



# Verkettung von Datenbankeinträgen

## k-Anonymität und darüber hinaus

Thomas Maier, Kai Sonnenwald, Tom Petersen

Universität Hamburg  
Fachbereich Informatik



Universität Hamburg

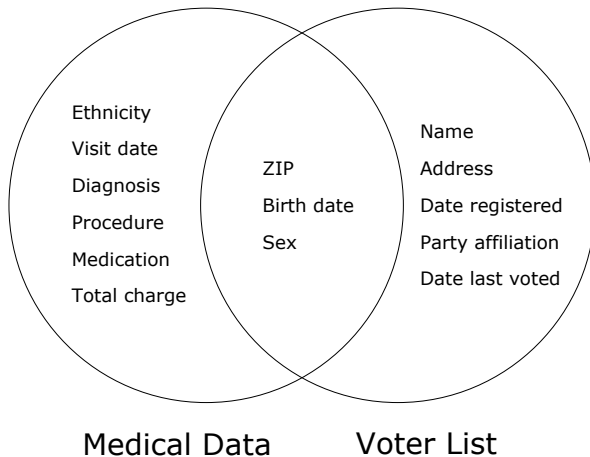
DER FORSCHUNG | DER LEHRE | DER BILDUNG

# Agenda

---

1. Motivation
2. k-Anonymität
  - Generalisierung
  - Suppression
  - Schwächen der k-Anonymität
3. l-Diversität
  - Verbesserung zu k-Anonymität
  - Schwächen der l-Diversität
4. t-Closeness
  - Earth Movers Distanz (EMD)
  - EMD Formeln
  - EMD Beispiel
5. Fazit
6. Literaturverzeichnis

# Anonym?



Massachusetts Group Insurance Commission (GIC) medical data and voter registration data. Entnommen aus [Swe02].

## Anonym? II

Sweeney [Swe00](1990) und Golle [Gol06](2000) überprüften die Eindeutigkeit von demographischen Faktoren in der Bevölkerung der USA.

	<b>T. M. J.</b>	<b>M. J.</b>	<b>J.</b>	<b>2 J.</b>
<b>PLZ</b>	87.1 %	3.7 %	0.04 %	0.01 %
<b>Ort</b>	58.4 %	3.6 %	0.04 %	0.01 %
<b>County</b>	18.1 %	0.04 %	0.00004 %	0.00000 %

Eindeutig identifizierbarer Individuenanteil an der U.S.-Bevölkerung 1990. Entnommen aus [Swe00].

**Ergebnis:** Durch {Geburtsdatum, Geschlecht, PLZ} könnten ~87% der Bevölkerung eindeutig identifiziert werden.

## Abgrenzung

---

Vermeintlich anonyme Daten stellen sich als nicht anonym heraus.

**Daher:** Wie können wir Aussagen über die “Güte” der Anonymisierung machen?

## Abgrenzung

---

Vermeintlich anonyme Daten stellen sich als nicht anonym heraus.

**Daher:** Wie können wir Aussagen über die “Güte” der Anonymisierung machen?

Worum es nicht gehen soll:

- Begrenzung des Zugriffs (Authentifikation, Multi-Level-Datenbanken)
- Statistische Datenbanken (Aggregation, Begrenzung von Selektionsarten, Logging und Abwägen von Anfragen, Hinzufügen von Zufall)

Darum geht es:

- Veröffentlichung von Daten als Individualdatensätze ohne Integritätsverlust unter Wahrung der Anonymität.

## Beispiel: Private Tabelle

<b>Identifikator</b>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<b>Name</b>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geb.dat.</b>	<b>Erkrankung</b>
Sofia Müller	w	22981	22.12.1944	Hepatitis
Emma Weber	w	22362	27.3.1945	Gicht
Sofia Koch	w	22669	3.9.1949	Arthrose
Emilia Wagner	w	22862	1.3.1985	Diabetes
Emma Meyer	w	22875	16.2.1992	Demenz
Noah Meyer	m	22997	19.3.1936	Arthrose
Elias Schäfer	m	22121	26.11.1949	Diabetes
Finn Fischer	m	22350	28.11.1963	Demenz
Leon Schmidt	m	22188	26.4.1964	Demenz
Elias Koch	m	22997	7.10.1975	Hepatitis

# Begriffe

---

**Expliziter Identifikator** Attribut, das ein Individuum (nahezu) eindeutig identifiziert. Beispiele: Name, Adresse, Steuernummer, ...

**Sensibles Attribut** Attribut, dessen Wert für ein Individuum in einer Datenmenge nicht öffentlich gemacht werden soll.

