# INFO 7470
# Session 12 Statistical Tools: Methods of Confidentiality Protection

John M. Abowd and Lars Vilhuber

April 25, 2016

# Outline Part I

- Most of today's lecture is based on

Abowd, John M. and Ian Schmutte "Economic Analysis and Statistical Disclosure Limitation" *Brookings Papers on Economics Activity* (Spring 2015): 221-293. Includes discussion. [free download] (Brookings does not use DOIs.) Online appendix is in the same place.

Curated URL (Labor Dynamics Institute, Cornell)

# First, a Live Demo!

- The demo will only be live at Cornell and Census. Other sites have too few enrollees for the exercise to work properly
- The local instructor is now distributing a sealed envelope to each person in the class

CHOOSE YOUR OWN ENVELOPE

DO NOT LET THE INSTRUCTOR CHOOSE

DO NOT OPEN THE ENVELOPE!!!!!!!!!

# Now, Let's Analyze the Data

# The Basic Economics

- Scientific data quality is a pure public good
- Quantifiable privacy protection is also a pure public good (bad, when measured as "loss") when supplied using the methods I will discuss shortly
- Computer scientists have succeeded in providing feasible technology sets relating the public goods: data quality and privacy protection
- These technology sets generate a quantifiable production possibilities frontier between data quality and privacy protection

# The Basic Economics II

- We can now estimate the marginal social cost of data quality as a function of privacy protection—a big step forward
- The CS models are silent (or, occasionally, just wrong) about how to choose a socially optimal location on the PPF because they ignore social preferences
- To solve the social choice problem, we need to understand how to quantify preferences for data quality v. privacy protection
- For this we use the Marginal Social Cost of data quality and the Marginal Social Benefit of data quality, both measured in terms of the required privacy loss

# Now, Back to the Live Example

# Four Principles of Ideal Data Publication - Privacy Protection Systems

- To the maximum extent possible, *scientific analysis* should be performed on the *original confidential data*
- Publication of *statistical results* should respect a quantifiable *privacy-loss budget* constraint
- Data *publication algorithms* should provably *compose*
- Data *publication algorithms* should be provably *robust to arbitrary ancillary information*

# Outline

- Why must users of restricted-access data learn about confidentiality protection?
- What is statistical disclosure limitation?
- What are formal privacy methods?
- Examples of SDL and formal privacy methods
  - Traditional SDL
  - Differential Privacy
  - The Inferential Disclosure Link

# Why Are We Covering This?

- The vast majority of data users have no exposure to the SDL techniques applied to the data they use
- The tradition in SDL is to protect the details of what was done as part of the protocol
  - Here's the complete description for the American Community Survey public-use micro sample: http://www.census.gov/acs/www/data_documentation/pums_confidentiality/

# Explosion of Research in Computer Science

- Differential privacy, developed by Cynthia Dwork and many collaborators fundamentally changed the nature of the discussion
- Generalizations of Dwork's approach are called "formal privacy" models
- The standards of modern cryptography apply:
  - An algorithm only provides protection if it can survive an attack by anyone armed with all the details of the protection algorithm except the actual random numbers, if any, used in the protection
- This is where the four principles above originate

# Restricted-access Data Users

- Must normally subject their analyses to statistical disclosure limitation, including limitation by differential privacy methods
- It is extremely important to understand what this means for the quality of the released research and its replicability

# Statistical Disclosure Limitation

- Protection of the confidentiality of the underlying micro-data
  - Avoiding identity disclosure
  - Avoiding attribute disclosure
  - Avoiding inferential disclosure
- Identity disclosure: who (or what entity) is in the confidential micro-data
- Attribute disclosure: value of a characteristic for that entity or individual
- Inferential disclosure: improvement of the posterior odds of a particular event (identity or attribute)
- Reference: Duncan, George T., Mark Elliot, and Juan-José Salazar-González (2011) *Statistical Confidentiality: Principles and Practice*, New York: Springer.

