Karlsruher Institut für Technologie

Datenschutz und Privatheit in vernetzten Informationssystemen

Kapitel 3: Anonymität und Anonymitätsmaße

Erik Buchmann (buchmann@kit.edu)

IPD, Systeme der Informationsverwaltung, Nachwuchsgruppe „Privacy Awareness in Information Systems“

Inhalte und Lernziele dieses Kapitels

- Quasi-Identifier
- Anonymitätsmaße
  - k-Anonymity
  - l-Diversity
  - t-Closeness
  - Differential Privacy
- Abschluss

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

Karlsruher Institut für Technologie

Quasi-Identifikatoren und Verknüpfung mit korrelierendem Wissen

IPD, Systeme der Informationsverwaltung, Nachwuchsgruppe „Privacy Awareness in Information Systems“

Motivation

- Daten mit Personenbezug müssen oft Dritten zugänglich gemacht werden
  - 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

- BDSG §3(6)
  - **Anonymisieren** ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse *nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand* an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.

- Wann ist ein Datensatz anonym?

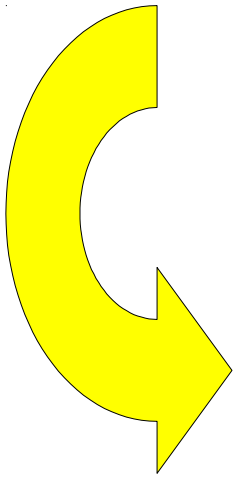
Wie kann man Anonymität erreichen?

- Rauschen hinzufügen
  - Wert_neu := Wert + Zufallszahl
- Dummy-Datensätze einfügen
  - plausible, echt aussehende Daten künstlich generieren
- Unterdrücken von Informationen
  - Tupel oder Attribute löschen
- Daten vertauschen
  - z.B. Geburtsdaten in einer medizinischen Statistik tauschen
- Generalisierung der Daten
  - z.B. „Ford Fiesta“ → „Auto“
    oder 973 → [900;1000]

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



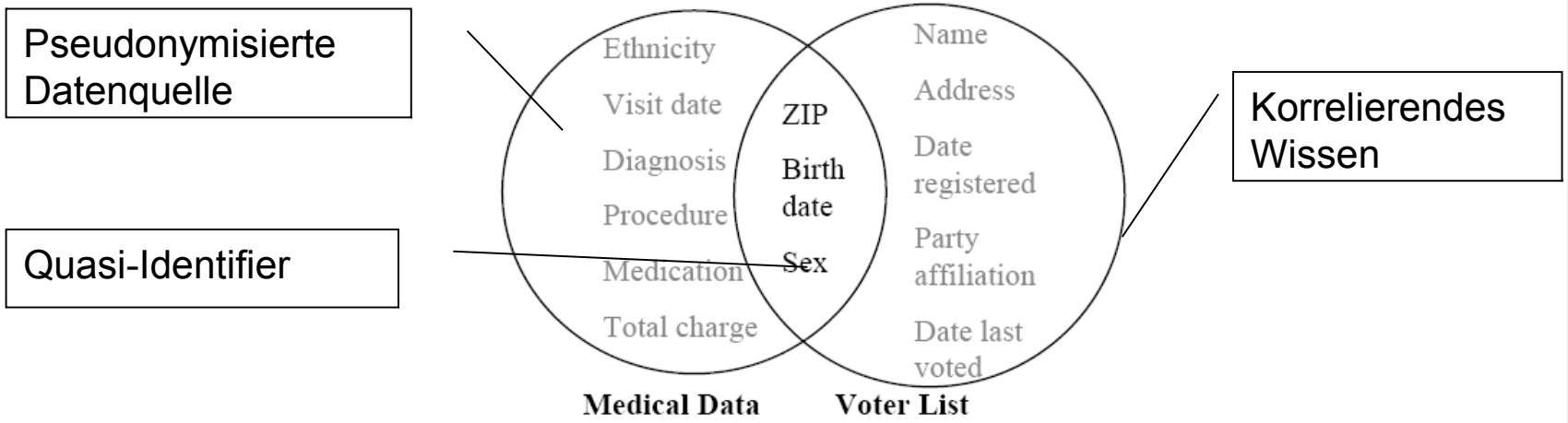
Attribut „Name“ gelöscht

Name	Geb.	Geschl.	PLZ	Krankheit
1	19.04.75	M	76227	Impotenz
2	05.07.75	M	76228	Hodenkrebs
3	17.01.75	M	76227	Sterilität
4	23.04.81	M	76139	Schizophrenie
5	30.12.81	M	76133	Diabetes
6	05.07.83	W	76133	Magersucht
7	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

