Differentially Private Publication of Data on Wages and Job Mobility

Ian M. Schmutte
Department of Economics
University of Georgia

60th ISI World Statistics Congree
Rio de Janeiro
28 July 2015

- Question for data providers
  - “How much privacy loss must be incurred to increase accuracy”
- Answer: Differential privacy
  - Privacy loss measured by parameter ϵ
  - Formal proofs yield marginal cost of privacy
    - …in foregone accuracy

- Can existing methods be applied to generate interesting DP synthetic data?
- Are the resulting synthetic data useful?
- What is the actual cost of increasing privacy?

Roadmap

- Application: Data on job-to-job transitions
  - by employer-specific wage premium
  - and residual wages
  - from Brazil (RAIS)
- Generate DP synthetic data using
  - Multiplicative Weights - Exponential Mechanism (MWEM) algorithm (Hardt et al. 2012)
  - ϵ-differentially private
  - formal accuracy guarantee
- Results:
  - Empirical accuracy far superior to theoretical guarantee
  - Synthetic data effective for training queries
  - pretty poor out of sample

Data

Longitudinal employer-employee data for Brazil

- *Relação Anual de Informações Sociais* (RAIS)
- years 2003–2010
- collected from plant managers for program administration
- covers all formal-sector jobs (50 million per year)

Longitudinal employer-employee data for Brazil

data items *all reported by employer*:

- **job characteristics:**
  - wage, hours, occupation, date of hire
- **plant characteristics:** industry, size, location ...
- **worker characteristics:** age, education, race, sex ...

Full Data:

- All RAIS jobs in plants with more than 1 employee
- 358,894,761 job-year observations

$$\ln w_{it} = x_{it}\beta + \theta_i + \psi_{G(i,t)} + \varepsilon_{it}$$

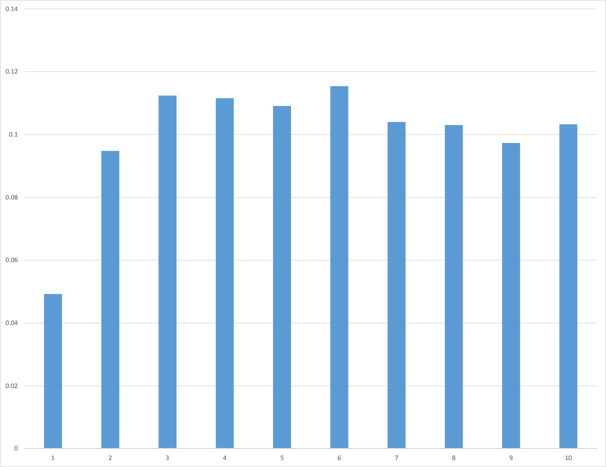
- $\ln w_{it}$, is the log hourly wage
- x_{it} are observed time-varying controls: experience and year effects
- indicator function $G(i,t) = g$ if worker i was employed in g in year t
- $\psi_{G(i,t)}$ measures unobserved employer-specific determinants of compensation
- θ_i captures unobserved worker-specific determinants of compensation

Summary of the components of log wage

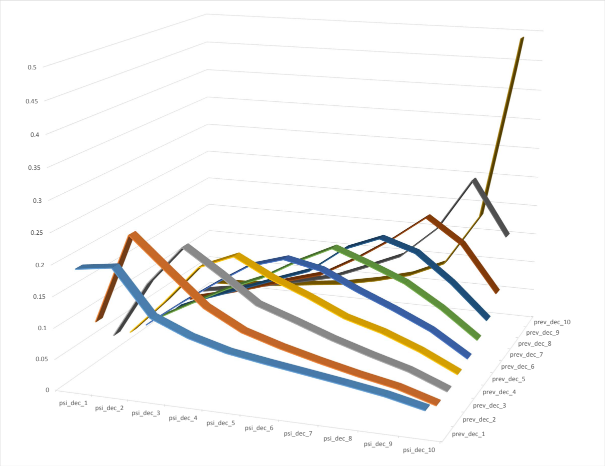
	Mean	Std. Dev.	Correlation				
			Log Wage	$X\beta$	θ	ψ	ε
Log Wage	1.30	0.760	1				
Time-varying characteristics	1.30	0.377	0.243	1			
Worker effect	−0.00	0.502	0.599	−0.476	1		
Estab.-Occup. effect	−0.00	0.397	0.800	0.118	0.333	1	
Residual	0.00	0.196	0.258	−0.000	0.000	0.000	1

