Disclosure Limitation

BLS Future in Disclosure Limitation

Daniell Toth

U.S. Bureau of Labor Statistics

The Penn State Team

Kaitlyn Ruth Dowden

John Durrell

Prottay Protivash

Aleksandra (Sesa) Slavkovic

Daniel Kifer

Danfeng Zhang

Why Disclosure Limitation?

- Purpose of collecting data is to make data available for use.

- However, we promise to keep your responses confidential.

- Goal: Choose a method that protects the individual users responses from being known, while providing useful data.

QCEW

Provides employment and wage data in tabular form

NAICS	e20101	e20102	e20103	e20104	total
Series 1	2600	2899	3022	2599	11120
Sub1	1981	2256	2382	1957	8576
Sub2	32	33	37	33	135
Sub3	587	610	603	609	2409

QCEW

Provides employment and wage data in tabular form

NAICS	e20101	e20102	e20103	e20104	total
Series 1	2600	2899	3022	2599	11120
Sub1	1981	2256	2382	1957	8576
Sub2	32	33	37	33	135
Sub3	587	610	603	609	2409

Need to protect sensitve cells

Sensitive Cells

$$T = X_1 + \ldots + X_n$$

where $X_i \geq X_{i+1}$

- Cell too small $n < 3$
- P% - Rule Fails

P%-Rule

$$R = X_3 + \ldots + X_n$$ remainder

$$T = X_1 + X_2 + R$$

Let $p \in (0, 1)$

Suppress if remainder is too small

$$R < pX_1$$

Motivation of P%-Rule

Suppose respondant 2, wants to know the value of respondent 1.

Estimate value $E_1 = T - X_2 = X_1 + R$

if $R < pX_1$

then $E_1 < (1+p)X_1$

so $E_1 \in \left(X_1, (1+p)X_1\right)$

Cell Suppression

	Q1	Q2	Q3	Q4	Annual Total
Industry 1	22	22	23	22	89
Industry 2	16	17	15	17	65
Industry 3	15	15	13	15	58
Total	53	54	51	54	212

sensitive cell

Remove Value

	Q1	Q2	Q3	Q4	Annual Total
Industry 1	22	[redacted]	23	22	89
Industry 2	16	17	15	17	65
Industry 3	15	15	13	15	58
Total	53	54	51	54	212

Can’t Remove Just One

	Q1	Q2	Q3	Q4	Annual Total
Industry 1	22	[redacted]	23	22	89
Industry 2	16	17	15	17	65
Industry 3	15	15	13	15	58
Total	53	54	51	54	212

Secondary Cell Suppression

	Q1	Q2	Q3	Q4	Annual Total
Industry 1	22	[redacted]	[redacted]	22	89
Industry 2	16	17	15	17	65
Industry 3	15	15	13	15	58
Total	53	54	51	54	212

Cox (1995) uses Complicated algorithm to find secondary suppressions

Quickly Looks Like "Swiss Cheese"

	Q1	Q2	Q3	Q4	Annual Total
Industry 1	22	[redacted]	[redacted]	22	89
Industry 2	16	17	15	17	65
Industry 3	15	[redacted]	[redacted]	15	58
Total	53	54	51	54	212

Suppression

Advantages

+ Provides accurate totals for cells that are published

Disadvantages

- No information for some cells
- QCEW suppresses over 60% of all possible cells
- No formal guarantee of protection
- Difficult to manage additional publications

Formal Privacy

Given the dataset D, let M(D) the released statistic after applying the disclosue limitaion method.

