Westat

Improving Lives Through Research®

What Can Agencies Do Right Now to Improve Privacy Protection?...

Tom Krenzke

June 7, 2019

NAS workshop: Challenges and new approaches for protecting privacy in Federal statistical programs

Session Topic

What can agencies do right now that achieves improved privacy protection?

What do they need to investigate in the short and long term?

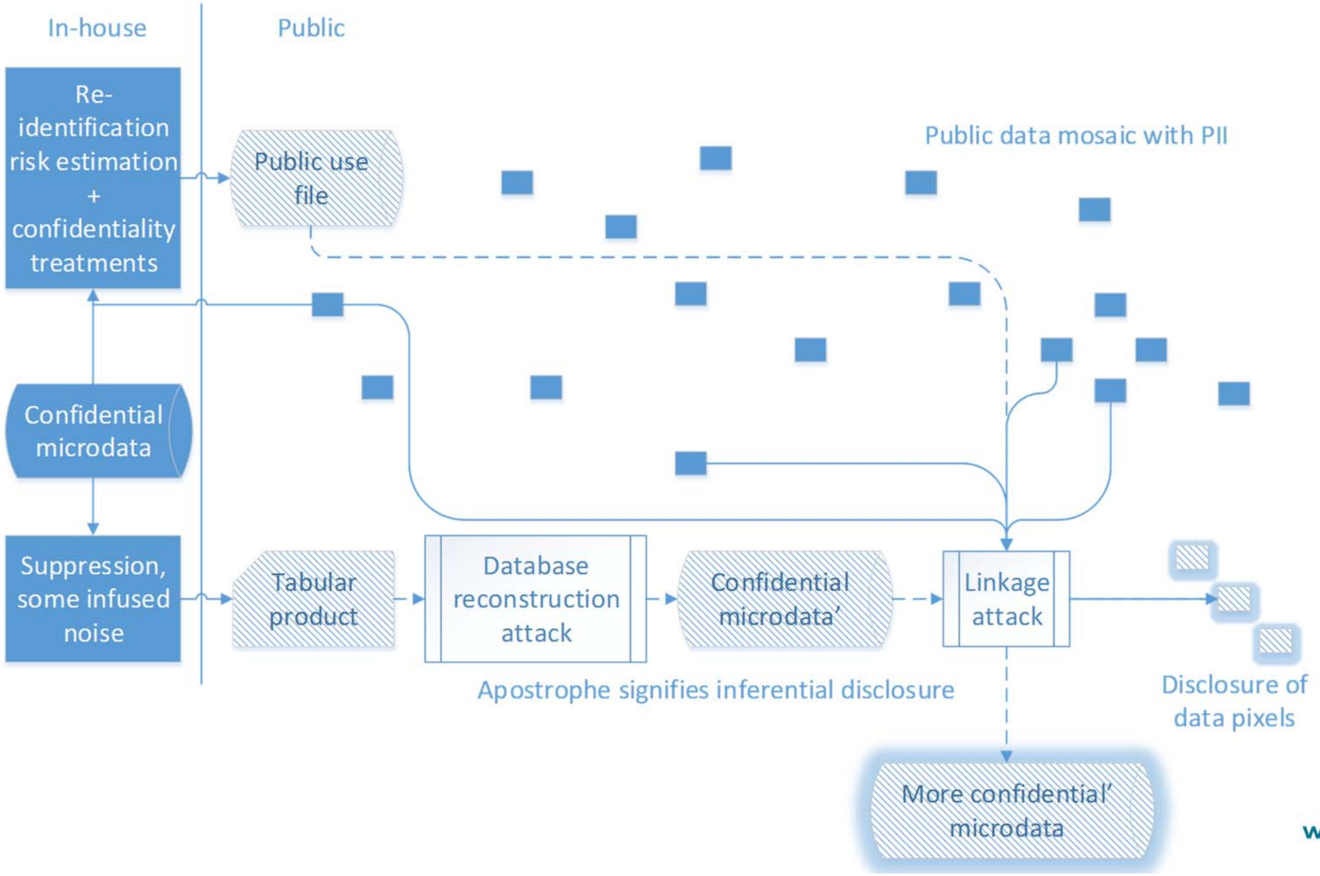
What are immediate solutions?

