Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Vorlesung

Datenschutz und Privatheit in vernetzten Informationssystemen

Kapitel 3: Anonymität

Erik Buchmann
buchmann@ipd.uka.de

Ich danke Herrn Burghardt für seine Hilfe bei diesem Foliensatz

Motivation

- Häufig werden private Informationen gespeichert oder sogar zwischen Organisationen getauscht.
- Ziel für die Speicherung und den Informationsaustausch sind (unter Anderem)
  - Dienstverbesserung, z.B., Optimierung häufig genutzter Funktionen
  - Statistik, z.B. im Gesundheitswesen
  - Erfüllung von Gesetzesanforderungen
  - Forschung
- Erhebung und Weitergabe der Daten gefährdet die Privatheit der erfassten Personen.

Motivation

- Der Gesetzgeber sieht zwei Möglichkeiten zum Schutz der Privatheit vor (Neben der Datensparsamkeit oder dem Verzicht auf Speicherung)
  - Pseudonymisierung BDSG §3(6a)
  - Anonymisierung BDSG §3(6)

Motivation

BDSG §3(6a)

Pseudonymisieren ist das *Ersetzen* des Namens und anderer Identifikationsmerkmale *durch ein Kennzeichen* zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren.

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.

Motivation

- Beispiel Pseudonymisierung
  - Avatare, Benutzernamen bei Spielen, etc.

- Beispiel Anonymisierung
  - Wann ist ein Datensatz so verändert, dass „die Einzelangaben […] nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft“ einer Person zugeordnet werden können?

Beispiel AOL Suchprotokoll

- 2006 [Pas06] wurde von AOL ein „Anonymisiertes“ Suchprotokoll von 650k Benutzern mit mehr als 38Mio Anfragen zu Forschungszwecken publiziert.

AnonID	Query	QueryTime	ItemRank	ClickURL
100218	tennessee department of transportation	2006-03-01 11:08:30	1	http://www.tdot.state.tn.us
100218	tennessee federal court	2006-03-01 11:53:44	1	http://www.constructionweblinks.com
100218	state of tennessee emergency communications boar	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	dixie youth softball	2006-03-02 10:36:48	2	http://www.dixie.org
100218	cdwg	2006-03-03 14:29:07	1	http://www.cdwg.com
100218	cdwg scam cdwge	2006-03-03 14:30:11	0	
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com

NYT [Bar06] zeigte, wie einfach AOL-Benutzer 4417749 identifiziert werden konnte: „There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.” It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn…

Motivation

- Was war passiert?
- Zwei Gründe, warum Anonymisierung gescheitert ist
  - Quasi-Identifikatoren im Datensatz
  - Korrelierendes Wissen

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Quasi-Identifikatoren und Verknüpfung von korrelierendem Wissen

IPD, Forschungsbereich *Systeme der Informationsverwaltung*

Beispiel

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

Schlüssel

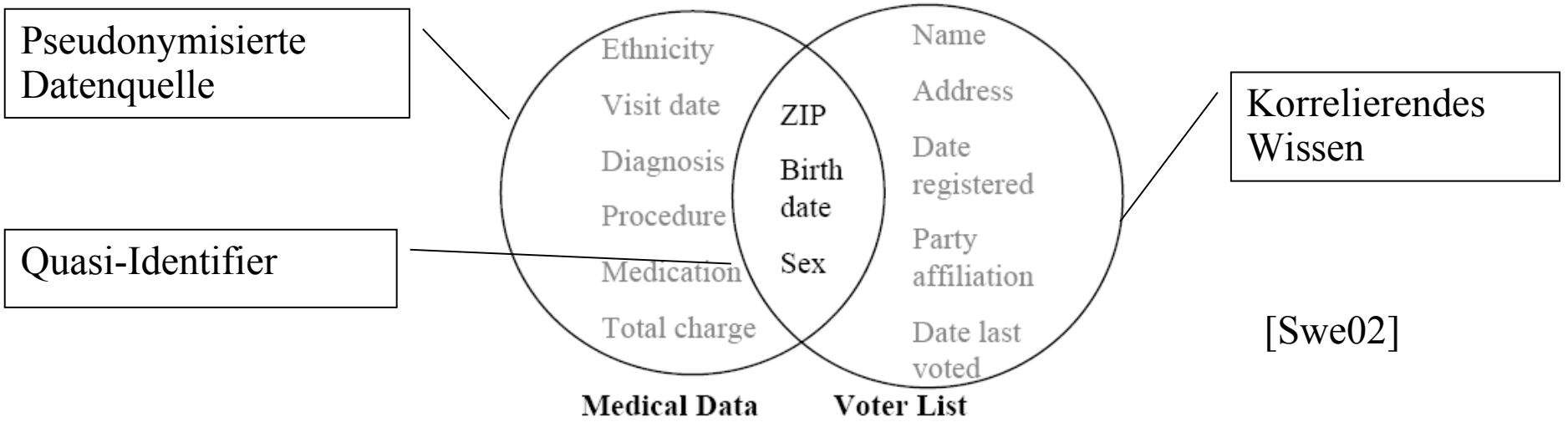
Name	Geb.	Geschl.	PLZ	Krankheit
1	19.04.75	M	76227	Impotenz
2	05.07.75	M	76228	Hodenkrebs
3	17.01.75	M	76227	Sterilität
4	23.04.81	M	76139	Schizophrenie
5	30.12.81	M	76133	Diabetes
6	05.07.83	W	76133	Magersucht
7	16.10.83	W	76131	Magersucht

Anonym ?