# Privacy-preserving Datamining and Differential Privacy

- Formally define the properties of "privacy"
- Introduce algorithmic uncertainty as part of the statistical process
- Prove that the algorithmic uncertainty meets the formal definition of privacy
- Differential privacy defines protection in terms of making the released information about an entity as close as possible to being independent of whether or not that entity's data are included in the tabulation data file
- Reference: Dwork, Cynthia, and Aaron Roth (2014) "The Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Computer Science 9, nos. 3–4: 211–407. [free download]

# Point of Commonality: Inferential disclosure

- Confidential data item Y={0,1}
- Published data item Z={0,1}
- Prior odds: Pr[Y=1]/Pr[Y=0]
- Posterior odds: Pr[Y=1|Z=z]/Pr[Y=0|Z=z]
- Inferential disclosure if

$$\frac{\frac{Pr[Y=1|Z=z]}{Pr[Y=0|Z=z]}}{\frac{Pr[Y=1]}{Pr[Y=0]}} > e^{\varepsilon}$$

- Where exp($\varepsilon$) is a predetermined limit
- Reference: Duncan, George T., and Diane Lambert (1986) "Disclosure-Limited Data Dissemination," *Journal of the American Statistical Association* 81, no. 393: 10–18.

# Point of Commonality: Inferential disclosure

- Consider two confidential datasets Y and Y*
- Z and Z* are neighbors if sup|Y-Y*|$\leq 1$ (think: differs in the value in one row)
- Randomized publication algorithm: M(Y) publishes Z
- Publication is $\varepsilon$–differentially private if and only if

$$sup_{\{Y,Y^*i\}}\left\{\frac{Pr[Z=M(Y)=i|Y=i]}{Pr[Z=M(Y^*)=i|Y^*=i]}\right\}<e^{\varepsilon}$$

  for all sup|Y-Y*|$\leq 1$ (neighbors) and *i*=0,1.

- Reference: Dwork, C. (2006) "Differential privacy," *Proceedings of the International Colloquium on Automata, Languages and Programming* (ICALP), pp. 1–12.

# The Impossibility Theorm

- $\varepsilon$-differential privacy implies a bound on the maximal inferential disclosure from a publication Z based on confidential data Y

$$sup_{\{Y,Y^*,i\}}\left\{\frac{Pr[Z=M(Y)=i|Y=i]}{Pr[Z=M(Y^*)=i|Y^*=i]}\right\}<e^{\varepsilon}$$

for all sup|Y-Y*|$\leq$ 1 (neighbors) and *i*=0,1

- implies

$$sup\left\{\frac{\frac{Pr[Y=1|Z=z]}{Pr[Y=0|Z=z]}}{\frac{Pr[Y=1]}{Pr[Y=0]}}\right\}<e^{\varepsilon}$$

where the supremum is taken over the rows of Y and Z.

- Therefore, there can be no informative data publication without some inferential disclosure

# Discussion

- Many of the details have been left out
- Notice that when $\varepsilon$=0, the posterior odds are equal to the prior odds, and the publication contains nothing informative
- This concept is called semantic security in the cryptography literature
- Reference: Goldwasser, S. and Micali, S. (1982) "Probabilistic encryption & how to play mental poker keeping secret all partial information," *Proceedings of the fourteenth annual ACM symposium on Theory of computing*, ACM, pp. 365–377.

# Now, Back to the Live Example

# General Methods for Statistical Disclosure Limitation

- At the Census Bureau SDL is called Disclosure Avoidance Review
- Traditional methods
  - Suppression
  - Coarsening
  - Adding noise via swapping
  - Adding noise via sampling
- Newer methods
  - Explicit noise infusion
  - Synthetic data
  - Formal privacy-preserving sanitizers

# Suppression

- This is by far the most common technique
- Model the sensitivity of a particular data item or observation (“disclosure risk”)
- Do not allow the release of data items that have excessive disclosure risk (primary suppression)
- Do not allow the release of other data from which the sensitive item can be calculated (complementary suppression)

# Suppression in Model-base Releases

- Most data analysis done in the RDCs is model-based
- The released data consist of summary statistics, model coefficients, standard errors, some diagnostic statistics
- The SDL technique used for these releases is usually suppression: the suppression rules are contained (up to confidential parameters) in the RDC Researcher's Handbook

