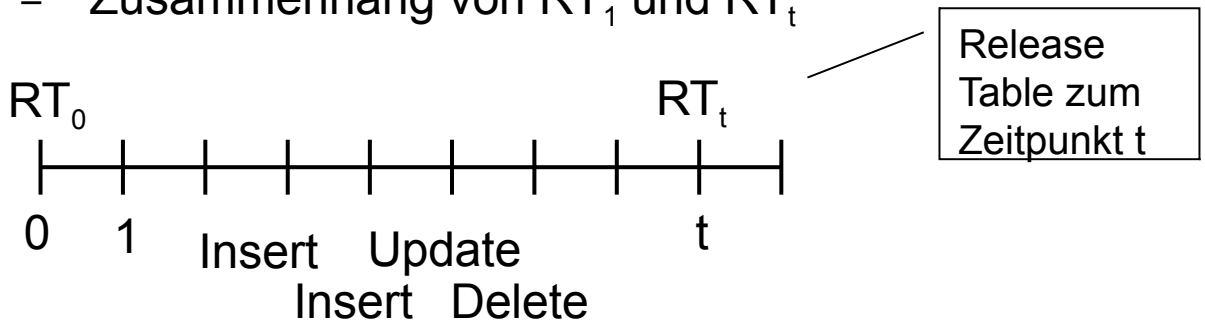
Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

→ *nicht spezifisch für die k-Anonymität!*

WEITERE HERAUSFORDERUNGEN (2/2)

- Angriffe bei dynamischen Datenbeständen (Temporal Attack)
  - Zusammenhang von RT_1 und RT_t



  - Möglicher Angriff
    - $RT_1 \bowtie RT_t$
    - $RT_1 \cap RT_t$
    - $RT_1 \setminus RT_t$

→ *nicht spezifisch für die k-Anonymität!*

L-DIVERSITY

- Wenn k-Anonymität kein „gutes“ Maß ist, was wäre denn dann besser?

PREISGABE VON INFORMATIONEN BEI DER K-ANONYMITY

- **Positive Preisgabe**
  - Wert eines sensiblen Attributes (mit einer bestimmten Wahrscheinlichkeit) für eine Person *bestimmbar*
    → Wer 83 geboren wurde hat Magersucht
- **Negative Preisgabe**
  - Wert eines sensiblen Attributes (mit einer bestimmten Wahrscheinlichkeit) für eine Person *ausgeschlossen*
    → Wer 81 geboren wurde hat keine Magersucht, Impotenz, etc.

Geb.	Sex	PLZ	Krankheit
..75	M	7622*	Impotenz
..75	M	7622*	Hodenkrebs
..75	M	7622*	Sterilität
..81	M	7613*	Schizophrenie
..81	M	7613*	Diabetes
..83	W	7613*	Magersucht
..83	W	7613*	Magersucht

PRINZIP VON L-DIVERSITY

- Gegeben
  - eine Tabelle **T** und die generalisierte Tabelle **T***
  - ein Attribut **q*** als generalisierter Wert von **q**
  - ein **q*-Block** ist eine Menge von Tupeln aus T*, deren nicht-sensiblen Attribute zu q* generalisiert wurden

- Ein q*-Block ist l-divers, wenn er mindestens l „*wohl-repräsentierte*“ Werte für das sensible Attribut S beinhaltet. Eine Tabelle ist l-divers, wenn jeder q*-Block l-divers ist.