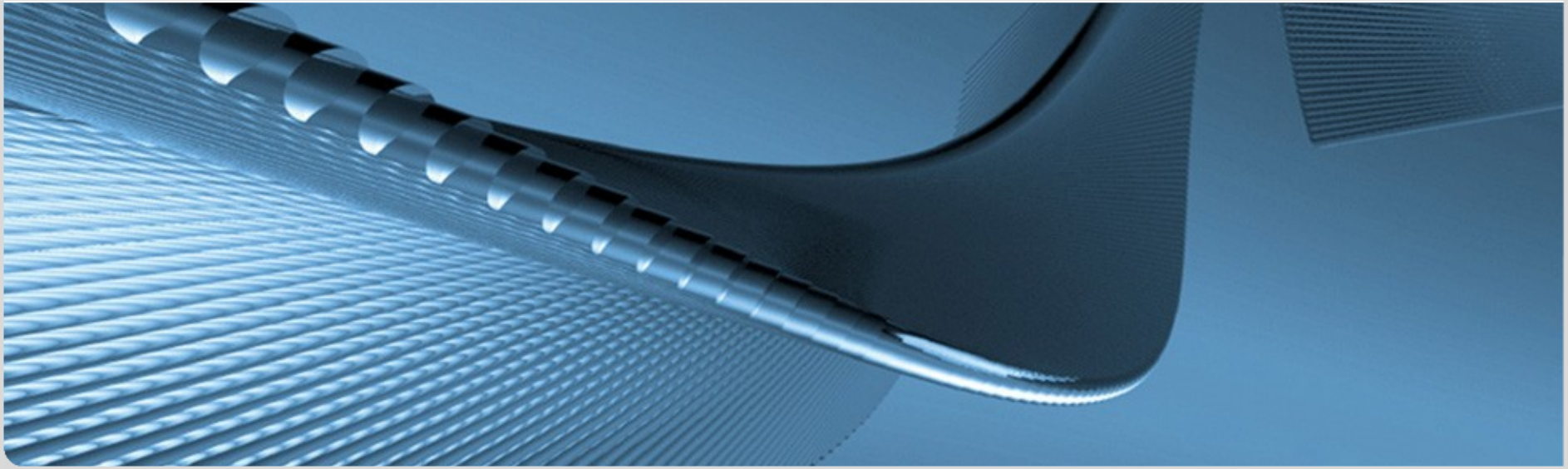
Definition Quasi-Identifier

- 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: $\exists p_i \in U: f_g(f_c(p_i)[Q_T]) = p_i$

Karlsruher Institut für Technologie

k-Anonymität

IPD, Systeme der Informationsverwaltung, Nachwuchsgruppe „Privacy Awareness in Information Systems“



Idee der k-Anonymität

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

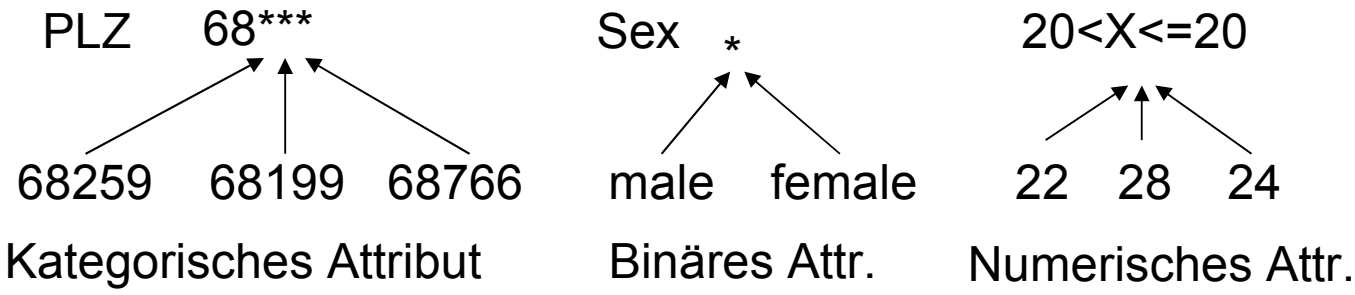
- k Datensätze formen eine Äquivalenzklasse
  - k-Anonymität schützt mit einer Konfidenz von 1/k vor einer ‚korrekten' Verknüpfung einer Person mit ihren sensitiven Attributen

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.

k-Anonymität durch Generalisierung



Beispiel einer generalisierten Tabelle für k=2

Name	Geb.	Sex	PLZ	Krankheit
1	**.**.75	M	7622*	Impotenz
2	**.**.75	M	7622*	Hodenkrebs
3	**.**.75	M	7622*	Sterilität
4	**.**.81	M	7613*	Schizophrenie
5	**.**.81	M	7613*	Diabetes
6	**.**.83	W	7613*	Magersucht
7	**.**.83	W	7613*	Magersucht

Anonym?

Problem: Homogenitätsangriff

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

Name	Geb.	Sex	PLZ	Krankheit
1	**.**.75	M	7622*	Impotenz
2	**.**.75	M	7622*	Hodenkrebs
3	**.**.75	M	7622*	Sterilität
4	**.**.81	M	7613*	Schizophrenie
5	**.**.81	M	7613*	Diabetes
6	**.**.83	W	7613*	Magersucht
7	**.**.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.

Name	Geb.	Sex	PLZ	Krankheit
1	**.**.75	M	7622*	Impotenz
2	**.**.75	M	7622*	Hodenkrebs
3	**.**.75	M	7622*	Sterilität
4	**.**.81	M	7613*	Schizophrenie
5	**.**.81	M	7613*	Diabetes
6	**.**.83	W	7613*	Magersucht
7	**.**.83	W	7613*	Magersucht

Beispiel einer generalisierten Tabelle für k=2

Weitere Herausforderungen (1/2)

- Sortierungsbasierte Angriffe (Unsorted Matching Attack)
  - Werden die generalisierten Tabellen GT1 und GT2 in gleicher Sortierung preisgegeben, kann der originale Datenbestand (PT) wieder hergestellt 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

