Towards a Bayesian Characterization of Privacy Protection & the Risk-Utility Tradeoff

Stephen E. Fienberg

Department of Statistics, Machine Learning Department, Cylab, and i-lab

Carnegie Mellon University

Pittsburgh, PA 15213-3890 USA

fienberg@stat.cmu.edu

Towards a Bayesian Characterization of ~~Privacy~~ Confidentiality Protection & the Risk-Utility Tradeoff

Stephen E. Fienberg

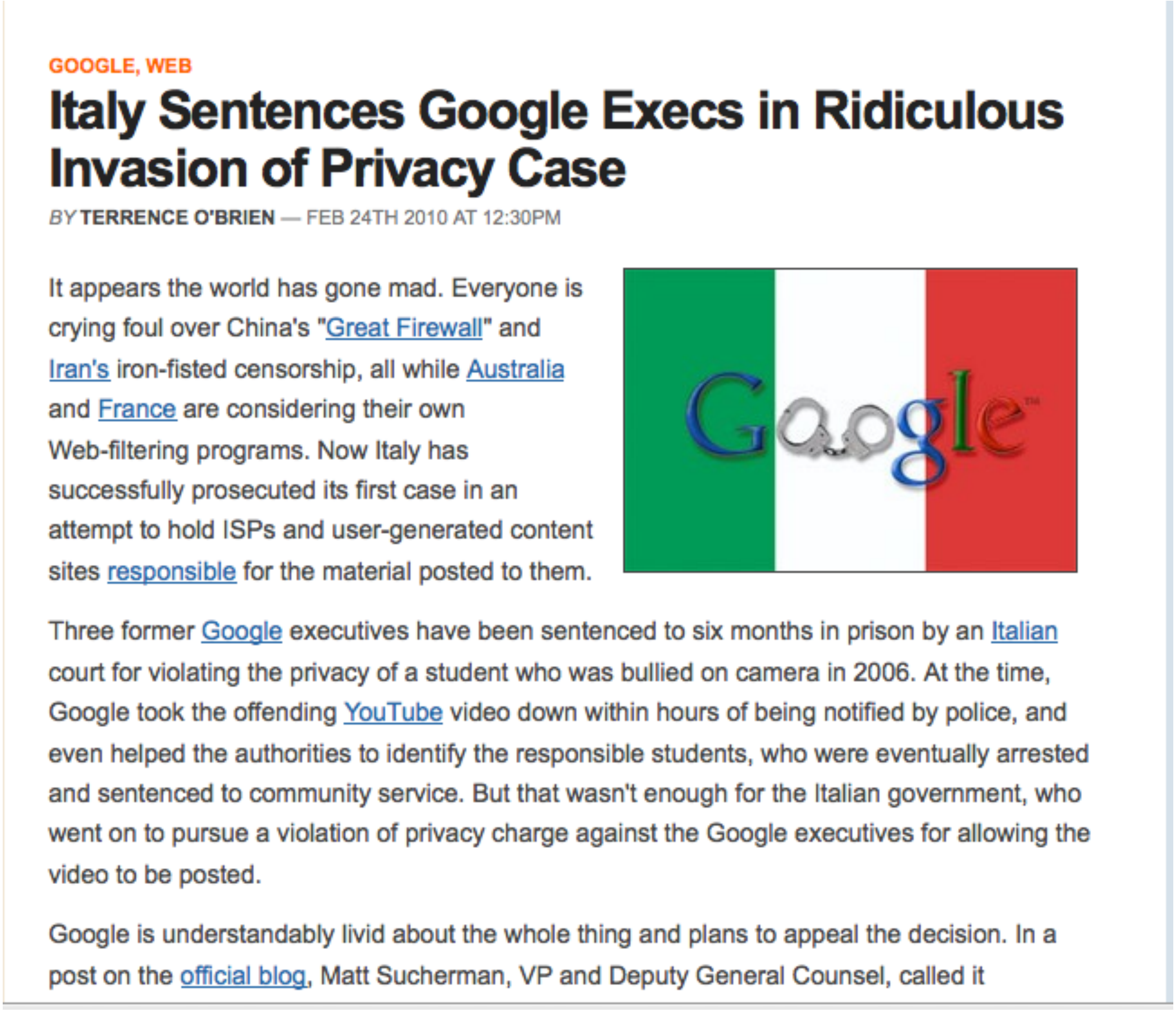
Department of Statistics, Machine Learning Department, Cylab, and i-lab

