Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support, Fall 2005
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo

Disclosure Control: Privacy

Staal A. Vinterbo

Harvard-MIT Division of Health Science and Technology

Decision Systems Group, BWH

Harvard Medical School

Nov 2005: HST 951/MIT 6.873 Class

Motivation

- Ethics
- US Regulations: HIPAA
- Practicalities: HIPAA and the IRB

Privacy
Defined

Definition (Privacy)

The non-disclosure of the relationship between any explicit identifier of an individual and private data items.

Definition (Private data items)

Attribute values (observations, measurements, comments, etc.) that an individual

- does not want disclosed, and
- the disclosure of would not be in the best interest of the individual.

Background
Ethics

- Hippocrates (ca. 460-377 BC) recognized rights to privacy.
- Warren and Brandeis (HLR, 1890) see the right to privacy as an extension of rights against bodily harm to a right against harm to one's intellectual and emotional life.

Background

Legal

- 1890 Justice Louis Brandeis extolled 'a right to be left alone.'
- Liberty of personal autonomy protected by the 14th amendment to the constitution.
- Privacy Act of 1974 - government
- Gramm-Leach Bliley Act of 1999 - financial institutions
- Fair Credit Reporting Act - consumer reporting agencies
- Children's Online Privacy Protection Act - parents
- Health Insurance Portability and Accountability Act (2000)

Background

Violation Consequences

Hypothetical Scenarios:

- Loss of public "face"
- Loss of employment
- Loss of health insurance

The secondary consequences are numerous.

Background

Anecdotes

From http://www.hipaaps.com/main/examples.html

- Joan Kelly, an employee of Motorola, was automatically enrolled in a "depression program" by her employer after her prescription drugs management company reported that she was taking anti-depressants. (R. O'Harrow, "Plans' Access to Pharmacy Data Raises Privacy Issue," The Washington Post, September 27, 1998, p. A1)
- A banker who also served on his county's health board cross-referenced customer accounts with patient information. He called due the mortgages of anyone suffering from cancer. (M. Lavelle, "Health Plan Debate Turning to Privacy: Some Call For Safeguards on Medical Disclosure. Is a Federal Law Necessary?" The National Law Journal, May 30, 1994, p. A1)

Disclosure Control

Components

Definition (Disclosure Control)

Mechanism by which we regulate the disclosure of information.

Disclosure control has two major aspects:

- Policy - management rules
- Technology - how

Disclosure Control

Policy

The first step to sound disclosure control is to define a policy. A policy can include:

- Access control: who and how. Need to know.
- Communication security: to whom and and to whom not.
- Limited application: what can be done with information.
- Destruction. Lifetime of information.
- Accountability. Who is accountable and what repercussions are there.
- Binding agreements.

Disclosure Control

Technology

Technology that supports a given policy can include:

- Cryptography.
- Access barriers: physical and electronic.
- Uniforms, badges: recognizability.
- Audit trails.
- Transformation of data.

Dissemination of Research Data

Background

Presumption:

Retrospective data is of importance for the advancement of the biomedical field, and health care in general.

Support:

- IOM report 1991
- Literature
- “data analysis” matched 12% of all PubMed publications indexed for the year 2004.

Dissemination of Research Data

Circles of Trust

Definition (Circle of Trust)

Set of entities that you can entrust specific information.

A circle of trust is characterized by

- what information
- trust level
- repercussions

and have associated

- mechanisms of disclosure
- agreements

Dissemination of Research Data

Circles of Trust:Simple Version

Simple version:

- Level 1: Full trust and dissemination, with adequate repercussions for breaches of trust.
- Level 2: No trust and no dissemination.

This version arguably protects against unwanted disclosure, but is also too limited to be practical.

Dissemination of Research Data

The Need for Disclosure Mechanisms

A more nuanced scenario requires mechanisms of disclosure control.

Example

HIPAA requires a *reasonable effort* to minimize breaches of privacy rights to be made before disclosure.

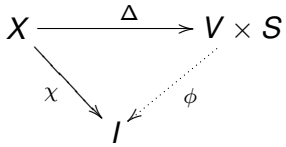
This points to two related issues:

- What constitutes a reasonable effort?
- How do we quantify risk of unwanted disclosure?

