Workshop on anonymization
Berlin, March 19, 2015

Basic Knowledge

Terms, Definitions and general techniques

Murat Sariyar
TMF

Outline

Background

Aims of Anonymization

Relevant terms

Anonymization Techniques

Further Issues

Background

Background

Large amount of person-specific data are collected, both by public institutions and by private entities

Laws and regulations require that some collected data must be made public, for example: Census data

Data sets

- Health-care: Clinical studies, hospital discharge databases
- Genetic datasets: 1000 genomes, HapMap, TCGA, …

Contracts alone cannot guarantee that sensitive data will not be carelessly misplaced. Can anonymization guarantees that?

Sweeney (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	[illegible]icity	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	02142	9/15/61	female	democrat	....
....	....	....	....	....	....	....	

Figure [illegible]e-identifying anonymous data by linking to external data

Public voter dataset

(5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.

There are different communities in which research regarding anonymization is done

- Database community
- Statistical disclosure community
- Cryptography community

Aims of Anonymization

One attempt to define Anonymization

ISO 29100:2011: "Anonymization is the **process** by which personally **identifiable** information (PII) is **irreversibly** altered in such a way that a PII principal can no longer be **identified** directly or indirectly, either by the PII **controller** alone or in collaboration with any other party."

Central aim and problem of anonymization

Aim: to produce **"open data"** whilst mitigating the risks for individuals concerned

Problem: Creating an anonymous dataset whilst retaining as much of the underlying information as required for the task (**usefulness**)

Minimal and optimal anonymization

A table is **minimal anonymous** if it satisfies the given privacy requirement and if the sequence of anonymization operations cannot be reduced without violating the requirement

A table is **optimal anonymous** if it satisfies the given privacy requirement and contains most information according to the chosen **information metric** among all satisfying tables

Finding the optimal anonymization is NP-hard…

Utility metrics

General purpose metric (principle of minimal distortion)

Information loss of generalization $\mathrm{G}\colon \{c_1, \dots, c_n\} \rightarrow p$

$$I(G) = Info\left(S_p\right) - \sum_i \frac{N_{ci}}{N_p} Info(S_{ci})$$

$Info(S) = -\sum_i p_i \log p_i,\ \ p_i$ is the percentage of label i

Special purpose metric: e.g. retain usefullness for classification

=> In general, list of data uses (e.g. regression models, association rules, other data mining techniques, etc.)

Trade-off Metric: maximizes the information gained per each loss of privacy

Relevant terms

Relevant terms: kind of Attributes

Kind of attributes:

(1) Unique Identifiers (e.g., social security number)

(2) Quasi-Identifiers (e.g., Zip-Code) => QIDs

(3) Sensitive attributes (exhibiting a special characteristic)

(4) Non-sensitive attributes

Relevant terms: Quasi-Identifier

OECD-Definition for a Quasi-Identifier:

Variable values or combinations of variable values within a dataset that are not structural uniques but might be empirically unique and therefore in principle uniquely identify a population unit.

Should contain an attribute A if an attacker could potentially obtain A from other external resources.

The choice of QIDs remains an open issue

Risks

What is disclosure risk?

Singling out: isolate records identifying an individual

Record Linkage: classify recs as belonging to the same individual

Attribute Linkage: Infer sensitive values from the existing attributes

Table Linkage: Infer presence of an individual

Probabilistic Inference: Change belief on sensitive information

Attacks are context-specific
Example: Attacks on k-Anonymity

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Anonymization techniques

Anonymization techniques

Randomization

- **Noise addition**
- **Permutation**

Generalization (replacing QIDs with more general values)

- **Aggregation**
- **K-Anonymity (inference attacks are still possible)**
- **L-Diversity (semantic meaning of attributes are not considered: Gastric ulcer, Gastritis)**
- **T-Closeness (mirroring the initial distribution in each equivalence class; skewness attack)**

Suppression

- **Tuple and cell suppression**

Anonymization techniques: Cave

These are criteria not techniques:

- **K-Anonymity**
- **L-Diversity**
- **T-Closeness**

And there is no hierarchy!

- **K-Anonymity protects against identity disclosure**
- **L-diversity and T-Closeness protect against attribute disclosure**

What about Fung et al. (2010) statement:

"…distinct l-diversity privacy model automatically satisfies k-anonymity, where k = l, because each qid group contains at least l records."

?