- Im Folgenden zwei konkrete Definitionen von „wohl-repräsentiert“
  - Anzahl einzigartiger Werte
  - Entropie

DISTINCT L-DIVERSITY

- Definition Distinct l-Diversity:
  - eine Tabelle **T** und die generalisierte Tabelle **T***
  - ein Attribut **q*** als generalisierter Wert von **q**
  - ein **q*-Block** ist eine Menge von Tupeln aus T*, deren nicht-sensiblen Attribute zu q* generalisiert wurden

    Eine Tabelle ist Distinct l-Diverse, wenn jeder q* Block mindestens l *unterschiedliche* sensiblen Werte aufweist

ENTROPY-L-DIVERSITY

- Definition Entropy-l-Diversity:
  - eine Tabelle **T** und die generalisierte Tabelle **T***
  - ein Attribut **q*** als generalisierter Wert von **q**
  - ein **q*-Block** ist eine Menge von Tupeln aus T*, deren nicht-sensiblen Attribute zu q* generalisiert wurden

    Eine Tabelle ist Entropy-l-Diverse, wenn für jeden q*-Block gilt:

$$-\sum_{s\in S} P_{(q*,s)} \log\left(P_{(q*,s)}\right) \geq \log(l) \quad \text{und} \quad P_{(q*,s)} = \frac{n_{(q*,s)}}{\sum_{s'\in S} n_{(q*,s')}}$$

  - $P_{(q*,s)}$ ist der Anteil der Datensätze in *q**, der den Wert *s* hat
  - l ist minimales Maß der Unordnung in den Blöcken

BEISPIEL ENTROPY-L-DIVERSITY

Beispiel einer generalisierten Tabelle für k=2 entropy-0-diversity

Geb.	Sex	PLZ	Krankheit
..75	M	7622*	Impotenz
..75	M	7622*	Hodenkrebs
..75	M	7622*	Sterilität
..81	M	7613*	Schizophrenie
..81	M	7613*	Diabetes
..83	W	7613*	Magersucht
..83	W	7613*	Magersucht

Beispiel einer generalisierten Tabelle für k=2 entropy-2.8-diversity

$$-3*\frac{1}{3}*\log(\frac{1}{3}) = 0.47$$

$$-[\frac{2}{4}*\log(\frac{1}{4}) + \frac{2}{4}*\log(\frac{2}{4})] = 0.45$$

$$\log(2.8) = 0.44$$

Geb.	Sex	PLZ	Krankheit
..75	M	7622*	Impotenz
..75	M	7622*	Hodenkrebs
..75	M	7622*	Sterilität
..8*	*	7613*	Schizophrenie
..8*	*	7613*	Diabetes
..8*	*	7613*	Magersucht
..8*	*	7613*	Magersucht

PROBLEME VON L-DIVERSITY

- Schwierig zu erreichen und unter Umständen unnötig
- Es kann gezeigt werden, dass die Entropie der gesamten Tabelle mindestens log(l) sein muss.
  - Kommen wenige Attribute sehr häufig vor, ist diese Anforderung sehr restriktiv.
  - Beispiel: Eine Tabelle, die auch den Zustand „gesund“ speichert
- Nicht ausreichend, um vor der Preisgabe von Attributen zu schützen
  - Skewness Attack
  - Similarity Attack

L-DIVERSITY IST SCHWIERIG ZU ERREICHEN

- Beispiel
  - Nur ein sensibler Wert: Infiziert:={positiv, negativ}
  - 10.000 Datensätze, 99% negativ, 1% positiv

- Problem:
  - Private Information nur negativ
  - 2-diversity für eine Klasse die nur negative Tupel abbildet unnötig
  - Bei nur 1% positiver Tupel kann es maximal 100 2-diverse Äquivalenzklassen geben → u.U. hoher Informationsverlust

SKEWNESS ATTACK

- Beispiel
  - Nur ein sensibler Wert: Infiziert:={positiv, negativ}
  - 10.000 Datensätze, 99% negativ, 1% positiv
  - A: Eine Äquivalenzklasse hat gleich viele positive wie negative Datensätze
  - B: Ein Äquivalenzklasse hat 49/1 positive und 1/49 negativen DS

- Skewness Attack
  - A: Jeder in dieser Klasse hätte zu 50% eine Infektion, auch wenn das im Kontrast zu dem originalen Datenbestand steht.
  - B: Obwohl deutlich unterschiedliche Privatheit ist die Diversity gleich.

