

Patient Record Data: Statistical Disclosure Control for Remote Data Access

Duncan Smith, Kingsley Purdam, Mark Elliot

CCSR

University of Manchester

Context

- Medical data is often be viewed as sensitive data.
- Therefore the protection of patient privacy is paramount.
 - Legal
 - Moral
- However the protection of privacy in medical records has inhibited service provision and research.
- There is a need to provide access to patient record data without compromising confidentiality.

CLEF Consortium

Approximately 40 Strong consortium from

- University of Manchester
- University of Sheffield
- University College London
- Open University
- Royal Marsden Hospital, London

CLEF Consortium

- Text mining
- HPC
- Medical Informatics
- Computer Security
- SDC
- Medicine

General Purpose

- To provide a system for allowing research access to patient data, whilst maintaining privacy.
- Electronic health records
- Texts such as referral letters and other clinical texts
 - Text mining system convert to microdata

More specifically...

- CLEF is focused on developing methods for managing and using pseudonymised repositories of long term patient histories which can be linked with genetic, genomic, and image data.
- Specifically, CLEF aim to establish a pseudonymised repository of cancer patient histories.
- The main users of CLEF are to be clinicians and researchers.
 - Clinicians will be able to request detailed patient records over a person's whole life.
 - Researchers will be able to select particular samples of patients by condition, symptoms, treatment and by particular demographics or genomic information.

Our own research

- Reviewing the risk context
 - Evaluating the externalities of the disclosure risk.
 - Reviewing current practice.
 - Building user models.
- Building an appropriate architecture for control the risk of disclosure.

A Model of Disclosure Control for Remote Access

Disclosure

- Disclosure
 - The making available of sensitive information about an entity to anyone whom the entity would reasonably not wish it to be made available
- SDC
 - Any set of methods designed to control (manage) the risk of disclosure

- Logical inference
 - An intruder can make inferences with certainty (given their prior knowledge about a target / the data, and no data / world divergence)
- Stochastic inference
 - An intruder might be ‘highly’ confident in some hypothesis regarding an individual (given their prior knowledge about a target / the data, and some assumptions about the data / world divergence)

Identification

- Identification is the association of an entity with a specific record in a dataset
- Legislation tends to concentrate on the risk of identification (e.g. the UK Data Protection Act)
- Identification does not imply disclosure

Attribution

- Attribution is the association or disassociation of a set of attributes with a member of some population
- Attribution with regard to previously unknown attribute levels IS disclosure

A 2-way table

		Department			Row sum
		A	B	C	
Profession	Lawyer	18	4	2	24
	Accountant	2	1	0	3
Col sum		20	5	2	27

Multiple table margins

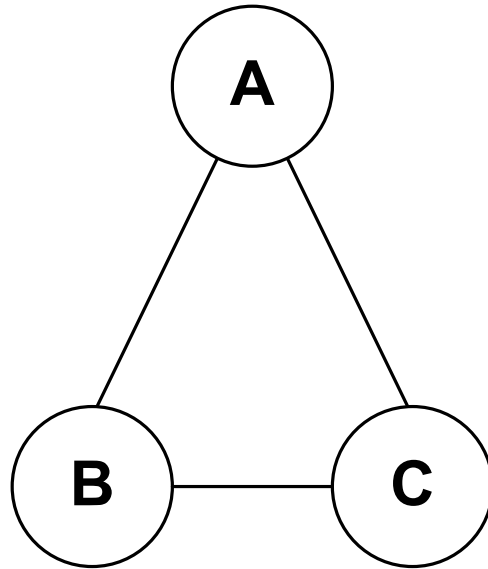
- Releasing only marginal tables can disguise the true counts in the ‘base’ table
- Integer Linear Programming methods can be used to generate lower and upper bounds on the counts in the base table (computationally demanding)
- For certain tabular releases bounds can be calculated very efficiently

- Bounds easily calculated for **decomposable graphical** cases
 - Pointwise addition and subtraction for lower bounds
 - Pointwise minimum for upper bounds

		Department			Row sum
		A	B	C	
Profession	Lawyer	17	2	0	24
	Accountant	0	0	0	3
Col sum		20	5	2	27

		Department			Row sum
		A	B	C	
Profession	Lawyer	20	5	2	24
	Accountant	3	2	2	3
Col sum		20	5	2	27

