

Confidential Data Access Using Grid Computing: An Outline of the Issues and Possible Solutions

Mark Elliot, Kingsley Purdam, Duncan Smith
CCSR, University of Manchester

Mark.Elliot@manchester.ac.uk

Cathie Marsh Centre for Census and Survey Research, University of
Manchester

Overview

- Confidentiality and Privacy issues and opportunities arising from the use of grid computing.
- Monitored remote access systems
- Data environment analysis

Some Key Questions

- 1) What new data possibilities does grid computing provide and what confidentiality implications do they have? (1st PDP)
- 2) How could the grid computing be used to enable disclosure risk assessment and control? (2nd PDP)
- 3) How could grid computing enable a *data intruder*?
- 4) What are the possibilities and issues provided by remote access? (CLEF project and further funding)

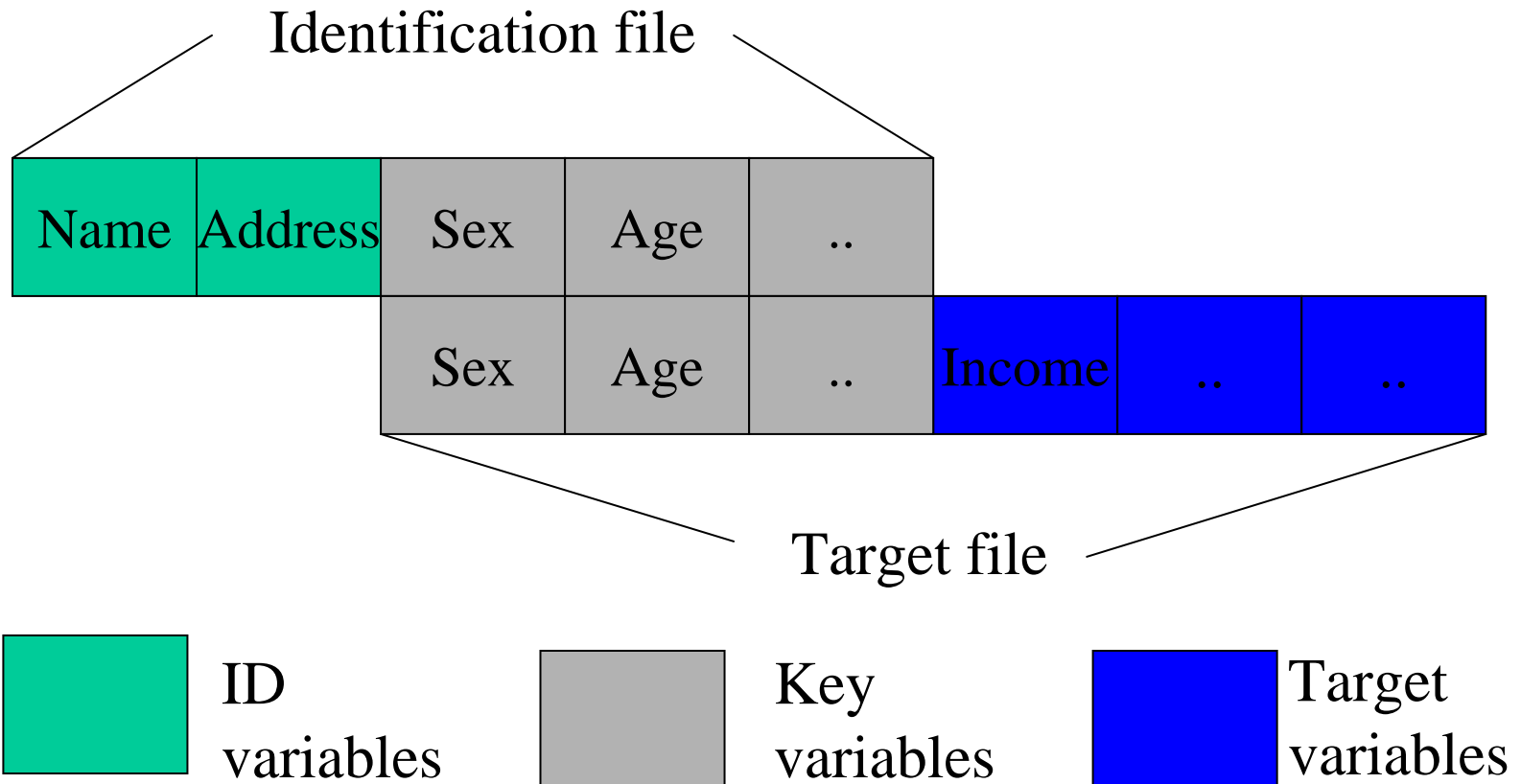
Data Data Everywhere...

