



k-Anonymität

Thomas Maier, Kai Sonnenwald, Tom Petersen

Universität Hamburg
Fachbereich Informatik



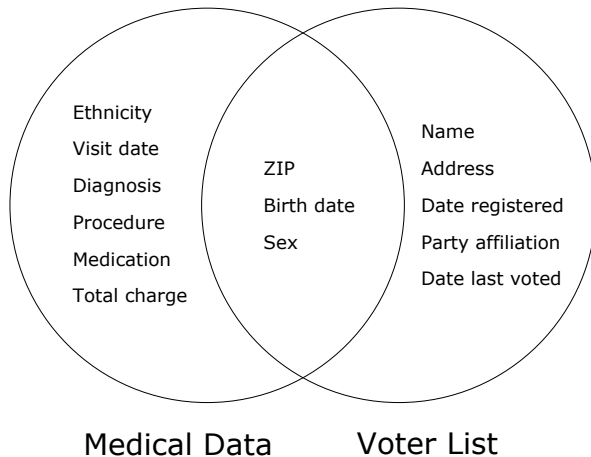
Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Agenda

1. Motivation & Abgrenzung
2. k-Anonymität
 - Generalisierung
 - Suppression
3. Schwächen der k-Anonymität
4. l-Diversity
5. Schwächen der l-Diversity
6. t-Closeness
7. 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.

	T. M. J.	M. J.	J.	2 J.
PLZ	87.1 %	3.7 %	0.04 %	0.01 %
Ort	58.4 %	3.6 %	0.04 %	0.01 %
County	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

Identifizier	Nicht-sensibel			Sensibel
Name	Geschl.	PLZ	Geb.dat.	Erkrankung
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	Gicht
Leon Schmidt	m	22188	26.4.1964	Demenz
Elias Koch	m	22997	7.10.1975	Hepatitis

Begriffe

Explicit identifier Attribut, das ein Individuum (nahezu) eindeutig identifiziert. Bsp: Name, Adresse, Steuernummer, ...

Sensitive attribute Attribut, dessen Wert für ein Individuum in einer Datenmenge nicht herausgefunden werden darf.

Quasi identifier Attributmenge, die ein Individuum in Kombination identifizieren kann. *Formal in [Swe02] p. 7 auch [MKG07] p. 3:* 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 identifizier” ist.

k-Anonymität

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

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 identifiziers beschreibt.

Beispiel: k-anonyme Tabelle

Identifizier	Nicht-sensibel			Sensibel
Name	Geschl.	PLZ	Geb.dat.	Erkrankung
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	Gicht
Leon Schmidt	m	22188	26.4.1964	Demenz
Elias Koch	m	22997	7.10.1975	Hepatitis

Beispiel: k-anonyme Tabelle

<i>Identifizier</i>	Nicht-sensibel			Sensibel
<i>Name</i>	Geschl.	PLZ	Geb.dat.	Erkrankung
-	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	Gicht
-	m	22188	26.4.1964	Demenz
-	m	22997	7.10.1975	Hepatitis

Beispiel: k-anonyme Tabelle

<i>Identifizier</i>	Nicht-sensibel			Sensibel
<i>Name</i>	Geschl.	PLZ	Geburtsjahr	Erkrankung
-	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	Gicht
-	m	22188	1964	Demenz
-	m	22997	1975	Hepatitis

Beispiel: k-anonyme Tabelle

<i>Identifizier</i>	Nicht-sensibel			Sensibel
<i>Name</i>	Geschl.	PLZ	Geburtsjahr	Erkrankung
-	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	Gicht
-	m	22188	1964	Demenz
-	m	22997	1975	Hepatitis

Beispiel: k-anonyme Tabelle

<i>Identifizier</i>	Nicht-sensibel			Sensibel
<i>Name</i>	Geschl.	PLZ	Geburtsjahr	Erkrankung
-	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	Gicht
-	m	22188	1964	Demenz
-	m	22997	1975	Hepatitis

Beispiel: k-anonyme Tabelle

<i>Identifizier</i>	Nicht-sensibel			Sensibel
<i>Name</i>	Geschl.	PLZ	Geburtsjahr	Erkrankung
-	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	Gicht
-	m	22188	1964	Demenz
-	m	22997	1975	Hepatitis

Beispiel: k-anonyme Tabelle

<i>Identifizier</i>	Nicht-sensibel			Sensibel
<i>Name</i>	Geschl.	PLZ	Geburtsjahr	Erkrankung
-	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	Gicht
-	m	22***	1964-64	Demenz
-	m	22997	1975	Hepatitis