**Quasi-Identifikator** Eine Menge nicht-sensibler Attribute  $\{A_i, \dots, A_j\}$  einer Tabelle, deren Attribute mit einer externen Datenquelle verknüpft werden können, um mindestens ein Individuum der Gesamtmenge eindeutig zu identifizieren.



## k-Anonymität

Eine Tabelle erfüllt  $k$ -Anonymität, wenn jede Zeile ununterscheidbar von mindestens  $k - 1$  anderen Zeilen im Bezug auf einen Quasi-Identifikator ist.

### k-Anonymität

Sei  $T(A_1, \dots, A_n)$  eine Tabelle und  $Q_T = \{A_i, \dots, A_j\}$  der zugehörige Quasi-Identifikator.

$T$  erfüllt  **$k$ -Anonymität** genau dann, wenn jede Belegung von Werten in  $T[Q_T]$  mindestens  $k$  mal auftritt, wobei  $T[Q_T]$  die duplikatenerhaltende Projektion von  $T$  auf die Attribute des Quasi-Identifikators beschreibt.

Die so entstandenen Äquivalenzklassen werden auch als  $q^*$ -**Blöcke** bezeichnet.

## Beispiel: k-anonyme Tabelle

<b>Identifikator</b>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<b>Name</b>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geb.dat.</b>	<b>Erkrankung</b>
Sofia Müller	w	22981	22.12.1944	Hepatitis
Emma Weber	w	22362	27.3.1945	Gicht
Sofia Koch	w	22669	3.9.1949	Arthrose
Emilia Wagner	w	22862	1.3.1985	Diabetes
Emma Meyer	w	22875	16.2.1992	Demenz
Noah Meyer	m	22997	19.3.1936	Arthrose
Elias Schäfer	m	22121	26.11.1949	Diabetes
Finn Fischer	m	22350	28.11.1963	Demenz
Leon Schmidt	m	22188	26.4.1964	Demenz
Elias Koch	m	22997	7.10.1975	Hepatitis

## Beispiel: k-anonyme Tabelle

<i>Identifikator</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geb.dat.</b>	<b>Erkrankung</b>
-	w	22981	22.12.1944	Hepatitis
-	w	22362	27.3.1945	Gicht
-	w	22669	3.9.1949	Arthrose
-	w	22862	1.3.1985	Diabetes
-	w	22875	16.2.1992	Demenz
-	m	22997	19.3.1936	Arthrose
-	m	22121	26.11.1949	Diabetes
-	m	22350	28.11.1963	Demenz
-	m	22188	26.4.1964	Demenz
-	m	22997	7.10.1975	Hepatitis

## Beispiel: k-anonyme Tabelle

<i>Identifikator</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22981	1944	Hepatitis
-	w	22362	1945	Gicht
-	w	22669	1949	Arthrose
-	w	22862	1985	Diabetes
-	w	22875	1992	Demenz
-	m	22997	1936	Arthrose
-	m	22121	1949	Diabetes
-	m	22350	1963	Demenz
-	m	22188	1964	Demenz
-	m	22997	1975	Hepatitis

## Beispiel: k-anonyme Tabelle

<i>Identifikator</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22***	1944-45	Hepatitis
-	w	22***	1944-45	Gicht
-	w	22669	1949	Arthrose
-	w	22862	1985	Diabetes
-	w	22875	1992	Demenz
-	m	22997	1936	Arthrose
-	m	22121	1949	Diabetes
-	m	22350	1963	Demenz
-	m	22188	1964	Demenz
-	m	22997	1975	Hepatitis

## Beispiel: k-anonyme Tabelle

<i>Identifikator</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22***	1944-45	Hepatitis
-	w	22***	1944-45	Gicht
-	*	22***	1949	Arthrose
-	w	22862	1985	Diabetes
-	w	22875	1992	Demenz
-	m	22997	1936	Arthrose
-	*	22***	1949	Diabetes
-	m	22350	1963	Demenz
-	m	22188	1964	Demenz
-	m	22997	1975	Hepatitis

## Beispiel: k-anonyme Tabelle

<i>Identifikator</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22***	1944-45	Hepatitis
-	w	22***	1944-45	Gicht
-	*	22***	1949	Arthrose
-	w	228**	1985-92	Diabetes
-	w	228**	1985-92	Demenz
-	m	22997	1936	Arthrose
-	*	22***	1949	Diabetes
-	m	22350	1963	Demenz
-	m	22188	1964	Demenz
-	m	22997	1975	Hepatitis