Carnegie Mellon University

Pittsburgh, PA 15213-3890 USA

fienberg@stat.cmu.edu

Journal of Privacy and Confidentiality http://repository.cmu.edu/jpc/



GOOGLE, WEB

Italy Sentences Google Execs in Ridiculous Invasion of Privacy Case

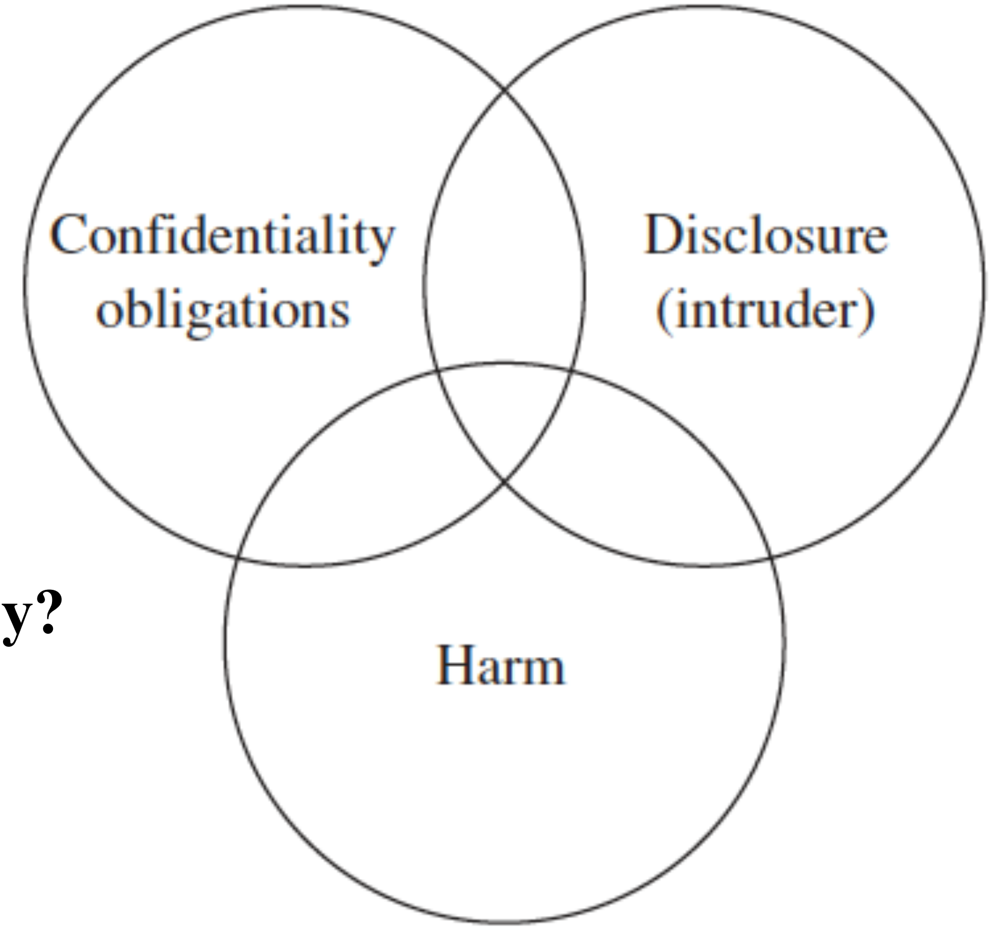
BY **TERRENCE O'BRIEN** — FEB 24TH 2010 AT 12:30PM

It appears the world has gone mad. Everyone is crying foul over China's "Great Firewall" and Iran's iron-fisted censorship, all while Australia and France are considering their own Web-filtering programs. Now Italy has successfully prosecuted its first case in an attempt to hold ISPs and user-generated content sites responsible for the material posted to them.

Three former Google executives have been sentenced to six months in prison by an Italian court for violating the privacy of a student who was bullied on camera in 2006. At the time, Google took the offending YouTube video down within hours of being notified by police, and even helped the authorities to identify the responsible students, who were eventually arrested and sentenced to community service. But that wasn't enough for the Italian government, who went on to pursue a violation of privacy charge against the Google executives for allowing the video to be posted.

