

E-Privacy – Privacy in the Electronic Society



Database Privacy

Günter Karjoth

Spring term 2009

The “digital universe” of data*

- 281 exabyte in 2007 (2.25×10^{21})
 - 45 gigabytes for every person on earth
- compound annual growth rate of almost 60 %
 - 1,800 exabyte in 2011 (predicted)
- Information explosion caused by
 - digital cameras, digital surveillance cameras, digital television
 - Internet access in emerging countries
 - sensor-based applications
 - social networks

* IDC Study 2008

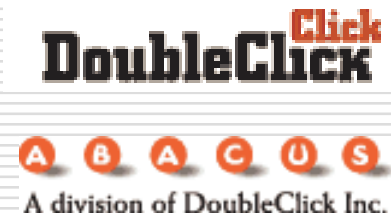


The digital shadow

- Only half of your digital footprint is related to your individual actions
 - taking pictures, sending e-mails, making digital voice calls, ...
- The other half – the “digital shadow” – is information about you
 - names in financial records, names on mailing lists, Web surfing histories, images taken of you by security, ...

Online and Offline Merging

- In November 1999, DoubleClick purchased Abacus Direct, a company possessing detailed consumer profiles on more than 90% of US households.



- In mid-February 2000 DoubleClick announced plans to merge “anonymous” online data with personal information obtained from offline databases
- By the first week in March 2000 the plans were put on hold
 - Stock dropped from \$125 (12/99) to \$80 (03/00)

[Source: Langheinrich, 2001]

JetBlue Violates Passenger Privacy

- In Sep 2003, *JetBlue Airways* gave 5 million passenger records including names, addresses, phone numbers and flight information to *Torch Concepts*, a private DoD contractor.
- *Torch* then purchased additional customer demographic information from data aggregator *Acxiom*.
- By matching the JetBlue passenger list with the Acxiom information, *Torch* developed passenger profiles to identify possible terrorist suspects.
 - *Torch* was able to extract demographic information including income information, social security number, occupations, and years at residence for approximately 40% of those passengers;
 - data transfer directly violated JetBlue's privacy policy;
 - lawsuits and investigations have been initiated.

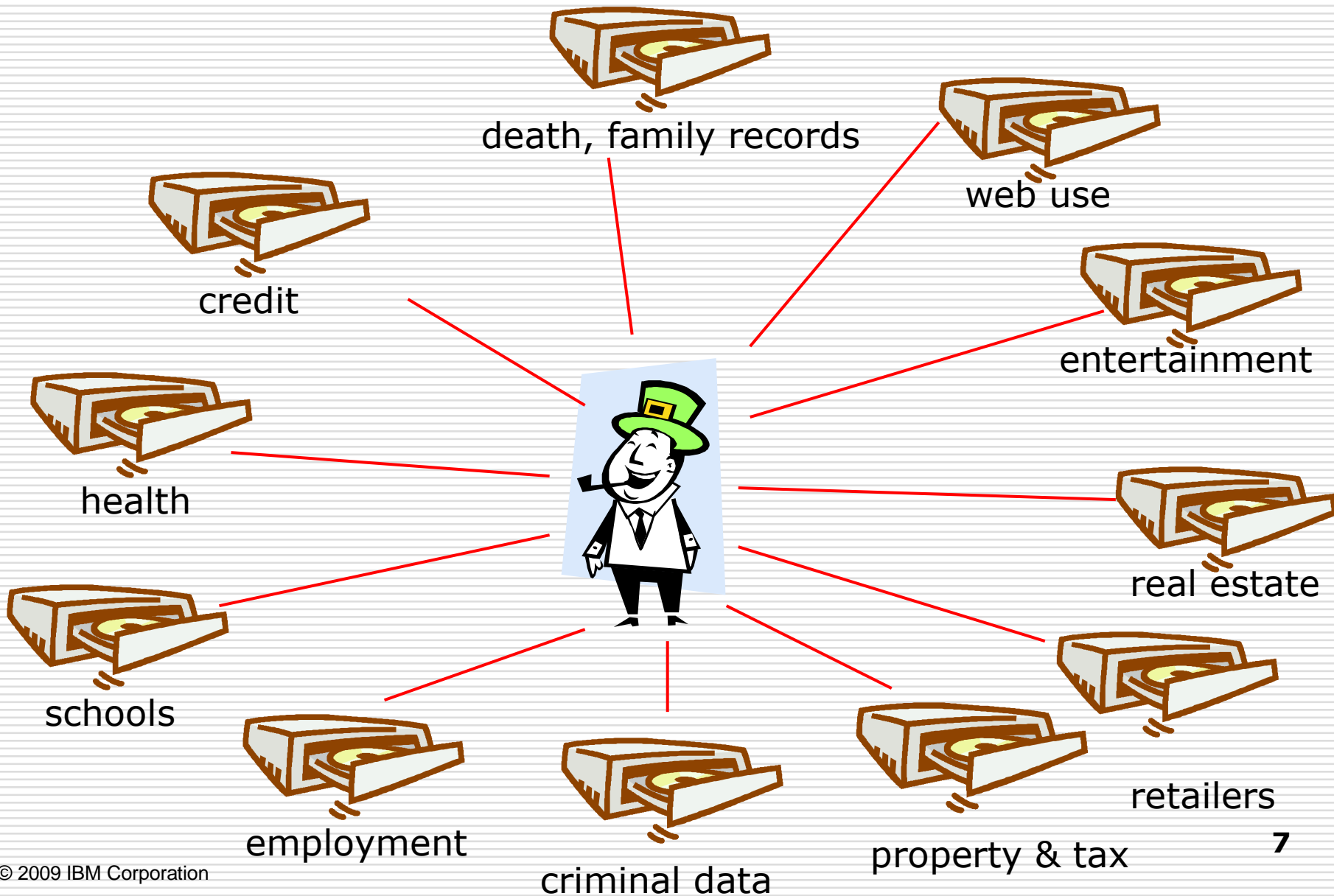
[see article of Anton, He & Baumer; 2004]