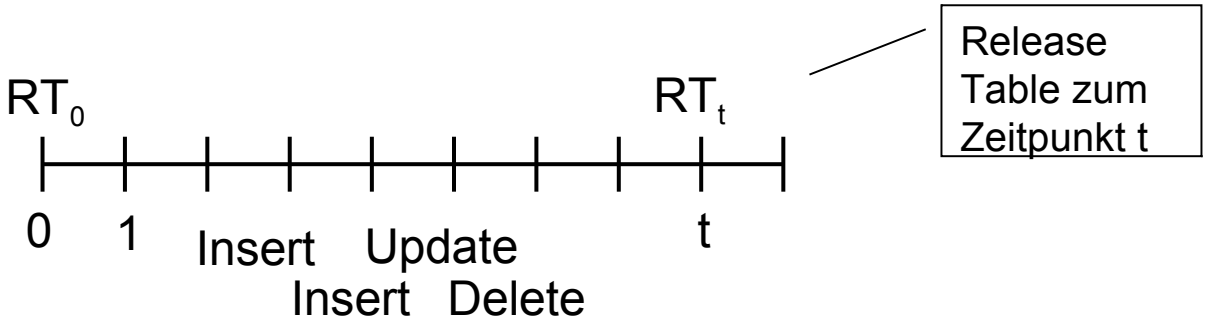
Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

→ *nicht spezifisch für die k-Anonymität!*

Weitere Herausforderungen (2/2)

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



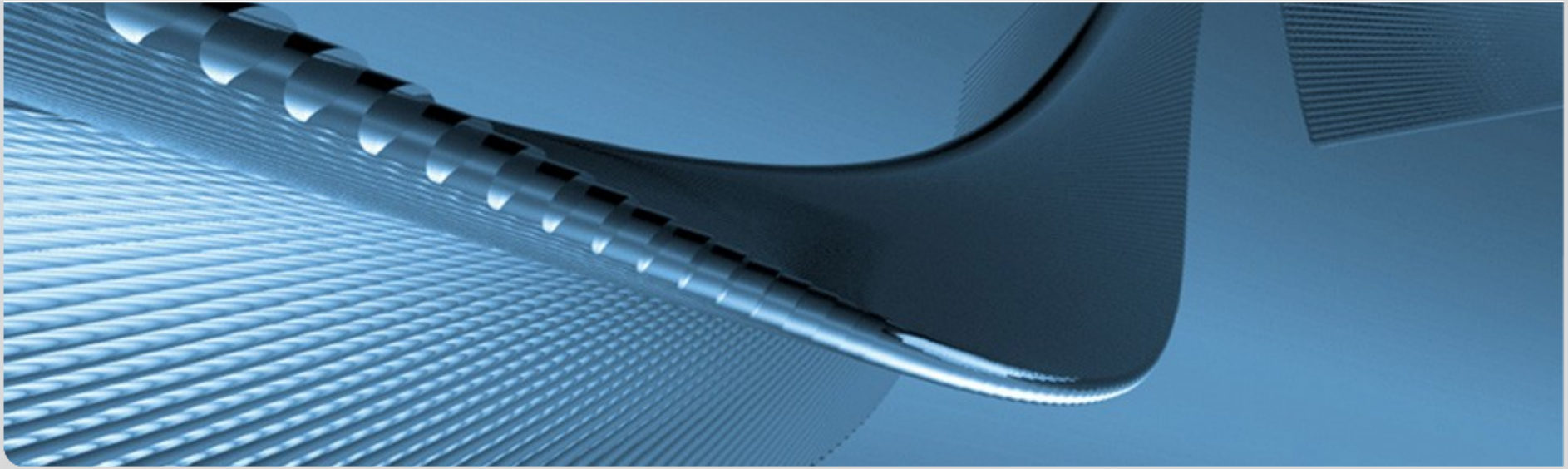
  – Möglicher Angriff
    - $RT_1 \bowtie RT_t$
    - $RT_1 \cap RT_t$
    - $RT_1 \setminus RT_t$

→ *nicht spezifisch für die k-Anonymität!*

Karlsruher Institut für Technologie

l-Diversity

IPD, Systeme der Informationsverwaltung, Nachwuchsgruppe „Privacy Awareness in Information Systems“



Idee der l-Diversity (*)

- Hintergrundwissen wird explizit modelliert
  - Wahrscheinlichkeitsfunktion über die Attribute
- Statistische Methoden zur Untersuchung auf Privatheit
  - Bayesian Inference
- Prinzip: Die Differenz aus Vorwissen (prior belief) und Wissen nach der Publikation eines Datensatzes (posterior belief) soll möglichst gering sein.

(*) vorgestellt als Bayes-Optimal Privacy

Hintergrundwissen

- Positive Preisgabe
  - Wert eines sensiblen Attributes mit hoher Wahrscheinlichkeit bestimmbar
    → Wer 83 geboren wurde hat Magersucht
- Negative Preisgabe
  - Wert eines sensiblen Attributes mit hoher Wahrscheinlichkeit ausgeschlossen
    → Wer 81 geboren wurde hat keine Magersucht, Impotenz, etc.

Name	Geb.	Sex	PLZ	Krankheit	
1	**.**.75	M	7622*	Impotenz	
2	**.**.75	M	7622*	Hodenkrebs	
3	**.**.75	M	7622*	Sterilität	
4	**.**.81	M	7613*	Schizophrenie	2
5	**.**.81	M	7613*	Diabetes	
6	**.**.83	W	7613*	Magersucht	1
7	**.**.83	W	7613*	Magersucht	

l-Diversity

- Problem von Bayes-Optimal Privacy
  - Nicht für jedes Attribut muss die Verteilung bekannt sein
  - Wissen des Angreifers ist unbekannt
  - Nicht jedes Wissen ist probabilistisch modellierbar
  - Zusammenschluss von Angreifern würde Modellierung jeder Kombination von Wissen erforderlich machen.