# Coarsening

- Coarsening is the creation of a smaller number of categories from the variable in order to increase the number of cases in each cell
- Computer scientists call this “generalizing”
- Geographic coarsening: block-block group-tract-minor civil division-county-state-region
- Top coding of income is a form of coarsening
- All continuous variables in a micro-data file can be considered coarsened to the level of precision (significant digits) released
- This method is often applied to model-based data releases by restricting the number of significant digits that can be released

# Swapping

- Estimate the disclosure risk of certain attributes or individuals
- If the risk is too great, attributes of one data record are (randomly) swapped with the same attributes of another record
- If geographic attributes are swapped this has the effect of placing the risky attributes in a different location from the truth
- Commonly used in household censuses and surveys
- Rarely used with establishment data

# Sampling

- Sampling is the original SDL technique
- By only selecting certain entities from the population on which to collect additional data (data not on the frame), uncertainty about which entity was sampled provides some protection
- In modern, detailed surveys, sampling is of limited use for SDL

# Rules and Methods for Model-based SDL

- Refer to Chapter 3 of the RDC Researcher's Handbook
- Suppress: coefficients on detailed indicator variables, on cells with too few entities
- Smooth: density estimation and quantiles, use a kernel density estimator to produce quantiles
- Coarsen: variables with heavy tails (earnings, payroll), residuals (truncate range, suppress labels of range)

# PROPERTIES OF STATISTICAL DISCLOSURE LIMITATION METHODS

Statistical disclosure limitation is *ignorable* if and only if the analysis designed for the confidential data yields the same result when applied to the published data.

**SDL is almost never ignorable.**

Nonignorable statistical disclosure limitation is *known* if and only if the analysis of the published data can be exactly corrected for the data alterations introduced by the SDL.

**Nonignorable SDL is known for a limited number of methods.**

Nonignorable statistical disclosure limitation is *discoverable* if and only if the analysis of the published data can be probabilistically corrected for the data alterations introduced by the SDL.

**Nonignorable SDL is discoverable for many methods.**

In modern SDL, the **relevant trade-off** for economists is between methods that are **easy to make SDL aware** (generalized randomized response when certain parameters are public) and those that are **not** (suppression and swapping as currently implemented).

# Example 1: Randomized Response

- Oldest SDL technique (Warner 1965; predates formal SDL itself)
- Hide the question asked from the interviewer
- Respondent's answer "yes" applies to either the sensitive question or a non-sensitive question
- Two SDL parameters: probability asked sensitive question, probability "yes" for non-sensitive question
- SDL affects the mean and standard error for parameter of interest: proportion of population "yes" for sensitive question

# Example 1: Randomized Response (continued)

- "Yes": Probability asked sensitive question ½
- Probability "yes" for innocuous question ½
- Ever xxxx:
- Inferential disclosure (Bayes factor):
- $\varepsilon$-differential privacy (*maximum* ln Bayes factor): ln () =
- SDL *nonignorable* and *known*

# Now, Back to the Live Example

# Example 2: Topcoding

- Published income topcoded at *T*
- Quantiles of published income = quantiles of confidential income for all quantiles less than the quantile of *T*
- SDL is *ignorable* for quantiles less than the quantile of *T*
- SDL is *nonignorable* but *discoverable* for quantiles at or above the quantile of *T* (discovery is via inspection of the data combined with the knowledge that SDL topcoding was applied)

# Example 3: Tabular Suppression

- Published data are a large contingency table with confidential data available for every cell
- Oldest formal SDL technique (Fellegi 1972)
- Most common SDL method in use worldwide
- Some values are suppressed because either the data in the cell were deemed "sensitive" or a complementary suppression was needed to protect another sensitive cell
- The missing data items are *not ignorable*; therefore the SDL is *nonignorable*
- Suppression is almost always *discoverable*

# Example 4: Regression Discontinuity and Regression Kink

- Running variable subjected to SDL from the generalized randomized response class (noise infusion, suppress and impute, synthetic data)
- SDL is *nonignorable* because the estimated treatment effect is confounded by the probability that SDL was applied to the running variable
- SDL is *known* if the agency publishes that probability
- SDL is *discoverable* if the agency publishes the fact that a generalized randomized response method was used

