Overview of Statistical Disclosure Limitation

Aleksandra Slavkovic
sesa@psu.edu
Departments of Statistics & Public Health Sciences
Pennsylvania State University

Sep 24, 2013 @ Harvard University
"Integrating approaches to Privacy across the Research Lifecycle" workshop

Acknowledgement: NSF BCS – 0941553 to Penn State University
NIH/NCATS UL1 TR000127 to Penn State CTSI

Statistical Disclosure Limitation/Control

- How can we use the collected data while respecting our promise of confidentiality?

- Preserving confidentiality in statistical data products for individuals and organizations -- **minimizing disclosure risk**

- Providing access to useful statistical data such that statistical inference is possible -- **maximize data utility**

What constitutes disclosure?

- ***Identity disclosure***: Risk of re-identifying someone in the real dataset based on the published data and outputs.
  - Measures often based on population uniques.
  - Or, one may simulate an adversary and estimate the probability that she can correctly link the individuals in the published data.

- ***Attribute disclosure:*** Risk of learning sensitive attribute about a specific record is revealed through the released file

- ***Inferential disclosure***: if from the released data one can determine the value of some characteristic of an individual more accurately than otherwise would have been possible, e.g., Dalenius (1977)
  - improvement of the posterior odds of a particular event (identity or attribute)

Ref: http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main

What constitutes data utility?

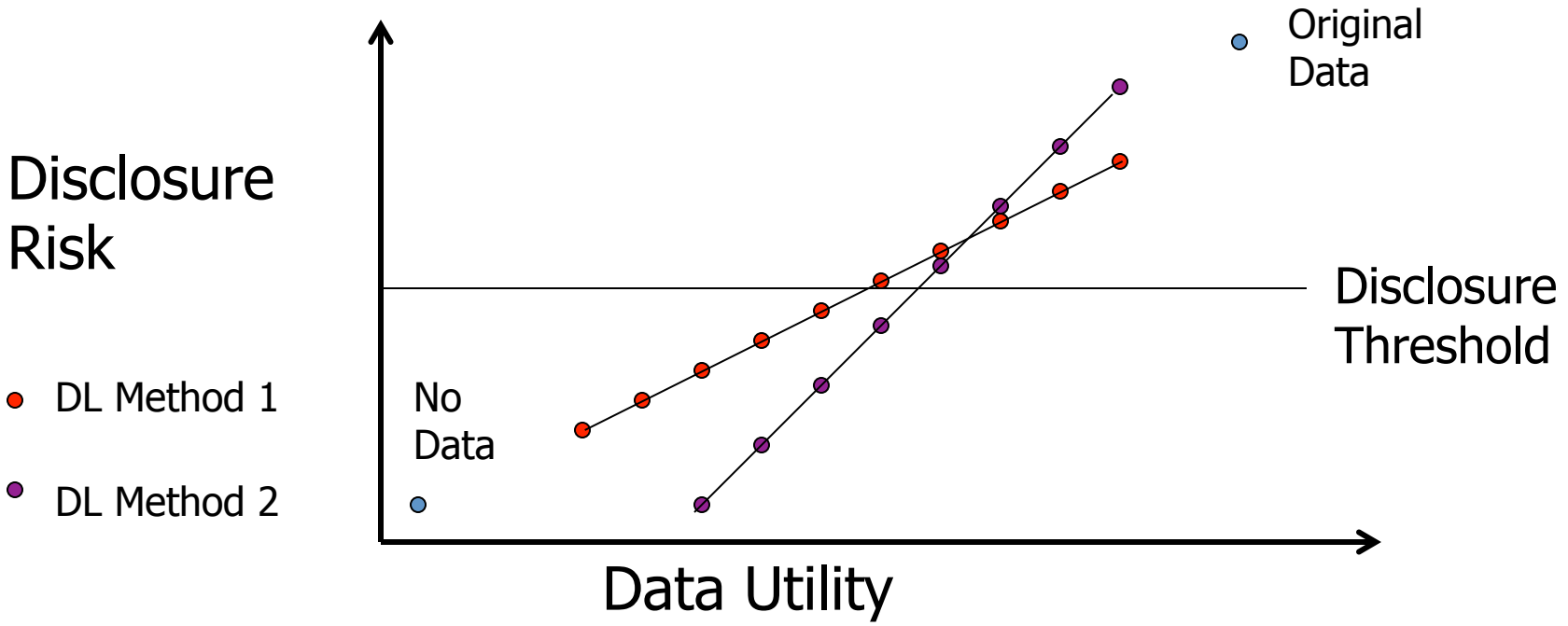
- What are the key characteristics and uses of the data?
- Sample design, nonresponse,…
- Providing access to useful statistical data, not just few numbers
  - Inferences should be the "same" as if we had original complete data
  - Support exploratory analysis
  - Sufficient variables and statistics to allow for proper multivariate analyses
  - Ability to assess goodness of fit of models
- Requires ability to reverse disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models (e.g., likelihood function for disclosure procedure)

The R-U Confidentiality Map

- Which *SDC* methods should be used?