Google is understandably livid about the whole thing and plans to appeal the decision. In a post on the official blog, Matt Sucherman, VP and Deputy General Counsel, called it

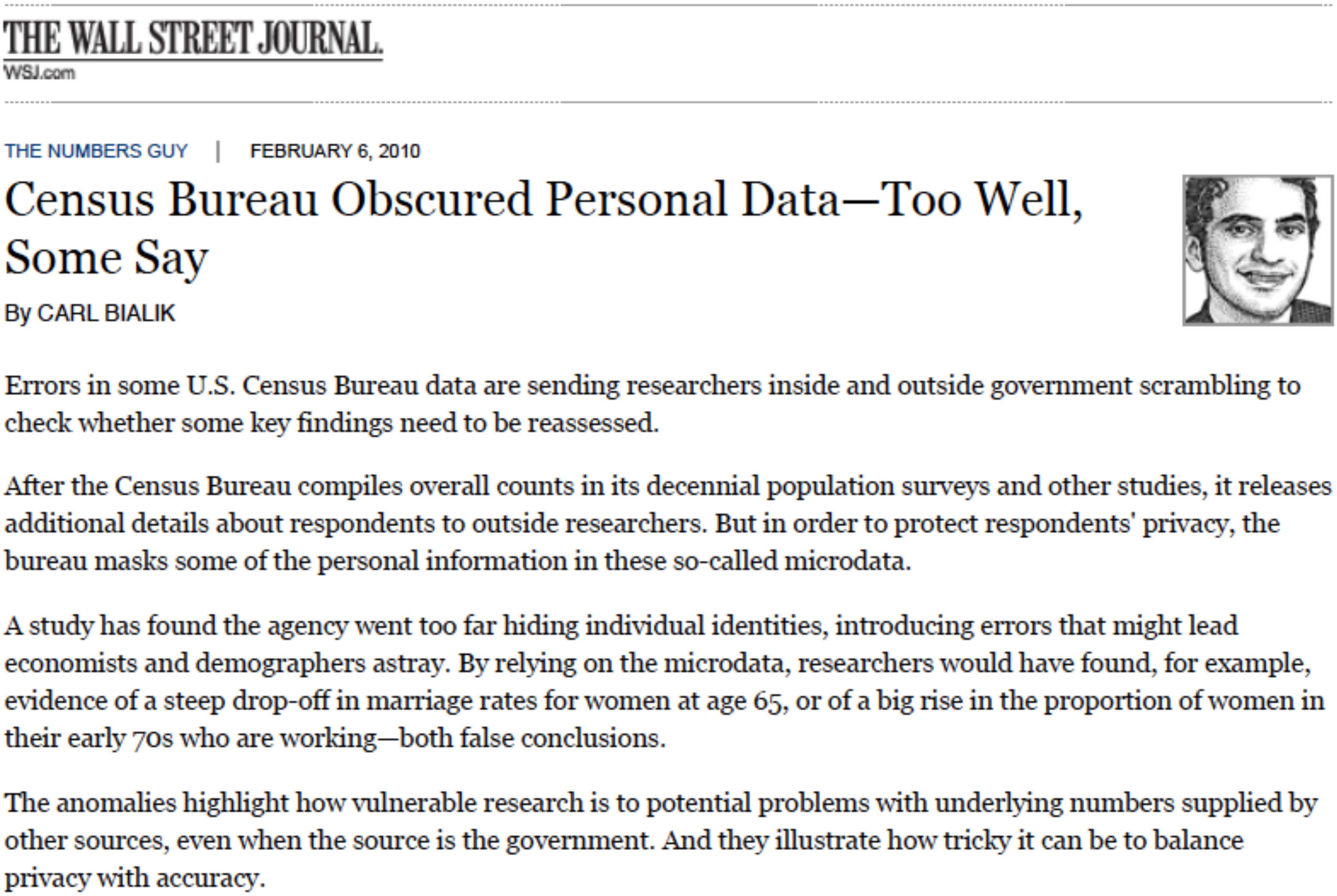
Disclosure Limitation, Confidentiality & Harm



Where is privacy?

Outline

- The Census Bureau snafu.
  - Principles of data sharing and statistical disclosure limitation.
  - Risk-Utility trade-off.
- Differential Privacy (DP) in a focused statistical problem:
  - Protecting contingency table data.
- Record Linkage as alternative to DP.
  - A partially baked idea!



THE WALL STREET JOURNAL.
WSJ.com

THE NUMBERS GUY | FEBRUARY 6, 2010

Census Bureau Obscured Personal Data—Too Well, Some Say

By CARL BIALIK

Errors in some U.S. Census Bureau data are sending researchers inside and outside government scrambling to check whether some key findings need to be reassessed.