Netflix Prize (*)

- In 2006, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge to come up with better recommendation systems
- data was anonymized by removing personal details and replacing names with random numbers
- some of the Netflix data was de-anonymized by comparing rankings and timestamps with public information in the Internet Movie Database (IMDb)
- research demonstrated how little information is required to de-anonymize information in the Netflix dataset

(*) www.netflixprize.com 6

Sources of data on individuals



Trends in Data Collection Behaviors

- **Collect more**
Expand an existing person-specific data collection.
- **Collect specifically**
Replace an existing aggregate data collection with a person-specific one.
- **Collect if you can**

Examples (#fields)	1983	1996
Each birth	15	226
Each hospital visit	0	50
Each grocery visit	0	1,272

Statistical Databases

- official statistics
 - statistical agencies must guarantee statistical confidentiality when data released
 - health information
 - HIPAA requires strict regulation of protected health information for use in medical research
 - e-commerce
 - no public profiling
 - subject to regulations
- how to protect static individual data (microdata)

Privacy-Enhancing Techniques

- Privacy-preserving data mining
 - lets businesses derive the understanding they need without collecting accurate personal information
 - Information sharing across private repositories
 - to allow businesses to compile aggregate models without having to merge the individual data
 - [secure multiparty computations](#)
 - Privacy-preserving search
 - data owner's privacy -
 - searcher's privacy - protecting the query search criteria
 - [private information retrieval](#)
- limiting the amount of data that users can acquire