# Example 4: Regression Discontinuity and Regression Kink (continued)

- In RD and RK designs where the probability of SDL is *known*, divide the treatment effect and its standard error by the probability
- In designs where the SDL is *discoverable* use the "compliance" function implied by generalized randomized response to estimate the treatment effect using the appropriate fuzzy RD/RK estimator; adjustment of standard error is part of the fuzzy RD/RK method
- All other RD/RK assumptions are identical to those made in the confidential data analysis
- Analysis similar to Lee and Card (2008)

# Example 5: Tabular Noise Infusion

- Same data structure as in tabular suppression: published data are a large contingency table with confidential data available for every cell
- Instead of suppression, all items are published by tabulating input data that have been infused with noise (every input value modified)
- SDL is *nonignorable*

# Example 5: Tabular Noise Infusion (continued)

- SDL is *known* if the agency publishes the statistical properties of the noise process—usually, *i.i.d.* with zero mean and constant known variance
- SDL is *discoverable* if there is another tabular summary with the same expected value but published using a different SDL method or different noise infusion parameters
- Example: Quarterly Workforce Indicators, Quarterly Census of Employment and Wages, County Business Patterns

# Example 6: Microdata Noise Infusion

- All of the following can be modeled as generalized random response:
  - Swapping
  - Input noise infusion (adding random values)
  - Suppress and impute (replacing some values with imputations; also called partial synthetic data)
  - Synthetic data (replacing some or all values with samples from the posterior distribution)
  - Case considered by Alexander, Davern and Stevenson (2010)
- SDL is *never ignorable*
- SDL is *discoverable*

# Example 6: Microdata Noise Infusion (continued)

- Regression analysis when the dependent variable has been subject to “suppress and impute” SDL
- No other SDL
- SDL is *nonignorable*
- SDl is discoverable if the agency publishes the suppression rate, the list of variables used in the imputation models, and uses the complete confidential data matrix (including observations with suppressions) for imputation
- Analysis similar to Bollinger and Hirsch (2006)

# Example 7: Synthetic Data with Validation

- Fully synthetic data (all values replaced with samples from their posterior predictive distributions)
- Models fit on synthetic data validated by the agency on the actual confidential data
- Custom tabulation SDL applied to output from the confidential data (usually some suppression of coefficients)
- SDL is *nonignorable*
- Synthetic data systems published with fully *SDL-aware analysis* tools

# Synthetic Data and Program Evaluation

- Fully synthetic data cannot identify RD/RK effects (program treatment effects)
- Do permit full development of the specification (semi-parametric estimation of the response and compliance functions)
- Certify design on synthetic data
- Validate certified design on the confidential data
- Prevents *ex post* adjustment of the evaluation (as in current best practices for clinical trials)
- Allows agency to fully restrict access to the confidential data without limiting evaluation options
- Publishes the intended evaluation

# The RD Evaluation Setup

$$
\begin{aligned}
y_{i1} &= \text{untreated outcome} \\
y_{i2} &= \text{treated outcome} \\
y_{i3} &= \text{RD running variable} \\
y_{i4} &= \text{indicator } 1\left[y_{i3} \geq \tau\right].
\end{aligned}
$$

- Variable $y_3$ has been subjected to formal privacy protection and is published as $z_3$.

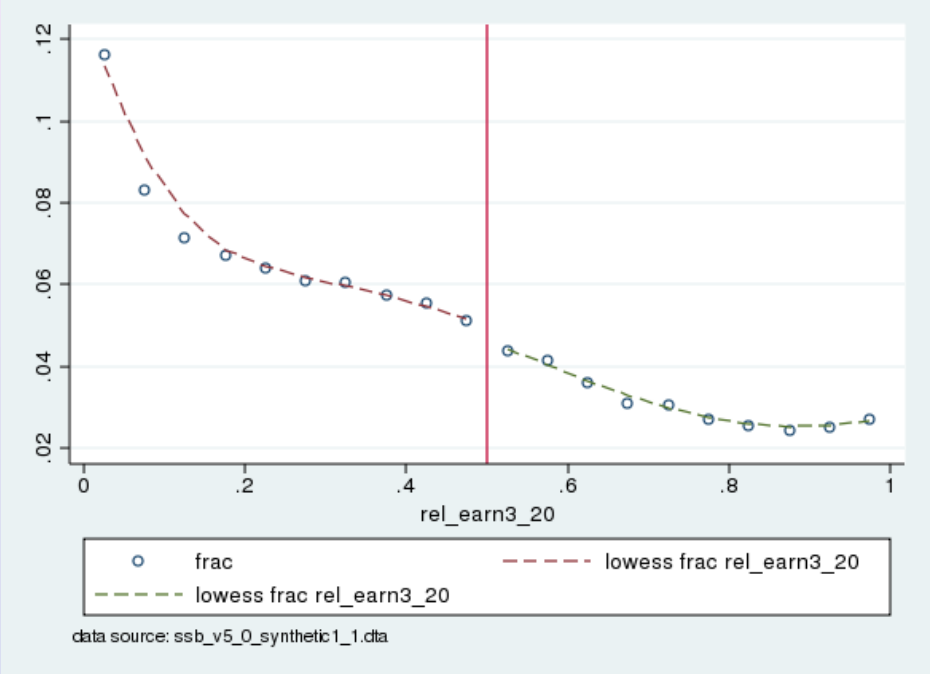
# Details of Evaluation with Synthetic Data