After the Census Bureau compiles overall counts in its decennial population surveys and other studies, it releases additional details about respondents to outside researchers. But in order to protect respondents' privacy, the bureau masks some of the personal information in these so-called microdata.

A study has found the agency went too far hiding individual identities, introducing errors that might lead economists and demographers astray. By relying on the microdata, researchers would have found, for example, evidence of a steep drop-off in marriage rates for women at age 65, or of a big rise in the proportion of women in their early 70s who are working—both false conclusions.

The anomalies highlight how vulnerable research is to potential problems with underlying numbers supplied by other sources, even when the source is the government. And they illustrate how tricky it can be to balance privacy with accuracy.

Census Costs & Products

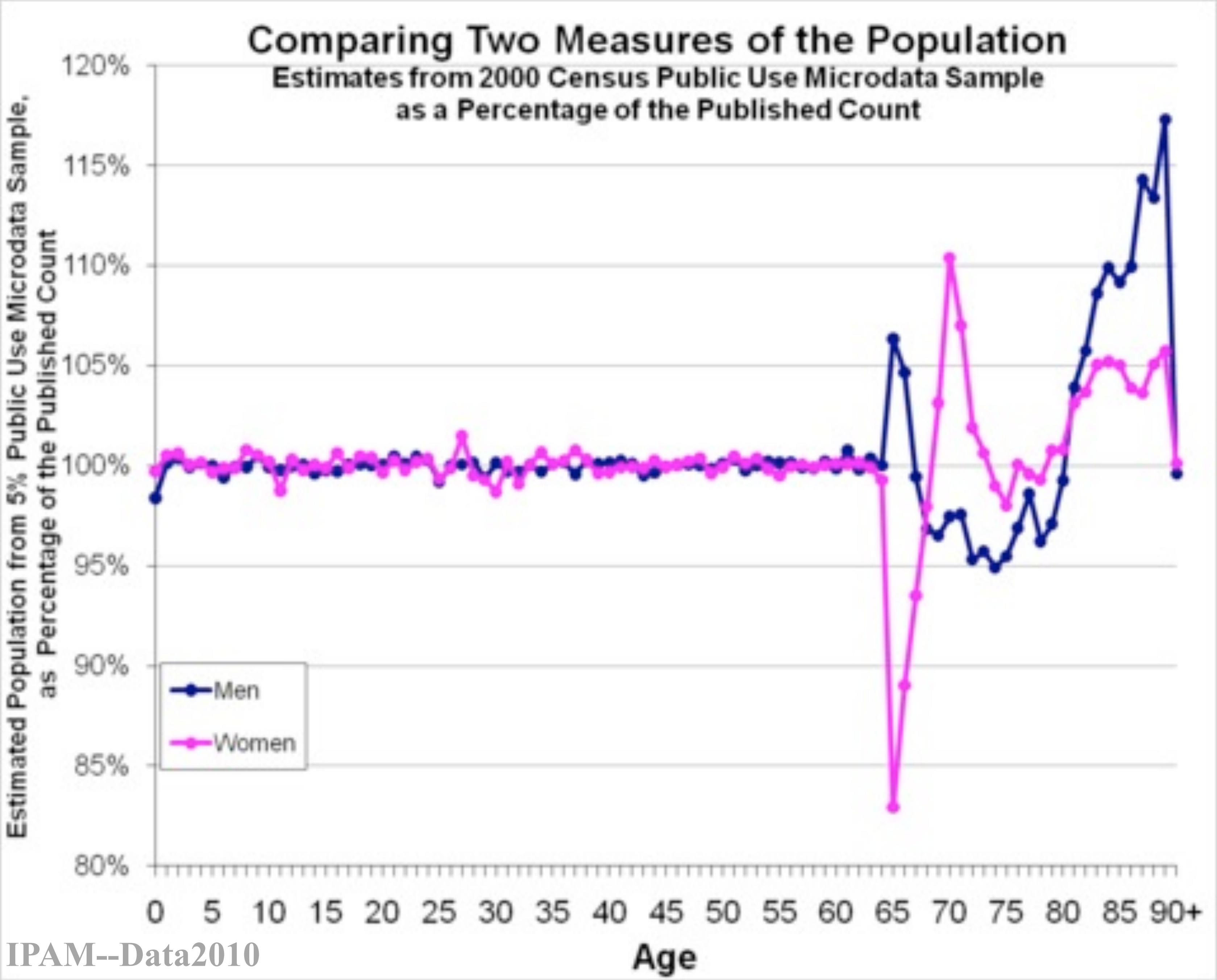
- Costs: $6.5 billion in 2000; $14 billion+ in 2010
- Short form data (100%)
  - State totals by Dec. 31 for reapportionment
  - Age (<18, ≥18) × Gender × Race for each census block to states for redistricting
- Long form data (sample of 1 in 6) via American Factfinder (replaced by ACS data in 2010):
  - Allocation of funds: $400 billion in 2010
  - Tables and special packages (e.g., travel-to work info for urban planners, etc.)
  - 1% and 5% PUMS