Dissemination of Research Data

Privacy Formalized

[Vinterbo(2004)]:



- X - population of interest (patients)
- Δ - data collection machinery (visits, labs, etc.)
- $V \times S$ - data (medical record)
- χ - explicit identifier function (social sec. card)
- I - set of explicit identifiers (social sec. numbers)
- ϕ - method by which adversary infers identity, i.e., $\phi(\Delta(x)) = \chi(x)$ for $x \in X$ (record linkage)

Dissemination of Research Data

Anonymization

Definition (Anonymization)

Anonymization is a process Ψ such that

$$\phi(\Psi(\Delta(x))) \neq \chi(x)$$

Dissemination of Research Data

Anonymization

The preceding definition of anonymization lets us envision two types of anonymization procedures.

- Generalization:
$$\Psi(\Delta(x)) = U \subseteq V \times S.$$
If $\Delta(x) \in U$, then the generalization is *truthful*.
- Property preserving transformations (PPT):
$$\Psi(\Delta(x)) \in W$$
for some W, but $p_i(\Delta(x)) = p_i(\Psi(\Delta(x)))$ for properties p_i.

Dissemination of Research Data

Anonymization

Problem:

We do not know what ϕ is.

This means that we don't know what the adversary is capable of. The two types of anonymization deal with this differently.

- Generalization induces *ambiguity*. The assumption is that if $|I'|$ is large enough, privacy is preserved.
- Property preserving transformations rely on the non-reversibility of Ψ. A simple way of ensuring this is to randomize Ψ.

Dissemination of Research Data

Anonymization

Problem

The utility of an analysis of released data is dependent on the quality of the data.

The consequence of this is that we want to "perturb" the data as little as possible. The consequences are

- Generalization: minimizing information loss while guaranteeing $|I'| > k$ for a given k is hard.
- PPT: Assuming it might possible to make Ψ non-invertible, the properties a are still fixed a priori. Such data is not suitable analyses using properties not on the a priori established list. Also some wanted properties are incompatible with non-invertibility.

Dissemination of Research Data

Modes of Dissemination

How do we disseminate the data?

- One shot dissemination. Drawback: not applicable to large amounts of data.
- Multiple disseminations of the same data. Applicable for large amounts of data. Usually in on-line data bases. Drawback: non-monotonicity of inferences with multiple disclosures,i.e, the conjunction of individually private data items can allow the inference of previously private data.

Dissemination of Research Data

Method Examples

Multiple disclosures data bases must keep track of what has been diclosed and censor subsequent disclosures dependent on this. A common approach is to only disclose aggregates so coarse that this is doable [Denning(1980), Brodsky et al.(2000), Boyens et al.(2004), see for example].

Dissemination of Research Data

Methods

For one shot disclosures there exists several methods and Systems:

- Data Fly [Sweeney(1997)],
- Pram [Kooiman et al.(1997)],
- Argus [Hundepool and Willenborg(1996)],
- cell suppression [Meyerson and Williams(2004)],
- k-ambiguity by clustering [Katirai et al.(2004)],
- decision tree based data swapping [Estivill-Castro and Brankovic(1999)],
- noise addition [Agrawal and Aggarwal(2001)],
- genetic algorithm based generalization [Iyengar(2002)]

Dissemination of Research Data

Example: Cell Suppression

HIV	Zip	Birth Date
Yes	2115	1/23/1974
No	2115	2/25/1965
Yes	2116	2/25/1965

Zip	Birth Date	SSN
2115	1/23/1974	1
2115	2/25/1967	2
2116	2/25/1967	3

Can be linked to produce:

HIV	Zip	Birth Date	SSN
Yes	2115	1/23/1974	1

Dissemination of Research Data

Example: Cell Suppression

HIV	Zip	Birth Date
Yes	2115	*
No	*	*
Yes	*	1/25/1973

Zip	Birth Date	SSN
2115	1/23/1974	1
2115	1/25/1973	2
2116	1/25/1973	3

Dissemination of Research Data

Cell Suppression: How

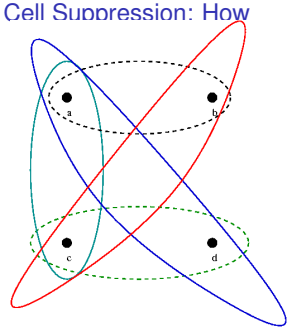
Patient	a	b	c	d	class
1	1	0	0	1	1
2	0	1	0	1	1
3	0	0	1	1	0
4	1	1	1	1	0
5	0	0	0	0	0
6	1	0	1	0	1