Anonymization techniques: another listing

Generalization and Suppression (hide some details in QID)

- Replace some values with a parent value in a taxonomy
- Full-domain and local (subtree, cell) generalization
- Suppression (see former slide)

Anatomization and Permutation (structural changes)

- Deassociate the relationship between QIDs and sensitive attributes
- Partition into groups and shuffle sensitive values within each group

Perturbation

- Additive Noise (Randomization; independent of other recs => data streams), Data swapping, synthetic data generation

Anonymization techniques: generalization

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053			
13053			
13068			
13068			

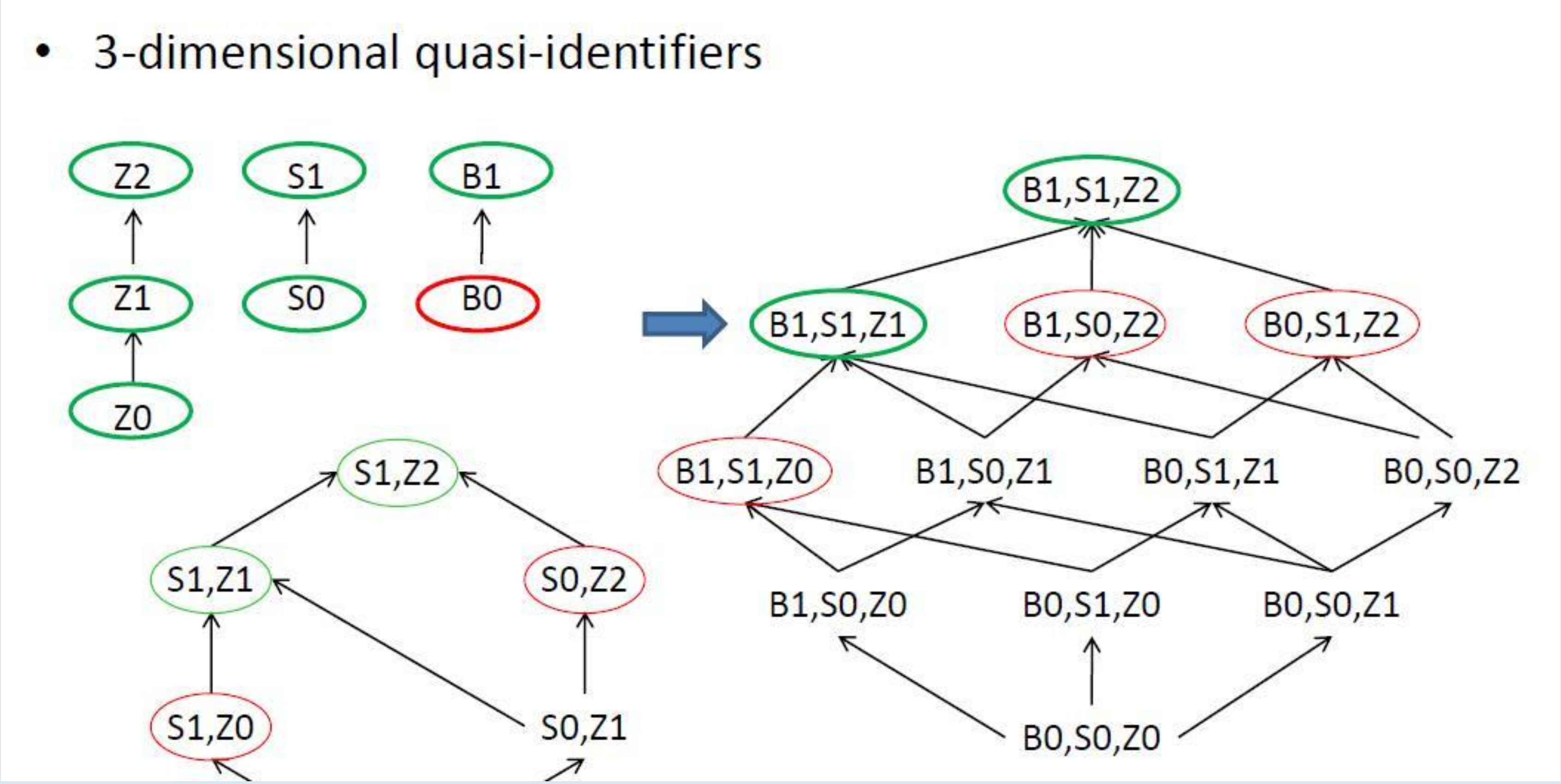
Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Equivalence Class: Group of k-anonymous records that share the same value for Quasi-identifier

Further issues

Anonymization algorithms, e.g. Incognito

Generates the set of all k-anonymous full-domain (multidimens.) generalizations. Bottom up aggregate computation



Genetic data, image data, and alternatives

Is anonymization feasible in this context?

Empirical data showed that a carefully chosen set of 45 SNPs is sufficient to provide matches with a type 1 error of 10^{-15} for most of the major populations across the globe (Pakstis et al. Candidate SNPs for a universal individual identification panel. 2007)

Alternatives:

- **secure computation techniques …**
  - **Secure multipart computation**
- **Fully homormorphic encryption**
- **…**

References

AJ Pakstis et al. Candidate SNPs for a universal individual identification panel. 2007 (Hum Genet.)

BCM Fung et al. Privacy-preserving data publishing: A survey of recent developments. 2010 (ACM Computing Surveys)

Y Erlich and A Narayanan. Routes for breaching and protecting genetic privacy. 2014 (Nature Reviews Genetics)

L Sweeney. K-anonymity: a model for protecting privacy. 2002 (International Journal on Uncertainty, Fuzziness and Knowledge-based Systems)

CC Aggarwal. Privacy-Preserving Data Mining: Models and Algorithms (Advances in Database Systems). 2008 (Springer)