Beispiel einer pseudonymisierten Tabelle

Quasi-Identifier

- [Swe00] hat gezeigt, dass 87% der amerikanischen Bevölkerung eindeutig anhand der Attribute {*Geburtsdatum*, *PLZ*, *Geschlecht*} identifiziert werden können. ([Gol06] korrigiert diese Zahl auf 63%).



[Swe02]

- → *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

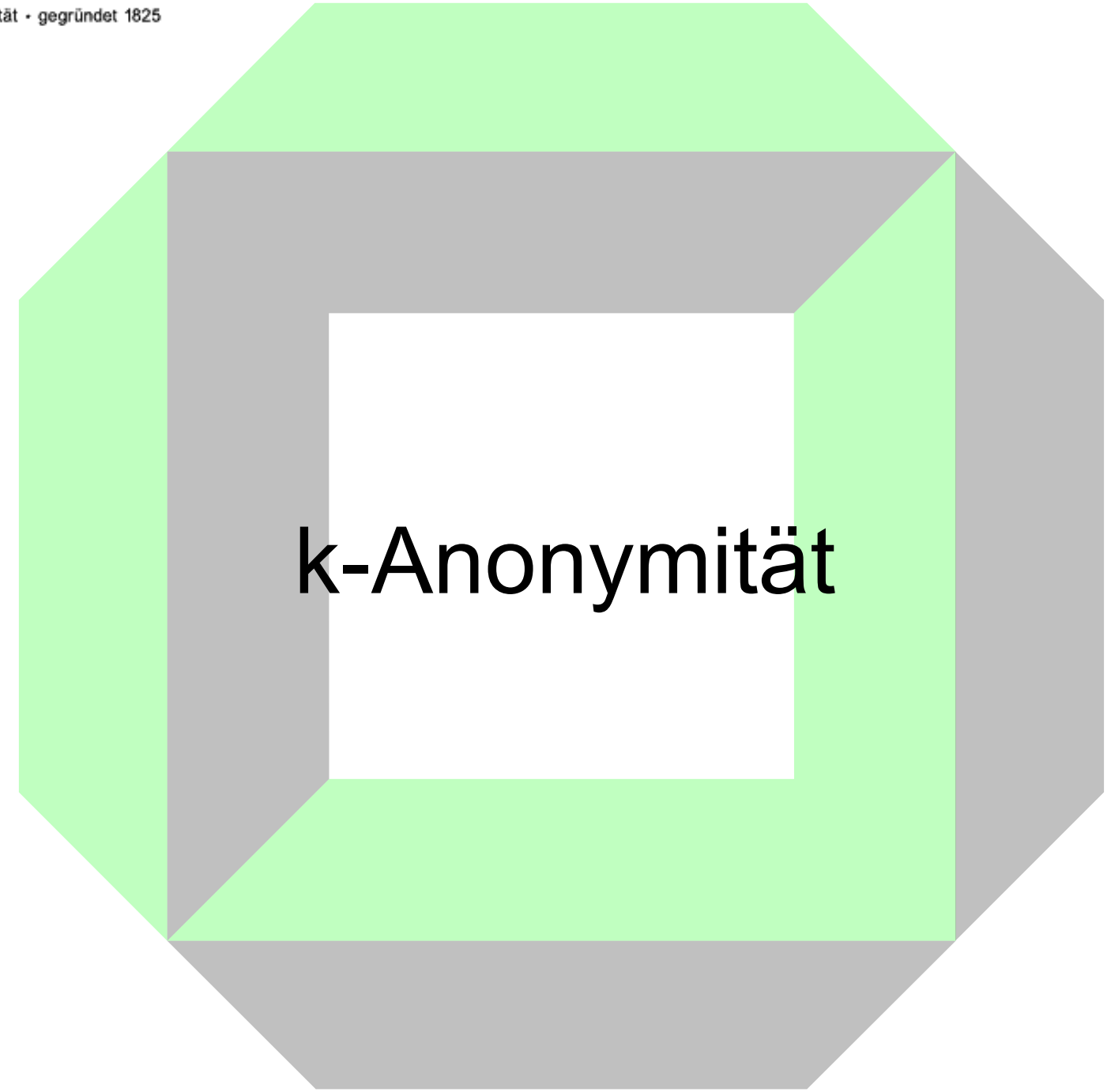
Gegeben sei

- 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 $f_g : T \rightarrow U'$ $U \subseteq U'$

Ein Quasi-Identifier von T; Q_T besteht aus einem Set von Attributen

$(A_i,...,A_j) \subseteq (A_1,...,A_n)$ für das gilt: $\exists p_i \in U : f_g(f_c(p_i)[Q_T]) = p_i$


k-Anonymität

Idee 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 von korrelierendem Wissen.