- Pragmatischerer Ansatz: l-Diversity
  - Basiert auf der vorgestellten Idee von Bayes-Optimal Privacy, umgeht jedoch die aufgezeigten Probleme

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“

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:

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

- Kurzfassung:
  - jeder q* Block besitzt mindestens l unterschiedliche sensiblen Werte
  - l ist minimales Maß der Unordnung in den Blöcken

Beispiel Entropy-l-Diversity

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

Name	Geb.	Sex	PLZ	Krankheit
1	**.**.75	M	7622*	Impotenz
2	**.**.75	M	7622*	Hodenkrebs
3	**.**.75	M	7622*	Sterilität
4	**.**.81	M	7613*	Schizophrenie
5	**.**.81	M	7613*	Diabetes
6	**.**.83	W	7613*	Magersucht
7	**.**.83	W	7613*	Magersucht

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

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

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

$$\log(2.8) = 0.44$$

Name	Geb.	Sex	PLZ	Krankheit
1	**.**.75	M	7622*	Impotenz
2	**.**.75	M	7622*	Hodenkrebs
3	**.**.75	M	7622*	Sterilität
4	**.**.8*	*	7613*	Schizophrenie
5	**.**.8*	*	7613*	Diabetes
6	**.**.8*	*	7613*	Magersucht
7	**.**.8*	*	7613*	Magersucht

Probleme von Entropy-l-Diversity

- 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
- Es ist schwierig, eine Tabelle zu erstellen, die den Eigenschaften von Entropy-l-Diversity genügt.

Recursive (c,l)-Diversity

- Ziel: häufige Werte sind nicht über-, seltene nicht unterrepräsentiert
- Formale Definition Recursive (c,l)-Diversity:
  - r_i ist die Häufigkeit, die der i-häufigste sensible Wert in einem q*-Block aufweist
  - gegeben eine Konstante c ist q*-Block (c,l)-Divers, wenn $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$
  - T* ist (c,l)-divers, wenn jeder q*-Block recursive-(c,l)-divers ist
  - Daraus folgt, dass nach der Eliminierung eines sensiblen Wertes der q*-Block immernoch (c,l-1) divers ist.

Weitere Konzepte für l-Diversity

- Positive Disclosure-Recursive (c.l)-Diversity
  - Erlaubt positive Preisgabe bestimmter weniger, z.B. nicht sensibler Attribute („gesund“)
- Negative/Positive Disclosure-Recursive(c_1,c_2,l)-Diversity
  - Schutz vor negativer Preisgabe von Attributen
  - Erfüllt Positive Disclosure-Recursive (c,l)-Diversity
  - Vor negativer Preisgabe zu schützende Attribute müssen in mindestens c_2% aller Tuper eines q*-Blocks vorkommen.
- Multi-Attribute l-Diveristy
  - für mehrere sensible Attribute
  - Ausschluss eines sensiblen Attributes soll nicht zur Preisgabe der anderen sensiblen Attribute führen.
  - Bsp: {(q*, s_1, v_1), (q*, s_1, v_2),(q*, s_2, v_3), (q*, s_3,v_3)}
    Gilt für eine Person nicht s_1 so gilt v_3

Probleme von l-Diversity

- Schwierig zu erreichen und unter Umständen unnötig
- 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 nur 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

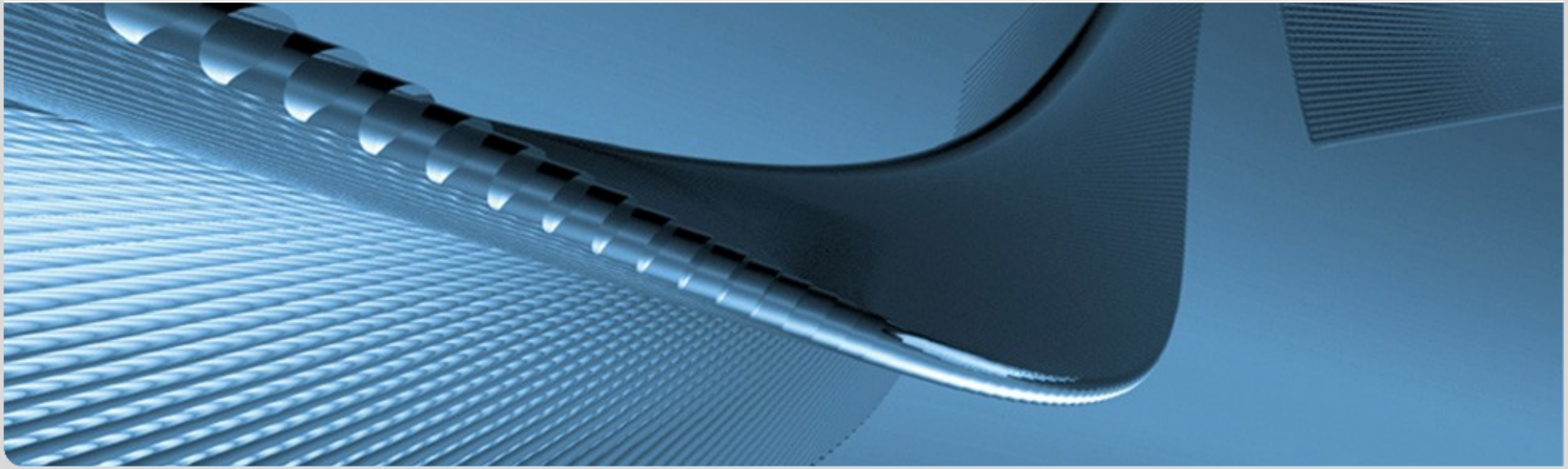
Name	Geb.	Sex	PLZ	Krankheit
1	**.**.75	M	7622*	Impotenz
2	**.**.75	M	7622*	Hodenkrebs
3	**.**.75	M	7622*	Sterilität
4	**.**.8*	*	7613*	Schizophrenie
5	**.**.8*	*	7613*	Diabetes
6	**.**.8*	*	7613*	Magersucht
7	**.**.8*	*	7613*	Magersucht

