



k-Anonymität

Thomas Maier, Kai Sonnenwald, Tom Petersen

Universität Hamburg
Fachbereich Informatik



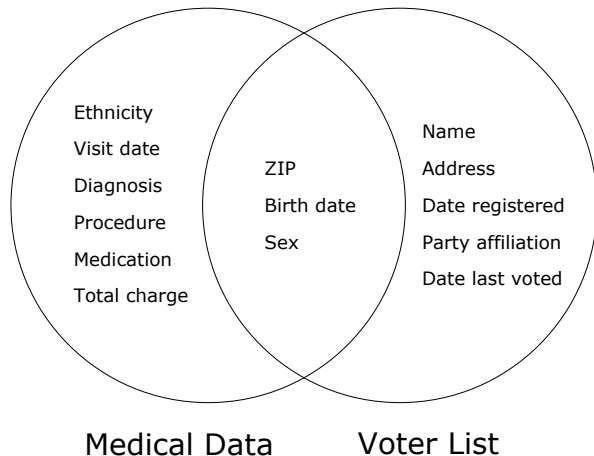
Universität Hamburg

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]

Sweeney - Beispiel [Swe02]

Anonym? II

-> [Swe00], [Gol06] Studien über die Eindeutigkeit von demographischen Faktoren in der U.S.-Bevölkerung

Sweeney für die 1990 US census data, Golle wiederholte das für 2000

	Geb.dat.	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

	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]

Durch Geburtsdatum, Geschlecht, PLZ konnten 87% der Bevölkerung eindeutig identifiziert werden

Abgrenzung

-> Vermeintlich anonyme Daten sind es nicht. Daher: wie können wir Anonymität formalisieren bzw. Aussagen über die "Güte" der Anonymisierung machen?

—

Worum geht es?

NICHT Begrenzung des Zugriffs (Authentifikation, Multi-Level-Datenbanken).

NICHT statistische Datenbanken (Aggregation, Begrenzung von Selektionsarten, Logging und Abwägen von Anfragen, Hinzufügen von Zufall). Oftmals Verlust der Integrität der Daten.

SONDERN Anonyme Veröffentlichung von Daten als Individualdatensätze.

Identifizier	Nicht-sensibel			Sensibel
Name	Geschlecht	PLZ	Geburtsdatum	Erkrankung
Mia Schulz	w	21989	20.5.1944	Osteoporose
Elias Wagner	m	21727	25.8.1983	Gicht
Hanna Weber	w	20817	28.3.1953	Osteoporose
Leon Schulz	m	21220	28.10.1994	Bronchitis
Sofia Koch	w	20270	21.1.1965	Gicht
Leon Schmidt	m	20188	5.5.1958	Hepatitis
Hanna Schäfer	w	21462	11.2.1999	Epilepsie
Elias Schneider	m	20388	3.8.1971	Multiple Skle
Mia Fischer	w	21896	14.12.1999	Diabetes
Ben Meyer	m	21024	8.1.1982	Diabetes

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 [MKGv07] 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

Informell: Eine Tabelle (Datensatz?) erfüllt k -Anonymität, wenn jede Zeile (jeder Eintrag) ununterscheidbar von $k - 1$ anderen Zeilen im Bezug auf jede “quasi identifizier“-Menge 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

Generalisierung

Generalisierung

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

Beispiele für Generalisierungshierarchien:

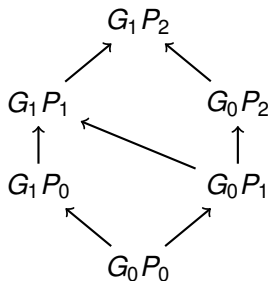
1. PLZ:

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

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:

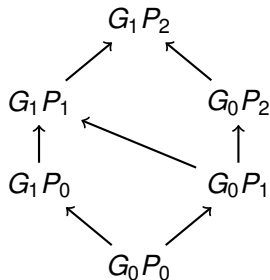
$$G_0 = \{\text{männlich, weiblich}\}$$

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

Generalisierungshierarchie für Attributmenge

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

Generalisierung II



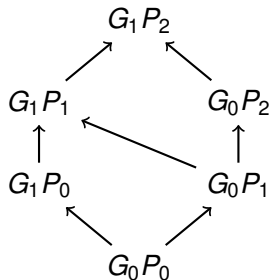
Generalisierungshierarchie für Attributmenge

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



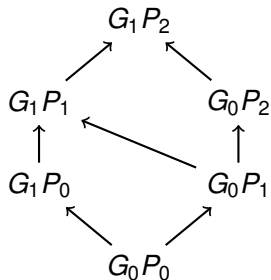
Generalisierungshierarchie für Attributmenge

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



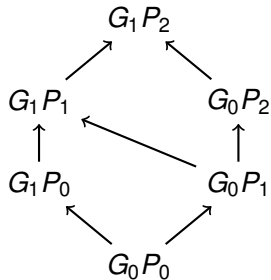
Generalisierungshierarchie für Attributmenge

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

Nicht jede Generalisierung ist gleichermaßen sinnvoll.
 Trivillösung: Für jedes Attribut die höchste Stufe der Generalisierung wählen -> Jedes Tupel bezogen auf den Quasi identifier enthält die gleichen Werte -> auf Kosten hoher Generalisierung und damit geringer Nutzbarkeit der Daten.

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

Berechnung von k-anonymer Tabelle NP-schwer, es wurden jedoch $\mathcal{O}(k)$ -Approximationsalgorithmen gefunden [?, ?].

	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 [?]

Zusätzlich (und hier nicht abgedeckt): gewichtete Attribute, Schwellwerte für maximale Anzahl an unterdrückten Tupeln, mehrere sensible Attribute, ...

Datafly μ -Argus Incognito

Schwächen der k -Anonymität

- Unsorted matching attack Veröffentlichung mehrerer k -anonymer Tabellen mit derselben Sortierung ausgehend von einer nicht-öffentlichen Tabelle. [Swe02] p.10
- Complementary release attack Veröffentlichung mehrerer k -anonymer Tabellen unterschiedlicher Generalisierung, die zusammengeführt die k -Anonymität verletzen. [Swe02] p.11
- Temporal attack Dynamische Tabellen können k -Anonymität verletzen. [Swe02] p.12
- Homogeneity attack Gleichheit der sensitive attributes einer Gruppe, die sich in den Werten des quasi identifiers gleicht, leakt das sensitive attribute eines Individuums. [MKG07] p. 2
- Background knowledge attack Nutzen von Hintergrundwissen, um mit hoher Wahrscheinlichkeit auf den Wert des sensitive attributes eines Individuums in einer Gruppe

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 verhindert diesen Angriff

Complementary release attack

TBD

Temporal attack

TBD?

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

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-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