Definition k-Anonymität

Gegeben sei

- eine personenspezifische Tabelle $T(A_1,...,A_n)$
- 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.

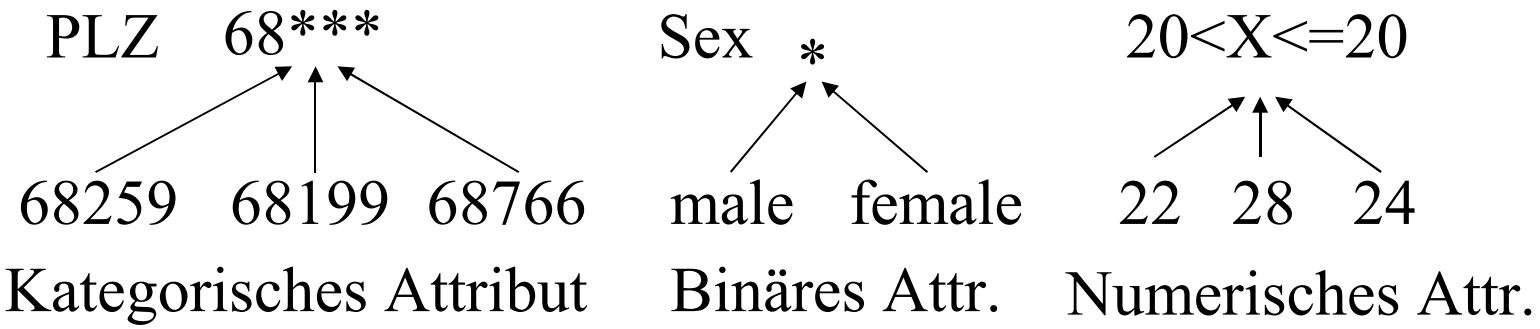
In anderen Worten:

Jedes Tupel ist von k-1 anderen Tupeln (bis auf das sensible Attribut) nicht unterscheidbar.

Wie k-Anonymität erreichen?

- Rauschen hinzufügen
- Dummy-Datensätze einfügen
- Unterdrücken von Informationen, d.h., Tupel löschen
- Daten vertauschen
- Generalisierung der Daten ← Unser Fokus

k-Anonymität durch Generalisierung



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 ?

Beispiel einer generalisierten Tabelle für k=2

Probleme von k-Anonymität

- Allgemeine Probleme beim Preisgeben von Datensätzen
  - Sortierungsbasierte Angriffe (Unsorted Matching Attack)
  - Angriffe bei dynamischen Datenbeständen (Temporal Attack)

- Spezielle Probleme von k-Anonymität
  - Homogenitätsangriff
  - Anwendung von korrelierendem (Hintergrund-) Wissen

Unsorted Matching

- 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

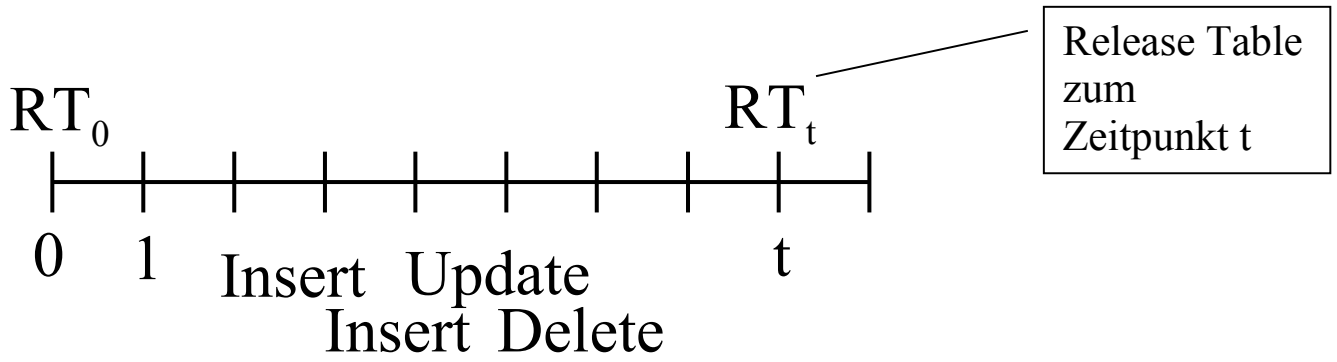
+

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

Dynamische Datenbestände

- Beispiel:



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

- Vorgeschlagenes Verfahren berücksichtigt nicht den Zusammenhang von RT_1 und RT_t