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

Karlsruher Institut für Technologie

t-Closeness

IPD, Systeme der Informationsverwaltung, Nachwuchsgruppe „Privacy Awareness in Information Systems“



t-Closeness

Wissen eines
potentiellen Angreifers

1) Initial

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

Belief	Wissen
B_0	Korrelierendes Wissen

t-Closeness

Wissen eines potentiellen Angreifers

1) Initial
2) Ohne Bezug auf die Person

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

t-Closeness

Wissen eines potentiellen Angreifers

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

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

t-Closeness

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 den Abstand?

Abstandsmaße für t-Closeness

- Gegeben
  - Verteilung P = {p_1, p_2, …, p_m}
  - Verteilung Q = {q_1, q_2, …, q_m}
- Maße
  - Variational Distance: $$D[P,Q] = \frac{1}{2}\sum_{i=1}^{m} | p_i - q_i |$$
  - Kullback-Leibler Distanz: $$D[P,Q] = \sum_{i=1}^{m} p_i \log(\frac{p_i}{q_j})$$

Sind das die richtigen Maße?

Problem von Variational und KL Distanz

- Gegeben
  - Q={3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}
  - zwei Einkommensverteilungen
    P_1={3k, 4k, 5k} und P_2={6k, 8k, 11k}
- Intuitiv hätten wir gerne
  - $D[P_1,Q] > D[P_2,Q]$, da in P_1 alle Elemente am unteren Ende sind
    → Mehr Information wird preisgegeben
- Die beiden Maße liefern das nicht, da alle Werte in P_1 und P_2 unterschiedlich sind und kein semantischer Bezug hergestellt wird.

Lösungsansatz: Earth Mover‘s Distance

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

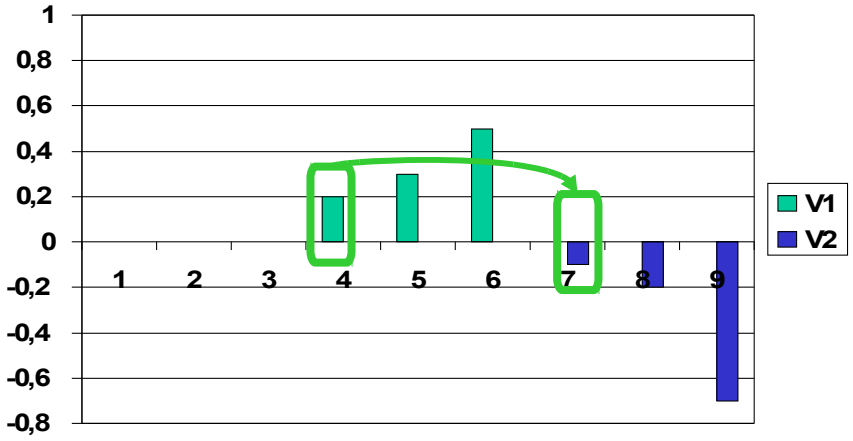
- Earth Mover‘s Distanz
  - $D[P,Q] = WORK(P,Q,F) = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij}$

- Idee (1D Fall)
  - Gegeben zwei Verteilungen V1 und V2
  - Fülle die nächstgelegenen Löcher

