Getting Back to Basics: The Why and How of Statistical Disclosure Limitation vs. Privacy Protection

Stephen E. Fienberg

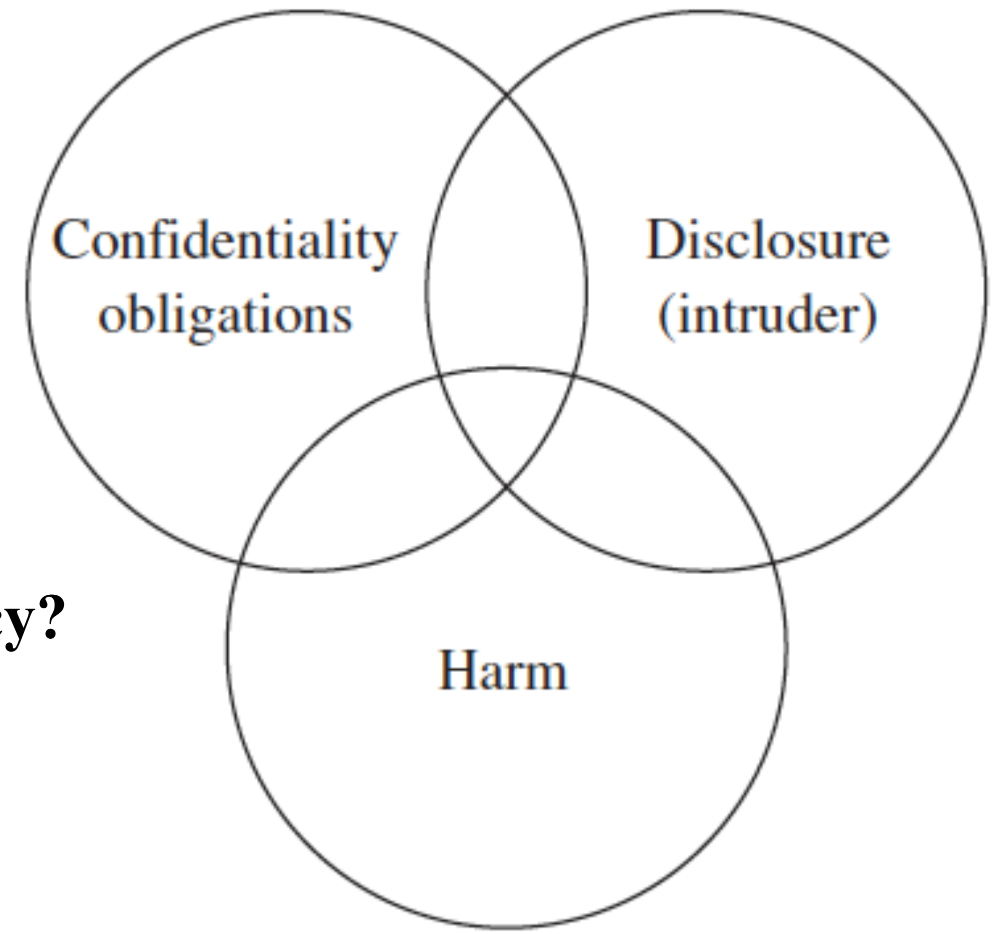
Department of Statistics, Heinz College, Machine Learning Department, and Cylab

Carnegie Mellon University

Pittsburgh, PA 15213-3890 USA

fienberg@stat.cmu.edu

Disclosure Limitation, Confidentiality & Harm



Where is privacy?

Privacy vs. Confidentiality

Privacy	**Don't ask.**
Confidentiality	**Don't tell.**

Prewitt, 2011

Do you agree/disagree with:

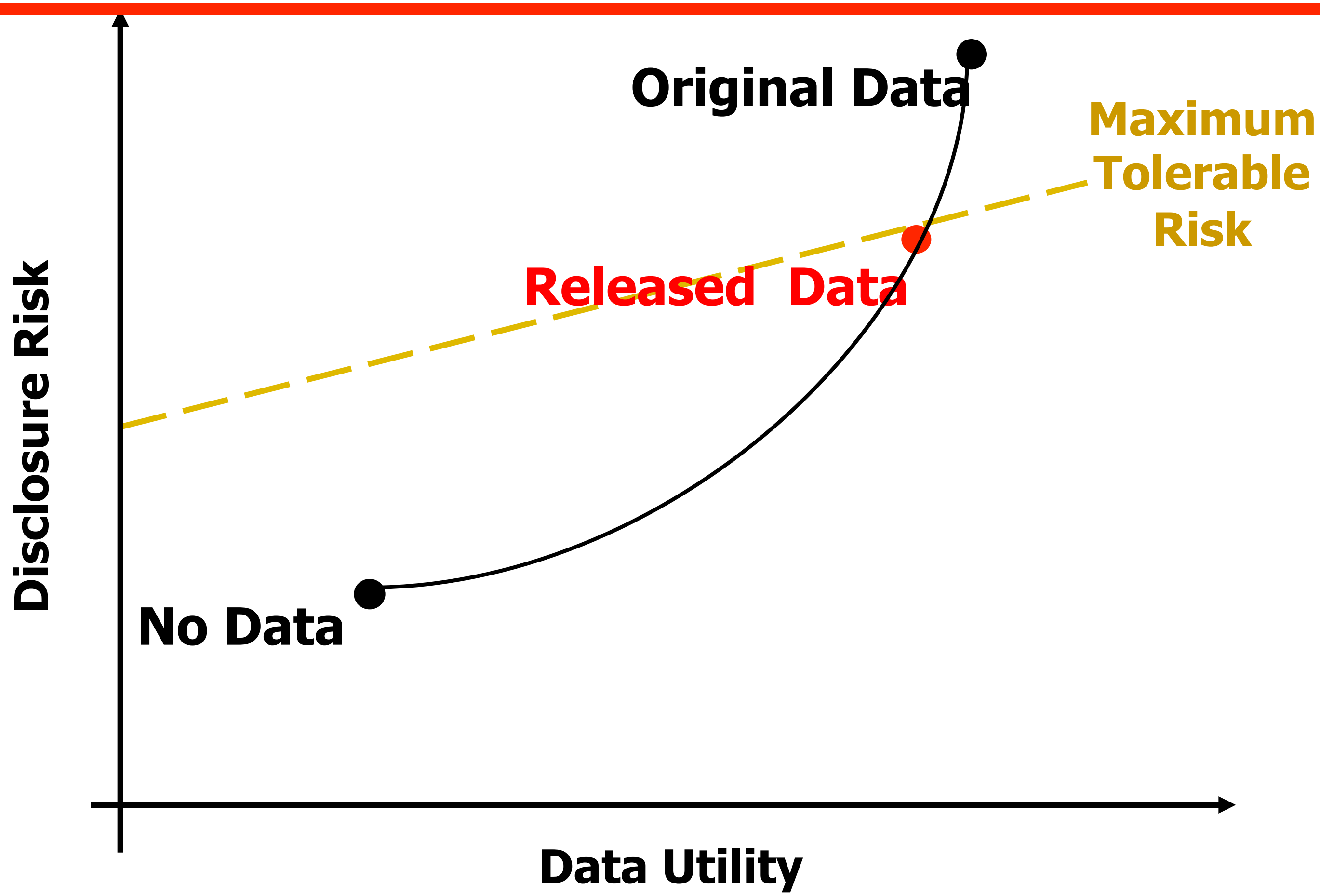
- The Census Bureau's promise of confidentiality cannot be trusted.
- My answers to the census could be used against me.
- The census is an invasion of my privacy.

Prewitt, 2011

Outline

- **Some Statistical ideas on confidentiality and privacy protection.**
- **Differential Privacy (DP) in a focused statistical problem:**
  - **Protecting contingency table data.**
- **Extensions to DP.**
- **Record Linkage as alternative to DP:**
  - **A partially baked idea!**

R-U Confidentiality Map



Usability, Transparency, & Duality in Privacy Protection

- **Usability:** extent to which released data are free from systematic distortions that impair inference.
- **Transparency:** extent to which methodology provides direct or implicit information on bias and variability resulting from disclosure limitation mask.
- **Duality:** extent to which methods aim at both disclosure limitation and making the maximal amount of data available for analysis.