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

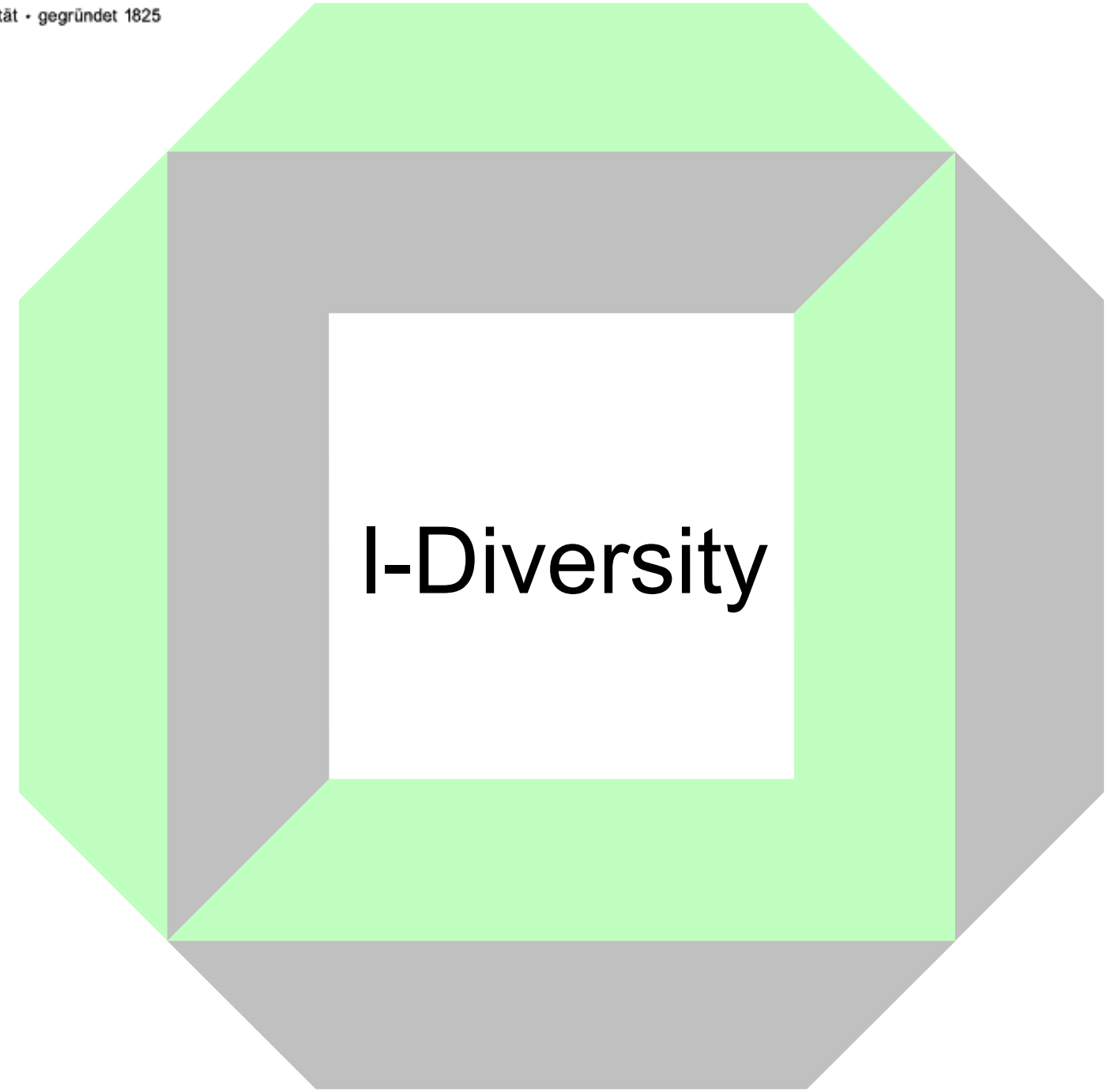
Beispiel einer generalisierten Tabelle für k=2

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


l-Diversity

Idee l-Diversity (*)

- Hintergrundwissen als Wahrscheinlichkeitsfunktion über den Attributen modelliert.
- Statistische Methoden (Bayesian Inference) zur Untersuchung auf Privatheit.
- 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 (positive disclosure)
  Die Veröffentlichung der Tabelle T* die aus T abgeleitet wurde resultiert in einer positiven Preisgabe, wenn der Angreifer den Wert eines sensiblen Attributes mit hoher Wahrscheinlichkeit bestimmten kann.

- Negative Preisgabe (negative disclosure)
  Die Veröffentlichung der Tabelle T* die aus T abgeleitet wurde resultiert in einer negativen Preisgabe, wenn der Angreifer mit hoher Wahrscheinlichkeit ein sensibles Attribut ausschließen kann.

Hintergrundwissen

- 1: Positive Preisgabe: Wer 83 geboren wurde hat Magersucht
- 2: 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 sei eine Tabelle **T** und die generalisierte Tabelle **T***, wobei **q*** der generalisierte Wert von **q** ist und ein **q*-Block** von Tupeln aus T*, deren nicht-sensiblen Werte 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

- Formale Definition Entropy-l-Diversity:

  Gegeben sei eine Tabelle T und die generalisierte Tabelle T*, wobei q* der generalisierte Wert von q ist und ein q*-Block von Tupeln aus T*, deren nichts-sensiblen Werte 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

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-0-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

Beispiel einer generalisierten Tabelle für k=3 entropy-2(.8)-diversity

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 die Anforderung u.U. zu restriktiv.
    Bsp: 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

- Formale Definition Recursive (c,l)-Diversity

In einem gegebenen q*-Block beschreibt r_i die Häufigkeit, die der i-häufigste sensible Wert in diesem q*-Block aufweist.