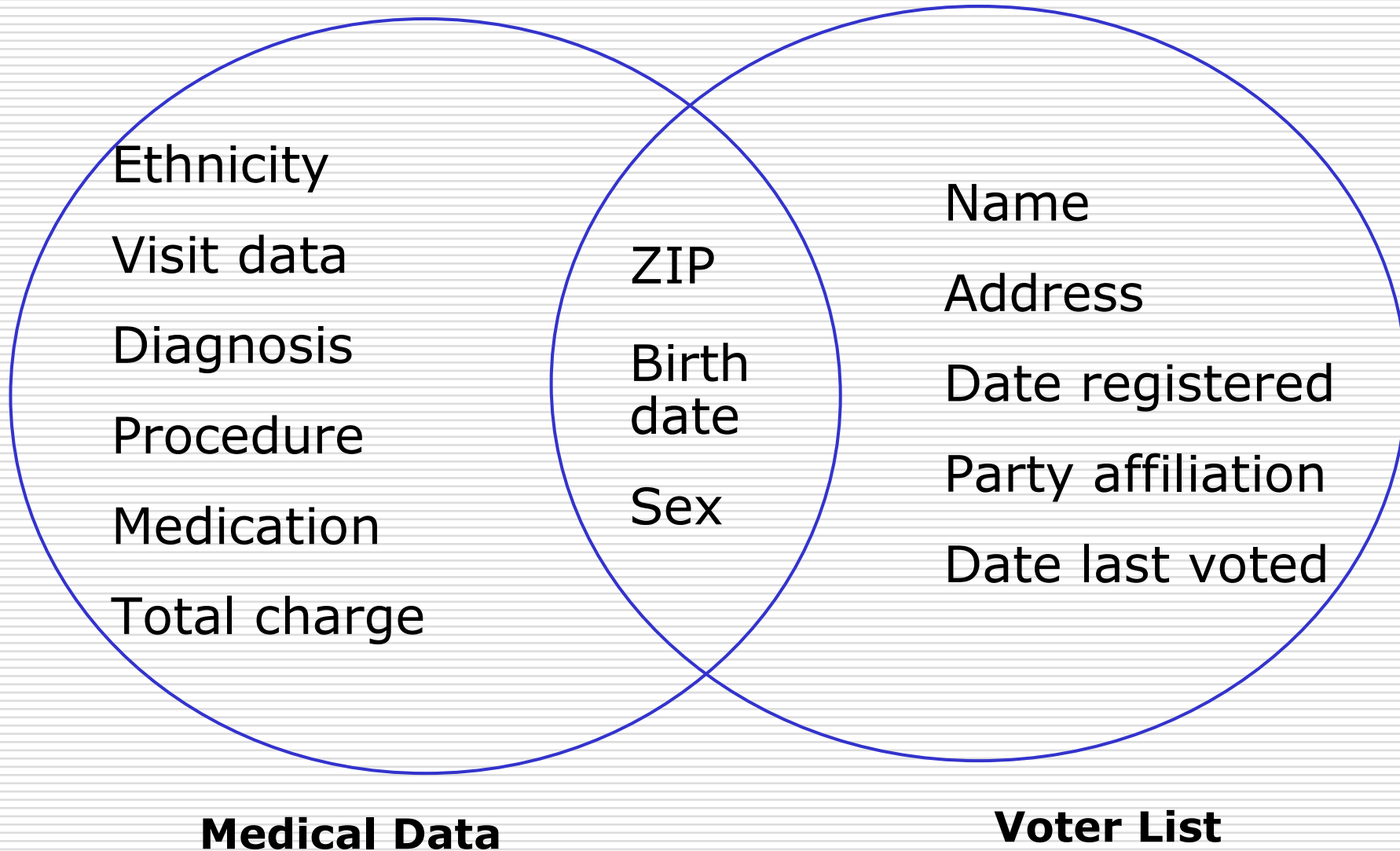
Medical Data Released as Anonymous

SSN	Name	Race	Date of Birth	Sex	ZIP	Marital Status	Health Problems
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		asian	04/15/64	male	94139	married	obesity
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		black	09/07/64	female	94141	married	obesity
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party
Sue J. Carlson	900 Market St	San Francisco	94142	9/15/61	female	democrat

Linking to re-identify data



Privacy via Interpretation

- **Interpretation** of request **R**, for data **D**, according to access control policy **P** defines privacy
 - May return interpreted data : $I(\mathbf{D})$
 - Nothing
 - **D**
 - A subset of **D**
 - Something derived from **D**
- Interpretation based on access control

Some Definitions

- ❑ **Reversibility** - hiding data by encryption
 - ❑ **Irreversibility** – hiding data by hashing
→ anonymization
 - ❑ **Inversibility** – impossible to re-identify the person except by applying an exceptional procedure restricted to highly trustworthy party → pseudonymization
-
- Linking allows associating one or several pseudonyms to the same person
 - reversion robustness
possibility to inverse the anonymization function
 - inference robustness
data disanonymization by means of unauthorized computation¹⁴

Inference Problem

Inferring sensitive data from non-sensitive data

- Direct attack

- Infer from few records retrieved
- “ n items over k percent” rule

- Indirect attack

- Using Sum, Count, Median to derive information
- Tracker attacks (Intersection of sets)
- Linear system vulnerability—
 apply algebra of multiple equations

Database Linkage Problem

How to prevents *users* to know the private information of an *individual* by linking some public or easy-to-know database with the data they receive legally from the data center.