General Methods for Protection

- **Removing obvious identifiers/near-identifiers**
  - **Names, geography, birthdate, etc.**
- **Data transformations:**
  - **Matrix masking** $X \rightarrow AXB + C$
    - **e.g., noise addition**
  - **Data suppression**
    - **Deleting cases / sampling**
    - **Cell suppression**
- **Synthetic data**

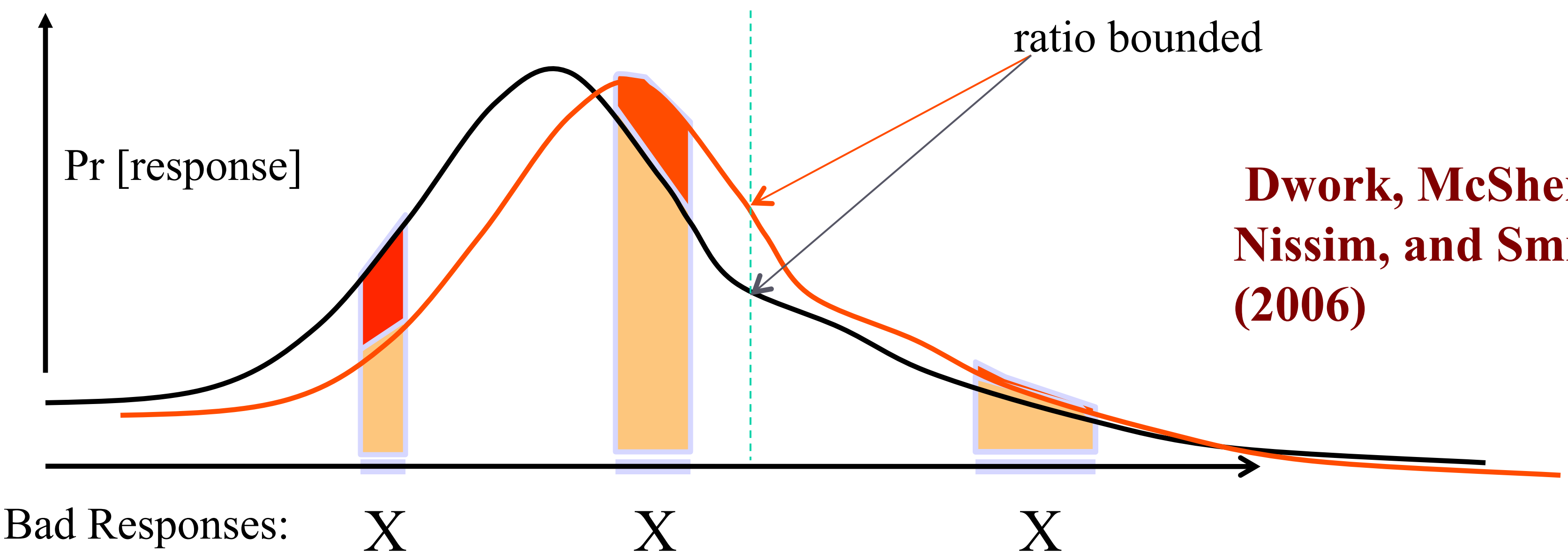
Inferential Utility

- **Want to achieve "Statistical reversibility" of data transformation:**
  - **Need (a) released data and (b) likelihood function including full information on transformation applied.**
  - **For noise addition this may involve using "measurement error model."**
- **Contrast with Naïve DP perspective and agency view of "just using" released data.**

Enter ε-Differential Privacy

Randomized function $\mathcal{K}$ gives ε-differential privacy if for all neighboring D_1 and D_2, and all $C \in \mathrm{range}(\mathcal{K})$:

$$e^{-\varepsilon} \leq \Pr[\mathcal{K}(D_1) \in C] / \Pr[\mathcal{K}(D_2) \in C] \leq e^{\varepsilon}$$



Dwork, McSherry, Nissim, and Smith (2006)

Differential Privacy

- Standard "DP mechanism" is addition of Laplace noise, with parameter ε.
  - The more data or statistics you protect the larger the noise required.
- Refinements such as "exponential mechanism," and perturbing an estimating equation, exploit a Lipchitz condition, and require less noise.