## Beispiel: k-anonyme Tabelle

<i>Identifikator</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22***	1944-45	Hepatitis
-	w	22***	1944-45	Gicht
-	*	22***	1949	Arthrose
-	w	228**	1985-92	Diabetes
-	w	228**	1985-92	Demenz
-	m	22997	1936	Arthrose
-	*	22***	1949	Diabetes
-	m	22***	1963-64	Demenz
-	m	22***	1964-64	Demenz
-	m	22997	1975	Hepatitis



## Beispiel: k-anonyme Tabelle

<i>Identifikator</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22***	1944-45	Hepatitis
-	w	22***	1944-45	Gicht
-	*	22***	1949	Arthrose
-	w	228**	1985-92	Diabetes
-	w	228**	1985-92	Demenz
-	m	22997	1936-75	Arthrose
-	*	22***	1949	Diabetes
-	m	22***	1963-64	Demenz
-	m	22***	1963-64	Demenz
-	m	22997	1936-75	Hepatitis

## Beispiel: k-anonyme Tabelle

<i>Identifikator</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22***	1944-45	Hepatitis
-	w	22***	1944-45	Gicht
-	*	22***	1949	Arthrose
-	w	228**	1985-92	Diabetes
-	w	228**	1985-92	Demenz
-	m	22997	1936-75	Arthrose
-	*	22***	1949	Diabetes
-	m	22***	1963-64	Demenz
-	m	22***	1963-64	Demenz
-	m	22997	1936-75	Hepatitis

**Ergebnis:** k-anonyme Tabelle mit  $k = 2$

# Generalisierung

## Generalisierung

Vergrößerung der Werte, die ein Attribut annehmen kann  
(Generalisierung auf Attributebene).

Beispiele für Generalisierungshierarchien:

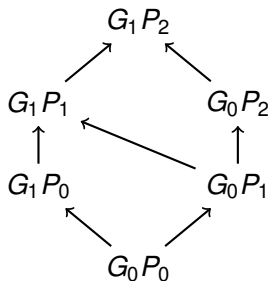
### 1. PLZ:

$$\begin{aligned}
 P_0 &= \{22765, 22769, 22529, 20246\} \text{ Grundwertebereich} \\
 \rightarrow P_1 &= \{2276^*, 2252^*, 2024^*\} \\
 \rightarrow P_2 &= \{2^{*****}\}
 \end{aligned}$$

### 2. Geschlecht:

$$\begin{aligned}
 G_0 &= \{\text{männlich, weiblich}\} \text{ Grundwertebereich} \\
 \rightarrow G_1 &= \{\text{nicht\_veröffentlicht}\}
 \end{aligned}$$

## Generalisierung II



### Generalisierungshierarchie für Attributmenge

Jeder Pfad von  $G_0 P_0$  zu  $G_1 P_2$  stellt  
einen möglichen Weg der  
Generalisierung dar.

PLZ:

$$P_0 = \{22765, 22769, 22529, 20246\}$$

$$\rightarrow P_1 = \{2276^*, 2252^*, 2024^*\}$$

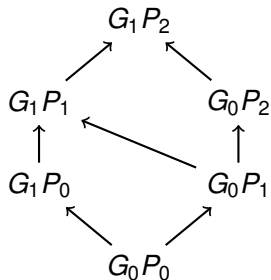
$$\rightarrow P_2 = \{2^{****}\}$$

Geschlecht:

$$G_0 = \{\text{männlich, weiblich}\}$$

$$\rightarrow G_1 = \{\text{nicht\_veröffentlicht}\}$$

## Generalisierung II



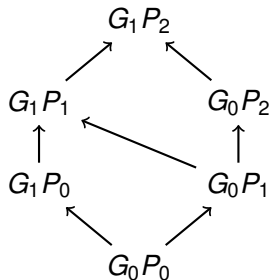
### Generalisierungshierarchie für Attributmenge

Jeder Pfad von  $G_0P_0$  zu  $G_1P_2$  stellt  
einen möglichen Weg der  
Generalisierung dar.

 $T_{G_0P_0}$ 

Geschlecht	PLZ
m	22765
m	22765
m	22769
m	22529
m	20246
w	22765
w	22765
w	22769
w	22529
w	22529
w	22529
w	20246

## Generalisierung II



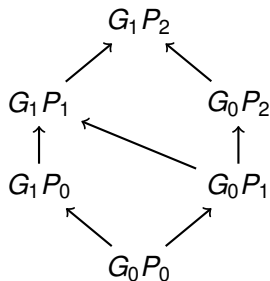
### Generalisierungshierarchie für Attributmenge

Jeder Pfad von  $G_0P_0$  zu  $G_1P_2$  stellt  
einen möglichen Weg der  
Generalisierung dar.