Beispiel: k-anonyme Tabelle

<i>Identifizier</i>	Nicht-sensibel			Sensibel
<i>Name</i>	Geschl.	PLZ	Geburtsjahr	Erkrankung
-	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	Gicht
-	m	22***	1963-64	Demenz
-	m	22997	1936-75	Hepatitis

Beispiel: k-anonyme Tabelle

<i>Identifizier</i>	Nicht-sensibel			Sensibel
<i>Name</i>	Geschl.	PLZ	Geburtsjahr	Erkrankung
-	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	Gicht
-	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:

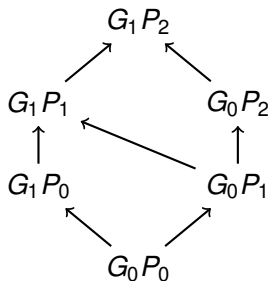
1. PLZ:

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

2. Geschlecht:

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

Generalisierung II



PLZ:

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

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

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

Geschlecht:

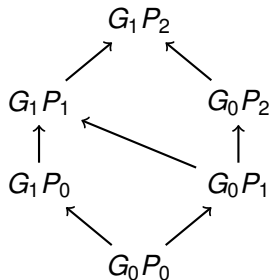
$$G_0 = \{\text{männlich, weiblich}\}$$

$$\rightarrow G_1 = \{\text{nicht_veröffentlicht}\}$$

Generalisierungshierarchie für Attributmenge

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

Generalisierung II



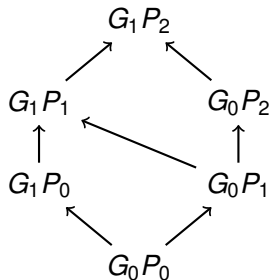
Generalisierungshierarchie für Attributmenge

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



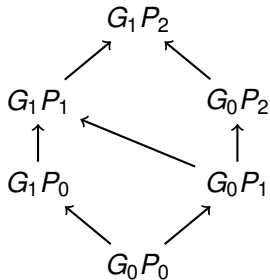
Generalisierungshierarchie für Attributmenge

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



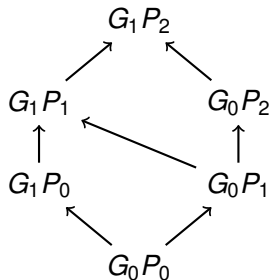
Generalisierungshierarchie für Attributmenge

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 k-anonymen Tabelle ist NP-schwer, ...

Implementierungen

Die Berechnung von k-anonymen Tabelle ist NP-schwer, ...

... es wurden jedoch $\mathcal{O}(k)$ -Approximationsalgorithmen gefunden
[?, ?].

Implementierungen

	Unterdrückung			
Generalisierung	Tupel	Attribut	Zelle	Keine
Attribut	AG_TS	AG_AS = AG	AG_CS	AG = AG_AS
Zelle	CG_TS	CG_AS	CG_CS = CG	CG = CG_CS
Keine	TS	AS	CS	-

Klassifizierung von Techniken für die Erstellung k-anonymer Tabellen. Entnommen aus [?]

Implementierungen

	Unterdrückung			
Generalisierung	Tupel	Attribut	Zelle	Keine
Attribut	AG_TS	AG_AS = AG	AG_CS	AG = AG_AS
Zelle	CG_TS	CG_AS	CG_CS = CG	CG = CG_CS
Keine	TS	AS	CS	-

Klassifizierung von Techniken für die Erstellung k-anonymer Tabellen. Entnommen aus [?]

- μ -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** [MKGV07]
- **Background knowledge attack** [MKGV07]

Unsorted matching attack

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

$k = 2$

Zufällige Sortierung der Tabellen verhindert diesen Angriff!

Homogeneity attack

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

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

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

$$k = 4$$

I-Diversity

Schwächen der I-Diversity

Skewness attack
 similarity attack

t-Closeness

Literaturverzeichnis I



GOLLE, Philippe:

Revisiting the uniqueness of simple demographics in the US population.


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




MACHANAVAJJHALA, Ashwin ; KIFER, Daniel ; GEHRKE, Johannes ; VENKITASUBRAMANIAM, Muthuramakrishnan:
l-diversity: Privacy beyond k-anonymity.


In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1 (2007), Nr. 1, S. 3

Literaturverzeichnis II

 SAMARATI, Pierangela ; SWEENEY, Latanya:
Protecting privacy when disclosing information: k-anonymity and
its enforcement through generalization and suppression /
Technical report, SRI International.
1998. —

Forschungsbericht

 SWEENEY, Latanya:
Simple Demographics Often Identify People Uniquely.
(2000)

 SWEENEY, Latanya:
k-anonymity: A model for protecting privacy.
In: *International Journal of Uncertainty, Fuzziness and
Knowledge-Based Systems* 10 (2002), Nr. 05, S. 557–570