Differential Privacy

- **DP offers strong privacy "guarantees," through all possible violations, but…**
  - **Strong privacy "guarantees" may destroy utility of the data.**
  - **Does not recognize iterative and possibly nature of statistical data analysis.**
- **Research users want data sets to analyze, not DP-protected coefficients.**

Differential Privacy

- **DP is fundamentally a *frequentist* notion:**
  - **Privacy resides in the method that generates the altered data, as well as extremal aspects of data themselves.**
  - **Has the flavor on minimax approaches.**

Protecting Contingency Tables

Barak et al. (2007)

- **Want to release a set of altered MSS marginals.**
  - **Use Fourier coefficient basis for noise addition.**
  - **This produces non-integer and inconsistent margins.**
  - **Consistency of margins doesn't guarantee existence of a table satisfying released margins.**
  - **Barak et al. find "nearby" set of consistent integer margins which preserve DP property.**
- **Assessment by Fienberg, Rinaldo and Yang (2010, 2011) show that the approach obliterates the data for large sparse tables.**

Lessons Learned

- **As ε increases, amount of noise added decreases**
  - **Deviance between DP generated tables and real MLEs gets smaller.**
  - **If we add a lot of noise, it has strong privacy guarantees but the statistical inference becomes infeasible.**
  - **When we add little noise, the statistical inference is better but no privacy guarantees.**
- **DP struggles with releasing useful information associated with large sparse contingency tables.**

Implications

- **Need to incorporate RU ideas into DP formulation for data releases to have real utility:**
    - **Learn how to draw inferences from privacy-protected releases.**
    - **Focus on model search processes, not simply reporting one set of summary statistics.**
- **Move from frequentist to Bayesian formulation.**

Extensions to DP: I

- **(ε, δ)-DP** (Dwork, et al. 2006)
  - **A randomized algorithm $\mathcal{K}$ gives (ε,δ)-DP if for all $S \subseteq Range(\mathcal{K})$,**

$$Pr[\mathcal{K}(D_1) \in S] \leq e^{\varepsilon} Pr[\mathcal{K}(D_2) \in S] + \delta,$$

where the probabilities are over the coin flips of the algorithm $\mathcal{K}$.

Extensions to DP: II

- **(ε, δ)-Probabilistic DP** **(Machanavajjhala et al., 2008)**

  $\Pr[\mathcal{K}(D) \in \text{Disc}(D, \varepsilon)] < \delta$.

  - **Claim:** **(ε, δ)-PDP lies strictly between (ε, δ)-DP and ε-DP.** **True?**
  - **How do we compute Disc(D, ε)? With respect to a prior, w.r.t. the joint distribution of the data and the prior (Abowd and Vilhuber, 2008), w.r.t. the randomizing function?**

Extensions to DP: III

(ε, δ)-Random DP (Hall, Rinaldo, Wasserman, 2011)

$Pr_D\{e^{-\varepsilon} \leq Pr[\mathcal{K}(D_1) \in C]/Pr[\mathcal{K}(D_2) \in C] \leq e^{\varepsilon}\} \geq 1-\delta$

- **Key here is that data are treated as random and deviations from DP are with respect to distribution of data.**
- **D_2 adds a randomly drawn new data element to database D_1.**
- **Get composition property w.r.t. ε and much better utility w.r.t. risk function.**

Related DP Issues

- **Should the bound on**

$$|\Pr[\mathcal{K}(D_1) \in C] / \Pr[\mathcal{K}(D_2) \in C]|$$

  be constant, ε , or depend on D?

- **Should perturbations of the data always involve adding continuous noise?**
  - **What about restricted swapping for count data?**

Statistical View of Record Linkage (Hall & Fienberg)

There exist two sets of observable records:

$$A = \{a_1 \ldots a_n\} \qquad B = \{b_1 \ldots b_m\}$$

Data are via **model** depending on Q

$$P_\theta(A, B; Q)$$

Goal of record linkage is to **estimate** the parameter Q

$$Q \in \{0,1\}^{n\times m} \qquad q_{i,j} = \begin{cases} 1 & a_i,\ b_j \text{ link} \\ 0 & \text{o/w} \end{cases}$$

There is an **unknown** matrix that contains **the true record linkage** information.