- main challenge is to achieve a balance between privacy protection and data availability
- check all possible kinds of knowledge that can be derived from the to-be-disclosed data
 - refuse the query
 - modify return data

Definitions

quasi-identifier

- a set of attributes that, in combination, can be linked with external information to re-identify the individuals
- depends on the external information available

k-anonymity

- if every record released cannot be related to fewer than k individuals
- set by the data holder, possibly as the result of a negotiation with other parties
- satisfaction requires knowing how many individuals each released tuple matches

⇒ How to produce a version of private table PT that satisfies k -anonymity wrt quasi-identifier QI ?

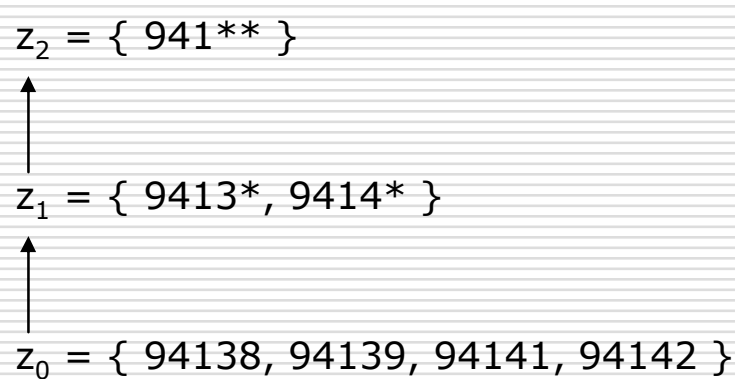
I. Generalization

- each attribute is associated with a *domain* to indicate the set of values that the attribute can assume
 - ground domains
- a set of (generalized) domains containing values and a mapping between each domain and domain generalizations of it
 - for instance, ZIP codes can be generalized by dropping, at each generalization step, the least significant digit

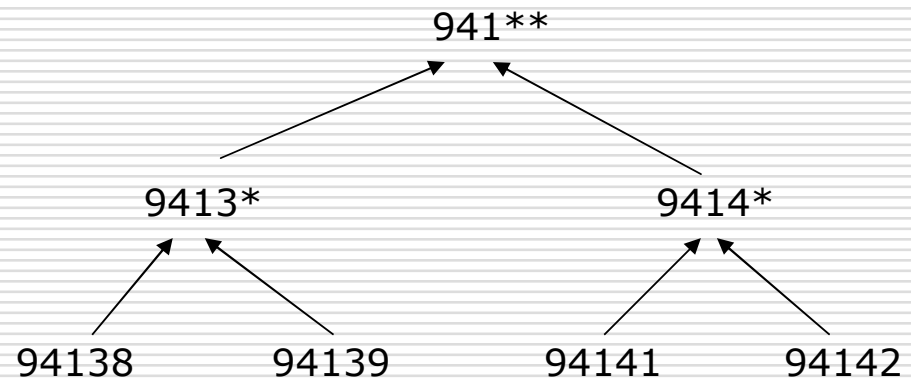
→ generalization relationship \leq_D

Generalization (cont'd)

