Enabling access to data:

Disclosure control and privacy – Tackling the Mosaic Effect

Keith Spicer (ONS)
Contents

• Disclosure Control – what, why......
• Balance of risk and utility
• Case study of Census
• Mosaic Effect
  – Intruder Challenges and Developments
  – Tackling
What is Disclosure Risk?

Disclosure risk occurs when information is published that could allow an intruder to deduce confidential information that may indicate the identity or particulars of:

• an individual
• a household
• a business
• or another statistical unit.

The risk cannot always be measured quantitatively.
What is Disclosure Control?

• Introducing sufficient ambiguity / damage into, or reducing level of detail of published statistics so that the risk of disclosing confidential information is reduced to an acceptable level and / or:

• Controlling access to data.
Why we need disclosure control 1

• Scenarios..........

• Note that just the perception that some of these scenarios take place is a key risk too!
Why might an intruder want to discover personal information?

- **Identity Theft**
- **Commercial Gain against competitors**
- **People who can identify themselves**
- **Nosy neighbour**
- **Journalist after a good ‘public interest’ story**
- **Find out sensitive information about people - salary, health**
- **Database enhancement**
- **Discredit government or GSS**
Why we need disclosure control 2

• Legislation
  – e.g. Data Protection Act (“risk of harm or distress”)
  – e.g. SRSA, relevant to ONS
  – e.g. May be specific to data source

• Code of Practice
  “no statistics will be produced that are likely to identify an individual unless specifically agreed with them”
Balance of risk and utility

• Important to protect data

• But just as important not to destroy them
Disclosure Risk: Information about confidential units

Data Utility: Information about legitimate items
<table>
<thead>
<tr>
<th>Disclosure Risk: Information about confidential units</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>The perfect result</td>
<td></td>
</tr>
<tr>
<td>Utility very high, Disclosure Risk very low.</td>
<td></td>
</tr>
<tr>
<td>If we could achieve this:</td>
<td></td>
</tr>
<tr>
<td>Philosopher’s stone of data</td>
<td></td>
</tr>
<tr>
<td>But how on earth do we do it??????</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Utility: Information about legitimate items</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
</tr>
</tbody>
</table>
Risk – Utility

Disclosure Risk: Information about confidential units

High

The worst case scenario
Data have little use
Data have high disclosure risk.

Satisfies no-one.

Users angry

Data providers worried

This is actually a very difficult (and unacceptable) place to reach

Low

Data Utility: Information about legitimate items

High
Risk – Utility

Disclosure Risk: Information about confidential units

Data Utility: Information about legitimate items

High

- No data – or very limited data
- Short-term: Safe – no breach possible
- Medium-term: Reputation of data ‘providers’
- Long-term: No data = No statistics

Policy based on ??????

NO JOBS
Risk – Utility

High Utility
High Risk

Users ecstatic............
............for a while

Identification and attribute disclosure very likely

• Public unhappy – Response rates plummet
• Reputation of output providers very low
• Quality and quantity of data reduce
• Potential legal sanctions against output providers

Long-term: Real damage to users, research community and statistics offices

Disclosure Risk:
Information about confidential units

Low
Data Utility: Information about legitimate items

High
Risk – Utility

Necessarily a trade-off between risk and utility so that

• the disclosure risk is acceptable
• the data utility is acceptable

Ideally – research analysis results not unduly affected

Disclosure Risk: Information about confidential units

Data Utility: Information about legitimate items
Risk – Utility

Disclosure Risk: Information about confidential units

Data Utility: Information about legitimate items
Case study: 2001/2011 Census

2001 Census
• Key risk – both identity and attribute disclosure
• Random record swapping
• + Small cell adjustment
• Additivity – except in Theme Tables
• Lack of consistency in cell counts and totals in different tables
• High level totals affected
• Tables don’t look “real”
• Some disclosure risk through differencing

2011 Census
• Key risk – attribute disclosure
• Record swapping targeted towards ‘risky’ records
• All tables additive
• All cell counts and totals consistent
• High level totals unaffected
• Tables still look disclosive
• Disclosure risk – “sufficient uncertainty” in small cells and apparent disclosures
  – Some real disclosures removed
  – Some false disclosures introduced
Risk – Utility

Disclosure Risk: Information about confidential units

High

Low

Data Utility: Information about legitimate items

2001

2011

High
Mosaic Effect (1)

- Assumption that each individual output has ‘acceptable disclosure risk’
- Combination of outputs may disclose information on individual case(s)
- Extension of disclosure by differencing
- Intruder needs to find same case in different outputs
Mosaic Effect (2) –
Linking cases between different releases within one dataset

• Self-contained problem for output provider
• Can track through ‘release log’
• Can assume same case where unique combination of variables
  – But....Do you know who is in the sample?
• 2001 UK Census Comm. Tables
  – Log of all tables (did not consider non-Census sources)
  – Thorough checking of risk of differencing
Mosaic Effect (3) – Linking cases between outputs from different datasets

Some natural ‘uncertainty’

– Same case?
  • Is fuzzy / probabilistic matching needed?

– Time-consistent?

– Question / Mode / Response?

– Classifications and categorisations?

– Data Quality
Intruder Challenges

- Sampling
- Non-response
- Imputed items and records
- Respondent error
- Capture error
- Coding error

For each dataset, protection provided by:

- Case in all samples?
- Time-consistent?
- Question / Mode / Response consistent?
- Classes / categories consistent?

For each match, additional protection provided by:
Intruder Developments

Computing Power
Sophisticated matching software
Availability of datasets (internet)
Timeliness of datasets

Greater motivations: Personal profit
More public concern

Greater challenge for output providers to be aware of all possible sources
Tackling the ‘Mosaic Effect’

For each dataset, protection provided by

Case in all samples?

Time-consistent?

Question / Mode / Response consistent?

Classes / categories consistent?

For each match, additional protection provided by

SO WHAT ELSE CAN WE DO?

But using these for protection creates tension with maximising data quality, harmonisation, utility
Tackling the ‘Mosaic Effect’

• Assess Intruder Challenges over Intruder Developments
• Consider output detail that is ‘needed’ rather than ‘nice to have’
• Risk: Consider what is ‘reasonable’ rather than what is ‘not impossible’ for an intruder
• Cannot protect completely against....
  – linking with all privately held datasets
  – use of private knowledge
• Type of outputs – Graphs / Summaries / Tables / Microdata
• Census case study – create ‘sufficient uncertainty’ – function of risk appetite of Head of Profession
• Consider Penetration Testing
• Access conditions? (frosted glass rather than transparent)
Final thoughts

• Balance of risk and utility
• Need to provide useful outputs while protecting individual cases
  – policy driven by ‘larger counts’ rather than individual cases
  – but ‘larger counts’ made up of groups of individual cases
• Outputs cannot be disclosure risk-free
• Risk must be manageable
• Must demonstrate that considered ‘means likely reasonably’ to be used by intruder
• This is not easy!
Any Questions

?