- Compute average residual on each job
  - "match effect"
- Restrict sample to observations with a job change
- Discretize employer effects to deciles
- Five percent simple random sample
- Final dataset has three categorical variables
  - origin employer type (10 deciles, plus non-employment)
  - destination employer type (10 deciles)
  - match type (10 deciles)
- Domain D has cardinality $|D|$=1,100.

Job-to-Job Mobility: True Data



Job-to-Job Mobility: True Data



Methods

Databases, Histograms, and Queries

- B database held by custodian with n entries
- each entry is iid draw from (discrete, finite) domain $D = D_1 \times \ldots \times D_K$
- H is a histogram representing B, $H \in R^{|D|}$
- **Queries**
  - A *linear query* is any database query that can be represented by a vector in $R^{|D|}$
  - Query answer: $a(q) = q'H$

Definition

(Differential Privacy) Let M be a random mechanism that maps histograms, H, to distributions over an output space, R.

M provides *ϵ-differential privacy* if

- for every $S \subset R$, and
- for all histograms H and K where $||H - K|| \leq 1$

$$\Pr[M(H) \in S] \leq \exp(\epsilon) \Pr[M(K) \in S].$$

That is:

$$\frac{\Pr[M(H) \in S]}{\Pr[M(K) \in S]} \leq \exp(\epsilon) \tag{1}$$

MWEM Mechanism – Hardt, Ligett, McSherry 2012 (NIPS)

Algorithm *Multiplicative Weights Exponential Mechanism*
Input: Data set, H, over a universe, D; a set Q of linear queries; total number of iterations $T \in N$; privacy parameter $\epsilon > 0$. The number of records in H is n.

1. Initialize the synthetic histogram, K_0, as n times the uniform distribution.
2. **for** $t \leftarrow 1$ **to** T
3. *Exponential Mechanism Step*: Select a query, $q_t \in Q$ using the Exponential Mechanism parameterized with $\epsilon/2T$ and score function

$$s_t(H, q) = |q' K_{t-1} - q' H| \qquad (2)$$

MWEM Mechanism – Hardt, Ligett, McSherry 2012 (NIPS) II

4. *Laplace Mechanism*: Set measurement $m_t = q_t' H + Lap(2T/\epsilon)$.
5. *Multiplicative Weights Step*: Let K_t be n times the distribution whose entries satisfy

$$K_t \propto K_{t-1} \times \exp(q_t \times \left(m_t - q_t' K_{t-1}\right)/2n) \quad (3)$$

6. **Output:** K as the simple average across all K_t for $t < T$.

Theorem

The MWEM satisfies ϵ-differential privacy.

Theorem

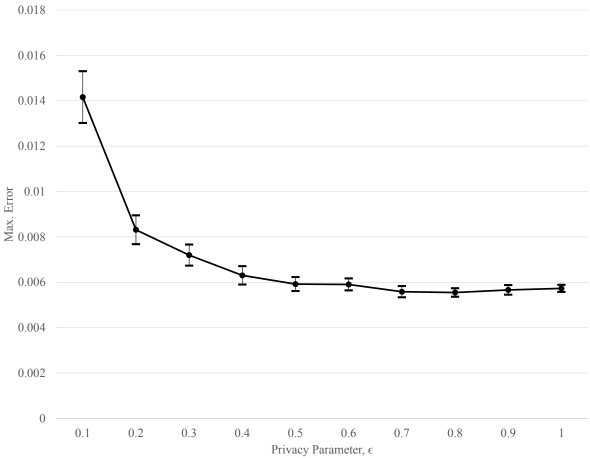
Given any dataset, H, with n records, together with a set of queries, Q, number of iterations T, and $\epsilon > 0$, with probability at least $q - 2T/|Q|$, MWEM produces synthetic histogram K that satisfies