$T_{G_1P_0}$

Geschlecht	PLZ
*	22765
*	22765
*	22769
*	22529
*	20246
*	22765
*	22765
*	22769
*	22529
*	22529
*	22529
*	20246

# Generalisierung II



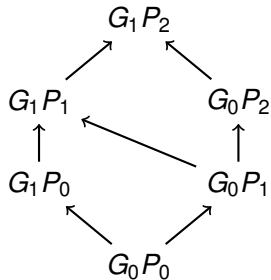
## Generalisierungshierarchie für Attributmenge

Jeder Pfad von  $G_0P_0$  zu  $G_1P_2$  stellt  
einen möglichen Weg der  
Generalisierung dar.

 $T_{G_0P_1}$ 

Geschlecht	PLZ
m	2276*
m	2276*
m	2276*
m	2252*
m	2024*
w	2276*
w	2276*
w	2276*
w	2252*
w	2252*
w	2252*
w	2024*

## Generalisierung II



### Generalisierungshierarchie für Attributmenge

Jeder Pfad von  $G_0P_0$  zu  $G_1P_2$  stellt  
einen möglichen Weg der  
Generalisierung dar.

$T_{G_1P_2}$

Geschlecht	PLZ
*	2****
*	2****
*	2****
*	2****
*	2****
*	2****
*	2****
*	2****
*	2****
*	2****
*	2****
*	2****



## Generalisierung III

---

**Aber:** Nicht jede Generalisierung ist gleichermaßen sinnvoll!

## Generalisierung III

**Aber:** Nicht jede Generalisierung ist gleichermaßen sinnvoll!

**$k$ -minimale Generalisierung.**

$T_i$  ist die  $k$ -minimale Generalisierung einer Tabelle  $T$  gdw.

- $T_i$   $k$ -Anonymität erfüllt und
- keine Tabelle  $T_j$  existiert, die ebenfalls  $k$ -Anonymität erfüllt und für die  $T_i$  eine Generalisierung darstellt.

# Unterdrückung

## Unterdrückung

Entfernen von Daten aus der Tabelle - hier auf Tupelebene, d.h. Tupel können nur komplett entfernt werden.

Unterdrückung ist jedoch auch auf Attributebene möglich (entspricht dann maximaler Generalisierung).

G.	PLZ
m	22765
w	22765
m	22769
w	22769
m	80043

Daten

G.	PLZ
m	*
w	*
m	*
w	*
m	*

Generalisierung

G.	PLZ
m	2276*
w	2276*
m	2276*
w	2276*

Unterdrückung  
& Generalisierung

## Implementierungen

---

Die Berechnung von optimalen  $k$ -anonymen Tabellen ist NP-schwer,  
...

# Implementierungen

---

Die Berechnung von optimalen  $k$ -anonymen Tabellen ist NP-schwer,

...

... es wurden jedoch  $\mathcal{O}(k)$ -Approximationsalgorithmen gefunden  
[AFK<sup>+</sup>05, MW04].

Implementationen

- $\mu$ -Argus
- Datafly
- Incognito
- Mondrian
- ...

## Schwächen der $k$ -Anonymität

---

- *Complementary release attack*: Veröffentlichung mehrerer  $k$ -anonymer Tabellen unterschiedlicher Generalisierung kann bei Kombination dieser Tabellen die  $k$ -Anonymität verletzen [Swe02].
- *Temporal attack*: Dynamische Tabellen können  $k$ -Anonymität verletzen [Swe02].
- **Unsorted matching attack** [Swe02]
- **Homogeneity attack** [MKG07]
- **Background knowledge attack** [MKG07]

## Unsorted matching attack

Veröffentlichung mehrerer  $k$ -anonymer Tabellen mit derselben Sortierung ausgehend von einer nicht-öffentlichen Tabelle identifiziert Individuen.

G.jahr	PLZ
1970-80	21985
1970-80	21986
1970-80	21724
1970-80	21725
1970-80	21985
1970-80	21986
1970-80	21724
1970-80	21725
1970-80	21985
1970-80	21986
1970-80	21724
1970-80	21725

$k = 3$

G.jahr	PLZ	Erkrankung
1970	2198*	Hepatitis X
1970	2198*	Hepatitis Y
1970	2172*	Hepatitis Z
1970	2172*	Hepatitis X
1975	2198*	Hepatitis Y
1975	2198*	Hepatitis Z
1975	2172*	Hepatitis X
1975	2172*	Hepatitis Y
1980	2198*	Hepatitis Z
1980	2198*	Hepatitis X
1980	2172*	Hepatitis Y
1980	2172*	Hepatitis Z

$k = 2$

## Homogeneity attack

Gleichheit des sensiblen Attributs einer Äquivalenzklasse verrät das sensible Attribut eines Individuums.