Häufigen Werte sind nicht über-, seltene nicht unterrepräsentiert

Gegeben eine Konstante c ist der q*-Block (c,l)-Divers, wenn gilt $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$.

Eine Tabelle T* ist (c,l)-Divers, wenn jeder q*-Block recursive-(c,l)-divers ist.

Daraus folgt, dass wenn ein nach der Eliminierung eines sensiblen Wertes der q*-Block immernoch (c,l-1) divers ist.

Weitere Konzepte für l-Diversity

(nur zur Vollständigkeit):

- **Positive Disclosure-Recursive (c.l)-Diversity**
  - Erlaubt positive Preisgabe bestimmter weniger, z.B. nicht sensibler Attribute („gesund“)
- **Negative/Positive Disclosure-Recursive(c1,c2,l)-Diversity**
  - Schutz vor negativer Preisgabe von Attributen
  - Erfüllt Positive Disclosure-Recursive (c,l)-Diversity
  - Vor negativer Preisgabe zu schützende Attribute müssen in mindestens c2% 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 → gilt v_3

Probleme von l-Diversity

- Nach [LiN07]:
  - 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.

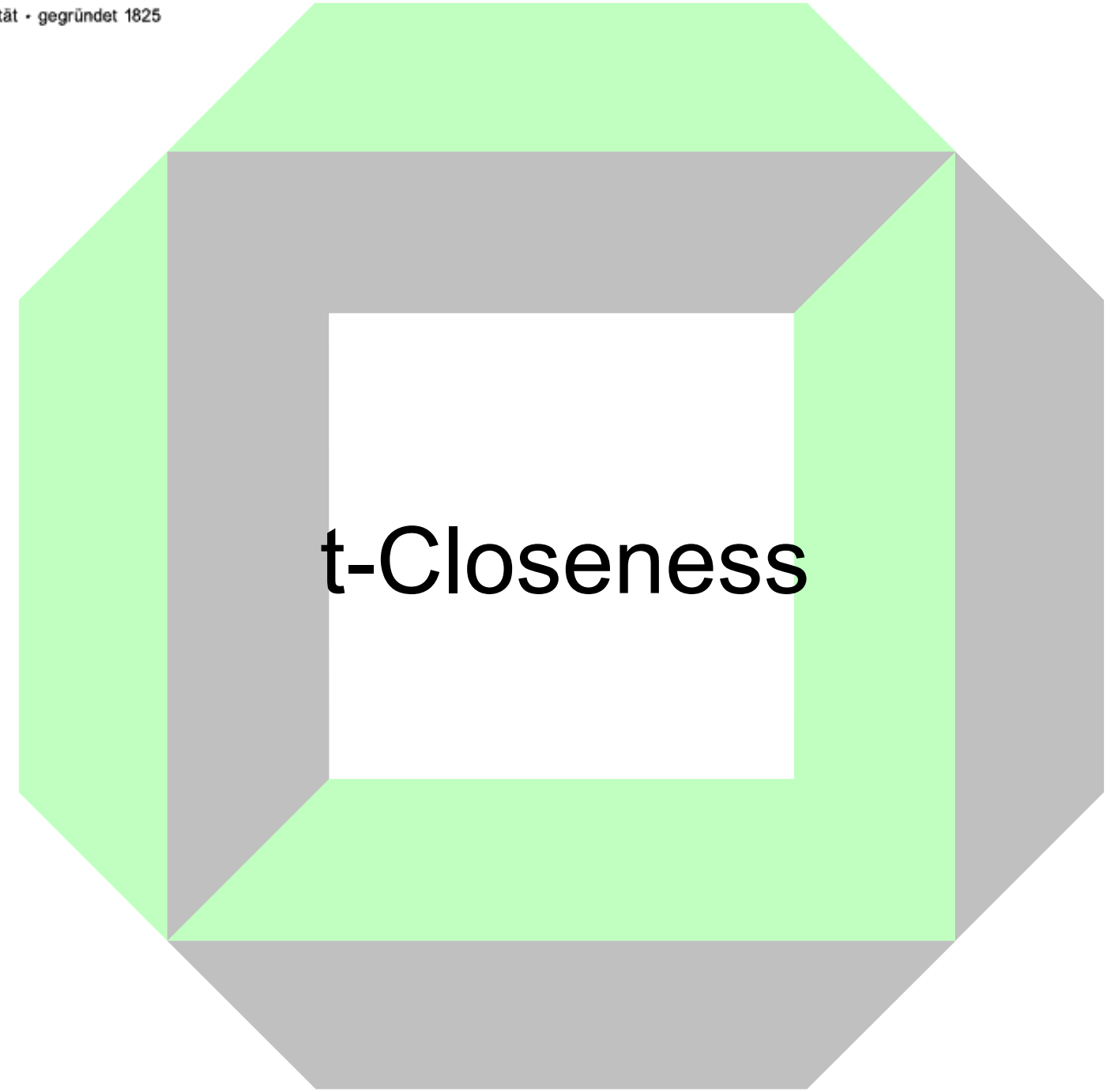
→ l-Diversity berücksichtigt nicht die Gesamtverteilung von sensiblen Attributen

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


t-Closeness

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

Beispiel einer vollständig generalisierten Tabelle

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

Beispiel einer preisgegebenen Tabelle

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

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 viel neue Information bzw. Neues im Vergleich zu einer weit verbreiteten Annahme
- $B_0 - B_2$
  - Die l-Diversity Idee ist es, die Differenz zwischen B_0 und B_2 durch die Diversity-Anforderung an P zu begrenzen
- $B_1 - B_2$
  - t-Closeness: Begrenz die Information 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 = $\{p_1, p_2, \ldots, p_m\}$
  - Verteilung Q = $\{q_1, q_2, \ldots, q_m\}$
- 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 unterschiedliche 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.