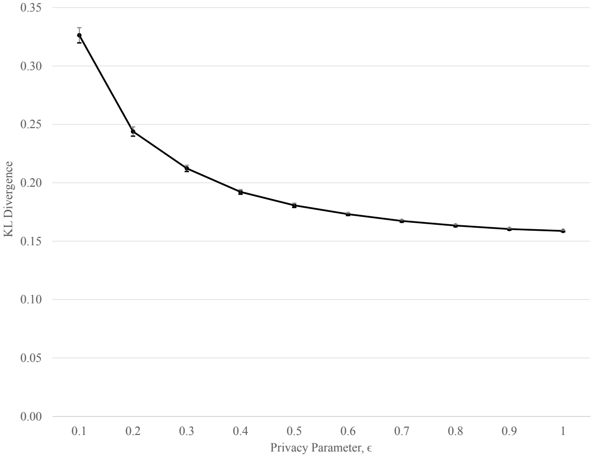
$$max_{q \in Q}|q'H - q'K| \leq 2n\sqrt{\frac{log|D|}{T}} + \frac{10Tlog|Q|}{\epsilon}. \tag{4}$$

Evaluation

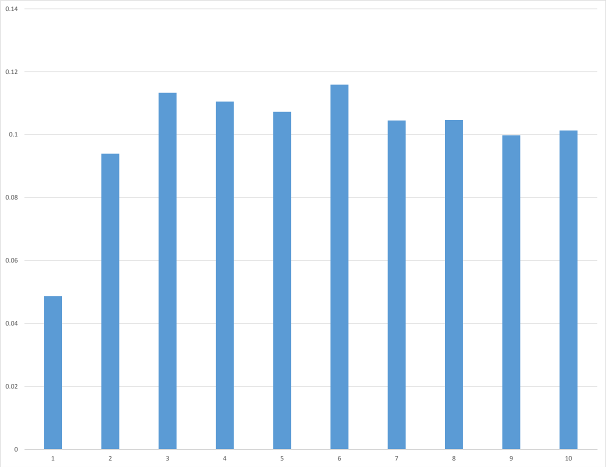
- Query set, Q: all first, second, third-order marginals
- Iterations, T: 300
- Replications: 3

Maximum Error

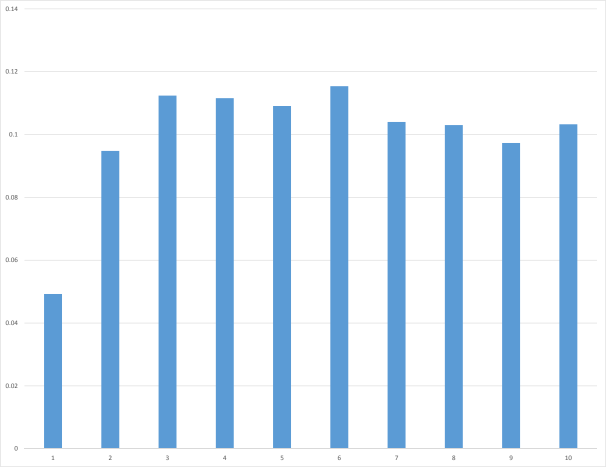



0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
KL Divergence
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Privacy Parameter, ϵ