- domain generalization hierarchy, DGH_D
- value generalization hierarchy, VGH_D

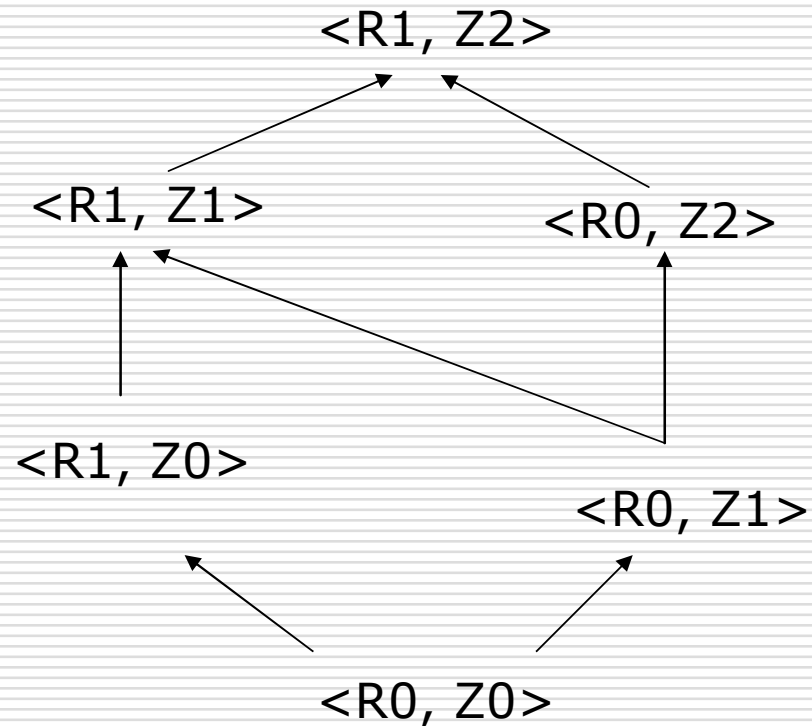


DGH_{z_0}



VGH_{R_0}

Domain Generalization Hierarchy DGH_{<R0,Z0>}



⇒ 3 domain and value generalization strategies

Examples of generalized tables

Race:R ₀	ZIP:Z ₀
asian	94139
asian	94139
asian	94139
asian	94139
black	94138
black	94138
black	94141
black	94141
white	94138
white	94138
white	94142

PT

Race:R ₁	ZIP:Z ₀
person	94139
person	94139
person	94139
person	94139
person	94138
person	94138
person	94141
person	94141
person	94138
person	94138
person	94142

GT_[1,0]

Race:R ₁	ZIP:Z ₁
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*

GT_[1,1]

Examples of generalized tables (cont'd)

Race: R_1	ZIP: Z_1
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*

$GT_{[1,1]}$

Race: R_0	ZIP: Z_1
asian	9413*
asian	9413*
asian	9413*
asian	9413*
black	9413*
black	9413*
black	9414*
black	9414*
white	9413*
white	9413*
white	9414*

$GT_{[0,1]}$

Race: R_0	ZIP: Z_2
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
black	941**
black	941**
white	941**
white	941**
white	941**

$GT_{[0,2]}$

Race: R_1	ZIP: Z_2
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

$GT_{[1,2]}$

k -minimal Generalization

Let $T_j(A_1, \dots, A_n)$ and $T_i(A_1, \dots, A_n)$ be two tables such that T_j is a generalization of T_i . The *distance vector* of T_j from T_i is the vector $DV_{i,j} = [d_1, \dots, d_n]$, where each d_z , $z = 1, \dots, n$ is the length of the unique path between $\text{dom}(A_z, T_i)$ and $\text{dom}(A_z, T_j)$ in the domain generalization hierarchy DGH_{Dz} .

A generalization $T_j(A_1, \dots, A_n)$ is k -minimal iff there does not exist another generalization $T_z(A_1, \dots, A_n)$

- satisfying k -anonymity
- with a distance vector smaller than that of T_j .

A table

Race	Date of Birth	Sex	ZIP	Marital Status
asian	09/27/64	female	94139	divorced
asian	09/30/64	female	94139	divorced
asian	04/18/64	male	94139	married
asian	04/15/64	male	94139	married
black	03/13/63	male	94138	married
black	03/18/63	male	94138	married
black	09/13/64	female	94141	married
black	09/07/64	female	94141	married
white	05/14/61	male	94138	single
white	05/08/61	male	94138	single
white	09/15/61	female	94142	widow

... and its minimal generalization

Race	Date of Birth	Sex	ZIP	Marital Status
asian	64	not released	941**	not released
asian	64	not released	941**	not released
asian	64	not released	941**	not released
asian	64	not released	941**	not released
black	63	not released	941**	not released
black	63	not released	941**	not released
black	64	not released	941**	not released
black	64	not released	941**	not released
white	61	not released	941**	not released
white	61	not released	941**	not released
white	61	not released	941**	not released

Race	Date of Birth	Sex	ZIP	Marital Status
person	[60-64]	female	9413*	been married
person	[60-64]	female	9413*	been married
person	[60-64]	male	9413*	been married
person	[60-64]	male	9413*	been married
person	[60-64]	male	9413*	been married
person	[60-64]	male	9413*	been married
person	[60-64]	female	9414*	been married
person	[60-64]	female	9414*	been married
person	[60-64]	male	9413*	never married
person	[60-64]	male	9413*	never married
person	[60-64]	female	9414*	been married