Public Use Microdata Samples (PUMS)

- **5% PUMS Files**
  - PUMS contain individual data for geographic units known as super-Public Use Microdata Areas (super-PUMAs) and Public Use Microdata Areas (PUMAs). Each PUMA must have a minimum of 100,000 population and each super-PUMA contains a minimum population of 400,000.

Census Disclosure Protection Approach

- Data swapping & Sampling & Imputation/ Editing
- PUMS files
  - Top-coding for variables like income
  - Population controls for geography
  - Some outlier values are averaged together, and that average is assigned to every one of those outliers.
  - Addition of statistical noise to the subset of older respondents
- No details on properties of each of these components, e.g. % of swapped files


Comparing Two Measures of the Population
Estimates from 2000 Census Public Use Microdata Sample as a Percentage of the Published Count
Estimated Population from 5% Public Use Microdata Sample, as Percentage of the Published Count
120%
115%
110%
105%
100%
95%
90%
85%
80%
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90+
Age
Men
Women

Census Bureau Response

- “We want to preserve confidentiality, and we want to maximize utility of our data. This tension is inherent in everything we do,” says Robert M. Groves, director of the Census Bureau.
- “Flawed software code designed to add the statistical noise to the subset of older respondents should have offset those changes with opposite adjustments made elsewhere in the data sample. This didn't happen as it should have, so that ages and other attributes were skewed.”
- Before the data were released in 2003, the Census Bureau's diagnostic tools flagged the problem, but it “didn't seem large enough in the judgment of our analysts to stop the release,” says Dr. Groves.

Morales of Story

- For the Census Bureau:
  - Ad hockery in DL can lead you astray.
  - Not releasing the details of DL methodology will likely get you in trouble in the long run.
- For us at this workshop:
  - “Accuracy” of “released” statistical data matters to both users and data owners.
  - Privacy protection is for the data at hand and not for possible replications that we will never see.