Some Context

Mosaic Effect and Breaches

- The concept of a mosaic effect is derived from the mosaic theory of intelligence gathering, in which disparate pieces of information become significant when combined with other types of information (Pozen 2005).
- Techworld (2019) & Wikipedia (2019) each log data breaches, mainly due to security breaches. Largest leak of data in January (Song, 2019)
  - Mainly registries of subgroups (e.g., consumer lists, admin data)
- Contribute to major amount of overall risk → PII in public
  - Makes the mosaic effect very real

One View of Potential Vulnerabilities


In-house
Public
Re-identification risk estimation + confidentiality treatments
Public use file
Public data mosaic with PII
Confidential microdata
Suppression, some infused noise
Tabular product
Database reconstruction attack
Confidential microdata'
Linkage attack
Apostrophe signifies inferential disclosure
Disclosure of data pixels
More confidential' microdata

What Can Agencies Do Right Now That Achieves Improved Privacy Protection?

What Can Agencies Do Right Now That Achieves Improved Privacy Protection?

A goal or challenge, while achieving improved privacy protection in future, is to release at least the same amount of data (to inform policy/improve society), at similar cost, resources, and time, as in past

Action: Identify what can be improved. Run risk assessments on current data releases to identify the risks that need to be addressed

- Review modes of dissemination with mosaic effect in mind
- What are sources of risk?
- Are there data releases that should be done differently?

Example: NCSES Review of Disclosure Risk by Westat

- Microdata assessed
  - SDR: 2013 on-line public use file (PUF), 2013 proposed PUF, 2015 proposed PUF (multiple), 2017 PUF cross-sectional and longitudinal
  - NSCG: 2013 PUF, 2017 PUF cross-sectional and longitudinal
  - SESTAT: 2013 PUF
  - SED: 2013 restricted use file (RUF)
- Tabular products (all open access) assessed
  - SDR 2013 Data Tables
  - SESTAT 2013 Data Tables & SESTAT Data Tool (2013 SDR, NSCG)
  - SED 2013 Detailed Statistical Tables, WebCASPAR (2013 SED), and SED Tabulation Engine (2013 SED)

NCSES Review of Disclosure Risk by Westat

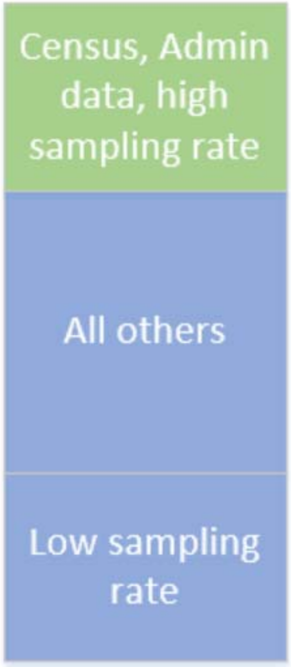
- Assessment methods
  - For microdata, generally estimated risk via loglinear modeling by Skinner and Shlomo (2008)
  - For tables, checks conducted on implementation of rules (e.g., suppression), used simple math logic, investigated across modes (e.g., fill-in suppressed data from another mode?)
- Some outcomes
  - Draft standards and guidelines
  - Awareness of vulnerabilities, proposed options for improved approaches toward risk reduction and ongoing modifications to data treatment/releases
  - Explored DP with other noise infusion approaches in Shlomo, et al (2019)

Privacy Day Seminar (February 2019) – see reference section for slides

- File and individual risk using risk metrics
- How to select variables and determine number of variables?
  - Is only checking indirect identifiers enough?
- What checks apply to the log-linear approach?
- How do we conduct a longitudinal risk assessment?
- How do clusters impact risk?
  - How to estimate cluster re-identification?
- What risk threshold values do we use?
- Is there risk with synthetic data?
- What are risks in a flexible table generator?