→ l-Diversity berücksichtigt nicht die Gesamtverteilung von sensiblen Attributen

SIMILARITY ATTACK

- Sensible Attribute sind unterschiedlich, jedoch semantisch ähnlich

Geb.	Sex	PLZ	Krankheit
..75	M	7622*	Impotenz
..75	M	7622*	Hodenkrebs
..75	M	7622*	Sterilität
..8*	*	7613*	Schizophrenie
..8*	*	7613*	Diabetes
..8*	*	7613*	Magersucht
..8*	*	7613*	Magersucht

Beispiel einer generalisierten Tabelle für k=3 entropy-2(.8)-diversity mit ähnlichen Repräsentanten in einer Äquivalenzklasse

T-CLOSENESS

- Wenn l-Diversity kein „gutes“ Maß ist, was wäre denn dann besser?

PREISGABE VON INFORMATIONEN (1/4)

Wissen eines
potentiellen Angreifers

1) Initial

„Junge Mädchen sind gefährdet für Magersucht“

Geb.	Sex	PLZ	Krankheit
..**	*	*****	*
..**	*	*****	*
..**	*	*****	*
..**	*	*****	*
..**	*	*****	*
..**	*	*****	*
..**	*	*****	*

Belief	Wissen
B_0	Korrelierendes Wissen

PREISGABE VON INFORMATIONEN (2/4)

Wissen eines potentiellen Angreifers

1) Initial
2) Ohne Bezug auf die Person

„28% im Datenbestand haben Magersucht“

Geb.	Sex	PLZ	Krankheit
..**	*	*****	Impotenz
..**	*	*****	Hodenkrebs
..**	*	*****	Sterilität
..**	*	*****	Schizophrenie
..**	*	*****	Diabetes
..**	*	*****	Magersucht
..**	*	*****	Magersucht

Belief	Wissen
B_0	Korrelierendes Wissen
B_1	Gesamtverteilung der sensiblen Werte Q

Eine große Differenz bedeutet viel neue Information bzw. Neues im Vergleich zu einer weit verbreiteten Annahme

PREISGABE VON INFORMATIONEN (3/4)

Wissen eines potentiellen Angreifers

1) Initial
2) Ohne Bezug auf die Person
3) Preisgabe der generalisierten Tabelle

„Magersucht ist ein Problem der 80er in Karlsruhe Ost“

Geb.	Sex	PLZ	Krankheit
..75	M	7622*	Impotenz
..75	M	7622*	Hodenkrebs
..75	M	7622*	Sterilität
..8*	*	7613*	Schizophrenie
..8*	*	7613*	Diabetes
..8*	*	7613*	Magersucht
..8*	*	7613*	Magersucht

Belief	Wissen
B_0	Korrelierendes Wissen
B_1	Gesamtverteilung der sensiblen Werte Q
B_2	Verteilung P_i der sensiblen Werte in Äquivalenzklasse i

Eine große Differenz bedeutet viel neue Information bzw. Neues im Vergleich zu einer weit verbreiteten Annahme

PREISGABE VON INFORMATIONEN (4/4)

Belief	Wissen
B_0	Korrelierendes Wissen
B_1	Gesamtverteilung der sensiblen Werte Q
B_2	Verteilung P_i der sensiblen Werte in der Äquivalenzklasse i

- $B_0 – B_1$
  - Wissensgewinn über die gesamte Population
  - eine große Differenz bedeutet viele neue Informationen
- $B_0 – B_2$
  - l-Diversity: Differenz zwischen B_0 und B_2 durch die Diversity-Anforderung an Population begrenzen
- $B_1 – B_2$
  - t-Closeness: Informationen begrenzen, die über ein bestimmtes Individuum gelernt werden kann

PRINZIP VON T-CLOSENESS

- Eine Äquivalenzklasse hat t-Closeness, wenn der Abstand der Verteilung eines sensiblen Attributes innerhalb der betrachteten Klasse und der Verteilung des Attributes in der gesamten Tabelle kleiner einer Schranke t ist.

- Eine Tabelle besitzt t-Closeness, wenn alle Äquivalenzklassen t-Closeness haben.