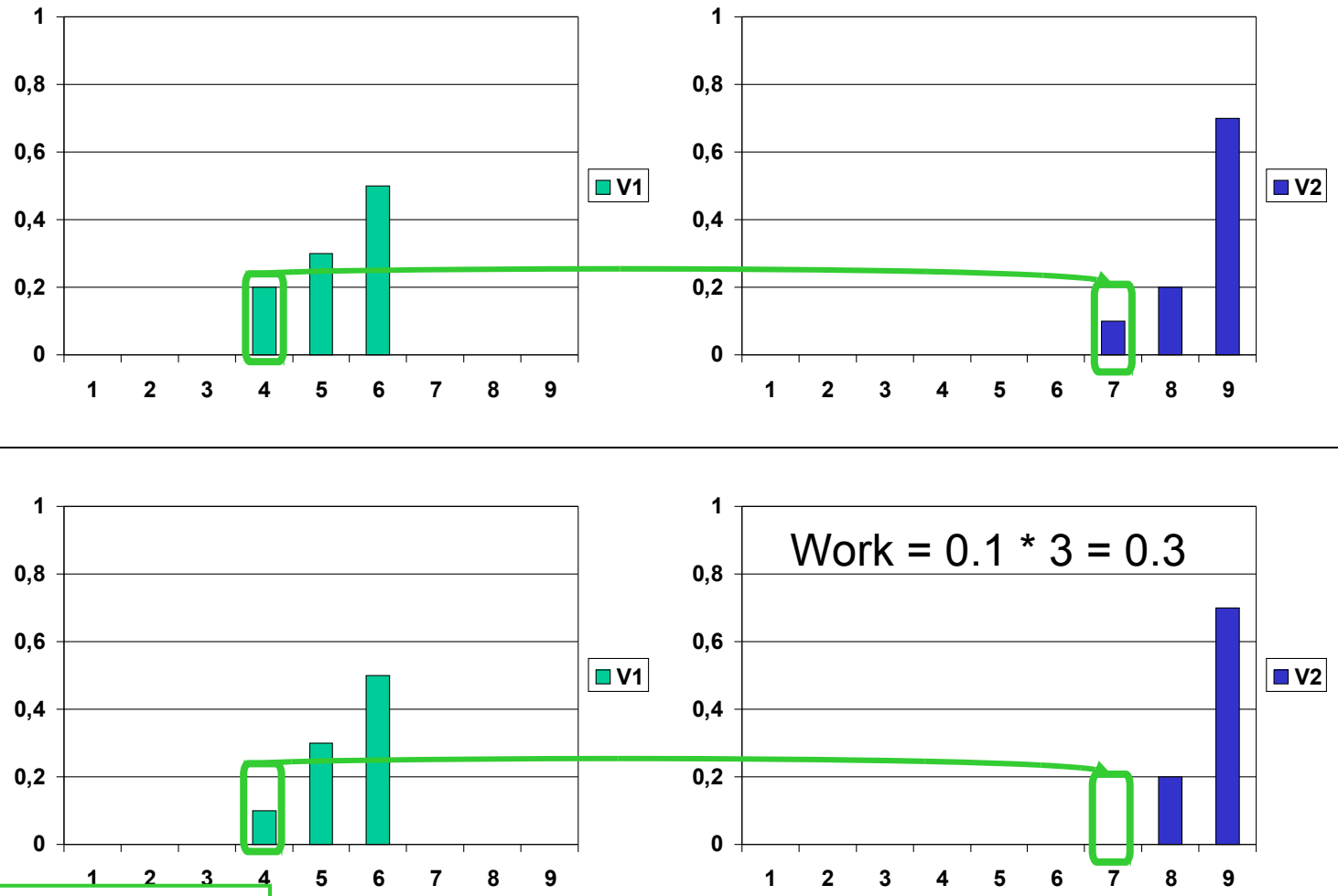
Work(f,i,j) = f * | i - j |;