Looking at row number 1 we can summarlze how this differs from the other rows as:

Differs	a	b	c	d	class
2	*	*			
3	*		*		*
4		*	*		*
5	*			*	*
6			*	*	

Dissemination of Research Data

Cell Suppression: How



- Find two sets in Figure such that their union is minimal, and at least one of them is drawn with a solid line
- Delete the cells in the row for patient 1 corresponding to the column names found in the union of the sets

Result:

Patient	a	b	c	d	class
	*	0	*	*	1

Dissemination of Research Data

Non-disclosable data?

Consider single nucleotide polymorphism data. These are essentially binary strings derived from our genetic material that allow the distinction between you and me. Currently there are around a million of them that are known. Considering that a lower bound for what we need to distinguish between all humans is 33, it might be problematic to release such data. Hence we should start looking at nuanced models of disclosure control that do not only rely on technical anonymization algorithms. Unfortunately, these might require controversial political instruments.

Institutional Internal Review Board

Function: make sure that research funded by or through the institution is according to legal and ethical standards.

Example Submission

- Summary of proposed study
- Information about co-investigators
- Information about data use, identifiable data in particular
- Risks to human subjects and plans on how to deal with these
- Enrollment, women and children, ethnicity
- ...

Not to be underestimated.

Breaches are serious: Research activities at institutions can, and are, shut down due to breaches of rules and regulations regarding human subject research.

References

Dakshi Agrawal and Charu C. Aggarwal.
On the design and quantification of privacy preserving data mining algorithms.
In *Proceedings of the Twenteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 21-23, 2001, Santa Barbara, California, USA*, pages 247–255. ACM, 2001.

Claus Boyens, Ramayya Krishnan, and Rema Padman.
On privacy-preserving access to distributed heterogeneous healthcare information.
In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 6*. IEEE, 2004.

Aleksander Brodsky, Csilla Farkas, and Sushil Jajodia.
Secure databases: Constraints, inference channels, and monitoring disclosures.
IEEE Transactions on Knowledge and Data Engineering, 12(6):900–919, 2000.

Dorothy E. Denning.
Secure statistical databases with random sample queries.
ACM Transactions on Database Systems, 5(3):291–315, 1980.

V. Estivill-Castro and L. Brankovic.
Data swapping: Balancing privacy against precision in mining for logic rules.
In M. Mohania and A.M. Tjoa, editors, *Data Warehousing and Knowledge Discovery DaWaK-99*, Lecture Notes in Computer Science, pages 389–398. Springer Verlag, 1999.

A. J. Hundepool and L. C. R. J. Willenborg.
Mu- and tau-argus: Software for statistical disclosure control.
In *Third International Seminar on Statistical Confidentiality at Bled*, 1996.

Vijay S. Iyengar.
Transforming data to satisfy privacy constraints.
In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 279–287, July 2002.

Hooman Katirai, Robert Fisher, and Staal A. Vinterbo.
A toolkit for the mathematics of privacy.
Technical Report DSG-TR-2004-14, Decision Systems Group, Brigham and Women's Hospital, Decision Systems Group, Dep. Radiology, Brigham and Women's Hospital. Boston, MA., December 2004.

P. Kooiman, L. Willenborg, and J. Gouweleeuw.
Pram: a method for disclosure limitation of microdata.
Rsm-80330, Statistics Netherland, 1997.

Adam Meyerson and Ryan Williams.
On the complexity of optimal k-anonymity.
In *PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228, New York, NY, USA, 2004. ACM Press.

L. Sweeney.
Guaranteeing anonymity when sharing medical data, the datafly system.
Proc AMIA Annu Fall Symp, :51–5, 1997.

Staal A. Vinterbo.
Privacy: A machine learning view.
IEEE Transactions on Knowledge and Data Engineering, 16(8):939–948, August 2004.