  Depends on the nature of the data, and potential data uses!

  For many disclosure limitation methods, we can choose one or more parameters that we will vary.



Ref: Duncan et al. (2001)), Bayesian framework (Trottini (2003))

Protecting confidentiality

- **Restricted access**
  - Conditions are imposed on who may access the data, for what purpose, at what location, what can be published, etc.
  - Sworn Employees at Census Bureau & Research Fellowships

- **Restricted data**
  - Remove "identifiers" (naïve anonymization), subsample, limit geography
  - Restrict the amount of information in files via various ***statistical disclosure control*** (SDL) techniques
  - Public-use data will typically come in two formats
    - Microdata files
    - Tabular data

- ***New:*** Combination of the two, e.g. online access
  - Online queries & Remote access servers
  - Secure computation & Distributed databases

Statistical Disclosure Limitation Methods

- Apply to microdata and/or tabulated data before release

- **Data masking:** Transform the original data (matrix ***X***) to the disseminated data **Y**
  - ***Y=AXB + C***
  - ***A***=record transformation, ***B***=attribute transformation, ***C***=noise addition

- **Traditional approaches**
  - Aggregation: Rounding, Topcoding & Tresholding
  - Suppression, e.g., cell suppression
  - Data Perturbations, Swapping
  - Post Randomization Method (PRAM)

- **Modern approaches:** Sampling and Simulation techniques
  - Synthetic data
  - Remote access servers
  - Partial information releases
  - Secure computation

Aggregation: Rounding & Topcoding

- A way of aggregating data values by key characteristics
- Rounding the reported income to the closest $1000
- Topcoding (coarsening, “generalizing”) or thresholding
  - Grouping similar income values together such as all individuals over $500,000

Aggregation: Rounding & Topcoding

- A way of aggregating data values by key characteristics
- Rounding the reported income to the closest $1000
- Topcoding (coarsening, “generalizing”) or thresholding
  - Grouping similar income values together such as all individuals over $500,000
- Ecological Inferences
- Works ok for large samples and middle of the distribution
- Problems in the upper tails due to high difference in income
- If intervals are not uniform, reporting the middle value will be biased

Suppress Sensitive Cells & Others

- The most common technique
- Not releasing sensitive fields of the microdata file or a table, or a model-based summaries
- Mostly applied to tabular data, e.g., cell suppression -- Cox (1980, 1995, 1999)

County	Low	Medium	High	Very High	Total
Alpha	15	1	3	1	20
Beta	20	10	10	15	55
Gamma	3	10	10	2	25
Delta	12	14	7	2	35
Total	50	35	30	20	135

Suppress Sensitive Cells & Others

- Not releasing sensitive fields of the microdata or a table
- Mostly applied to tabular data, e.g., cell suppression -- Cox (1980, 1995, 1999)

County	Low	Medium	High	Very High	Total
Alpha	15	p	s	p	20
Beta	20	10	10	15	55
Gamma	p	s	10	p	25
Delta	s	14	s	p	35
Total	50	35	30	20	135

- *"Swiss cheese" -- Issue with correlation structure; data loss only 7 values published and in the presence of external info not adequate.*
- Difficult to guarantee adequate protection; what ever that means!
- No extension to higher dimensional tables

