



The Disclosure Risk Issues Posed by the Grid (ESRC)

Mark Elliot, Stephen Pickles, Kingsley Purdam and Duncan Smith
Funded by the ESRC E-Science Pilot Project
Nov 2003-Oct 2004
www.ccsr.ac.uk/research/griddisclosure.htm

The Disclosure Risk Issues Posed by the Grid (ESRC)

- Defining the Grid
- Grid technology opens up a range of opportunities to enhance existing data sources and data quality to inform research, policy and service delivery
- Possibilities at present include: linking data and data sets online, distributed data storage and data processing for large-scale data sets and complex analyses, data mining across different data sets and real time data updates

- The use of the Grid in handling personal data raises a number of new issues in respect of privacy and disclosure control
 - Different data sets are likely to have been collected under different terms of use
 - Different data sets are also likely to contain variables which have different levels of sensitivity and different levels of disclosure risk.
 - Remote processing access also raises new issues
- The main aim of the pilot research is the development of methods for estimating the risk impact of the use of overlapping datasets. This will require both theoretical work and simulation

Methods

- 1. The Policy Issue:** The collection and analysis of confidentiality agreements across a number of case study surveys and an overview of the approach to disclosure risk across the ESRC pilot projects. This includes semi-structured interviews with database managers.
- 2. Investigating the data:** A review of publicly available data in the UK using online search techniques and form field analysis.
- 3. Developing the methodology:** An investigation of linkage between distinct databases.

1. Emerging Policy Issues

1. Diversity of grid applications and grid structures under development

2. Wide range of data types: health, finance, consumption, plant biology

3. Disclosure issues emerging

Wider access to high-powered computing increases the scope for linking data and statistical disclosure

Definition of what is sensitive data is problematic

Grid is perceived as secure environment

Data access and sharing proving a challenge

Trust\Personal relationships proving key but backed up by legal agreements

Ownership\access of linked datasets

4. Engagement with Grid technology limited in key data agencies

2. Investigating the data

- Increase in the collection and use of individual level data and possibly disclosive linking of data across public and private sector.
- New types of data are being collected: DNA, CCTV images, transaction data, consumption data and communication data.
- Individual data is increasingly becoming a commodity.
- Definition and regulation of individual data under the Data Protection Act is proving a challenge to data protection agencies and data holding organisations.

Your Help - Pilot Project Input

- Grateful for ongoing input from pilot projects on:
 - Policy issues as they emerge
 - Technological solutions used to prevent disclosure
 - Challenges over come
 - Good practice in reaching data access agreements