$GT_{[0,2,1,2,2]}$

$GT_{[1,3,0,1,1]}$

Suppression

- remove data from the table so that they are not released
 - applied at the record level
- to “moderate” the generalization process when a limited number of tuples with less than k occurrences would force a great amount of generalization

A table and its minimal generalization

Race	Date of Birth	Sex	ZIP	Marital Status
asian	09/27/64	female	94139	divorced
asian	09/30/64	female	94139	divorced
asian	04/18/64	male	94139	married
asian	04/15/64	male	94139	married
black	03/13/63	male	94138	married
black	03/18/63	male	94138	married
black	09/13/64	female	94141	married
black	09/07/64	female	94141	married
white	05/14/61	male	94138	single
white	05/08/61	male	94138	single

PT

Race	Date of Birth	Sex	ZIP	Marital Status
asian	09/64	female	94139	divorced
asian	09/64	female	94139	divorced
asian	04/64	male	94139	married
asian	04/64	male	94139	married
black	03/63	male	94138	married
black	03/63	male	94138	married
black	09/64	female	94141	married
black	09/64	female	94141	married
white	05/61	male	94138	single
white	05/61	male	94138	single

GT_[0,1,0,0,0]

Attacks against k -anonymity

- Unsorted matching attack
 - subsequent release of another k -anonymity table may allow direct matching of tuples
- Complementary release attack
 - joining tables on non-Quasi-identifiers
- Homogeneity attacks
 - all individuals have same attribute value
- Background Knowledge Attack

Inpatient microdata

Non-sensitive

Sensitive

Age	Nationality	ZIP	Condition
28	Russian	13053	Heart Disease
29	American	13068	Heart Disease
21	Japanese	13068	Viral Infection
23	American	13053	Viral Infection
50	Indian	14853	Cancer
55	Russian	14853	Heart Disease
47	American	14850	Viral Infection
49	American	14850	Viral Infection
31	American	13053	Cancer
37	Indian	13053	Cancer
36	Japanese	13068	Cancer
35	American	13068	Cancer

4-anonymous Inpatient Microdata

Non-sensitive

Sensitive

Age	Nationality	ZIP	Condition
<30	*	130**	Heart Disease
<30	*	130**	Heart Disease
<30	*	130**	Viral Infection
<30	*	130**	Viral Infection
≥40	*	1485*	Cancer
≥40	*	1485*	Heart Disease
≥40	*	1485*	Viral Infection
≥40	*	1485*	Viral Infection
3*	*	130**	Cancer
3*	*	130**	Cancer
3*	*	130**	Cancer
3*	*	130**	Cancer

Background Knowledge Attack

Homogeneity Attack

k-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute.

l-Diversity Principle

A q^* -block is a set of tuples in T^* whose non-sensitive attribute values generalize to q^* .

A q^* -block is *l*-diverse if it contains at least *l* “well-represented” values for the sensitive attribute *S*.

A table is *l*-diverse if every q -block is *l*-diverse.

- if there are *l* “well-represented” sensitive values in a q^* -block then the attacker needs *l*-1 damaging pieces to infer a positive disclosure
- There are different instantiations of the *l*-diversity principle, e.g. Entropy-*l*-Diversity (information-theoretic notion).

3-Diverse Inpatient Microdata

Non-sensitive

Sensitive

Age	Nationality	ZIP	Condition
≤40	*	1305*	Heart Disease
≤40	*	1305*	Viral Infection
≤40	*	1305*	Cancer
≤40	*	1305*	Cancer
>40	*	1485*	Cancer
>40	*	1485*	Heart Disease
>40	*	1485*	Viral Infection
>40	*	1485*	Viral Infection
≤40	*	1306*	Heart Disease
≤40	*	1306*	Viral Infection
≤40	*	1306*	Cancer
≤40	*	1306*	Cancer

The larger l is the higher is the protection of the sensitive attribute.