Develop Proof of Concept

- Do you have a census or a survey with a high sampling rate, or admin data where tables need to be generated for the public?
  - If so, these are prime candidate situations for applying differential privacy
- Goal: Proof of concept flexible table generator using data typically reserved for restricted use



Census, Admin data, high sampling rate

All others

Low sampling rate

Example: BLS’ Occupational Requirements Survey (ORS) Query Tool for SSA Purposes

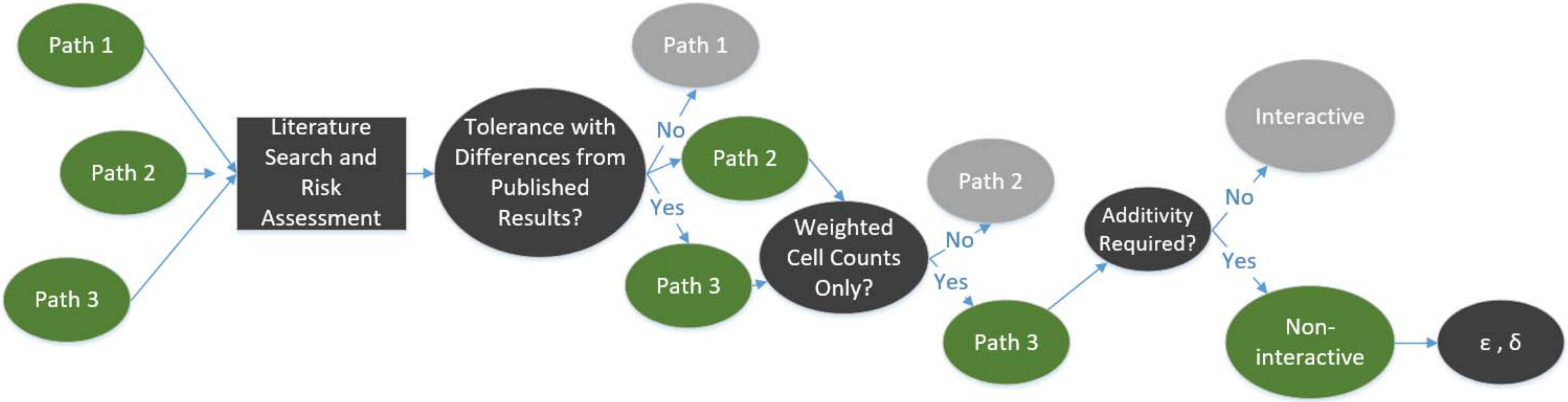
Task: Write specifications for a query tool to output…

- Weighted tabulations of Standard Occupational Codes (SOC)(indirect identifier) by several types of physical requirements for the jobs (sensitive variables)
  - Establishment-assisted sample data
  - Estimated total employment (weighted counts) in each cell
  - Any dimensional cross-tabs to be allowed (could be max of around 20-way tables)