G.jahr	PLZ	Erkrankung
1970	21***	Hepatitis X
1970	21***	Hepatitis Y
1970	21***	Hepatitis Z
1970	21***	Hepatitis Y
1975	21***	Hepatitis X
1975	21***	Hepatitis X
1975	21***	Hepatitis X
1975	21***	Hepatitis X

$$k = 4$$



## Background knowledge attack

Nutzen von Hintergrundwissen, um auf den Wert des sensiblen Attributs eines Individuums in einer Gruppe zu schließen.

G.jahr	PLZ	Erkrankung
1970	21***	Hepatitis X
1970	21***	Hepatitis Y
1970	21***	Hepatitis Z
1970	21***	Hepatitis Y
1975	21***	Hepatitis X
1975	21***	Hepatitis X
1975	21***	Hepatitis Y
1975	21***	Hepatitis Y

$$k = 4$$

**Hintergrundwissen:** Hepatitis X tritt nur bzw. mit hoher Wahrscheinlichkeit lediglich bei Männern auf.

## I-Diversität - Prinzip

**Prinzip** Eine Tabelle erfüllt *I*-Diversität, wenn in jedem *k*-anonymen Block mindestens *I* verschiedene Werte für das sensible Attribut vorkommen.

Bsp.-Tabelle

G.jahr	PLZ	Erkrankung
1970	21***	Hepatitis X
1970	21***	Hepatitis Y
1970	21***	Hepatitis Z
1975	21***	Hepatitis X
1975	21***	Hepatitis X
1975	21***	Hepatitis X

$$k = 3, I = 1$$

## I-Diversität - Definitionen

### Entropie basierte I-Diversität [MKGV07]

Eine Tabelle ist ***l*-divers**, wenn für jeden  $q^*$ -Block die folgende Ungleichung erfüllt wird:

$$-\sum_{s \in S} p_{(q^*, s)} \log(p_{(q^*, s)}) \geq \log(l)$$

Dabei stellt  $p_{(q^*, s)}$  den Anteil des Werts  $s$  in dem  $q^*$ -Block dar.

### rekursive ( $c, l$ )-Diversität [MKGV07]

Innerhalb eines  $q^*$ -Blocks sei  $r_i$  die Anzahl des  $i$ -häufigsten sensiblen Attributs. Mit einer gegebenen Konstante  $c$  erfüllt dieser  $q^*$ -Block **rekursive ( $c, l$ )-Diversität**, wenn  $r_1 < c(r_l + r_{l+1} + \dots + r_m)$  gilt. Eine Tabelle  $T^*$  erfüllt ( $c, l$ )-Diversität, wenn jeder  $q^*$ -Block ( $c, l$ )-Diversität erfüllt. 1-Diversität ist immer erfüllt.

## Beispiel: 2-diverse Tabelle

<i>Identifizier</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22***	1944-45	Hepatitis
-	w	22***	1944-45	Gicht
-	*	22***	1949	Arthrose
-	w	228**	1985-92	Diabetes
-	w	228**	1985-92	Demenz
-	m	22997	1936-75	Arthrose
-	*	22***	1949	Diabetes
-	m	22***	1963-64	Demenz
-	m	22***	1963-64	Demenz
-	m	22997	1936-75	Hepatitis

k-anonyme Tabelle mit  $k = 2$ , aber nur l-divers mit  $l = 1$

## Beispiel: 2-diverse Tabelle

<i>Identifizier</i>	<b>Nicht-sensibel</b>			<b>Sensibel</b>
<i>Name</i>	<b>Geschl.</b>	<b>PLZ</b>	<b>Geburtsjahr</b>	<b>Erkrankung</b>
-	w	22***	1944-45	Hepatitis
-	w	22***	1944-45	Gicht
-	*	22***	1949	Arthrose
-	w	228**	1985-92	Diabetes
-	w	228**	1985-92	Demenz
-	m	22***	1936-75	Arthrose
-	*	22***	1949	Diabetes
-	m	22***	1936-75	Demenz
-	m	22***	1936-75	Demenz
-	m	22***	1936-75	Hepatitis

**Ergebnis:** k-anonyme Tabelle mit  $k = 2$  und l-divers mit  $l = 2$

## Verbesserung zu k-Anonymität

---

- $l$ -Diversität sichert verschiedene Ausprägungen der sensiblen Attribute in den  $q^*$ -Blöcken zu.
  - Die *Homogeneity Attack* ist nicht mehr möglich.
  - *Background Knowledge Attacks* werden erschwert.