[Bund] Bekanntgabe Passagierdaten an die USA

Die folgenden Passagierdaten können die US-Behörden vor dem Abflug vom Reservierungssystem der Fluggesellschaft selbständig abrufen:

1. PNR-Buchungscode (Record Locator Code)
2. Datum der Reservation
3. Geplante Abflugdaten
4. Name
5. Andere Namen im PNR
6. Adresse
7. Alle Zahlungsinformationen
8. Rechnungsadresse
9. Telefonnummern
10. Gesamter Reiseverlauf des besagten PNR
11. Vielflieger-Eintrag (auf geflogene Meilen und Adress(en) beschränkt)
12. Reisebüro
13. Sachbearbeiter Reisebüro
14. Codeshare-Information des PNR
15. Reisestatus des Passagiers
16. Informationen über Splitting/Teilung einer Buchung
17. E-Mail-Adresse
18. Ticketing-Informationen
19. Allgemeine Bemerkungen
20. Ticketnummer
21. Sitzplatznummer
22. Datum der Ticketausstellung
23. Aufstellung nicht angetretener Flüge (no show)
24. Nummer der Gepäckanhänger
25. Angaben zu Ticket ohne Reservation (go show)
26. Andere Service-Angaben (OSI)
27. Besondere Service-Angaben und -Anforderungen (SSI/SSR)
28. Angaben zum Informanten
29. Änderungen am PNR
30. Zahl der Reisenden im PNR
31. Sitzplatzstatus
32. One-way-Tickets
33. Allfällige APIS-Informationen
34. ATFQ (Automatic Ticketing Fare Quote)-Felder (automatische Tarifabfrage)

Anonymous Data Analysis

Record #100031
Khalid Al-Midhar
Saudia Arabia
DOB: 07/12/76

one-way hash
→

Source: Agency #101
Record #100031
Name: cbd034409c22929518fa494f99dc9964
Citizen: b835b521c29f399c78124c4b59341691
DOB: 799709b2e5f26f796078fd815bebf724

#VX1RU9
Khaleed Al-midhar
San Francisco
DOB: 12/07/76
ID: 33000102334

?



[James X. Dempsey and Paul Rosenzweig, 2004]

Anonymous Data Analysis (cont'd)

Data Standardization

“Robert”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29
“ROBERT”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29
“Rob”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29
“Bob”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29
“Bobby”	“Robert”	4ffe35db90d94c6041fb8ddf7b44df29

Variations

07/12/76	07/12/76	799709b2e5f26f796078fd815bebf724
	12/07/76	8ceb0fe202b794c27694a83a5ad91df4
	1976	dd055f53a45702fe05e449c30ac80df9

→ dictionary attacks

Summary

- It is often desirable to make data public for various purposes.
- De-identifying data provides no (strict) guarantee of anonymity
 - released information often contains other data that can be linked to publicly available information to re-identify individuals and inferring information that was not intended for disclosure
- Disclosure limitation techniques
 - encryption, suppression, generalization
 - swapping values, perturbation, rounding, additive noise, ...

Literature & References

- P. Samarati: Protecting Respondents' Identities in Microdata Release. *IEEE Trans. on Knowledge and Data Engineering*. 13(6), 2001; 1010–1027.
- L. Sweeney: “k-anonymity: a model for protecting privacy.” *Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570.
- A. Machanavajjal, J. Gehrke, D. Kifer, M. Venkitasubramaniam: I-Diversity: Privacy beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2007.
- J.X. Dempsey and P. Rosenzweig: Technologies That Can Protect Privacy as Information Is Shared to Combat Terrorism. Legal Memorandum #11, The Heritage Foundation, May 2004
www.heritage.org/Research/HomelandDefense/Im11.cfm
- R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Hippocratic Databases. VLDB 2002.
- R. Agrawal, R. Bayardo, C. Faloutsos, J. Kiernan, R. Rantzaou, R. Srikant: [Auditing Compliance with a Hippocratic Database](#), VDLB 2004
- K. LeFevre, R. Agrawal, V. Ercegovic, R. Ramakrishnan, Y. Xu, D. DeWitt: [Limiting Disclosure in Hippocratic Databases](#), VLDB 2004
- R. Agrawal, R. Srikant: [Privacy-preserving Data Mining](#), SIGMOD 2000
- D. Asonov, J.-C. Freytag: [Almost Optimal Private Information Retrieval](#), PET 2002
- J. He, M. Wang: [Cryptography and Relational Database Management Systems](#), IDEAS 2001
- Tom Rosamalia. Privacy of Data, a business perspective.
www.almaden.ibm.com/institute/pdf/2003/TomRosamalia.pdf