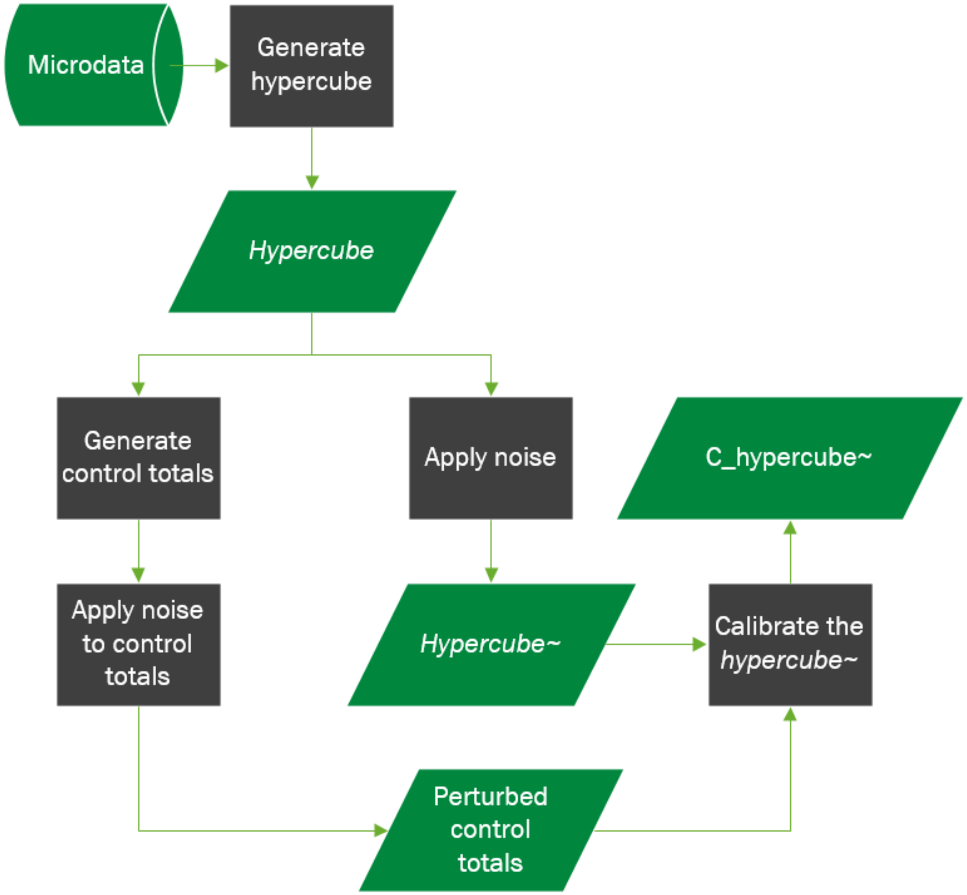
Potential solutions

- Path 1. Generate results from original microdata
- Path 2. Generate results from perturbed or synthetic microdata, or
- Path 3. Generate results from original microdata, then perturb the output before displaying the results

ORS Query Tool Systematic Approach With Select Decision Points



ORS Query Tool Calibrated Perturbed Hypercube – A Non-interactive Approach



- Susceptible to large noise accumulation for low dimensional tables
- Census considering top-down approach (Abowd, 2019)
- Draft specifications for the Query Tool written to be calibrated bottom-up
  - Calibrate the perturbed hybercube to low dimensional perturbed table
  - Objective: To reduce variance in moderate dimensional tables

Evaluation has started

Perturbation Vector Examples for Differential Privacy – Probability of Perturbation

Perturbation amount	$\varepsilon = 2$ and cap of ±7	$\varepsilon = 5$ and cap of ±2
-7 or +7	.0000006	
-6 or +6	.0000047	
-5 or +5	.0000356	
-4 or +4	.0002555	
-3 or +3	.0018878	
-2 or +2	.013949	.0000448
-1 or +1	.10307	.0066477
0	.76159	.98661

ORS Query Tool Simulation Plan

Objective: To determine the impact of the calibration on mid-dimensional tables

- Dataset has about 160 small cells -- many singletons with 10 variables
- Calibrate to a two-way table
- Run for different values of ε and δ
- Apply noise to unweighted counts, then apply average weight
- Determine the impact on tables with varying number of aggregated cells by processing all 1-way to 10-way tables
- Metrics – Example… Variance by # of aggregated cells, for tables with both, at least one, or no calibration variables

What Do Agencies Need to Investigate in the Short Term – Immediate Solutions?

What Am I Hearing?