Example: The QCEW employment table with suppressed cells

Let be D^* a copy of the dataset with one of the observed values x, changed to $x^* = (1 \pm p)x$

A formally private method uses a ***stochastic mechanism*** M and its protection is guarteed by the fact that for all[‡] D^*

$$P(M(D^*) = M(D)) > 0$$

‡ or at least most of the relevant values in the range of M

Cell Suppression

Deterministic Method

$P(M(D^*) = M(D))=1$ or $P(M(D^*) = M(D))=0$

If value x is in a suppressed cell then

$$M(D^*) \quad = \quad M(D)$$

	Q1	Q2	Q3	Q4
Industry 1	22	[redacted]	23	22
Industry 2	16	17	15	17
Industry 3	15	15	13	15

	Q1	Q2	Q3	Q4
Industry 1	22	[redacted]	23	22
Industry 2	16	17	15	17
Industry 3	15	15	13	15

not true if we publish annual totals

Cell Suppression

Deterministic Method

$P(M(D^*) = M(D))=1$ or $P(M(D^*) = M(D))=0$

If value x is not in a suppressed cell then

$$M(D^*) \neq M(D)$$

	Q1	Q2	Q3	Q4
Industry 1	22	[redacted]	**27**	22
Industry 2	16	17	15	17
Industry 3	15	15	13	15

	Q1	Q2	Q3	Q4
Industry 1	22	[redacted]	23	22
Industry 2	16	17	15	17
Industry 3	15	15	13	15

Formal Privacy

Formally Private Method

$$P(M(D^*) = M(D)) > 0$$

If M adds random noise N(0, 1) to each cell value then rounds. Then with Probabity >> 0

D = M(D)

	Q1	Q2	Q3	Q4
Industry 1	22	22	23	22
Industry 2	16	17	15	17
Industry 3	15	15	13	15

	Q1	Q2	Q3	Q4
Industry 1	23	21	23	21
Industry 2	16	19	13	17
Industry 3	15	15	14	15

Formal Privacy

Advantages

+ Allows publication of most cells with small relative error
+ Guaranteed protection under very weak assumptions
+ Provides an easy way to manage new publications of data
+ Protection of one response is independent of others

Disadvantages

- Cell totals will have error
- Must use for other non-optimized applications

Randomized Response

- Warner (1965) proposed using random mechanism to change responses with known probability.
- Fuller (1993) proposed using additive noise to mask true values.

$$\tilde{y}_i = y_i + \epsilon_i$$

- Dwork (2008) developes differentially private definition

$$\mathbb{P}(M(D) \in S) \quad \leq \quad e^{\varepsilon}\, \mathbb{P}(M(D') \in S)$$

and framework for choosing noise level and protection guarantees.

- Wasserman & Zhou (2010) relates protection guarantee of ε-δ

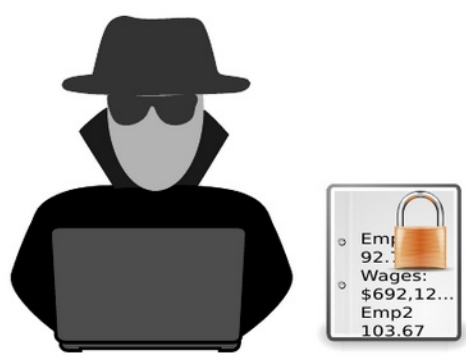
$$\mathbb{P}(M(D) \in S) \quad \leq \quad e^{\varepsilon}\, \mathbb{P}(M(D') \in S) + \delta$$

to hypothesis testing.

Protection Guarantee

- Difficulty of inference is expressed as point hypothesis test. E.g
  - Null: employment = 100 (reported value)
  - Alternate: employment = 110
  - Evidence: published confidentiality protected data 

100 vs. 110



BLS Approach in Development

- M(Employment by Establishment)
  - Noise is added to each establishment's employee data independently.
  - Uncertainty interval parameter β.
  - Privacy budget to spend: μ

- Establishment i:
  - M adds additive noise $N(0, \sigma^2)$ with $\sigma = \beta/\mu$ to $\sqrt{\text{employment}}$.
  - This is converted to unbiased employment estimate: $(\sqrt{employment} + N(0, \sigma^2))^2 - \sigma^2$
  - Attacker sees noisy employment: $\widetilde{E}$.
  - Can attacker distinguish between whether noise was added to E_1 vs. E_1'?
    - For any given significance level α, power in deciding E_1 vs. E_1' has slightly less than power in deciding between $N(0, 1)$ vs. $N(\mu, 1)$.