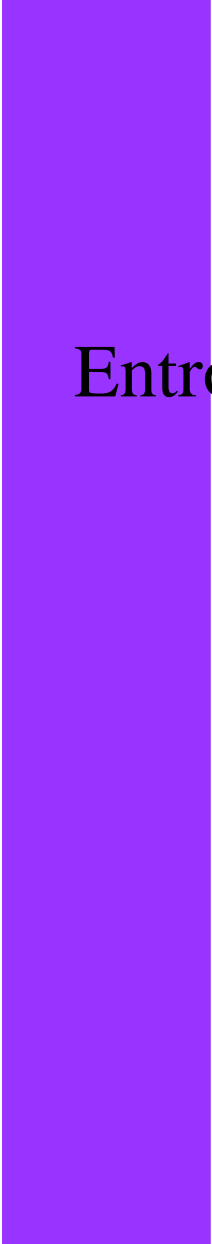
Contact: kingsley.purdam@man.ac.uk

3. Developing A Disclosure Control Methodology

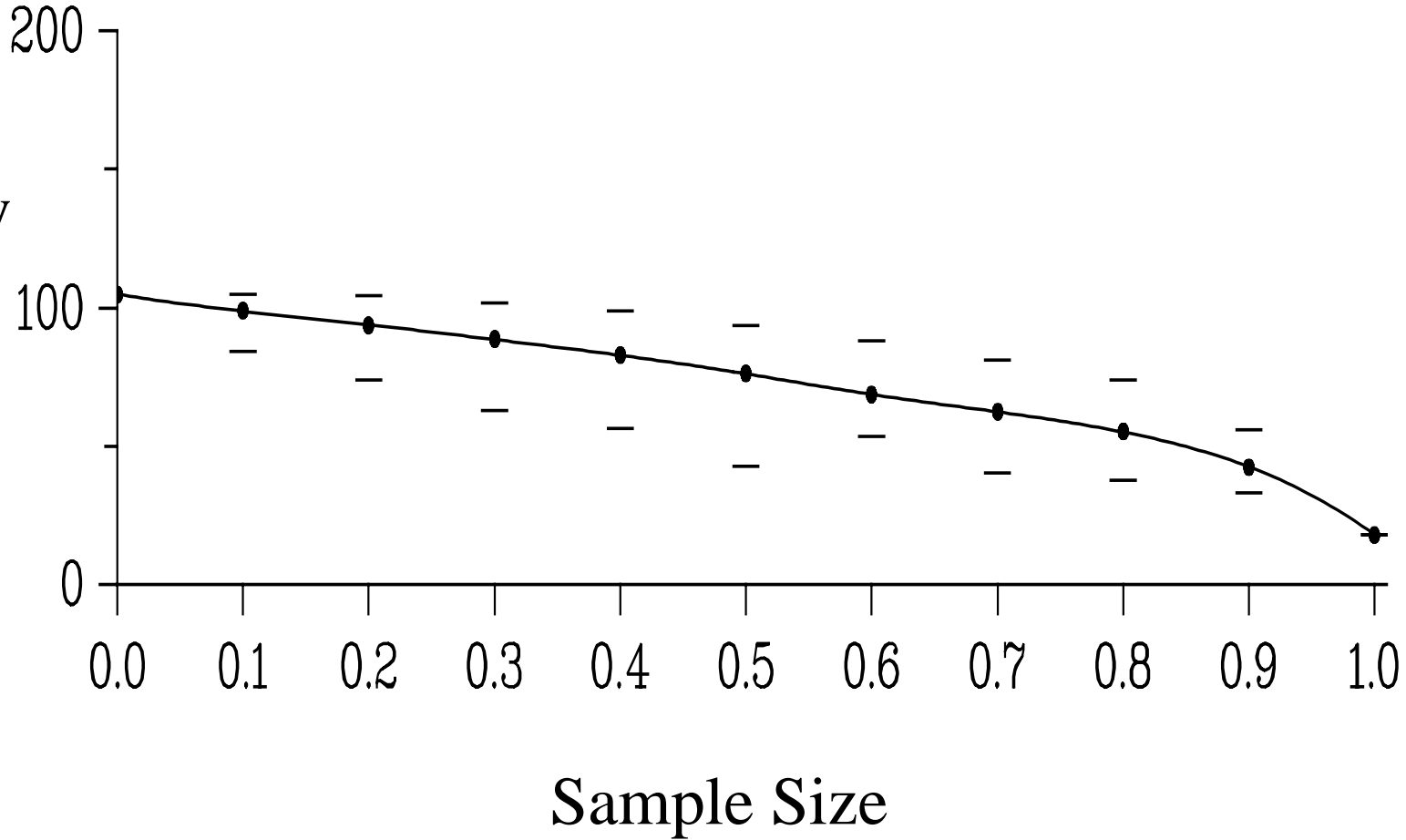
Experiment 1

- **Recovery of information via ‘subtraction’**
- **Known values in tabular data can be disclosive**
 - **E.g. A particular value is the turnover within a certain industry in a given geographical area**
 - **Only two relevant firms exist in the area, and each can recover the competitors turnover by subtracting their own turnover**

- **Known counts in tabular data can be disclosive**
 - Zero counts in tabular data can lead to attribution
 - Individuals in the relevant population who are known to an intruder can be subtracted, leaving a disclosive table that applies to the residual population
- **Bounds on cell values (given 3 marginal tables) calculated**
- **An entropy-based measure of intruder uncertainty is used**
- **Effect on entropy of various levels of intruder knowledge (a known sample) is investigated**



Entropy



Experiment 2

- **Linkage of records between distinct samples of microdata**
- **Only unique matches were considered (as these are generally most risky)**
- **Various marginal / sample tables from lifestyle / aggregated data were incrementally added to the data (some tables were perturbed)**

- **A simple Bayesian approach is used to generate posterior probabilities of a correct match**
- **Unique matches ranked according to true / estimated probabilities**
- **Highly significant rank correlations for some sets of tables**
- **The analysis was repeated for various perturbation methods**

Interim Conclusions

Here are a few notes to edit\comment

Implications

- New disclosure challenges posed by new technology
- Need for legal regulation and codes of practice
- Disclosure control should be seen as integral to grid design
- Scope for in-time on line disclosure control?
- More research needed

Completion Oct 2004

- Grateful for ongoing input from pilot projects on issues
- Dissemination Plans
- Input into development of Grid policy and technology