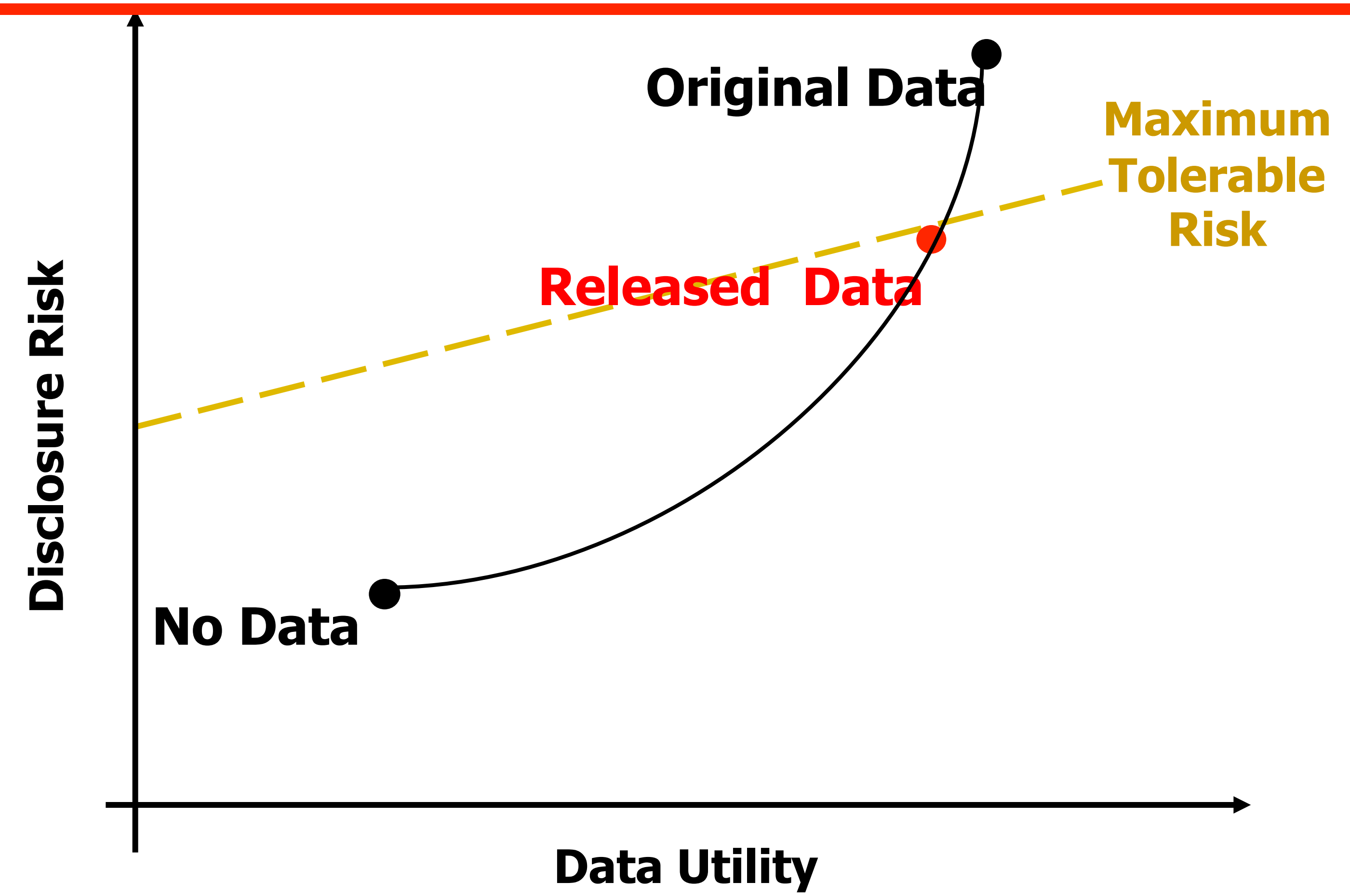
Usability, Transparency, & Duality in Privacy Protection

- **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.

Inferential Utility

- “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” since most (all) of variables are measured with error.
  - Williams-McSherry probabilistic inference?

R-U Confidentiality Map

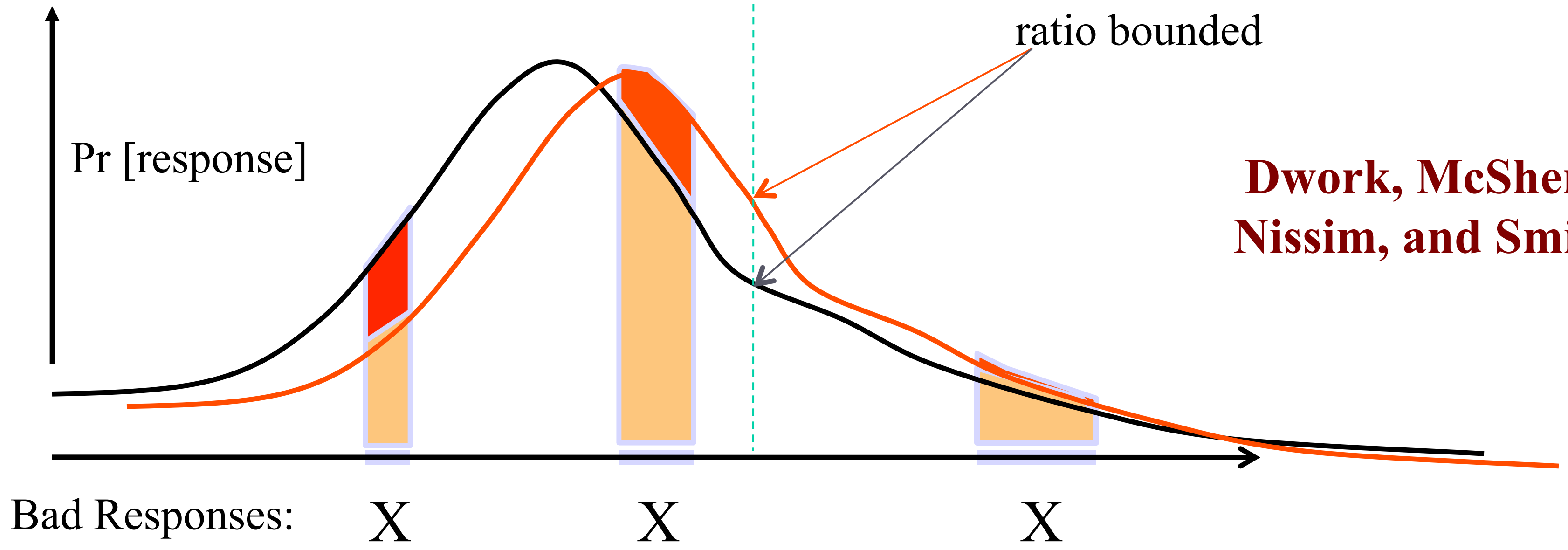


(Duncan, et al. 2001, 2004, Trottini, 2002; Ting et al. 2008)