f = Anzahl Werte

i,j die Werte

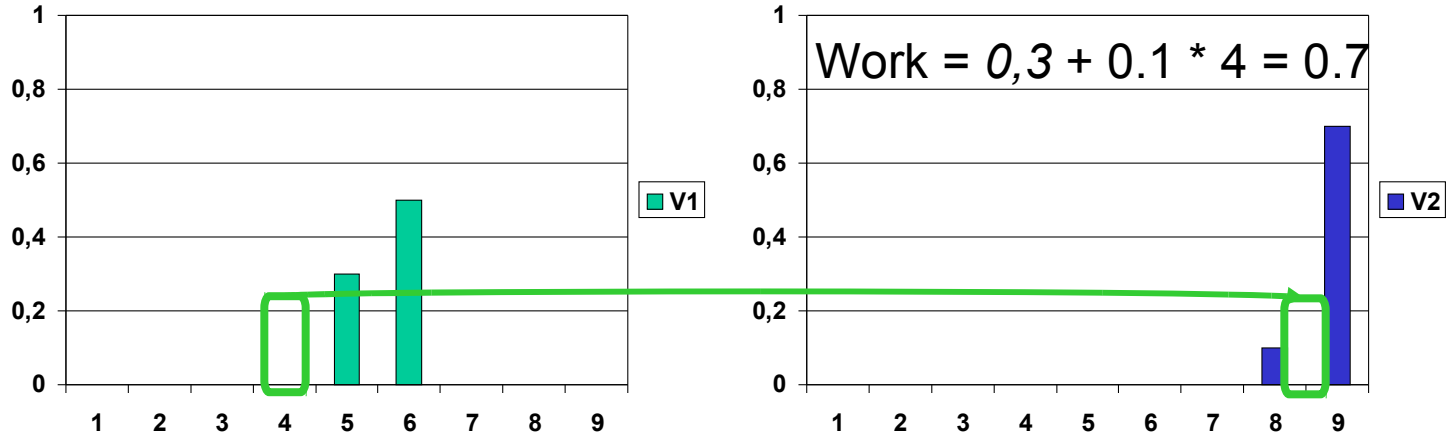


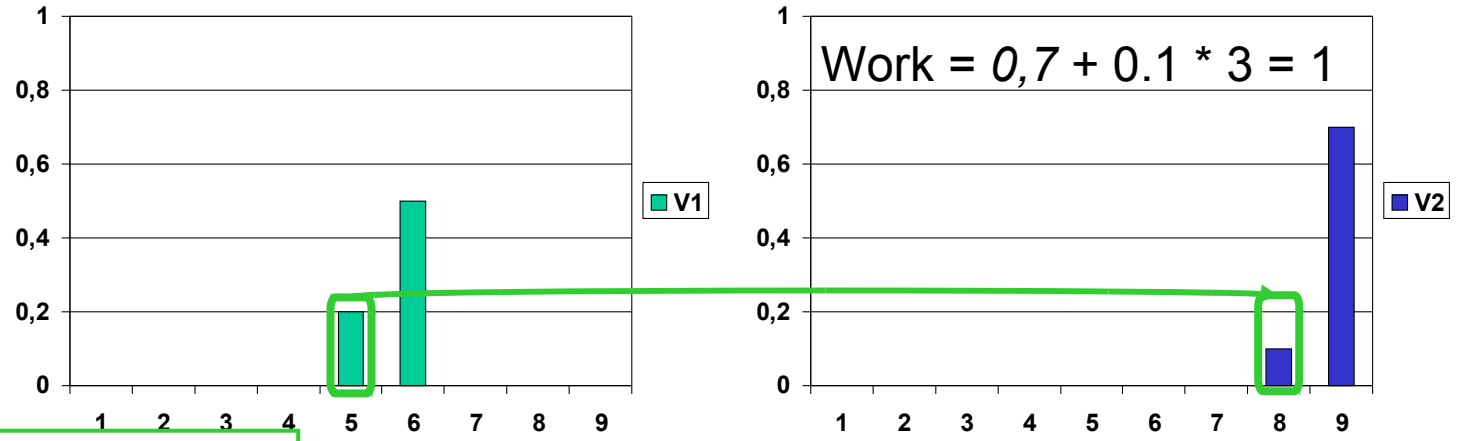
Beispiel Earth Mover‘s Distance



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

Beispiel Earth Mover‘s Distance





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

… 1 + 0,2 * 4 + 0,5 * 3 = 3,3

t-Closeness

- Verwendung der Earth Mover‘s Distanz

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

- basierend auf den Bedingungen

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m,\ 1 \leq j \leq m \qquad (c_1)$$

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i \qquad 1 \leq i \leq m \qquad (c_2)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{j=1}^{m} q_j = 1 \qquad (c_3)$$

t-Closeness

- Sind die Distanzen d_{ij} normalisiert ($0 \leq d_{ij} \leq 1$),
  dann ist auch $0 \leq D[P,Q] \leq 1$ unter den Bedingungen c_1, c_3
  D.h., die Schranke t kann zwischen 0 und 1 gewählt werden.

- Generalisierungseigenschaft
  - Gegeben die Tabelle T und zwei Generalisierungen A und B, mit A stärker generalisiert als B. Erfüllt T t-closeness bzgl. A, dann auch bzgl. B

- Teilmengeneigenschaft
  - Gegeben eine Tabelle T und eine Menge von Attributen C aus T. Erfüllt T t-closeness bzgl. C, dann erfüllt es dies auch bzgl. jede Teilmenge D.

t-Closeness

- EMD für numerische Attribute
  - Sortierdistanz

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

- EMD für kategorische Attribute
  - Äquivalenzdistanz

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

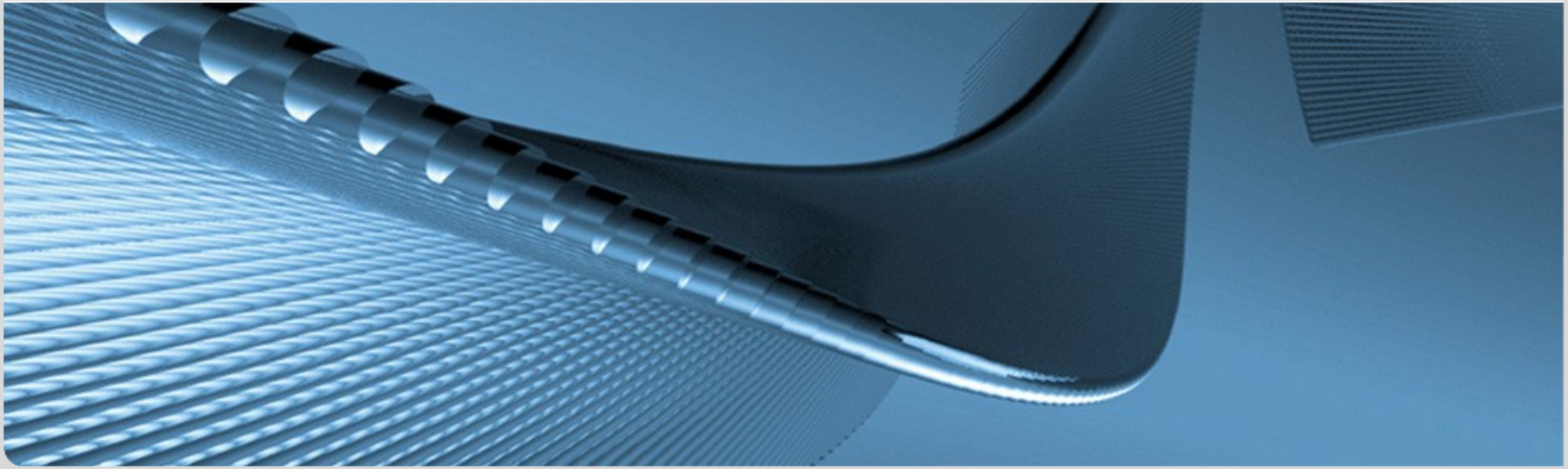
  - Hierarchische Distanz

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

Karlsruher Institut für Technologie

Differential Privacy

IPD, Systeme der Informationsverwaltung, Nachwuchsgruppe „Privacy Awareness in Information Systems“