Data perturbation: Data swapping

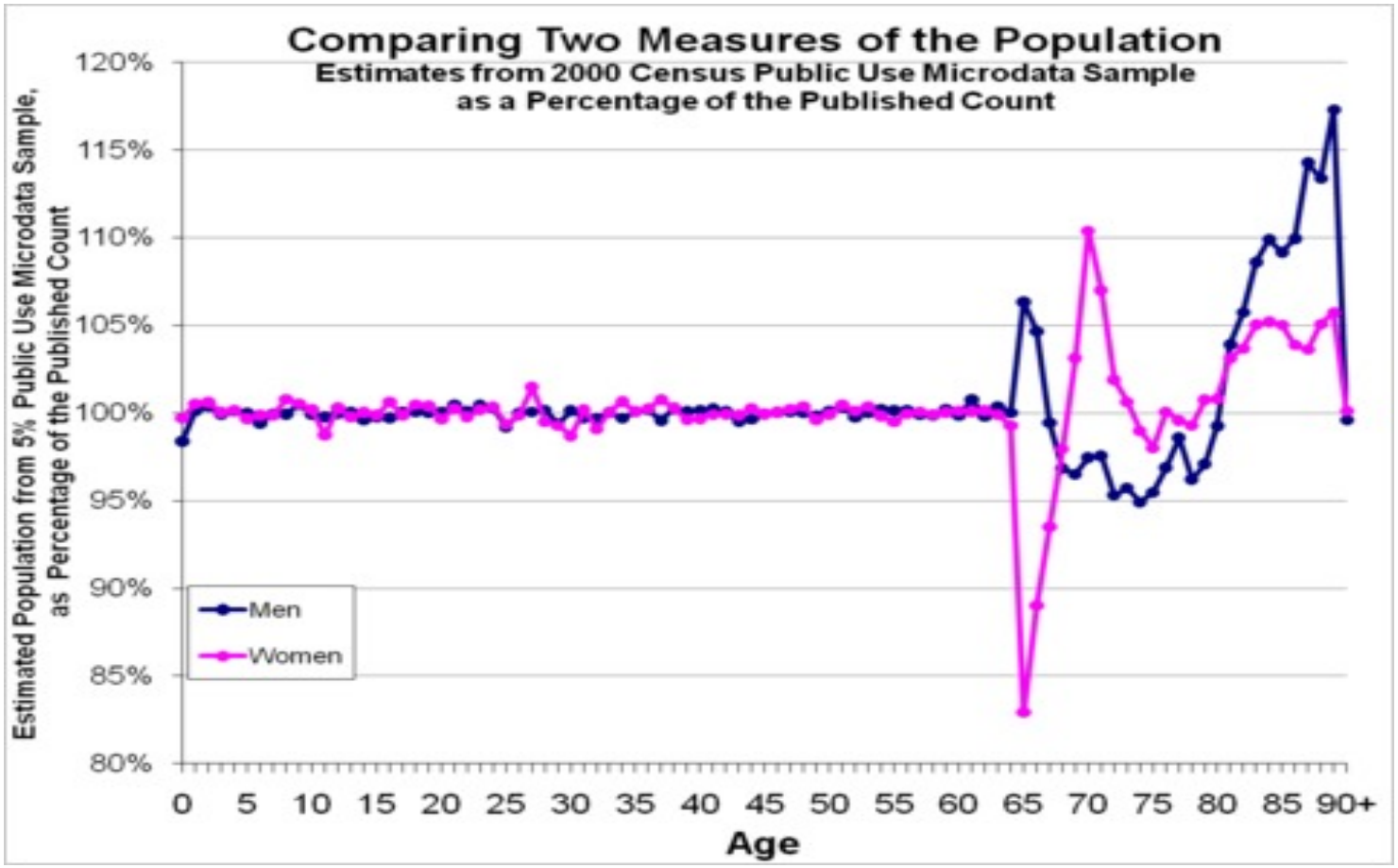
- Perturbation generally means altering values before releasing
  - E.g., Adding random noise, Evans, T., Zayatz, L., and Slanta, J. (1998)

- Proposed by Dalenius & Reiss (1982), and is used by several agencies
  - Used on household census and surveys (e.g., decennial census)

- Exchange a fraction of data at random between two respondents
  - Small fraction swapped to preserve some statistics
  - A maximum percent is set, but unknown to public
  - While it preserves marginal distributions, it reduces or destroys correlations and joint distributions of swapped and non-swapped variables (Dreschsler and Reiter (2010))

Too much protection?

- Source: *Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications* by T. Alexander, M. Davern and B. Stevenson (2010)



Recent methods: Sampling & Simulation

- World "without original micordata"
- Synthetic & Partially synthetic data uses Bayesian methodology
  - Raghunathan, Reiter, and Rubin (2003), Reiter (2005)
- Partial data releases for tabular data & logistic regression with algebraic tools
  - Dobra et al. (2008), Slavkovic & Lee (2009)
  - Links to synthetic data and swapping
- Remote access servers
  - Gomatam, Karr, Reiter, Sanil (2005)
- Secure computation
  - Karr et al. (2005), Slavkovic et al. (2007), Hall et al.(2013)
- Differential Privacy & Statistical Utility

Synthetic Data

- Sampling & imputation technique so that released data look like actual data
  - Values simulated from posterior predictive distribution
  - Fully synthetic and partially synthetic

- Does it guarantee confidentiality?
  - Significantly lowers the risk of identity disclosure
  - Attribute disclosure is still possible
  - With extreme values, it may still be possible to re-identify a source record
  - Some simulated individuals may have data values virtually identical to original sample individuals (Fienberg 1997, 2003)

- Is it valid for inferences?
  - It depends on the model used to generate the data
  - Not unless we are careful in how it is synthesized
  - Choosing the models and posterior distributions can be tricky

- Many open questions: missing data, longitudinal, non-parametric models
- Raghunathan, Reiter, and Rubin (2003, JOS), Reiter (2003, Surv. Meth.; 2005, JRSSA)

"Secure" Analysis

- Setting & Goals:
  - A "global" database partitioned among multiple agencies
  - Share "valid" statistical analysis as if had the "global" database but without any actual data integration in order to improve each own analysis
  - Do not reveal identities or sensitive attributes

- Tools from the Computer Science literature based on secure multiparty computation

- Issues:
  - How to choose a model a priori?
  - How to account for data quality, e.g., reconcile measurement errors?
  - Partially overlapping databases with measurement error
  - What do results reveal?

- Karr, Lin, Sanil, and Reiter (2005.)
- Fienberg et al. (2006)
- Slavkovic et al. (2007)

Remote Access Servers

- Submit queries for output from statistical analyses of microdata, but does not give direct access to microdata.
- Queries: Users request a table, or results from fitting a statistical model to the data
- Response from the server:
  - Answerable query: model output & ideally model diagnostics.
  - Unanswerable query: no results.
- Challenges
  - Non-statistical: Operation costs, server security, etc.
  - Statistical:
    - Disclosure risks from smart queries (e.g., subsets, transformations).
    - Inferential disclosure risks
    - Enabling complex model fitting.
  - Interactive vs. non-interactive
  - Limit to the number of releases
- Gomatam, Karr, Reiter, Sanil (2005, *Stat. Science*)

Differential Privacy & Statistical Inference

- Recent theoretical & methodological developments on connections between DP and traditional statistical inference

  - Parametric estimation, Smith (2008).
  - Robust statistics, Dwork & Lei (2009).
  - Approximation of smooth densities, Wasserman & Zhou (2010).
  - Efficiency of "private estimators" and hypothesis testing, Duy & Slavkovic (2009)
  - OnTheMap, Machanavajjhala et al. (2008)
  - Degree sequence with differentially private graphical degree sequence, Karwa & Slavkovic (2012)
  - Private data sharing of GWAS, Uhler et al. (2012)

Some General Principles for Developing SDL Methods

- Need to understand/model potential intruder behavior
  - All data are informative for intruder, including non-release or suppression.
- Need to define and understand potential statistical uses of data in advance
  - Leads to useful reportable summaries (e.g., MSSs).
- Identifies sensitive values a priori
- Methods should allow for reversibility for inference purposes
  - Missing data should be "ignorable" for inferences.
  - Assessing goodness of fit is important.

Overall trends & limitations

- Mostly developed for survey samples and “flat files” and in part rely on randomness in the sample for protection

- New Trends
  - Synthetic data methods currently look promising and are being implemented
  - More focus on alterations in microdata
  - Need to combine methods: Crypto, PPDM, SDL
  - Some recent successes with Differential Privacy and Statistical Utility

- Extensions and effectiveness to big data are not clear
  - Partial releases and synthetic data combined with CS privacy guarantees (e.g., secure computing, distributed diDP,...)

The issues that remain

- Most methods that agencies use are still ad-hoc
- Implementations of new (and old) methods are nontrivial
- Increasing the sample size does NOT necessarily decrease the risk
- Definitions of utility and risk, and disclosure
- The usual hard problems remain hard
  - Preserving the structure of sampling design
  - Multi-wave surveys
  - Geographical data
  - Longitudinal data
  - Capturing the multivariate statistical characteristics
  - Modeling the joint distribution of multivariate categorical data (especially in presence of sparse data)
  - BIG Data! Non-flat world!

Further General Reading

- Statistical Policy Working Paper 22 (2005) by Federal Committee on Statistical Methodology: http://www.fcsm.gov
- Checklists: http://www.fcsm.gov/checklist/
- American Statistical Association: http://www.amstat.org/comm/cmtepc/index.cfm
- For EU see Computational Aspects of Statistical Confidentiality (CASC): http://neon.vb.cbs.nl/casc
- Hundepool et al. (2012) *Statistical Disclosure Control.* Wiley
- Info7470 VirtualRDC @ Cornell University
- Journal of Official Statistics
- Journal of Privacy and Confidentiality

Bertinoro, 2006: Grappa to the rescue