## Verbesserung zu k-Anonymität

---

- *l*-Diversität sichert verschiedene Ausprägungen der sensiblen Attribute in den  $q^*$ -Blöcken zu.
  - Die *Homogeneity Attack* ist nicht mehr möglich.
  - *Background Knowledge Attacks* werden erschwert.
  
- Ein Vorteil von *l*-Diversity ist, dass vorhandene Algorithmen für *k*-Anonymität leicht angepasst werden können.

## Schwächen der l-Diversität

### Skewness attack [Li07]

- Tabelle mit einem sensiblen Attribut, 2 Ausprägungen.
- Wahrscheinlichkeit für Ausprägung 1 ist sehr hoch.
- Wahrscheinlichkeit für Ausprägung 2 entsprechend niedrig.
- Es liegt 2-diverse Tabelle mit Block  $q^*$  vor.
- $q^*$  beinhaltet zu 50% Ausprägung 1 und zu 50% Ausprägung 2.
- Die Wahrscheinlichkeit, dass ein Tupel aus  $q^*$  Ausprägung 2 besitzt, liegt nun bei 50%.

**Beispiel:** Angenommen das sensible Attribut hat die Werte: krank / gesund. In der Bevölkerung sind 1% krank und 99% gesund. Die Wahrscheinlichkeit, dass eine Person aus dem Block  $q^*$  krank ist liegt nun bei 50% und nicht mehr bei 1%.



## I-Diversität - Schwächen

### Similarity attack [Li07]

I-Diversität garantiert, dass in jedem Block unterschiedliche sensible Werte stehen. Es kann jedoch vorkommen, dass sich diese Werte ähneln.

G.jahr	PLZ	Erkrankung
1970	21***	Diabetes
1970	21***	Syphilis
1970	21***	Gicht
1975	21***	Tripper
1975	21***	Syphilis
1975	21***	Chlamydien

$$k = 3, l = 3$$

Kann man eine Person dem zweiten Block zuordnen, so weiß man auch, dass diese eine Geschlechtskrankheit hat.

## t-Closeness

t-closeness stellt ein Maß für minimalen Wissensgewinn, der durch Betrachtung eines  $q^*$ -Blocks im Vergleich zur gesamten Distribution entsteht, dar [Hau07].

### Definition t-closeness

Eine **Äquivalenzklasse** ( $q^*$ -Block) hat die Eigenschaft **t-closeness**, wenn die (semantische) Distanz zwischen der Verteilung der Werte eines sensiblen Attributes innerhalb der Äquivalenzklasse und der Verteilung der Werte des sensiblen Attributes innerhalb der Tabelle nicht größer als  $t$  ist.

Eine **Tabelle** hat die Eigenschaft **t-closeness**, wenn diese Eigenschaft für alle Äquivalenzklassen erfüllt ist.

**Problem:** Wie bestimmt man die (semantische) Distanz?

## Earth Movers Distanz (EMD)

Die Earth Movers Distanz (EMD) basiert auf der minimalen Arbeit, die zu verrichten ist, um eine Verteilung in eine andere zu überführen.

### Definition von EMD [Li07]

Gegeben sei  $P = (p_1, \dots, p_m)$ ,  $Q = (q_1, \dots, q_m)$ ,  $d_{ij}$  ist die Grunddistanz zwischen  $p_i$  und  $q_j$ .  $f_{ij}$  ist die minimale Masse, die transportiert werden muss, um  $p_i$  in  $q_j$  zu verwandeln. EMD ist dann die gesamte Arbeit, die verrichtet werden muss

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

unter den folgenden Bedingungen

- i)  $f_{ij} \geq 0 \mid 1 \leq i \leq m, 1 \leq j \leq m$
- ii)  $p_i - \sum_{j=1}^m f_{ij} + \sum_{j=1}^m f_{ji} = q_i \mid 1 \leq i \leq m$
- iii)  $\sum_{i=1}^m \sum_{j=1}^m f_{ij} = \sum_{i=1}^m p_i = \sum_{i=1}^m q_i$

## Earth Movers Distanz (EMD)

Aus den drei Bedingungen folgen die zwei Fakten [Li07]:

**Fakt 1:** If  $\forall i, j | 0 \leq d_{ij} \leq 1$  then  $0 \leq D[P, Q] \leq 1$ . Das bedeutet, dass wenn die Grunddistanz normalisiert ist, auch die EMD normalisiert ist. **Somit kann ein einheitliches Maß für t bestimmt werden.**