"Privacy" Overview

Goal: To release a sanitized database that includes potentially sensitive data elements, while maintaining individual privacy.

Police Records

Name	Address	Criminal?
Robert	123 Fake St	N
Dave	456 Fake St	Y

Sanitized Police Records

Name	Zip Code	Criminal?
REDACTED	15232	N
REDACTED	15232	Y

In general, we must **sanitize** the data somehow.

Adversary's Data

Name	City	…
Robert	Pittsburgh	…
…	…	…

Envision an adversary attempting to **infer the sensitive information** via **record linkage.**

Setting/Assumptions

The columns of the data partition into the sensitive attributes, and the quasi-identifiers:

Name	Address	Criminal?
Robert	123 Fake St	N
Dave	456 Fake St	Y

"Quasi-identifiers" aka "key variables" (Name, Address) **"Sensitive attribute"** (Criminal?)

complete record, ***sensitive attributes***, **quasi-identifiers**:

$$\underbrace{a_i}_{\text{complete record}} = (\underbrace{a'_i}_{\text{quasi-identifiers}}, \underbrace{s_i}_{\textit{sensitive attributes}})$$

The goal is to release a set of sanitized records:

$$b_i = (b'_i, \tilde{s}_i)$$

"Privacy" and Record Linkage

- Suppose adversary knows exact values for quasi-identifiers for subset of records in private database:

Complete database

$$A = \{a_1 \ldots a_n\}$$

Adversary's database

$$A' = \{a'_{i_1} \ldots a'_{i_m}\}$$

$$P_\theta(A, B; Q)$$

Choose a permutation *Q* uniformly at random, and a model *P*, then draw *B*|*A;Q*

$$B = \{b_1 \ldots b_n\}$$

Sanitized database

Adversary faces record linkage problem, where model is specified by the data owner.

Fully Bayesian “Privacy”?

- Suppose that the choice of model *P* is made public knowledge:
- Then the “correct” way to do inference about *S* is to maintain uncertainty about the record linkage:

$$\pi(S \mid B) \propto \sum_{Q_i \in Q} P_\theta((A', S), B; Q_i)\pi(S)$$

(sum over all possible linkage structures)

- A possible criterion for privacy protection would be to require the “statistical distance” between the posterior and prior is small for all prior distributions: $D_H(\pi(\cdot), \pi(\cdot \mid B)) \leq \tau$
- Adversaries and legitimate statisticians are treated the same.
- Choice of D_H and τ gives tradeoff between utility and privacy.

Fully Bayesian "Privacy"?

- **Some Context:**
  - ***k-anonymity, l-diversity, t-closeness*** **may be viewed as successively improving approximations to this idea, but they also unnecessarily restrict the model class.**

    P(A,B;Q) concentrated on {B: B is *k*-anonymized}
- **"Protect" sensitive values?**
  - **We could output exact identifiers, allow adversary perfect record linkage, but apply double exponential or other kind of perturbtions to sensitive attributes.**
  - **Expanded options to explore.**
- **We need to understand the formal properties.**

Relationship to DP

- **Differential privacy from BP perspective:**
  - **Adversary has *n*-1 complete records and belief about *n*th record doesn't change much when seeing data.**
  - **DP criterion implies Hellinger distance (*f*-information).**
  - **In BP approach, use *n*-1 quasi-identifiers, and point mass prior on *n* true sensitive values.**
    - **Adversary's prior on *n*th sensitive value doesn't change much re inferring quasi-identifiers for *n*th record.**
    - **Choice of distance function, e.g., KL-information.**
    - **BP scheme doesn't protect the identifiers.**

Summary

- **Some Statistical ideas on confidentiality and privacy protection.**
- **Differential Privacy (DP) in a focused statistical problem:**
  - **Protecting contingency table data.**
- **Extensions to DP.**
- **Record Linkage as alternative to DP:**
  - **A partially baked idea!**

End

- **My CMU privacy collaborators:**
  - **Rob Hall, Jiashin Jin, Alessandro Rinaldo, Xiaolin Yang, Larry Wasserman**
- **Joint CMU/PSU/Cornell collaboration**