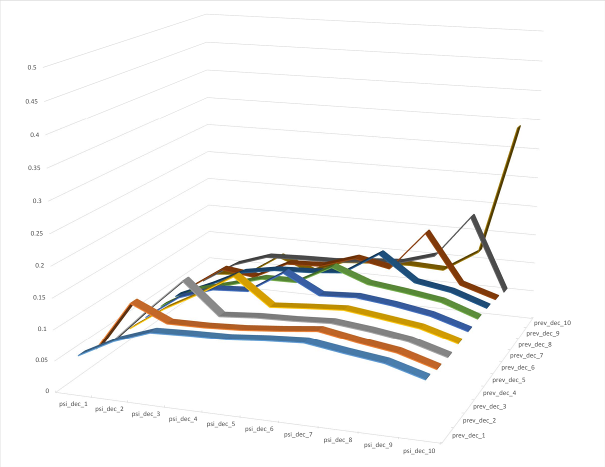
Synthetic Job-to-Job Transitions



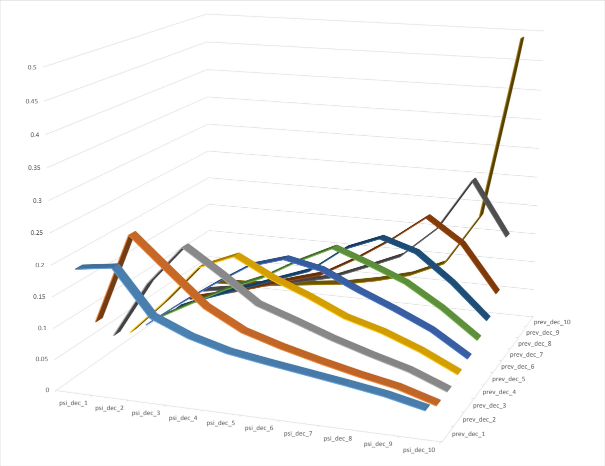
True Job-to-Job Transitions



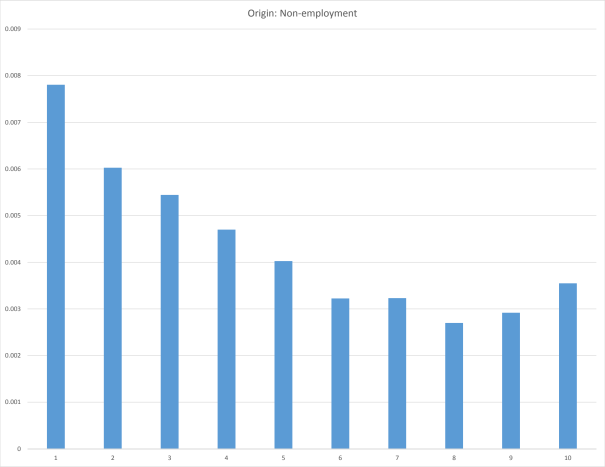
Synthetic Job-to-Job Transitions

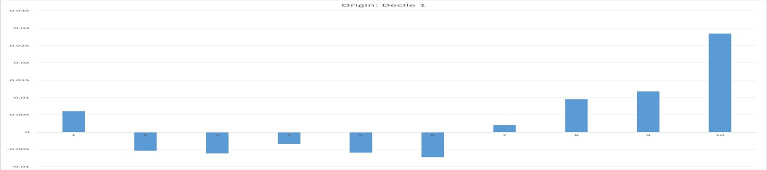


True Job-to-Job Transitions

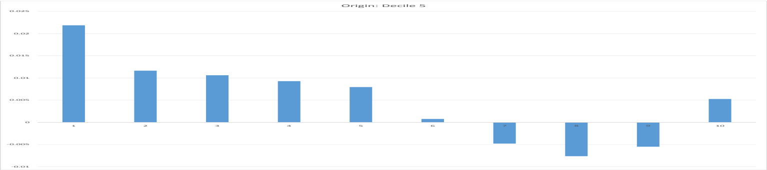


Average Residual by Transition Cell: True Data



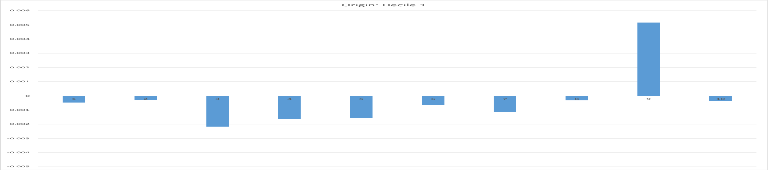


(a) Origin Employer Decile 1

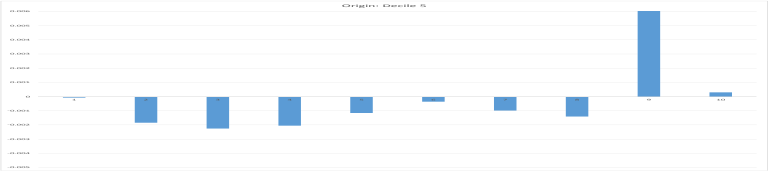


(b) Origin Employer Decile 5

Average Residual by Transition Cell: Synthetic Data



(c) Origin Employer Decile 1



(d) Origin Employer Decile 5