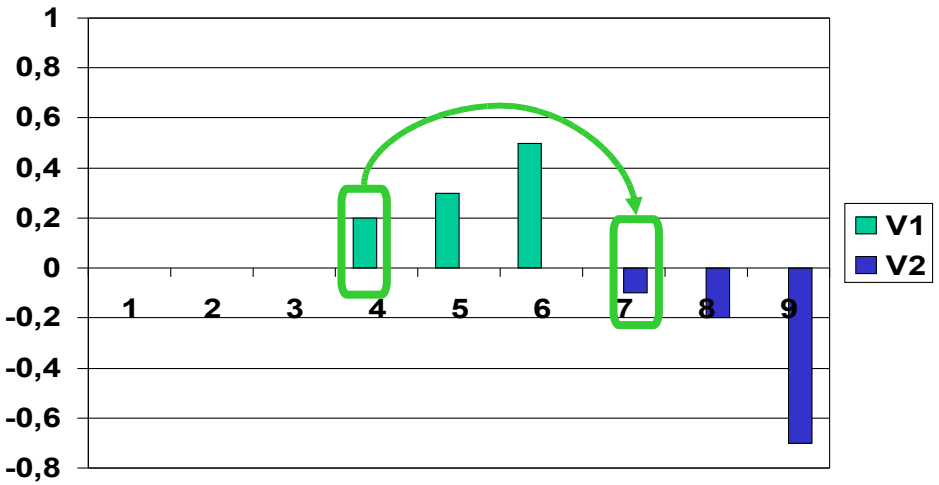
Definition

- 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

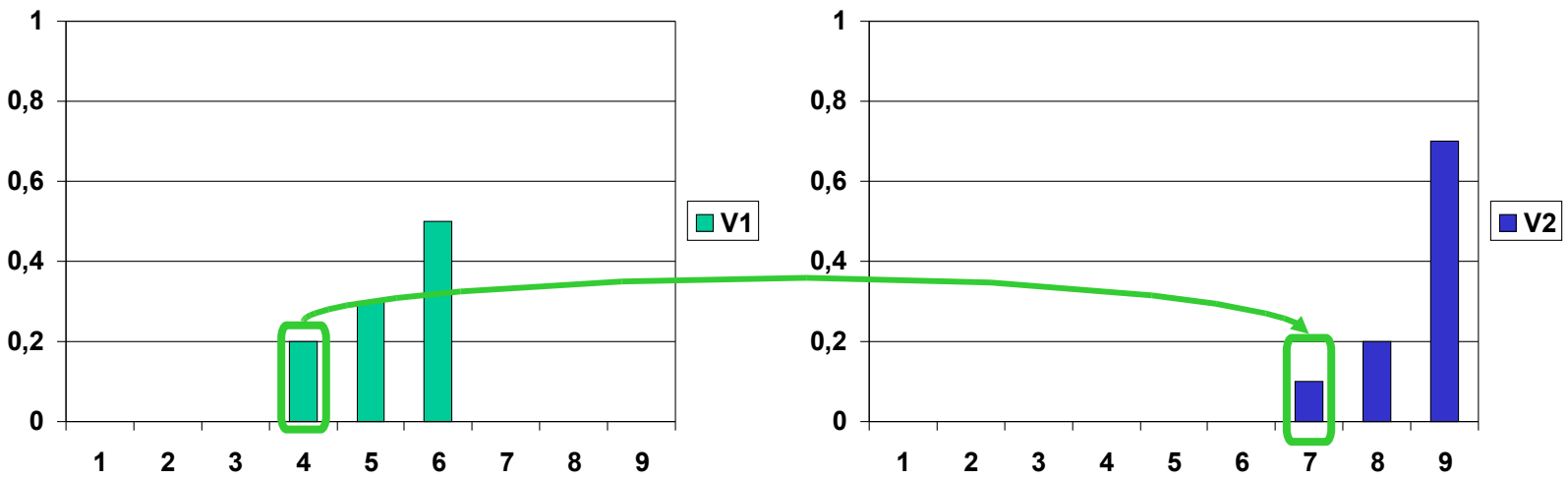
Work(f,i,j) = f * | i - j |;