- **Massive and exponential increase in data;** Mackey and Purdam(2002); Purdam and Elliot(2002).
 - These studies have led to the setting up of the data monitoring service.
- **Singer(1999) noted three behavioural tendencies:**
 - Collect more information on each population unit
 - Replace aggregate data with person specific databases
 - Given the opportunity collect personal information
- **Purdam and Elliot (2003) add:**
 - Link data whenever you can

New data

- One of the key potentials for e-social science is the possibility of bringing together different data sources through linking and fusing.
- However, this is precisely the disclosure risk situation.

The Disclosure Risk Problem: Type I: Identification



The Disclosure Risk Problem: Type II: Attribution

Income levels for two occupations				
	High	Medium	Low	Total
Accademics	0	100	50	150
Pop Stars	100	50	5	155
Total	100	150	55	305

The Disclosure Risk Problem: Type II: Attribution

Income levels for two occupations				
	High	Medium	Low	Total
Academics	1	100	50	150
Pop Stars	100	50	5	155
Total	100	150	55	305

New Access modalities: a solution?

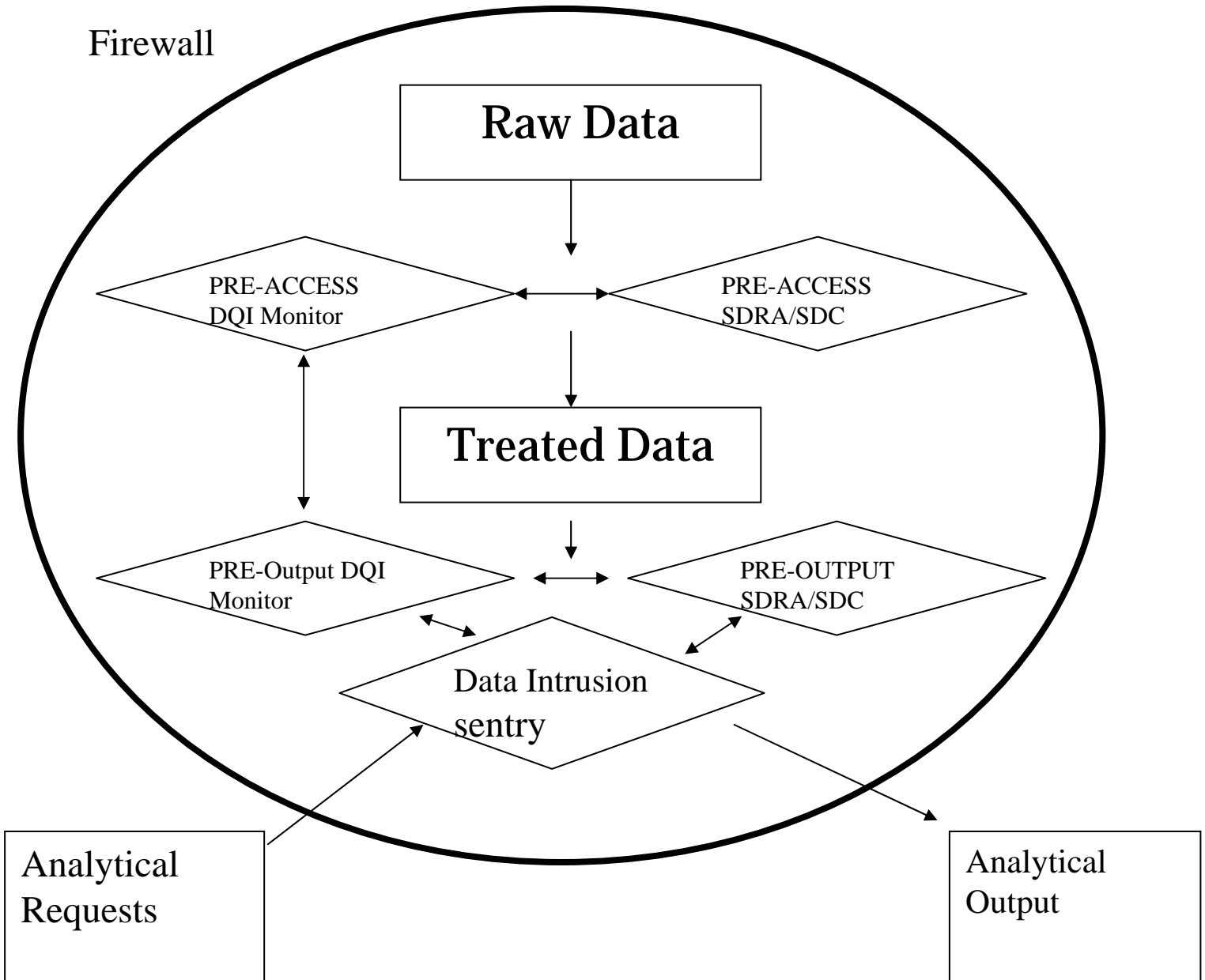
- Virtual remote access has the potential to provide a safe setting model for data access.
- New question:
 - how safe is that output?