Non-graphical case



- All 2×2 marginals of a $2 \times 2 \times 2$ table
- A complete subgraph (*clique*) without an individual corresponding table

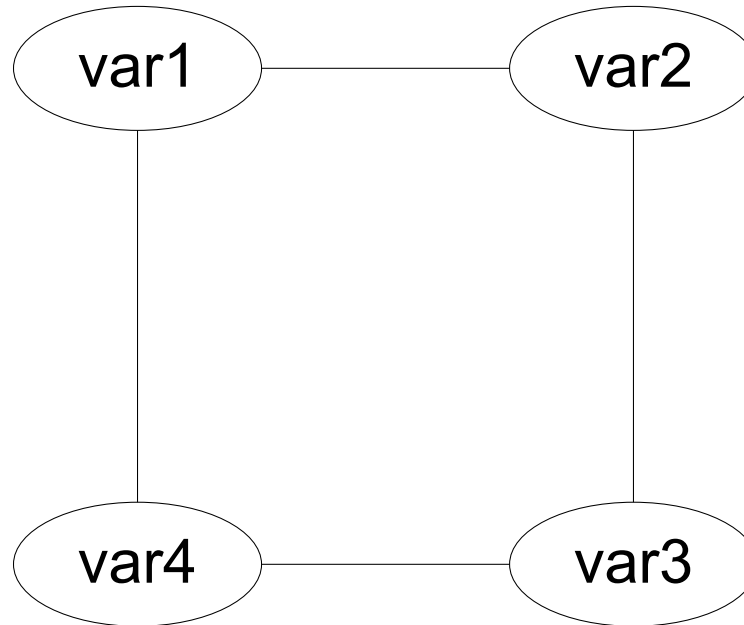
	Var1	
Var2	A	B
C	3	9
D	2	2

	Var1	
Var3	A	B
E	1	10
F	4	1

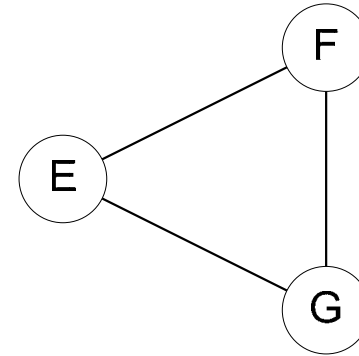
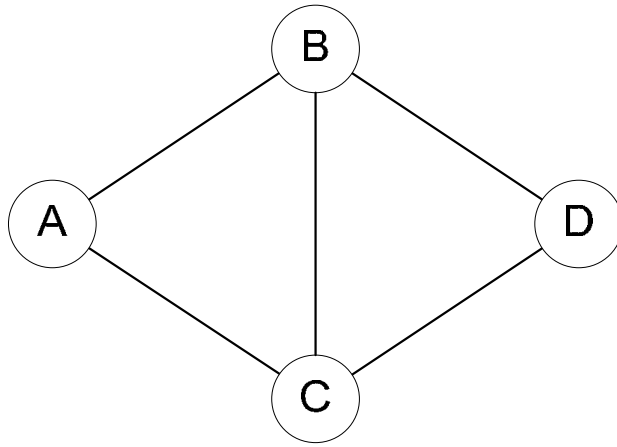
	Var2	
Var3	C	D
E	8	3
F	4	1

	Var1 and Var2			
Var3	A, C	A, D	B, C	B, D
E	0	1	8	2
F	3	1	1	0

Non-decomposable case



- A graph is decomposable if, and only if, it contains no unchorded cycles of length >3