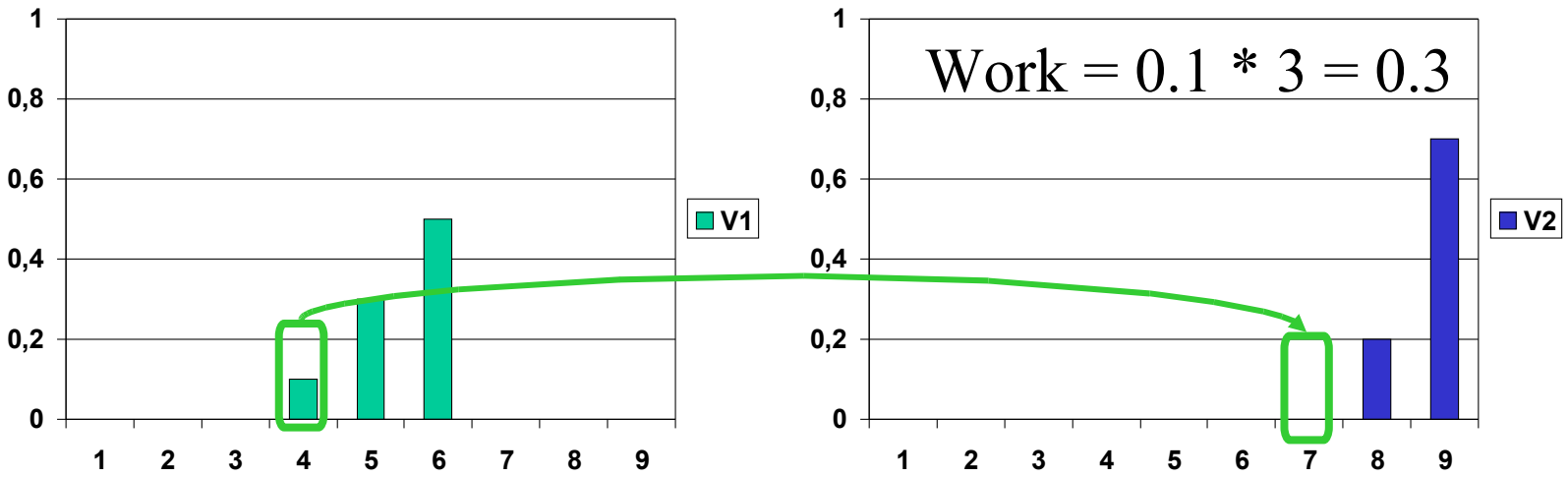
f = Anzahl Werte

i,j die Werte



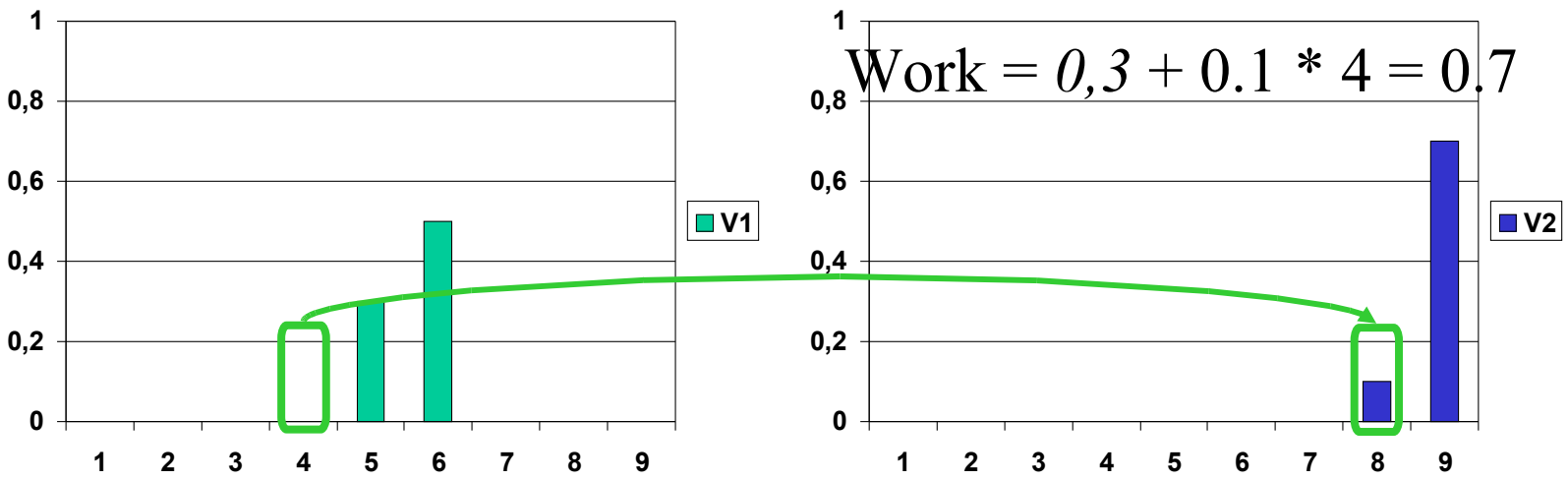
Beispiel Earth Mover‘s Distance



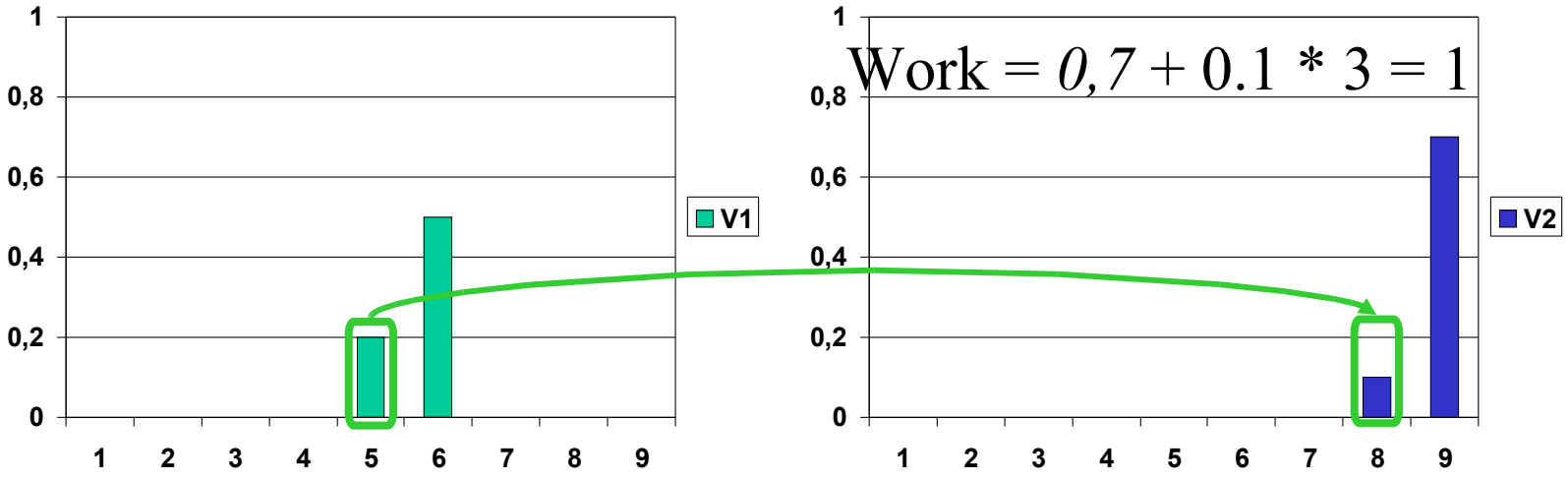


$$Work(f,i,j) = f * | i - j |;$$

Beispiel Earth Mover‘s Distance



Work = *0,3* + 0.1 * 4 = 0.7



Work = *0,7* + 0.1 * 3 = 1

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

- Daraus kann abgeleitet werden
  - Sind die Distanzen d_{ij} normalisiert ($0 \leq d_{ij} \leq 1$), dann ist auch $0 \leq D[P,Q] \leq 1$ (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 $$hierarchical - dist(v_i, v_j) = \frac{level(v_i, v_j)}{H}$$

* Vielen Dank an Dietmar Hauf für seine Seminarausarbeitung

Zusammenfassung

- **Identity Disclosure**
  - Ein Individuum kann mit einem preisgegebenen Datensatz verknüpft werden.
  - k-Anonymity [Swe02]
- **Attribute Disclosure**
  - Sensible Attribute einer Person werden preisgegeben.
  - l-Diversity [Mac06]
  - t-Closeness [LiN07]
  - Anatomy [Xia06]
- **Membership Disclosure**
  - Information, ob eine Person in einem publizierten Datensatz ist oder nicht wird preisgegeben.
  - δ-presence [Ner07] nicht behandelt

Literatur I

[Pas06] Pass, G.; Chowdhury, A. & Torgeson, C.
A picture of search, *InfoScale '06: Proceedings of the 1st international conference on scalable information systems, ACM,* **2006**