Wie messen wir die Distanz zwischen Verteilungen?

DISTANZMASS: EARTH MOVER‘S DISTANCE

- Earth Mover‘s Distance misst Distanz zwischen zwei Verteilungen in einer definierten Region

- Gegeben
  - Verteilung P = $\{p_1, p_2, \ldots, p_m\}$
  - Verteilung Q = $\{q_1, q_2, \ldots, q_m\}$
  - d_{ij}: Die Ground Distance zwischen Element i aus P und Element j aus Q.

- Idee
  - Finde einen Fluss F=$[f_{ij}]$ bei dem f_{ij} der Fluss der Masse von Element i aus P zu Element j aus Q ist, der die gesamte Arbeit minimiert.

DEFINITION

- Earth Mover's Distanz

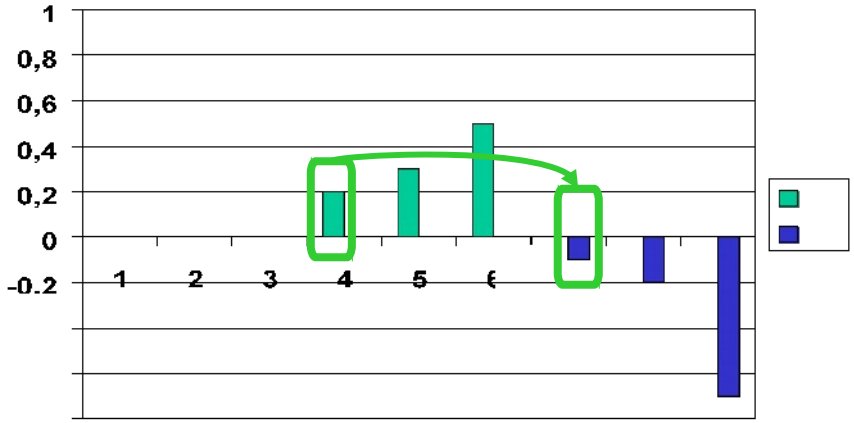
$$D[P,Q] = WORK(P,Q,F) = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij}$$

- Idee (zur Vereinfachung 1-dimensional dargestellt)
  - Gegeben zwei Verteilungen V1 und V2
  - Fülle die nächstgelegenen Löcher

