

Disclosure Control in Disseminated Relational Medical Data

Staal A. Vinterbo, Ph.D.



Decision Systems Group
Brigham and Women's Hospital,
Harvard Medical School

Overview

- Example
- Why now?
- Context
- Overview of approaches
- Example of a formal approach to disclosure control

Example

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

=

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

Example

- Cell suppression
 - * denotes suppressed value

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

Example

- Generalization + cell suppression
 - Zip-generalization:
 - abcd becomes abc*
 - Hierarchy:
 - ****
 - a***
 - ab**
 - abc*
 - abcd

HIV	Zip	Birth Date
Yes	211*	*
No	211*	*
Yes	211*	1/25/1973

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

Example

- Removal of semantics + Cell Suppression
 - "Birth Date" becomes "Y"
 - Y(MM/DD/YYYY) = character(YYYY - 1970)

HIV	Zip	Y
Yes	2115	d
No	*	c
Yes	*	c

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

Regulations

- 1996 - Health Insurance Portability and Accountability Act (HIPAA)
 - Authorizes federal national standard for medical record privacy
- 1999 – Department of Health and Human Services
 - Proposes rules
- Dec. 2000 – Federal Register
 - Issues regulations on safeguarding confidentiality of patient data

HST 951 Spring 2002

Staal A. Vinterbo

Protection Contexts I

- Controlled
 - Circle of trust
 - Proper application
 - Proper storage
 - Proper transmission
 - Mechanisms for maintaining circle integrity
 - Policy (“need to know” (HIPAA/IOM))
 - Technology (encryption, monitoring, trails)
 - Retroactive “punishment” for breach of policy/trust

HST 951 Spring 2002

Staal A. Vinterbo

Protection Contexts II

- Uncontrolled
 - No trust
 - No effective retroactive “punishment”
 - Mechanisms
 - Disclosure control
 - Policy: minimal need to know (HIPAA/IOM)
 - Minimal need to know: hard to establish (research)
 - Policy enforced: “definitely should not know”: patient identity

HST 951 Spring 2002

Staal A. Vinterbo

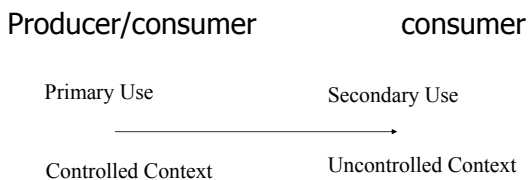
Data Use Contexts (IOM)

- Primary use of data
 - Care delivery
 - Care management
 - Administration of billing and resources
- Secondary use of data
 - Education
 - Regulation
 - Research
 - Policy
 - Industry

HST 951 Spring 2002

Staal A. Vinterbo

Connecting Control and Use



HST 951 Spring 2002

Staal A. Vinterbo

Controlling: What

- Inferences about data items we want to keep hidden
 - Control disclosure of protected data items (cells)
 - Control linking to outside data

HST 951 Spring 2002

Staal A. Vinterbo

Disclosure Control: Problem

- Disclosure control generally incurs a loss of information.
- This loss must be balanced by the utility of the resulting data
- Proof of anonymity

Disclosure Control Mechanisms I

- Data suppression (removal)
 - Column – suppression of attribute (Su 91)
 - Row – suppression of case (Sweeney 97)
 - Cell – suppression of particular attribute value for a particular case (Øhrn 99; Hundepool 96)
 - Column and row suppression are special cases of cell suppression
- Often presumed: a suppressed value represents the absence of restrictions with respect to the actual value.

Disclosure Control Mechanisms II

- Generalization of attribute (Sweeney 97; Hundepool 96)
 - Dependent on a hierarchy on attribute values where a value represents the possibility of all values below in the hierarchy
 - Ambiguity is increased by transforming an attribute to assign values higher in the hierarchy to chosen elements
 - Cell suppression could be thought of as being generalization in a binary depth hierarchy with a single top element

Disclosure Control Mechanisms III

- Removal of semantics (Armstrong 99; Kooiman 97)
 - The attribute is transformed in order to obscure the meaning of the attribute
 - Transformation needs to be non-invertible for the recipient of transformed data
 - Transformation needs to preserve some utility of the attribute (e.g., statistical properties allowing clustering)
 - Can encompass change of attribute label

Disclosure Control Mechanisms: Properties

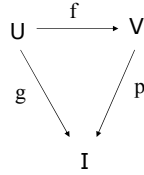
- Cell suppression:
 - maximal information loss in each suppressed cell
- Generalization:
 - Generally smaller information loss per changed entry over cell suppression
 - Needs predetermined hierarchy

Disclosure Control Mechanisms: Properties

- Removal of semantics
 - Needs knowledge about intended use
 - Potential loss of utility due to
 - Lost semantics
 - Lost data characteristics
- All: Potentially computationally expensive