Pilot Demonstrator project I

- Our pilot project work shows that adding a third data set tends to increase the linkability of two other datasets.
- i.e the more data concerning a given population existing in a given data environment the greater the disclosure risk

Monitored remote access systems

Tentative architecture for complete system for disclosure control in remote access systems.



Data Intrusion Detection

- Virtual access allows the possibility of monitoring use.
- Usage can be analysed for patterns resembling intrusion (similar to fraud detection).

Data Environment Analysis

- The increasing availability of personal information has impacts on the disclosure control problem in 4 ways:
 1. Decrease in sensitivity of information
 2. Decrease in value to an intruder
 3. Increase in probability of intruder access to key data
 4. Increase in amount of key data intruder has access to

Data Environment Analysis

- Need to move with the technology from:
 - One shot analyses of individual datasets
 - Ongoing analyses of the data environment
- The question is not “How safe is my data” but “How disclosive is the data environment?”.
- A process of data monitoring is one aspect of this.

- **DEA provides a measure of the amount and type of individual information in**
 - the public domain
 - restricted access datasets and
 - commercially available data
- **Metadata are generated through**
 - form analysis
 - metadata questionnaires
 - web-crawling software.
- **Ultimately the process could be automated and tailored to specific grid computing systems.**

DEA Interface

The screenshot shows a web browser window displaying the CAPRI Confidentiality And Privacy Group website. The page features a navigation menu on the left with links for Home, Background, Service Development, Example uses, Datasets, and Contact. The main content area is titled "Data Environment Analysis — Variable Database Search Tool" and contains a detailed description of the tool's purpose and a search interface. The search interface includes a "1. Select Dataset" section with radio buttons for "General" (selected), "employment", and "health". Below this is a "2. Select subject" section with a dropdown menu showing "age" and a "show" button. A link for "General - Age example results page" is provided. At the bottom, there is a note that the data sets can be downloaded and a footer with the University of Manchester logo and name.

CAPRI Confidentiality And Privacy Group

Home
Background
Service Development
Example uses
Datasets
Contact

Data Environment Analysis — Variable Database Search Tool

Data Environment Analysis allows statistical agencies, data providers and data protection organisations an up to date assessment of the availability of individual level data. This is a prototype data environment variable search interface. It allows specific searches of the data environment metadatabase of individual level data. The individual level metadatabase has been compiled as part of a scoping study conducted by the CAPRI group at the [Centre for Census and Survey Research](#) at the University of Manchester. The data has been identified through a series of case study form field analyses of services across the public and private sector in the UK. In the long term it is proposed the service will be automated.

1. Select Dataset General
 employment
 health

2. Select subject age

[General - Age example results page](#)

The data sets behind this analysis can be [downloaded](#).

These pages are maintained by [Mark Elliot](#).

MANCHESTER
1824
The University of Manchester

Done Internet

- The DEA meta-data provides an understanding of:
 - what variables are available
 - under what coverage,
 - which could be linked with the anonymised release sets
- The subsequent analyses of disclosure risks with for example, **lifestyle microdata, qualitative patient records, public data sources such as the media and aggregate census or neighbourhood statistics data** are very computationally demanding and so are likely to benefit from the resources available through grid computing

- **The potential value of DEA:**
 1. it provides a potential to enable more appropriate understanding and classification of the total real risk of disclosive events
 2. it gives description of the *de facto* attitude of our culture towards personal data, thus enabling us to make more informed decisions on such subjects as privacy and data protection law

What sort of society?

- Informational Transparency?
- Human- Computer Interdependence?
- Individualism vs Collectivism

- Choices:
 - More legislation or less?
 - Personal information a commodity or public good

Concluding Remarks

- Grid computing provides the potential for unprecedented access to high quality individual level data
- However, as the amount of data on individual population units stored on computing systems increases, so does the threat to anonymised data releases

- The possibility that such data release may come to a halt as it becomes impossible to maintain sufficient data quality whilst meeting ever more stringent disclosure control constraints, means that it is vital that creative data access solutions are developed
- This presentation has described two possible partial solutions, data environment analysis and controlled remote access systems.