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

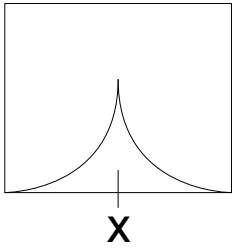
- Es gilt:
  - Verteilung aller Attribute ist bekannt, d.h.
    Pr[E]: Eintrittswahrscheinlichkeit für Ereignis E ist bekannt
  - X: Datensatz einer Person
  - Zwei Datenbestände DB_1 und DB_2: $DB_2 = DB_1 \cup \{ X \}$

- 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)



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

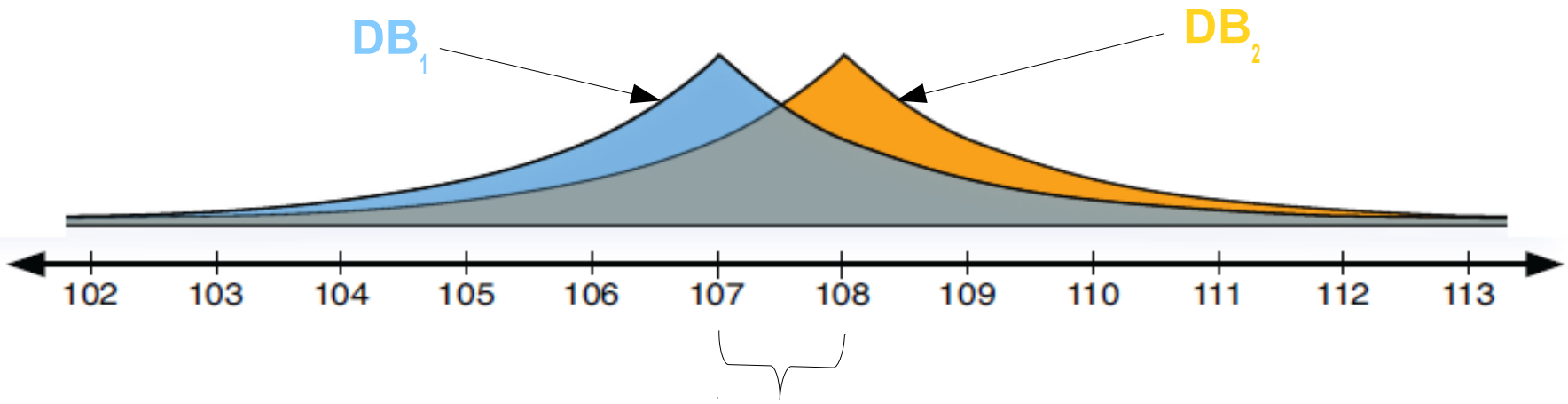
  - Definition von K (n Tupel genügen P):

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

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

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

- Formal nachweisbare Privatheitsgarantie für statistische Datenbanken
  - Stellt sicher und quantifiziert wie „groß“ das Risiko eines einzelnen bei der Veröffentlichung der Daten in einer statistischen Datenbank ist
  - Updates der Datenbank sind kein Datenschutzproblem
    - jedenfalls solange sich Verteilung der Attribute nicht ändert

- Nachteile
  - für komplexere Anfragen ist Nachweis schwierig
    - wie ist die Verteilung der Attribute im Anfrageergebnis über alle theoretisch möglichen Ergebnisse?
  - Grad der Verfälschung der Datenbank hängt von Gruppengröße c (in Form von $e^{(c\,\epsilon)}$) und Komplexität der Anfrage ab
  - anfragebezogen, keine Veröffentlichung einer generalisierten Tabelle
  - „Verteilung der Welt“ Pr(E) muss bekannt sein und darf sich nicht ändern

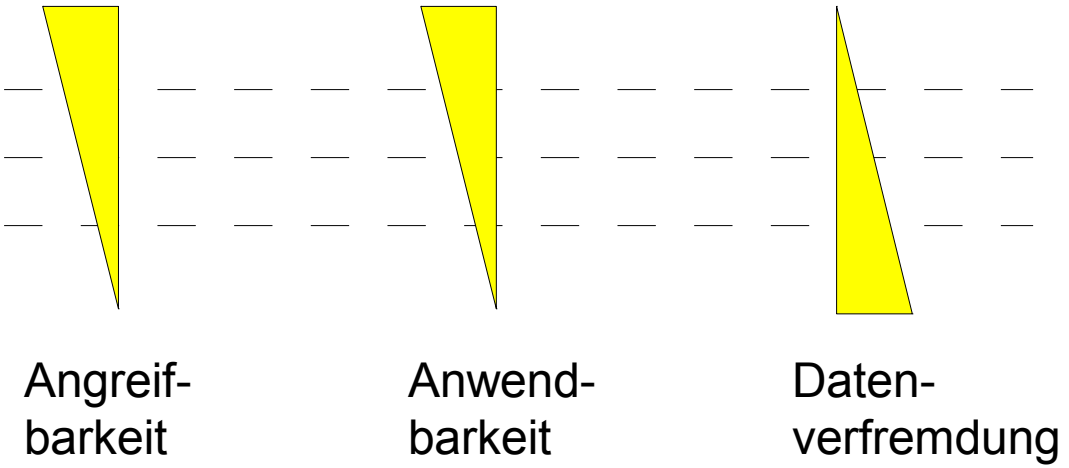
Karlsruher Institut für Technologie

Abschluss*

* Vielen Dank an Stephan Kessler und Dietmar Hauf für ihre Aufarbeitung dieses Themas

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