**Fakt 2:** Gegeben sind zwei Äquivalenzklassen  $E_1$  und  $E_2$ .  $P_1$  ist die Verteilung eines sensiblen Attributes aus  $E_1$ .  $P_2$  ist die Verteilung eines sensiblen Attributes aus  $E_2$ .  $P$  ist die Verteilung eines sensiblen Attributes aus  $E_1 \cup E_2$ . Dann gilt die folgende Ungleichung:

$$D[P, Q] \leq \frac{|E_1|}{|E_1|+|E_2|} D[P_1, Q] + \frac{|E_2|}{|E_1|+|E_2|} D[P_2, Q]$$

$$\Rightarrow D[P, Q] \leq \max(D[P_1, Q], D[P_2, Q])$$

## Earth Movers Distanz (EMD)

**Fakt 2:**  $D[P, Q] \leq \max(D[P_1, Q], D[P_2, Q])$  Dies bedeutet, dass die maximale Distanz zwischen einer Äquivalenzklasse und der Tabelle beim Zusammenführen zweier Äquivalenzklassen nicht steigt. **Somit bleibt die t-closeness-Eigenschaft beim Zusammenführen erhalten.**

**Generalisation Property:** Sei  $T$  eine Tabelle,  $A$  und  $B$  sind Generalisierungen von  $T$ , wobei  $A$  mehr generalisiert ist als  $B$ . Wenn  $B$  die Eigenschaft t-closeness hat, dann hat auch  $A$  die Eigenschaft t-closeness.

*Beweis:* Die Äquivalenzklassen aus  $A$  bestehen aus der Vereinigung mehrerer Äquivalenzklassen aus  $B$ . Nach Fakt 2 kann somit die maximale Distanz nicht größer werden. Somit hat auch  $A$  die Eigenschaft t-closeness.

# EMD Beispiel

PLZ	Alter	Einkommen
4767*	$\leq 40$	3K
4767*	$\leq 40$	4K
4767*	$\leq 40$	5K
4790*	$\geq 40$	6K
4790*	$\geq 40$	8K
4790*	$\geq 40$	11K
4760*	$\leq 40$	7K
4760*	$\leq 40$	9K
4760*	$\leq 40$	10K

Tabelle: Einkommenstabelle

$Q =$

$\{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$

$P_1 = \{3k, 4k, 5k\}$

$P_2 = (\{6k, 8k, 11k\})$

$P_3 = (\{7k, 9k, 10k\})$

Beispiel aus [Li07]

## EMD Formeln

---

### **Nummerische Attribute:**

Domäne:  $\{v_i, \dots, v_m\}$ , wobei gilt  $i < j \Rightarrow v_i \leq v_j$

$$\text{geordnete-Distanz}(v_i, v_j) = \frac{|i-j|}{m-1}$$

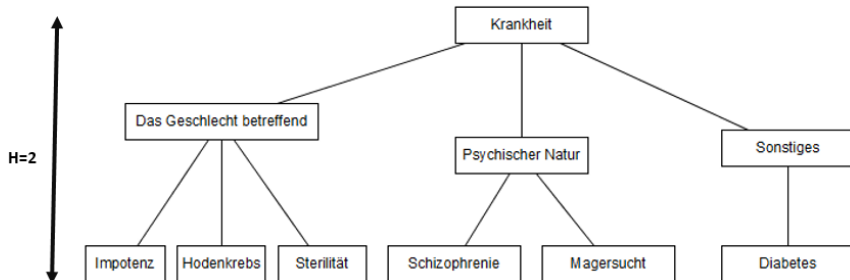
# EMD Formeln

## Hierrarchische Attribute:

$$\text{hierrarch-Distanz}(v_i, v_j) = \frac{\text{level}(v_i, v_j)}{H}$$

$$\text{hierr. -Dist}(\text{Impotenz}, \text{Hodenkrebs}) = 1/2$$

$$\text{hierr. -Dist.}(\text{Impotenz}, \text{Diabetes}) = 2/2$$





## EMD Beispiel

$$Q = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$$

$$P_1 = \{3k, 4k, 5k\}$$

Wahrscheinlichkeit  $\frac{1}{9}$  für folgende Transition:

$(5k \rightarrow 11k), (5k \rightarrow 10k), (5k \rightarrow 9k), (4k \rightarrow 8k), (4k \rightarrow 7k), (4k \rightarrow 6k), (3k \rightarrow 5k), (3k \rightarrow 4k), (3k \rightarrow 3k)$ .

$$\Rightarrow D[P_1, Q] = \frac{1}{9} \cdot \frac{6+5+4+4+3+2+2+1+0}{9-1} = 27/72 = 3/8 = 0.375$$