ε-Differential Privacy

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

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



Dwork, McSherry, Nissim, and Smith

Differential Privacy

- DP offers strong privacy "guarantees," through all possible violations, but…
  - Strong privacy "guarantees" may destroy utility of the data.
  - Does not recognize the iterative and possibly unstructured 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.
- But for "my problems," data are in hand when we begin to consider data release and disclosure limitation (not privacy).

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.**
- **What about**
  - **releasing *n*? Known in all of my applications!**
  - **utility?**

Ongoing Work

Poster by Yang, et al.:

- **Edwards 2^6 genetics table, with *n*=70.**
- **Czech auto workers 2^5 heart attack risk table, with *n*=1,841.**
- **American Community Survey 4 × 4 × 16 travel to work table (need extension to Barak et al. method).**

- **NLTCS**
  - **2^{16} disability table with *n*=21,574.**
  - **2^{96+5} version based on 6 waves (plus mortality), *n*~45,000. Our models have no MSSs!**

Our Approach

- We have been using an ad hoc approach to utility by looking at
  - For each noise level, we compute the deviance (KL-distance) between the MLE and 100 tables perturbed at this noise level.
  - Really want something more like the probabilistic inference described by Williams & McSherry, but it's too complex given dimensionality.

Specific Implications

- 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.

Possible Implications

- **We need to:**
  - **Incorporate RU ideas into DP formulation so that data releases have real utility.**
    - **Learn how to draw inferences from privacy-protected releases–Williams&McSherry again!**
    - **Focus on model search processes, not simply reporting one set of summary statistics.**
  - **Move from frequentist to Bayesian formulation:**
    - **Provide protection for actual data at hand.**
    - **Identify inferences from “record linkage”?**

Record Linkage Overview

File 1

Name	Address	…
Rob	123 Fake St	…
…	…	…

“Quasi-identifiers”