- To develop or obtain DP capability – discussed with reps from three agencies…
  - Education needed about DP
  - Need real examples of applications, implementation – less theory at this point
  - Need ways to maintain similar costs, resources, timelines as in past
  - Suggestion… pool resources with agencies to develop open source software
  - Issues/concepts
    - Consistent estimates for all modes vs tolerance for differences
    - Official statistics
  - Other challenges were discussed

Investigations Toward Immediate Solutions

- Identify what can be improved
  - Agency-wide risk assessment
- Develop a way forward
  - Discuss/settle on concepts (e.g., tolerance of differences, additivity, interactive/noninteractive) toward a framework as an agency
    - Can the agency tolerate small differences in estimates with raw data or between dissemination modes?
    - Can the agency tolerate loss of additivity?
    - If *No* and *No* to the above questions, then DP is not an option
    - If *Yes* and *Yes* to the above questions, then an Interactive approach (Shlomo, 2019) may be best
    - If *Yes* and *No*, then a Non-interactive approach may be best
  - Review all options (verification servers, synthetic data/remote access)

Investigations Toward Immediate Solutions (continued)

- Develop a research plan
  - Cross-agency workgroup
  - Examples: Impact on risk/utility (e.g., increasing ε), survey weights, variance estimation, multiple types of estimates
- Develop a proof of concept
  - Purposively select a small project and implement DP
  - Trains staff and gains insights
  - Develops operational process

Toward Longer Term Goals

Toward Longer Term Goals -- Major Solutions Will Exist When We…

- Goal: Unite concepts and practices
  - Action: Develop renewed standards and guidelines through collaboration among agencies
    - Example: cell suppression methods do not work well in flexible table generators
- Goal: Develop an operational road map for the same amount of data, cost, resources, and time as in past
  - Action: Develop a toolkit (software) through collaboration among agencies and researchers
    - Generate tables and microdata, account for noise in variances
    - Analyze results (e.g., on the order of SAS® Proc Survey*)
  - Action: Conduct trainings and demonstrations
- Can the ASA Privacy and Confidentiality Committee help?

References

References

- Abowd (2019). Differential Privacy in the Real World: The 2018 End-to-End Census Test. American Association for the Advancement of Science Annual Meeting Sunday, February 17, 2019 8:00-9:30 Presentation slides in https://www2.census.gov/programs-surveys/decennial/2020/resources/presentations-publications/2019-02-17-abowd-differential-privacy.pdf?

- Pozen, D.E. (2005). The Mosaic Theory, National Security, and the Freedom of Information Act. *The Yale Law Journal*, December 2005, pp. 628–679.

- Privacy Day Seminar (2019). Toward Protecting the Privacy of Individuals When Disseminating Data: Challenges in Disclosure Risk Assessment. ASA Privacy and Confidentiality Committee. February 6, 2019. Slides by Krenzke, T. and Li, J. in https://higherlogicdownload.s3.amazonaws.com/AMSTAT/284c0271-770e-46ef-975c-d99a87486bd3/UploadedImages/PrivacyDayPresentation.pdf

References

- Shlomo, N. (2019). Statistical Disclosure Limitation and Differential Privacy. Washington Statistical Society President's Invited Seminar, May 1, 2019. Washington, DC.
- Shlomo, N., Krenzke, T. and Li, J. (2019) Confidentiality Protection Approaches for Survey Weighted Frequency Tables. Submitted.
- Skinner, C. J. and N. Shlomo (2008). "Assessing Identification Risk in Survey Microdata Using Log-linear Models." Journal of American Statistical Association. 103, no. 483 (2008): 989–1001.

References

- Song, V. (2019) "Mother of All Breaches Exposes 773 Million Emails, 21 Million Passwords". Gizmodo. Retrieved 2019-01-18
- Techworld (2019). The UK's most infamous data breaches: The most notorious data breaches that have impacted UK citizens, from FIFA to Facebook. Published January 17, 2019. Available at https://www.techworld.com/security/uks-most-infamous-data-breaches-3604586/.
- Wikipedia (2019). List of data breaches. https://en.wikipedia.org/wiki/List_of_data_breaches

Thank You

Photos are for illustrative purposes only. All persons depicted, unless otherwise stated, are models.