Formal View

- U – population
- X – patients
- V – feature space
- I – identifiers
- Known g and p



$$g(x) = p(f(x)) \text{ for all } x \text{ in } X$$

$$h = \{(x,v) \mid g(x) = p(v)\}$$

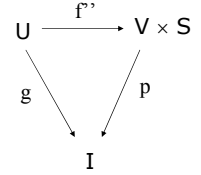
HST 951 Spring 2002

Staal A. Vinterbo

Problem

$$f''(x) = (f(x), s(x))$$

secret: $(x, s(x))$ for any x



g is 1-to-1
=>
h is a 1-to-1 function on X

for $(f(x), s(x))$ we can then
find x such that $h(x) = f(x)$

HST 951 Spring 2002

Staal A. Vinterbo

Approaches

- Find f' from X to V such that no useful h can be constructed and disseminate $(f'(x), s(x))$ instead
- $f'(x) \neq f(x)$
 - Permutation of first coordinate in f
 - $f'(x) \notin f(X)$
- Utility of $f'(X)$ problem remains

HST 951 Spring 2002

Staal A. Vinterbo

Measuring Utility

- Function i from V to \mathbf{N} measuring the utility of a point in v
- Utility $i(f'(X))$ of f' is then

$$i(f'(X)) = \sum_{x \in X} i(f'(x))$$

- Find the f' that maximizes utility
- Can data utility be defined in terms of utilities of individual points?

HST 951 Spring 2002

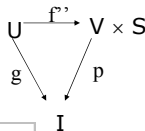
Staal A. Vinterbo

Utility: Example

$$V = \{0,1,*\} \times \{0,1,*\}$$

$$S = \{0,1\}$$

$$f(X) = \{00, 01, 10, 11\}$$



	g	f		s
1	1	0	0	1
2	2	0	1	0
3	3	1	0	1
4	4	1	1	0

Problem: find f' yielding minimum information loss

HST 951 Spring 2002

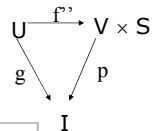
Staal A. Vinterbo

Utility: Example

$$V = \{0,1,*\} \times \{0,1,*\}$$

$$S = \{0,1\}$$

$$f(X) = \{00, 01, 10, 11\}$$



	g	f		s
1	1	*	0	1
2	2	*	1	0
3	3	*	0	1
4	4	*	1	0

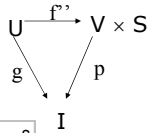
Problem: $X/f' = X/s$

HST 951 Spring 2002

Staal A. Vinterbo

Utility: Example

$V = \{0,1,*\} \times \{0,1,*\}$
 $S = \{0,1\}$
 $f(X) = \{00, 01, 10, 11\}$

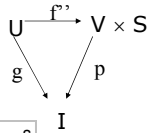


	g	f		s
1	1	0	*	1
2	2	*	1	0
3	3	1	*	1
4	4	*	1	0

Problem: Lack of redundancy

Utility: Example

$V = \{0,1,*\} \times \{0,1,*\}$
 $S = \{0,1\}$
 $f(X) = \{00, 01, 10, 11\}$



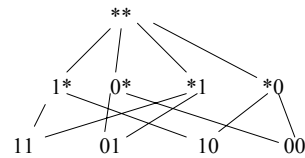
	g	f		s
1	1	0	*	1
2	2	0	*	0
3	3	1	*	1
4	4	1	*	0

Problem: ?

Utility: Example Formalized

- Attributes $a_i: U \rightarrow V_i$
- $V = V_1 \times V_2 \times \dots \times V_m$
- V_i contains special element $*$
- $i(v) = i(v_1, v_2, \dots, v_m) = |\{j \mid v_j \neq *\}|$
- $(v_1, v_2, \dots, v_m) \leq (w_1, w_2, \dots, w_m)$
iff $v_j = w_j$ or $w_j = *$

Semantics



- $V = \{0,1,*\} \times \{0,1,*\}$, $f(X) = \{00, 01, 10, 11\}$
- "meaning" of v is $m(v) = f(X) \cap \{w \mid w \leq v\}$
- $m(0*) = \{01, 00\}$
- $i(0*) = 1$, $i(11) = 2$

Requirements

- Existence of k number of y such that
 - $s(y) \neq s(x)$ and $f(y) \leq f(x)$ (*necessity*)
- $f(x) \leq f'(x)$ in order to disallow cell swapping type solutions (*generalization*)

Some Remaining Problems

- Theoretical:
 - Analyze parameters in necessities (e.g., k)
 - Find sufficiency requirements that minimize loss
 - Measures of utility and anonymity
- Practical:
 - Find efficient algorithms (low exponent polynomial)

Wish List

- Theory of privacy needs accompanied by methods of ensuring these
- Theory of data application requirements
- Theory connecting the above

References

- 1. Armstrong MP, Rushton G, Zimmerman DL. Geographically Masking Health Data To Preserve Confidentiality. *Statistics in Medicine*. 1999;497-525.
- 2. Su T-A, Ozsoyoglu G. Controlling FD and MVD Inferences in Multilevel Relational Database Systems. *IEEE Transactions on Knowledge and Data Engineering*. 1991;3:474-485.
- 3. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly System. *Proc AMIA Annu Fall Symp*. 1997:51-5.
- 4. Sweeney L. Weaving technology and policy together to maintain confidentiality [see comments]. *J Law Med Ethics*. 1997;25:98-110, 82.
- 5. Hundepool AJ, Willenborg LCRJ. MU- and TAU-Argus: Software for Statistical Disclosure Control. . *Third International Seminar on Statistical Confidentiality at Bled*. Bled; 1996.
- 6. Kooiman P, Willenborg L, Gouweleeuw J. PRAM: a method for disclosure limitation of microdata. : Statistics Netherland; 1997:19.
- 7. IOM. The computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition, 1997.