## EMD Beispiel

PLZ	Alter	Einkommen
4767*	$\leq 40$	3K
4767*	$\leq 40$	4K
4767*	$\leq 40$	5K
4790*	$\geq 40$	6K
4790*	$\geq 40$	8K
4790*	$\geq 40$	11K
4760*	$\leq 40$	7K
4760*	$\leq 40$	9K
4760*	$\leq 40$	10K

Tabelle: Einkommenstabelle

$$D[P_1, Q] = \frac{27}{72} = 0,375$$

$$D[P_2, Q] = \frac{12}{72} = 0,167$$

$$D[P_3, Q] = \frac{17}{72} = 0,236,$$

$$\Rightarrow t = 0,375$$

Die Einkommenstabelle hat die Eigenschaft 0,375-closeness

# EMD Beispiel

PLZ	Alter	Einkommen
4767*	$\leq 40$	3K
4767*	$\leq 40$	5K
4767*	$\leq 40$	9K
4790*	$\geq 40$	6K
4790*	$\geq 40$	8K
4790*	$\geq 40$	11K
4760*	$\leq 40$	4K
4760*	$\leq 40$	7K
4760*	$\leq 40$	10K

Tabelle: Einkommenstabelle

$$D[P'_1, Q] = \frac{12}{72} = 0,167$$

$$D[P_2, Q] = \frac{12}{72} = 0,167$$

$$D[P'_3, Q] = \frac{6}{72} = 0,083,$$

$$\Rightarrow t = 0,167$$

Die Einkommenstabelle hat die Eigenschaft 0,167-closeness

# Fazit

---

- **k-Anonymität**
  - mindestens  $k$  Tupel mit identischem Quasi-Identifikator
- **l-Diversität**
  - mindestens  $l$  verschiedene sensible Werte in jeder Äquivalenzklasse
- **t-Closeness**
  - Distanz zwischen Verteilung der sensiblen Attribute einer Äquivalenzklasse und der Gesamtverteilung unterscheidet sich maximal um einen Schwellwert  $t$


# Fazit

---

- **k-Anonymität**
  - mindestens  $k$  Tupel mit identischem Quasi-Identifikator
- **l-Diversität**
  - mindestens  $l$  verschiedene sensible Werte in jeder Äquivalenzklasse
- **t-Closeness**
  - Distanz zwischen Verteilung der sensiblen Attribute einer Äquivalenzklasse und der Gesamtverteilung unterscheidet sich maximal um einen Schwellwert  $t$
- **Ausblick**
  - gewichtete Attribute
  - Schwellwerte für maximale Anzahl an unterdrückten Tupeln
  - mehrere sensible Attribute
  - ...

## Literaturverzeichnis I

---

 Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu.

**Approximation algorithms for k-anonymity.**

*Journal of Privacy Technology (JOPT)*, 2005.

 Valentina Ciriani, S De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati.

**k-Anonymity.**

In *Secure data management in decentralized systems*, pages 323–353. Springer, 2007.

## Literaturverzeichnis II

---



Philippe Golle.

Revisiting the uniqueness of simple demographics in the US population.

In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 77–80. ACM, 2006.



Dietmar Hauf.

Allgemeine Konzepte K-Anonymity, I-Diversity and T-Closeness.


[https:](https://dbis.ipd.kit.edu/img/content/SS07Hauf_kAnonym.pdf)

[//dbis.ipd.kit.edu/img/content/SS07Hauf\\_kAnonym.pdf](https://dbis.ipd.kit.edu/img/content/SS07Hauf_kAnonym.pdf),  
2007.

Zugriff am 16.5.2016.

## Literaturverzeichnis III

---

-  Venkatasubramanian Li, Li.  
**t-Closeness: Privacy Beyond k-Anonymity and I-Diversity.**  
*2007 IEEE 23rd International Conference on Data Engineering,*  
 pages 106–115, 2007.
  
-  Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and  
 Muthuramakrishnan Venkitasubramaniam.  
**I-diversity: Privacy beyond k-anonymity.**  
*ACM Transactions on Knowledge Discovery from Data (TKDD),*  
 1(1):3, 2007.



## Literaturverzeichnis IV

---



Adam Meyerson and Ryan Williams.

On the complexity of optimal k-anonymity.

In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM, 2004.



Pierangela Samarati and Latanya Sweeney.

Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Technical report, Technical report, SRI International, 1998.



Latanya Sweeney.

Simple demographics often identify people uniquely.

*Health (San Francisco)*, 671:1–34, 2000.

## Literaturverzeichnis V

---



Latanya Sweeney.

k-anonymity: A model for protecting privacy.

*International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.