File 2

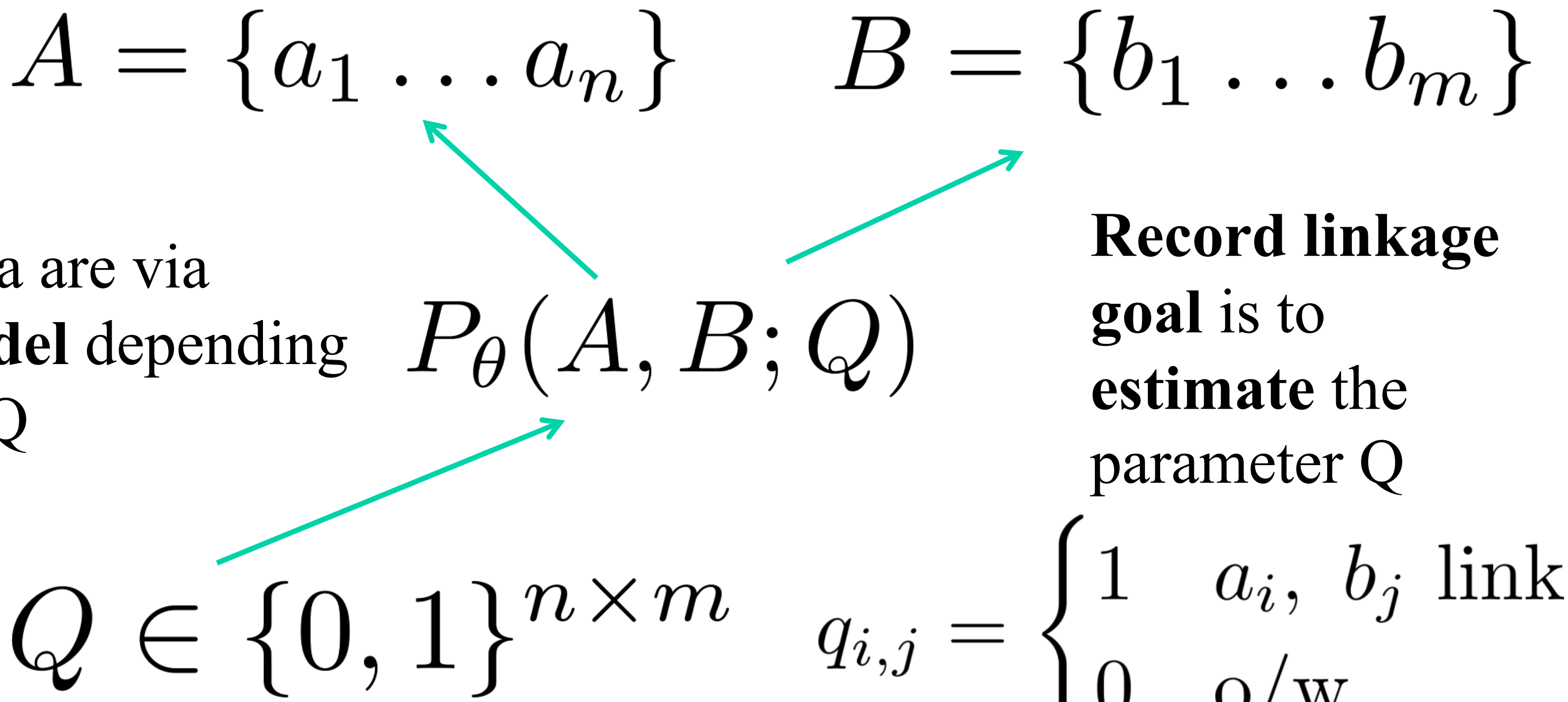
Name	City	…
Robert	Pittsburgh	…
…	…	…

Inherent uncertainty due to:

- **Sampling**
- **Typographical variation**
- **Measurement error**
- **Different survey times**

Statistical View of Record Linkage

There exist two sets of observable records:



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

Data are via **model** depending on Q

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

Record linkage goal 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 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

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

Sanitized Police Records

Name	Zip Code	Criminal?
REDACTED	15232	N
REDACTED	15232	Y

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"

"Sensitive attribute"

complete record

sensitive attributes

$$a_i = (a_i', s_i)$$

quasi-identifiers

The goal is to release a set of sanitized records:

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

“Privacy” and Record Linkage

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

Complete database

Adversary’s database

$$\begin{tikzcd}
A = \{a_1 \ldots a_n\} \arrow[d] & A' = \{a'_{i_1} \ldots a'_{i_m}\} \arrow[ddl, leftrightarrow] \\
P_\theta(A, B; Q) \arrow[d] & \\
B = \{b_1 \ldots b_n\} &
\end{tikzcd}$$

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

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:
- The "correct" way to do inference about S is to maintain uncertainty about the record linkage:

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

(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: $D_H\left(\pi(\cdot), \pi(\cdot|B)\right) \leq \tau$
- 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 output exact identifiers, allow adversary perfect record linkage, but apply double exponential or any other kind of noise to sensitive attributes.
  - Expanded options to explore.
- Role of calibration and refinement? (Kifer)

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

- The Census Bureau snafu.
  - Principles of data sharing and statistical disclosure limitation.
  - Risk-Utility trade-off.
- Differential Privacy (DP) in a focused statistical problem:
  - Protecting contingency table data.
- Record Linkage as alternative to DP.
  - A partially baked idea!

End

- My collaborators on this ongoing work:
  - Alessandro Rinaldo and Xiaolin Yang
  - Rob Hall

References

- Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: a holistic solution to contingency table release. *PODS* 2007: 273–282.
- Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., Fienberg, S.E. (2003). Adaptive name matching in information integration. IEEE Intell. Syst. 5, 16–23.
- DeGroot, M.H. and Fienberg, S.E. (1982). The comparison and accuracy of forecasters. The Statistician, 32, 12–22.
- Dobra, A., Fienberg, S.E., Rinaldo, A., Slavkovic, A.B. and Zhou, Y. (2008). In *Emerging Applications of Algebraic Geometry* (M. Putinar and S. Sullivant, eds., Springer, New York, 63–88.
- Fellegi, I. and Sunter, A. (1969). A theory for record linkage. *JASA,* 64, 1183–1210.
- Lahiri, P. and Larson, M.D. (2005). Regression Analysis with Linked Data.. *JASA*, 100, 222–230.