A Disclosure Avoidance Research Agenda

presented at FCSM research conf. ; Nov. 5, 2013
session E-3; 10:15am: Data Disclosure Issues

Paul B. Massell
U.S. Census Bureau
Center for Disclosure Avoidance Research
paul.b.massell@census.gov

Disclaimer: Ideas expressed are those of the author; not necessarily those of the U.S. Census Bureau

Meta-questions about D.A. [1]

- What are the types of disclosure risk that statistical agencies need to be concerned with (for each data release) ?
- TYPES: 1. re-identification (able to associate direct identifiers w. a micro record);
  2. positive attribution (person's age in [80-84])
  3. negative attribution (age not in [80-84])
  (for categorical and continuous variables)

Meta-questions about D.A. [2]

- What are good and practical ways to estimate disclosure risk?
- How accurately can an agency estimate disclosure risk?
- How does agency estimate what data intruders know about data before a release?

Meta-questions about D.A. [3]

- Is it possible for agency to know all the ways in which data users will use released data ?

- For given D.A. method, for each known use of released data, (ideally) agency should assess the impact on data quality.

Meta-questions about D.A. [4]

- Comparing D.A. methods

- For each D.A. method, how easily can agency know if it can be used to protect a given data product ?
- If two or more D.A. methods are applicable to a given data product, are there standard ways to compare the methods ?

Disclosure Risk of Microdata [1]

- Agency needs to avoid releasing variables that have great precision (e.g., Date of Birth)…i.e., small % of pop per category
- Ensure pop sizes of identified regions are not small
- Example: zipcodes have an average of 7,400 people.
- If record is only (zip, DOB, sex), over 60% of the records will be unique.

Disclosure Risk of Microdata [2]

- If a set of variables, S, has a ‘good chance’ of being unique in the region supplied in record, and if S is commonly found on online microdata files, the uniqueness of S can lead to many attribute disclosures

- This situation arises often enough that there is now more concern about releasing unmodified microdata

Creating Public Microdata, [1]

- Traditional methods: coarsen data (lessen precision); e.g. allow only big geographies; broad categories for age and other vars; used for ACS PUMS files

- Synthetic data; theory was developed in academia ; has been applied in recent years to special Census data products (e.g. census of group quarters, LEHD)

Creating Public Microdata [2]

- Drawback of traditional methods ; coarsening data causes information loss
- Drawback of synthetic data; complex process; expertise required

- Research Goal: Find ways that are computationally easy but preserve data quality (e.g., with noise)

When Standard Tables are Risky [1]

- Setup: Suppose a set of tables …
  - all reference the same region…and that region contains only a single entity (e.g., person, hhld)
  - E.g., a set of tables for a region in which there is only one household

- Risk Problem: A data intruder can construct a microdata record for all the variables used in the tables

When Standard Tables are Risky [2]

- There are thousands of blocks in the U.S. on which only one household is located (a ‘size 1’ block)
- Decision for agency to make: should a set of tables be published for a size 1 block ?
- If yes, does the corresponding microdata record constitute a violation of pledge of confidentiality ?
- If yes, need to ‘add noise’ to the data

When Standard Tables are Risky [3]

- What are simple ways to ‘add noise’ ?
- What quantities must be preserved ? (e.g., size of household, # of voting age…)

- How much perturbation of other vars is needed to provide adequate disclosure protection ?
- What is the effect on data quality ?

When Standard Tables are Risky [4]

- A key method currently used to protect Decennial and ACS tables is:
- Data Swapping (exchanging two similar records in same state)
- Swapping has some nice features:
- Computation is fast
- If all swapping is within county, then data analysis at county level is not affected

When Special Tables are Risky [1]

- When a large set of special tables is requested from an agency, and
- Certain variables are used repeatedly, causing the tables to be linked
- If many 0's, 1's, and 2's appear in certain marginal positions, a clever data user may be able to generate microdata from the tables

When Special Tables are Risky [2]

- There is a way to make the tables less risky but still preserve their usefulness
- Synthetic microdata can be created and then have the tables generated
- Drawback: high computation cost and expertise required

- Another approach: use cell suppression; drawback : loss of information in count tables may be too high for user's needs

When Special Tables are Risky [3]

- Possible Research agenda:
- Try to automate the detection aspect of the problem: i.e., whether the requested set of tables has too high a risk to be released
- Try 2 or more methods for protecting data
- Compare methods for ease of applicability and quality of data in product

Remote Access Systems

- Lightly modified microdata is used
- Nice feature: user can define a large set of tables (from a list of universes, spanning variables of various levels of detail)
- Geographical precision may be high
- The granularity of the variables is limited (e.g., age in 5 year increments)

Measuring Disclosure Risk [1]

- With magnitude data tables, the agency could say a cell is ‘sensitive’ if it would (if publ.) allow a data user to compute an estimate that is quite precise
- Level of precision is computed using the p% rule (see ref: WP22)
- Agency assumes data intruder has only rough estimates of values, before publ.

Measuring Disclosure Risk [2]

- For measuring risk of public microdata, agency measures the uniqueness of various sets of variables that appear in other public datasets

- Assessment of risk is challenging because knowing which sets of variables and which databases to consider is difficult

Measuring Disclosure Risk [3]

- In past decade, new ways to measure risk have come from computer science

- 'Differential privacy'…used to measure risk of an online database subject to queries
- 'Privacy leakage'…gradual loss of privacy as responses are given to queries of databases or from agency data products.

References [1]

- Books that provide an overview of D.A.

- Anco Hundepool et al, (D.A. researchers at European stat agencies), 'Statistical Disclosure Control', Wiley, 2012
- George Duncan, M. Elliot, JJ Salazar Statistical Confidentiality: Principles and Practice, Springer, 2011

References [2]

- FCSM report on D.A. methods used by federal statistical agencies
- **Statistical Policy Working Paper 22 (Revised 2005)- Report on Statistical Disclosure Limitation Methodology**
  http://www.fcsm.gov/working-papers/spwp22.html

- A Practical Beginners' Guide to Differential Privacy
  http://www.youtube.com/watch?v=Gx13lgEudtU

- Erica Klarreich; Privacy by the Numbers ; A New Approach to Safeguarding Data
- http://www.scientificamerican.com/article.cfm?id=privacy-by-the-numbers-a-new-approach-to-safeguarding-data