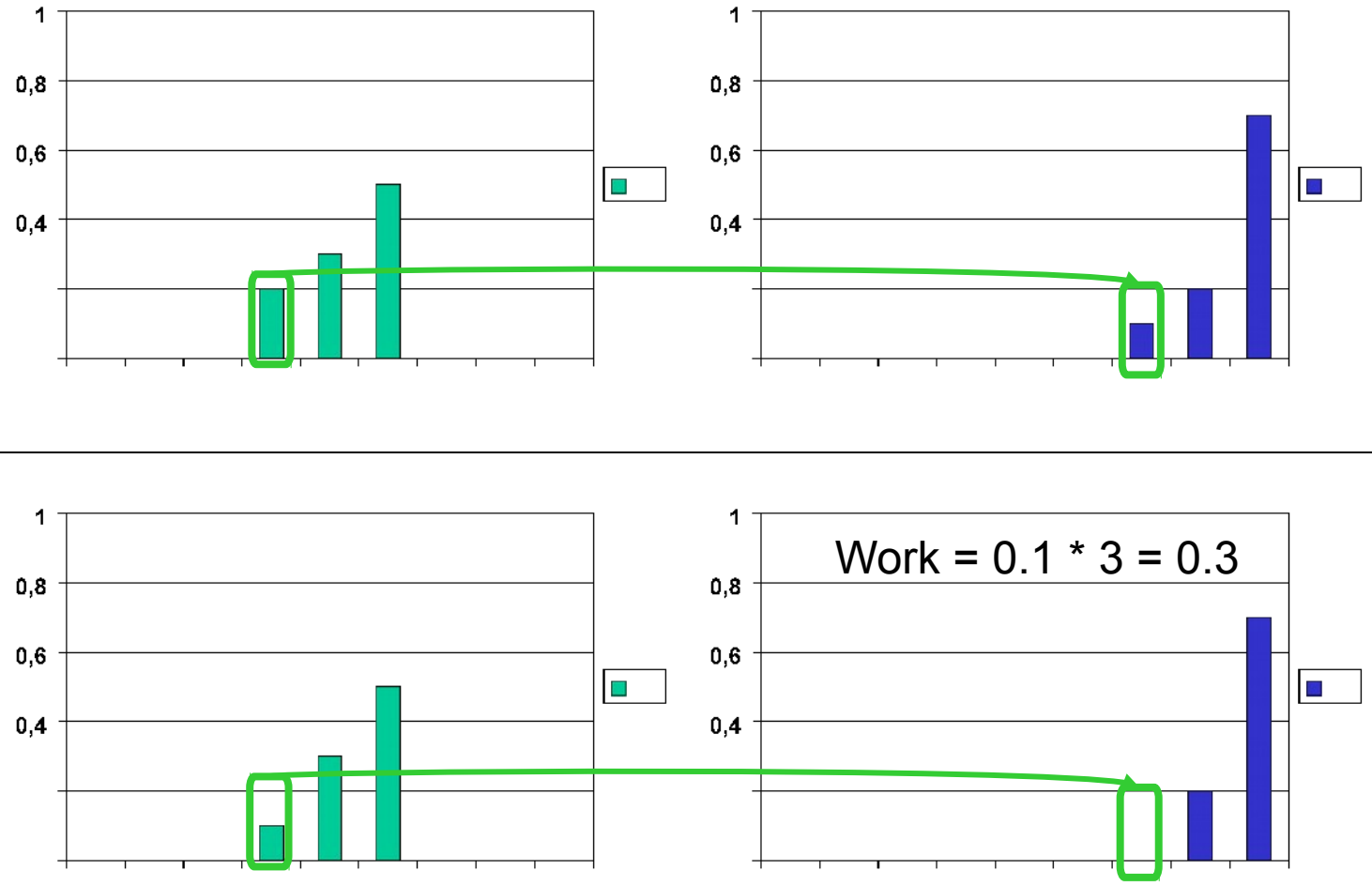
Work(f,i,j) = f * | i - j |;

f = Anzahl Werte (Y-Achse)

i,j die Werte selbst (X-Achse)



BEISPIEL EARTH MOVER‘S DISTANCE



Work(f,i,j) = f * | i - j |;

BEISPIEL: EMD FÜR UNTERSCHIEDLICHE ATTRIBUTTYPEN

- EMD für numerische Attribute
  - Sortierdistanz

$$ordered - dist(v_j, v_j) = \frac{|i - j|}{m - 1}$$

- EMD für kategorische Attribute
  - Äquivalenzdistanz

$$equal - dist(v_i, v_j) = 1$$

  - Hierarchische Distanz

$$hierarchical - dist(v_i, v_j) = \frac{level(v_i, v_j)}{H}$$

DISKUSSION

- t-Closeness führt zu sehr starken Veränderungen im Datenbestand
  → Nützlichkeit der Daten?