[Bar06] BARBARO, M. & ZELLER, T., *A Face Is Exposed for AOL Searcher No. 4417749*, *New York Times,* **2006**

[Swe00] Sweeney, L., *Uniqueness of simple demographics in the US population LIDAP-WP4,* **2000**

[Gol06] Golle, P., Revisiting the uniqueness of simple demographics in the US population, *WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society, ACM,* **2006**, 77-80

[Swe02] Sweeney, L., *k-anonymity: a model for protecting privacy Int. J. Uncertain. Fuzziness Knowl.-Based Syst., World Scientific Publishing Co., Inc.,* **2002**, *10*, 557-570

[Mac06] Machanavajjhala, A.; Gehrke, J.; Kifer, D. & Venkitasubramaniam, M. *l-Diversity: Privacy Beyond k-Anonymity*, *22nd IEEE International Conference on Data Engineering,* **2006**

[LiN07] Li, N.; Li, T. & Venkatasubramanian, S., t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on,* **2007**, 106-115

Literatur II

[Xia06] Xiao, X. & Tao, Y., Anatomy: simple and effective privacy preservation
VLDB'2006: Proceedings of the 32nd international conference on Very large data bases, VLDB Endowment, **2006**, 139-150

[Ner07] Nergiz, M. E.; Atzori, M. & Clifton, C.
Hiding the presence of individuals from shared databases
SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, **2007**, 665-676