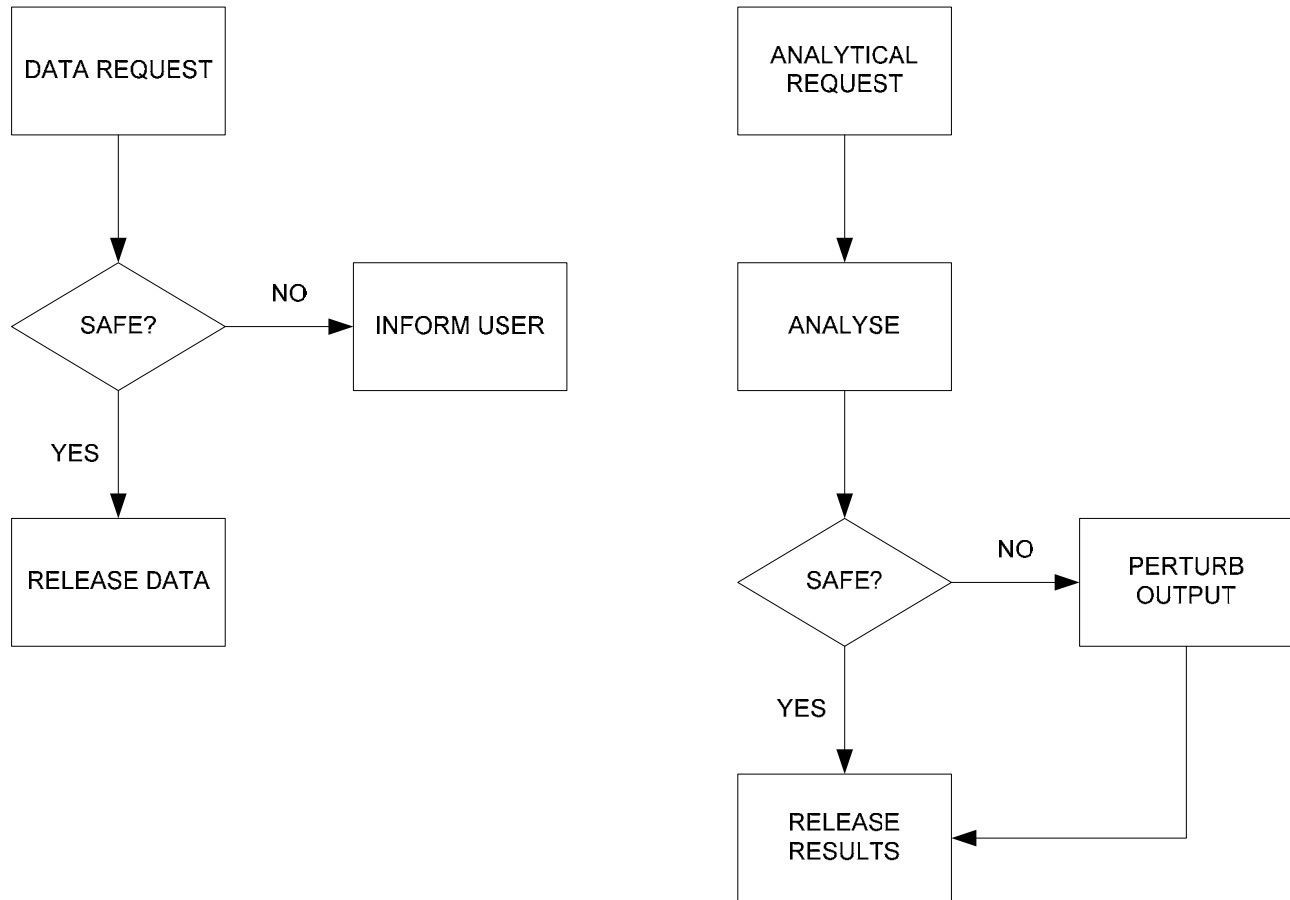


General approach

- Use risk measures that are based on cell bounds
 - Decomposable graphical cases are far easier to assess
 - Other cases can be assessed by considering the bounds for decomposable graphical ‘supersets’
- Risk assess distinct user groups separately (assume they will not share information)
 - User groups might be individual research groups or wider groups such as ‘the media’
 - Use different risk measures for different groups

- Allow users a degree of flexibility in querying the repository
 - Users can investigate the sets of tables that are allowed
 - Helps to prevent the user ‘painting themselves into a corner’
- Avoid imposing our utility function on the user
 - Distinct user groups and the ability to query the repository hypothetically

Architecture



Intruder detection

- ‘Suspicious’ requests flag the possibility of an intrusion attempt
- Bayesian classifier
- Model determination using Metropolis-Hastings and / or knowledge elicitation

Concluding remarks

- Outline of the issues and possible solutions for the disclosure risk issues associated with remote access to medical data repositories.
- functional model for disclosure control within medical data repositories
 - pilot software that instantiates that model