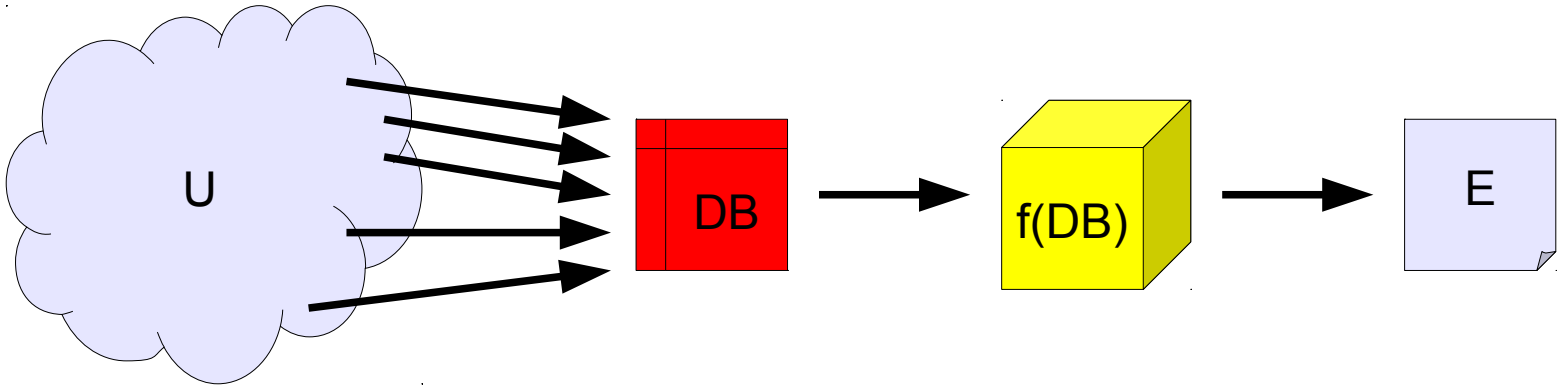
- Ein Angreifer kann aus einem t-closeness Datenbestand noch lernen
  - Ist eine Person im Datenbestand enthalten?
    - z.B. schließen Quasi-Identifier meinen Nachbarn aus? (negative disclosure)
  - Passt eine Person auf ein bestimmtes Profil?
    (muss nicht mal im Datenbestand enthalten sein!)
    - z.B. kann ein medizinischer Datensatz

- Negative Disclosure
  - Kann ausgeschlossen werden, dass eine Person in der Tabelle ist?

DIFFERENTIAL PRIVACY

- Der aktuelle Stand der Forschung

ZURÜCK ZUM AUSGANGSSZENARIO



- E := select count(*) from DB where Krankheit = 'Impotenz'

- Was wünscht man sich wirklich? Und lässt sich das realisieren?
  - Mein Datensatz hat keinerlei Auswirkungen
    f(DB) = f(DB – {ich})
  - Angreifer lernt nichts über mich
    Pr[Geheimnis(ich) | E] = Pr[Geheimnis(ich)]

WIEVIEL PRIVATHEIT LÄSST SICH REALISIEREN?

- Das Hinzufügen/Löschen eines beliebigen Datensatzes X hat kleine Auswirkungen auf Ergebnis E

$$\frac{Pr[\ f(DB) = E\ ]}{Pr[\ f(DB +/- \{X\}) = E]} \approx 1 \quad \text{für alle DB, X und E}$$

- f ist ein beliebiger Algorithmus,
  der Ergebnis E + zufälliges Rauschen produziert
  - z.B: E := select count(*)+rand() from DB where Krankheit = 'Impotenz'

→ zwei „Welten“, die sich kaum unterscheiden

  - *Wieviel Zufall ist angemessen, um Privatheit zu wahren?*

DIFFERENTIAL PRIVACY

- Idee: Wenn Hinzufügen/Entfernen einer Person die Verteilung eines Anfrageergebnisses nicht signifikant ändert, bleibt Privatheit gewahrt.
  - keine Unterscheidung in Quasi-Identifier und sensitive Attribute

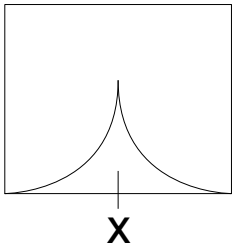
- Es gilt:
  - Verteilung aller Attribute ist bekannt, d.h.
    Pr[E]: Eintrittswahrscheinlichkeit für Ereignis E ist bekannt
  - X: Datensatz einer Person
  - Zwei Datenbestände DB_1 und DB_2: $DB_2 = DB_1 \cup \{X\}$