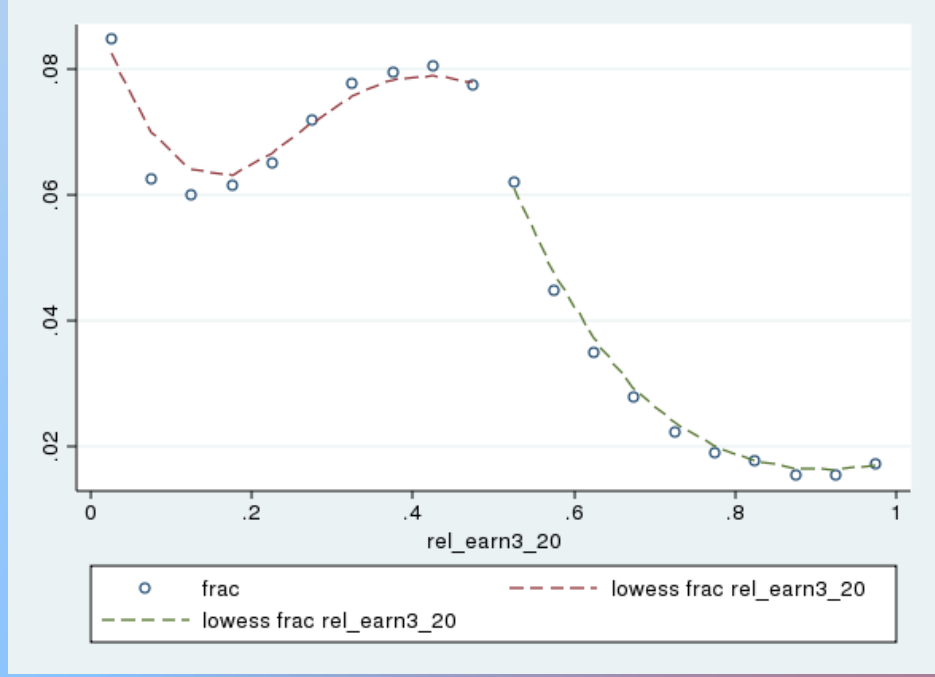
$$\hat{\theta}_{SRD} = \frac{\lim_{z_{i3}\downarrow\tau} \hat{f}_2(\tau) - \lim_{z_{i3}\uparrow\tau} \hat{f}_1(\tau)}{\rho_0}.$$

$$\hat{\theta}_{FRD} = \frac{\lim_{z_{i3}\downarrow\tau} \hat{f}_2(\tau) - \lim_{z_{i3}\uparrow\tau} \hat{f}_1(\tau)}{\lim_{z_{i3}\downarrow\tau} \hat{g}(\tau) - \lim_{z_{i3}\uparrow\tau} \hat{g}(\tau)}$$

$$g(z_{i3}) = \rho\, 1[z_{i3} \geq \tau] + (1-\rho)\, \Phi\left(\frac{z_{i3}-\tau}{\delta}\right)$$

# From Bertrand *et al.*




Timeline: SDS application November 2012, gold standard results January 2013. Bertrand, Marianne, Emir Kamenica and Jessica Pan (QJE 2015) "Gender identity and relative income within households."