Protection vs Accuracy

Protection and accuracy is decided by choice of parameters

- Level of protection $|\sqrt{E_1} - \sqrt{E_1'}| \leq \beta$
- Variance of noise added to value $\sigma = \beta/\mu$
- Power of test deciding between $N(0,1)$ vs. $N(\mu,1)$ is $\leq \Phi(\Phi^{-1}(\alpha) + \mu)$

Protection vs Accuracy

Let $|\sqrt{E_1} - \sqrt{E_1'}| \leq \beta = 1$

	sqrt ± 1	
E_1	**Uncertainty Interval for E_1'**	**Relative Size**
1	[0, 4]	400.0%
100	[81, 121]	40.0%
1,000	[937, 1065]	12.7%
10,000	[9801, 10201]	4.0%
100,000	[99368, 100634]	1.3%

Protection vs Accuracy

Let $|\sqrt{E_1} - \sqrt{E_1'}| \leq \beta = 1$ and $\alpha = 0.05$

μ	σ	power
0.5	2	0.1261
1.0	1	0.2595
1.5	0.67	0.4424
2.0	0.5	0.6387

$\sqrt{E_1}| \leq \beta$

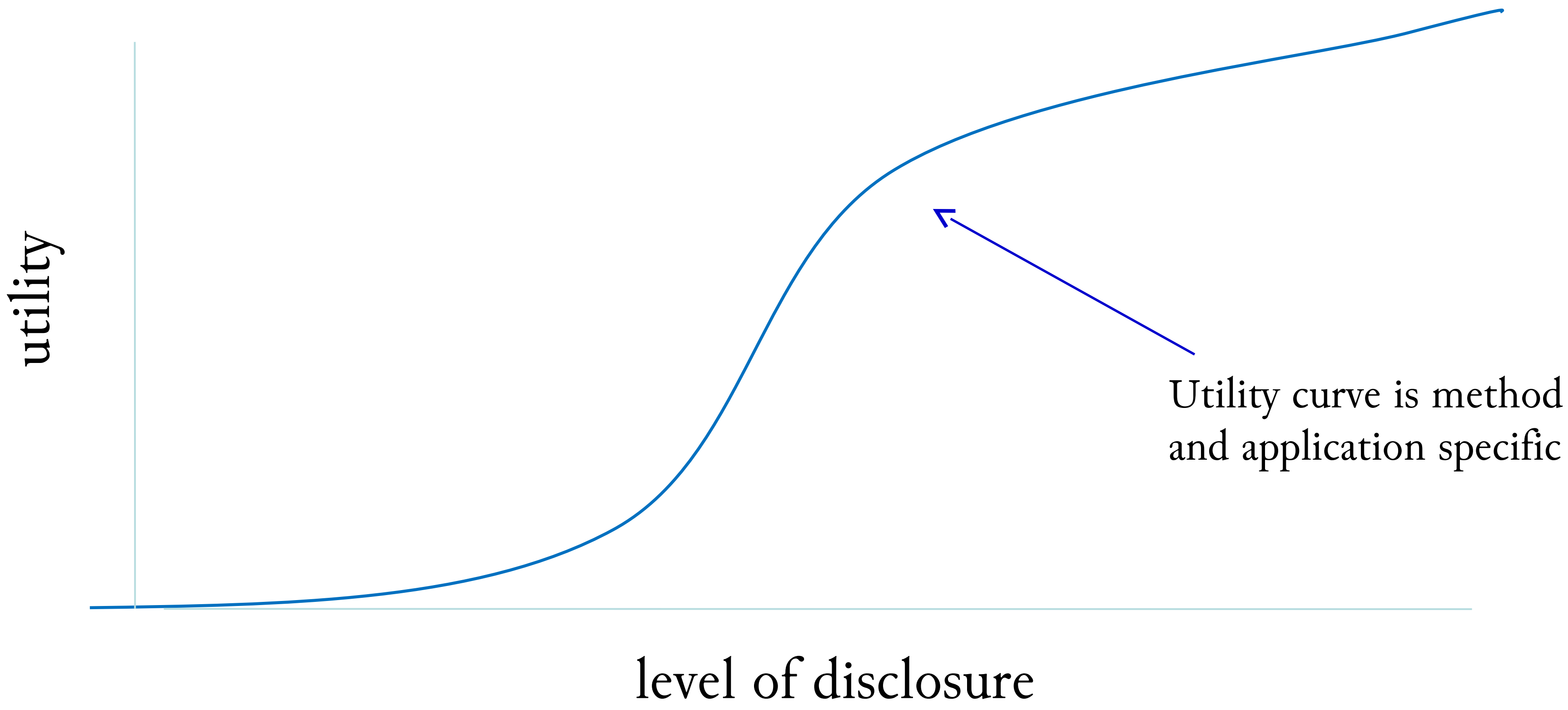
Protection vs Accuracy

Let $|\sqrt{E_1} - \sqrt{E'_1}| \leq \beta = 1$ and $\alpha = 0.05$

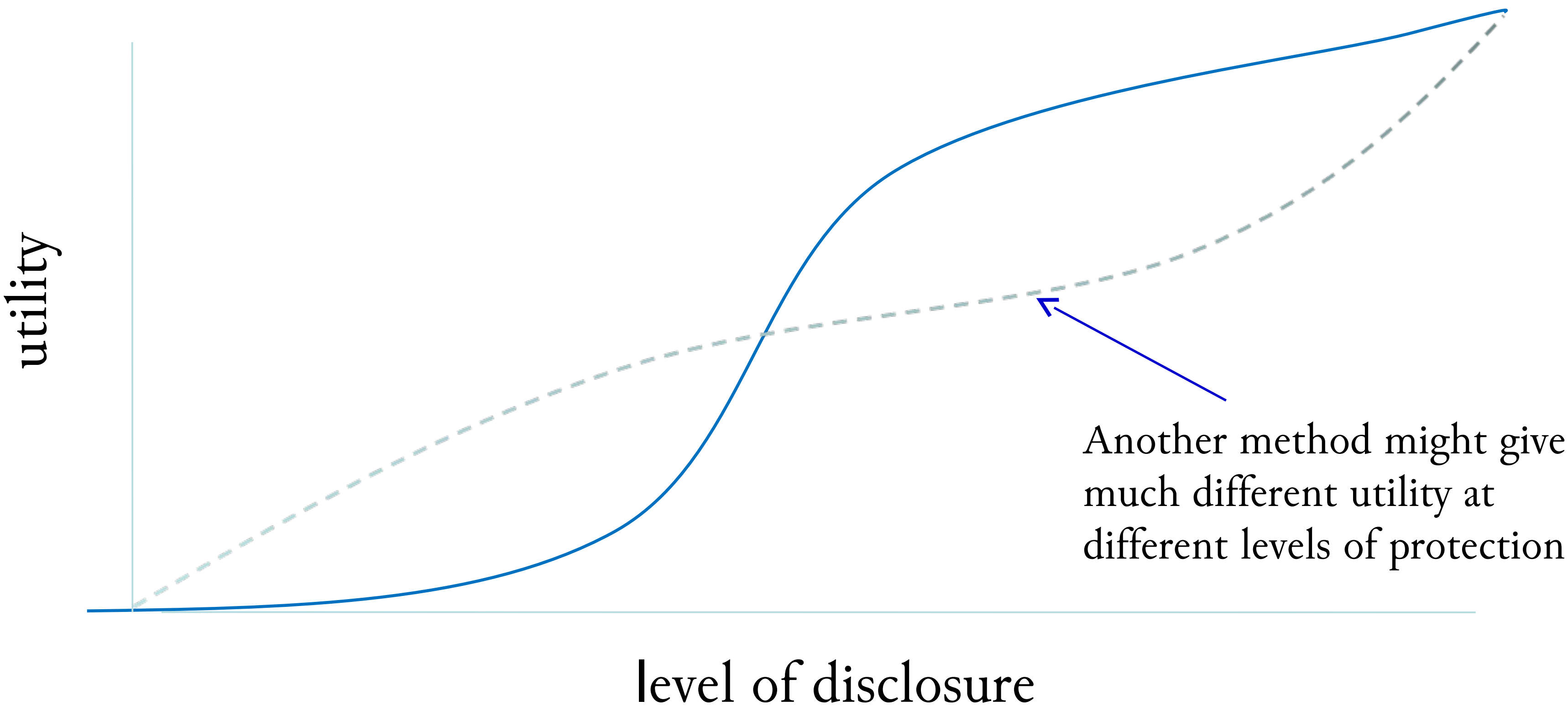
μ	σ	power
0.5	2	0.1261
1.0	1	0.2595
1.5	0.67	0.4424
2.0	0.5	0.6387