- Eine Funktion *K* genügt der ϵ-differential privacy, wenn für alle DB_1 und DB_2 und alle $S \subseteq Wertebereich(K)$ gilt:

$$Pr[K(DB_1) \in S] \leq e^{\epsilon} * Pr[K(DB_2) \in S]$$

BEISPIEL FÜR DIFFERENTIAL PRIVACY (1/2)

- Anfrage: select count(*) from database where P
  - Anzahl der Datensätze für die Prädikat P gilt
- Anonymisierungsfunktion:
  - Addiere auf Ergebnis einen zufälligen (Laplace verteilten) Wert mit Wahrscheinlichkeitsdichte (Schwerpunkt x)



$$p(x) \propto e^{(-|x|/\epsilon)}$$

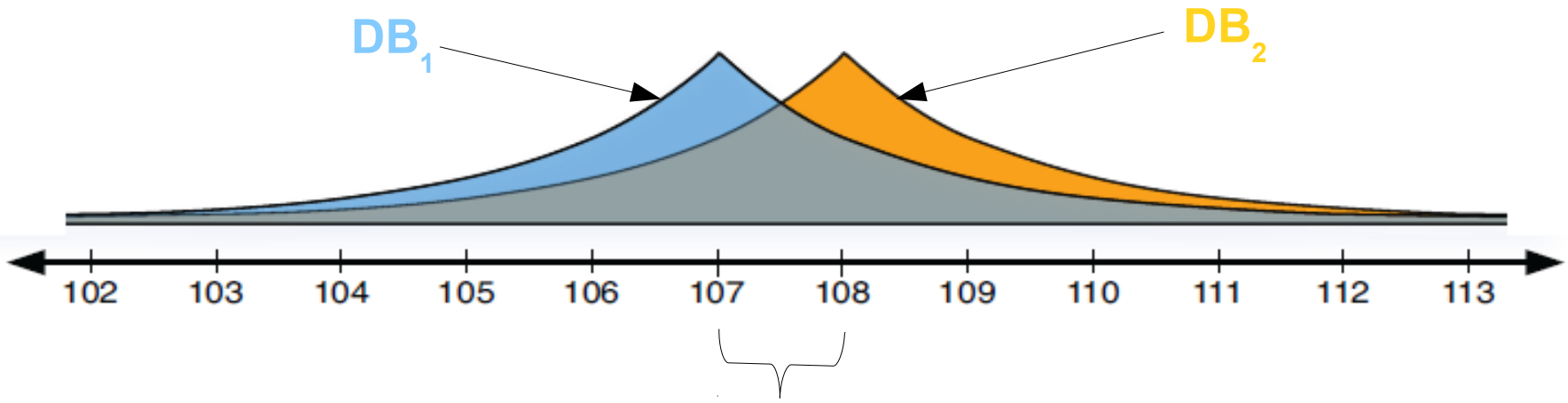
  - Definition von K (n Tupel genügen P):

$$K(\{x_1, .., x_n\}) = |(\{x_1, .., x_n\})| + Laplace(1/\epsilon)$$

- Є-differential privacy garantiert:
  - Ändert sich der Schwerpunkt der Verteilung um höchstens 1, verändert sich das Ergebnis um (multiplikativ) um höchstens e^{ϵ}.

BEISPIEL FÜR DIFFERENTIAL PRIVACY (2/2)

- Ergebnis „Count“ Anfrage an DB_1: 107
- Ergebnis „Count“ Anfrage an DB_2: 108
- Werte der Funktion K als Wahrscheinlichkeitsdichte:



definierte Maximaldifferenz

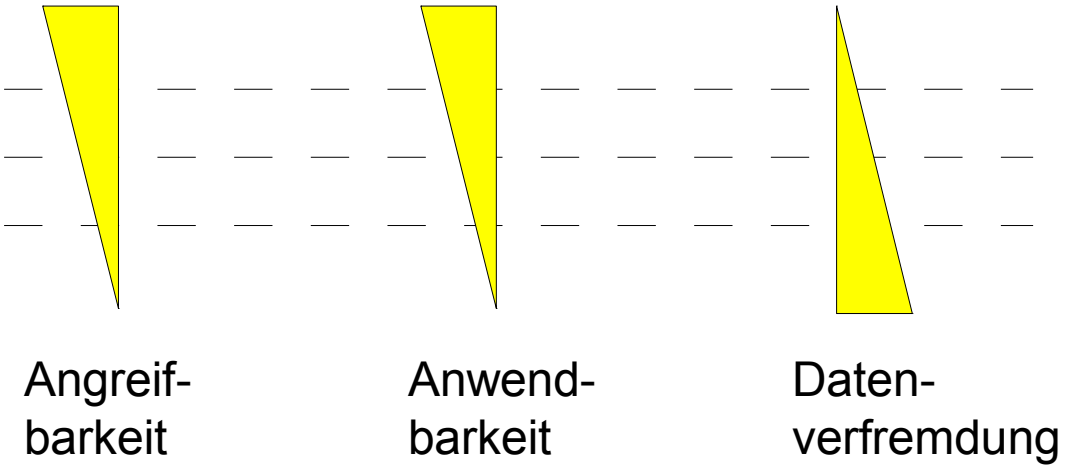
DIFFERENTIAL PRIVACY - DISKUSSION

- Formale Privatheitsgarantie für statistische Datenbanken
  - Stellt sicher und quantifiziert wie „groß“ das Risiko eines einzelnen bei der Veröffentlichung der Daten in einer statistischen Datenbank ist
  - Zugehörigkeit zur Datenbank abstreitbar
  - Updates der Datenbank sind kein Datenschutzproblem
    - jedenfalls solange sich Verteilung der Attribute nicht ändert
- Herausforderungen
  - für komplexere Anfragen ist Nachweis schwierig
    - wie ist die Verteilung der Attribute im Anfrageergebnis über alle theoretisch möglichen Ergebnisse?
  - anfragebezogen, keine Veröffentlichung einer generalisierten Tabelle
  - Angreifer lernt etwas, selbst wenn eine Person nicht Bestandteil von DB
    - Mit Hintergrundwissen:
      Japaner wenig Herzkrankheiten + Hintergrundwissen(Person = Japaner)
    - Bei Gruppenzugehörigkeit
      Wenn ich immer dasselbe tue wie meine 10 besten Freunde wird e^{ε} größer

ABSCHLUSS

ZUSAMMENFASSUNG

- Anonymität: Entfernen von Identifikatoren zu wenig!
  - Quasi-Identifier
- mehrere verschiedene Anonymitätsmaße mit unterschiedlichen Eigenschaften
  - k-Anonymity
  - l-Diversity
  - t-Closeness
  - Differential Privacy



LITERATUR

[Swe02] Sweeney, L.: K-*Anonymity: A Model for Protecting Privacy* Uncertainty and Fuzziness in Knowledge.-Based Systems, 2002, 10

[Mac06] Machanavajjhala, A.; Gehrke, J.; Kifer, D. & Venkitasubramaniam, M.: *l-Diversity: Privacy Beyond k-Anonymity*, International Conference on Data Engineering, 2006

[LiN07] Li, N.; Li, T. und Venkatasubramanian, S.: *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*, International Conference on Data Engineering, 2007

[Dw06] Dwork, C.: *Differential Privacy*. International Colloquium on Automata, Languages and Programming, 2006

MÖGLICHE PRÜFUNGSFRAGEN

- Auf welche Weise würden Sie einen Datensatz anonymisieren, bei dem
  - die Daten später für eine Durchschnittsberechnung benötigt werden (d.h., der Durchschnitt soll für durch die Art der Anonymisierung nicht stark verändert werden)
  - die Ursprungsdaten einer Normalverteilung entsprechen
  - die Ursprungsdaten alle sehr ähnlich sind?

- Geben Sie eine Beispiel-Tabelle mit 5 Spalten und 7 Zeilen an, die k=3-Anonym ist.

- Erfüllt eine Tabelle, die die t-Closeness erfüllt, auch immer die k-Anonymität? Begründen Sie Ihre Antwort.