How to use privacy budget effectively?

Utility / Protection Tradeoff



Different Method = Different Utility



Splitting the Budget

Use budget μ_1 for protection of individual establishment values

Use budget μ_2 for protection of cell totals

Then the overall budget as far as the accuary/protection tradeoff is

$$\mu \leq \sqrt{\mu_1^2 + \mu_2^2}$$

Examples: $\mu_1 = 1$, $\mu_2 = 1.5$ then $\mu \leq 1.80$

$\mu_1 = .75$, $\mu_2 = 1.9$ then $\mu \leq 2.04$

Example

- Original:

Total Employment

📈 👷‍♀️ 🥦 🏗️ 🍎 $=$ 1,000,000

Employment by County

A	B	C
📈 $= 1,000$	👷‍♀️ 🥦 $= 639,000$	🏗️ 🍎 $= 360,000$

- Add noise
  - $\beta = 1$,
  - $\mu_1 = 0.3$ for total, $\mu_2 = 0.4$ for county
  - overall $\mu = \sqrt{0.3^2 + 0.4^2} = 0.5$

Total Employment

📈 👷‍♀️ 🥦 🏗️ 🍎 $= 1,002,394.88$

Employment by County

A	B	C
📈 $= 1,226.92$	👷‍♀️ 🥦 $= 640,506.56$	🏗️ 🍎 $= 359,329.31$

Calibrate Protected Values

Total Employment

📈 👷‍♀️ 🥦 🏗️ 🍎 $= 1,002,394.88$

Employment by County

A	B	C
📈 $= 1,226.92$	👷‍♀️ 🥦 $= 640,506.56$	🏗️ 🍎 $= 359,329.31$

- Find values for 📈 👷‍♀️ 🥦 🏗️ 🍎 that minimize

$$\frac{(\text{📈} + \text{👷‍♀️} + \text{🥦} + \text{🏗️} + \text{🍎} - 1,002,394.88)^2}{\text{variance(Total Employment)}}$$

$$+\frac{(\text{📈} - 1,226.92)^2}{\text{variance(County A)}} + \frac{(\text{👷‍♀️} + \text{🥦} - 640,506.56)^2}{\text{variance(County B)}}$$

$$+\frac{(\text{🏗️} + \text{🍎} - 359,329.31)^2}{\text{variance(County C)}}$$

Advantages of Protected Micro-Data

- Just use the protected data to produce tables
- No need for cell suppressions
- Users can define areas of interest
- Use protected micro-data for new publication/analysis (no disclosue review needed!)

Selected References

Cox, L. (1995), "Network Models for Complementary Cell Suppression," *Journal of the American Statistical Association*, 90, 1453-1462.

Dwork (2008), "An ad omnia approach to defining and achieving private data analysis" F. Bonchi et al. (Eds.): PinKDD 2007, LNCS 4890, 1–13.

Fuller, W. (1993), "Masking procedures for microdata disclosure limitation," *Journal of Official Statistics*, 9, 383-406.

Warner, S. L. (1965), "Randomized response: A survey technique for eliminating evasive answer bias", *Journal of the American Statistical Association*, 60(309):63–69.

Wasserman, L. and Zhou, S. (2010), "A statistical framework for differential privacy." *Journal of the American Statistical Association*, 105(489):375–389.

Thank You

toth.daniell@bls.gov

This research was supported by a Broad Agency Announcement award from the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF).