Partially-synthetic linked employer-employee data

Gary Benedetto and Simon Woodcock

US Census Bureau and Simon Fraser University

August 2009

Introduction

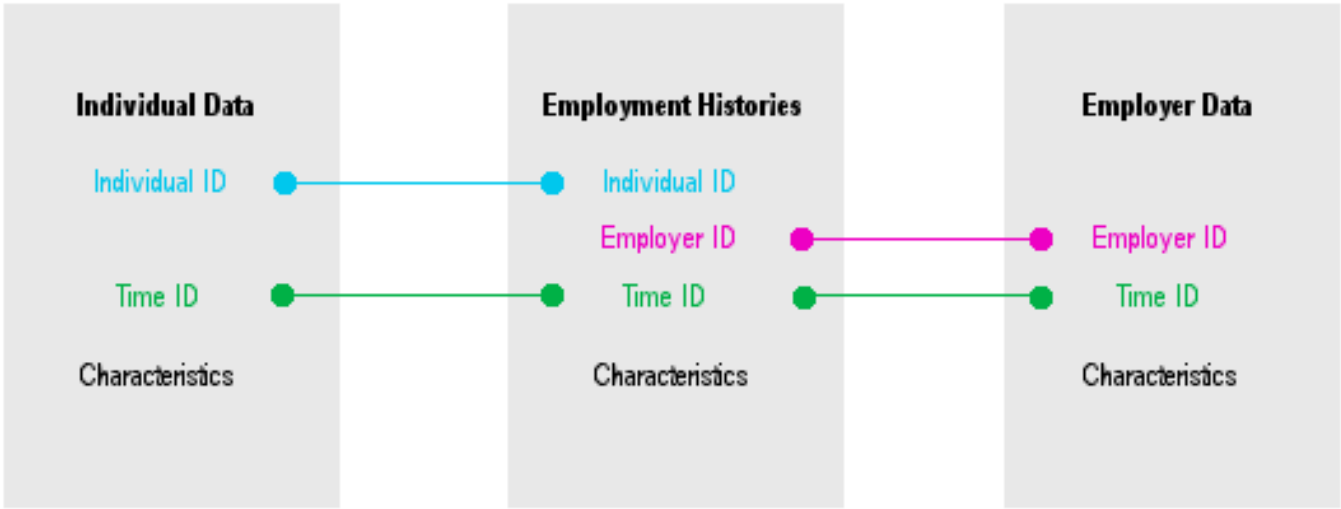
This talk: an overview of recent and ongoing work to limit disclosure risk in linked employer-employee data via partial synthesis.

Particular emphasis on applications involving the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program database:

- Approach and methods
- Current work
- Future directions

Linked Employer-Employee Data

A convenient representation:



Possible disclosure concerns:

- identity disclosure: individuals and employers
- attribute disclosure: attributes of individuals, employers, employment relationships

Potentially many quasi-identifiers: characteristics of individuals, employers, and employment relationships, including *relations between frames*

Synthetic Data

An alternative to traditional disclosure limitation methods that permits valid statistical inferences using standard software and methods is to release data sets comprised of synthetic records sampled from an estimate of the joint distribution of the confidential database.

- Rubin (1993), Raghunathan, Reiter, Rubin (2003): multiple imputation
- Fienberg (1994): bootstrap methods.

Under either approach, the released data pose little disclosure risk: they contain no actual data on actual respondents.

However, this requires knowledge, or a good estimate, of the joint distribution of the data. This is impractical in our case.

- Would require modeling which individuals are employed at which firms and when – i.e., relations between sampling frames. This remains intractable.

Partial Synthesis

We adopt an alternative approach: partial synthesis.

Partially synthetic data are data on actual respondents. Confidential characteristics are replaced with synthetic values sampled from an estimate of the joint distribution of the confidential data conditional on disclosable data.

Reiter (2003): multiply-imputed partially synthetic data allow valid statistical inferences about population quantities.

Estimates on each implicate are combined using simple formulae. Variance estimates reflect uncertainty due to imputation (for synthesis, possibly also to complete missing data).

Our Basic Approach

In our work with LEHD data, we replace confidential characteristics of workers, firms, and jobs with multiple synthetic values sampled from the posterior predictive distribution of an imputation model.

We do not synthesize relations between sampling frames (the *employment graph*: the history of which individuals were ever employed at which firms).

This solves the tricky problem of modeling who works where.

But it has implications for disclosure risk: some summaries of individuals' and firms' employment history are preserved, and this may allow an intruder to link records across partially synthetic implicates.

The Partial Synthesis Problem

Let $D = (X, Y)$ represent the database. Here:

- X are disclosable elements of the database
- Y are confidential elements
- Characteristics of individuals, employers, and jobs (including the relations between sampling frames) could be in either X or Y.

The partial synthesis approach is to replace confidential values with synthetic values $\tilde{Y}$ sampled from the posterior predictive distribution:

$$p\left(\tilde{Y}|X, Y\right) = \int p\left(\tilde{Y}|X, \theta\right) p\left(\theta|X, Y\right) d\theta$$

where $p\left(Y|X, \theta\right)$ is the likelihood, $p\left(\theta|X, Y\right)$ is a prior, and θ are parameters.

Repeat M times, producing M partially-synthetic versions of D ("implicates"). Inference is based on combining rules in Reiter (2003).

The Joint Likelihood: Early Work

In typical applications, specifying the joint likelihood $p(Y|X,\theta)$ is a challenge.

In early work, Abowd and Woodcock (2001) approximated $p(Y|X,\theta)$ by a sequence of conditional distributions defined by generalized linear models, using the Sequential Regression Multivariate Imputation (SRMI) method of Raghunathan et. al. (2001).

- Advantages:
  - variable-by-variable conditional imputation models: flexible & simple
  - handle wide variety of data types
  - preserve first two moments of joint distribution
- Disadvantages:
  - mis-specification of the imputation models can distort distribution of confidential data, and invalidate inferences

The Joint Likelihood: More Recently

1. Still focus on variable-by-variable imputation, but rely on factorization of the joint likelihood rather than approximation:

$$p\left(Y|X,\theta\right) = p_1\left(Y_1|X,\theta_1\right) p_2\left(Y_2|X,Y_1,\theta_2\right) \cdots p_K\left(y_K|X,Y_1,y_2,...,y_{K-1},\theta_K\right)$$

where $Y_k \in Y$ is a collection of confidential elements of the database (e.g., confidential values of one or several variables), and $k = 1,...,K$ indexes imputation order. Synthetic values are sampled from the posterior predictive distribution:

$$p_k\left(\tilde{Y}_k|X,Y\right) = \int p_k\left(\tilde{Y}_k|X,\tilde{Y}_1,...,\tilde{Y}_{k-1},Y_{k+1},...,Y_K,\theta_k\right) p_k\left(\theta_k|X,Y\right) d\theta_k$$

where $p_k\left(\tilde{Y}_k|X,\tilde{Y}_1,...,\tilde{Y}_{k-1},Y_{k+1},...,Y_K,\theta_k\right)$ is the likelihood of an imputation model for Y_k, and $p_k\left(\theta_k|X,Y\right)$ is the corresponding prior.

2. Use more flexible models to specify $p_k\left(Y_k|\cdot\right)$, e.g., Woodcock & Benedetto (in press) for continuous variables, or resampling.

Woodcock & Benedetto (in press) Procedure

Idea: pair a simple parametric imputation model (e.g., regression) with a nonparametric transformation.

Why? Agencies may prefer simple models: reduce modeling and computational burden, easier to diagnose and interpret, easier to describe to users, because the correct imputation model is unknown, etc.

However, synthetic data generated using a simple imputation model may fail to reproduce complex features of the confidential data, such as nonlinear relationships between variables, skewness, tail thickness, and the number and location of modes.

By pairing a simple model with a nonparametric transformation, we can preserve the distribution of Y_k on subdomains of primary interest, only relying on the simple imputation model to preserve relationships of *secondary interest* within those subdomains.

Sketch of Procedure

1. Divide the data into subdomains of primary interest.
2. In each subdomain, transform the variable under imputation to have a standard distribution that is compatible with a simple imputation model.
3. Generate synthetic values using a simple model on the transformed data. The role of the simple imputation model is to preserve relationships of secondary interest within subdomains.
4. Apply an inverse transformation that returns the synthetic values to the native scale and distribution of the underlying confidential variable. This preserves the distribution of the confidential variable on the subdomains of primary interest.

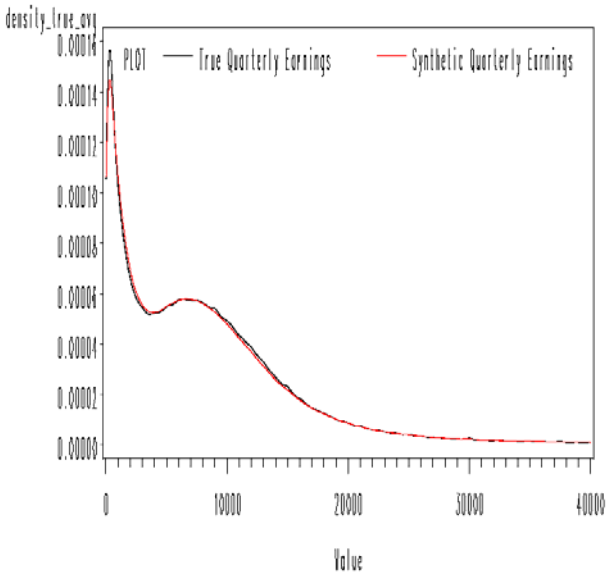
Less subject to mis-specification than a simple imputation model alone, because we only rely on the simple model to capture relationships of secondary interest.

Example

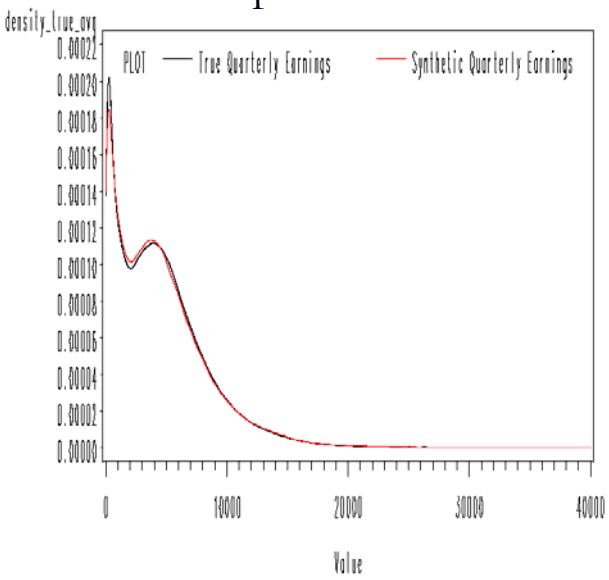
- Consider synthesis of $Y_k|W_1, W_2$ using a linear regression for $Y_k|W_2$ on subdomains $W_1 = w_1$.
- On each subdomain, estimate $\hat{F}_{Y|W_1=w_1}$ on an approximate Bayesian bootstrap sample of observations (e.g., using a KDE).
- Define $Z_k = \Phi^{-1}\left(\hat{F}_{Y|W_1=w_1}\left(Y_k|W_1 = w_1\right)\right)$. Note $Z_k \sim N(0,1)$ on each subdomain.
- Sample $\tilde{Z}_k$ from the posterior predictive distribution defined by the normal linear regression of Z_k on W_2 and an uninformative prior.
- In general, the distribution of $\tilde{Z}_k$ is unknown so construct a sample estimate $\hat{F}_{\tilde{Z}|W_1=w_1}$.
- Define the synthetic values $\tilde{Y}_k = \hat{F}^{-1}_{Y|W_1=w_1}\left(\hat{F}_{\tilde{Z}|W_1=w_1}\left(\tilde{Z}_k\right)\right)$.
- This replicates the distribution of Y_k in the synthetic data (up to sampling error), i.e., $\tilde{Y}_k \sim \hat{F}_{Y|W_1=w_1}$. The transformations are monotone, so montone relationships between Y_k and W_2 are preserved (in practice, more is preserved).

Densities of True and Synthetic Age and Quarterly Earnings on Selected Subdomains

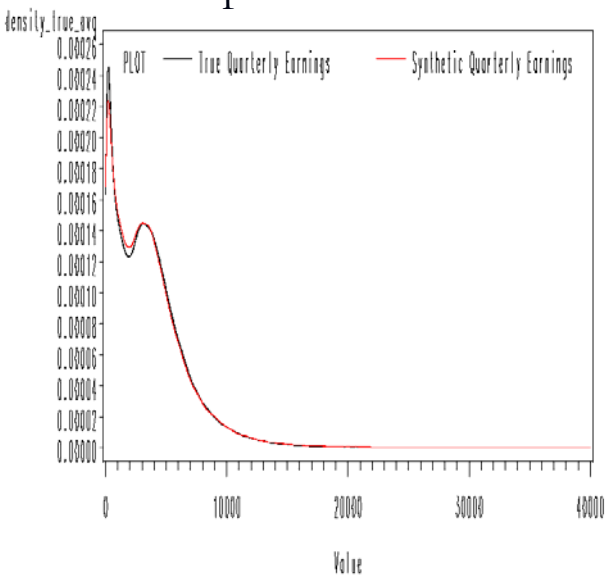
White Males



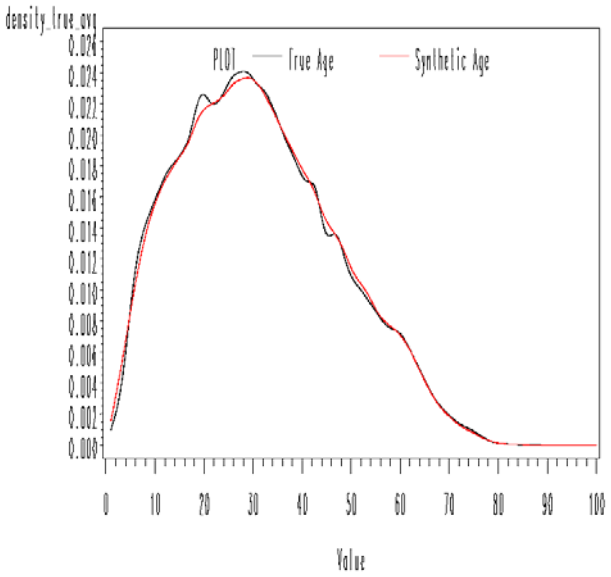
Hispanic Males



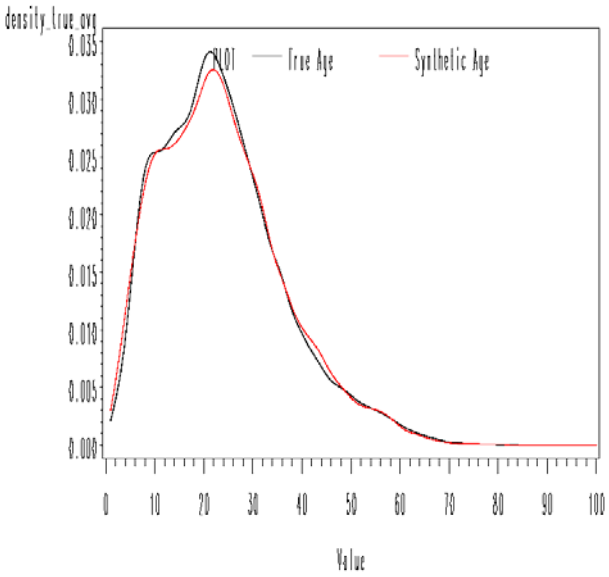
Hispanic Females



White Males



Hispanic Males



Hispanic Females

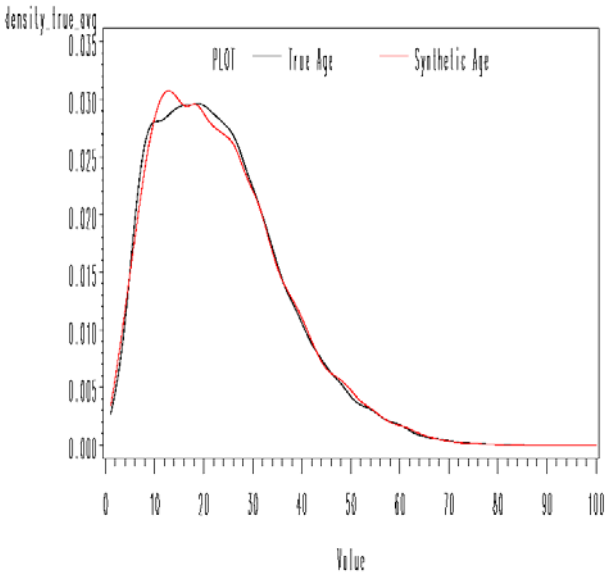


Table 5: Estimated Coefficients in Log Earnings Regression

	Population Value	Avg. Synthetic Estimate	95% CI Coverage	
			Observed	Synthetic
Years of experience	.086	.086	91.7	95.1
Experience^2/100	−.321	−.327	89.2	94.4
Experience^3/1000	.055	.057	86.9	94.9
Experience^4/10000	−.004	−.004	84.8	95.2
Initial Experience < 0	−.176	−.167	94.6	94.1
Years of Education	.012	−.002	94.1	97.8
Education^2/100	−.541	−.461	93.9	98.3
Education^3/1000	.733	.783	94.3	98.2
Education^4/10000	−.196	−.226	94.7	97.7
Race = Black	−.271	−.264	96.5	98.7
Race = Hispanic	−.205	−.186	97.3	92.3
Foreign born = 1	−.078	−.054	95.1	91.1
ln(Employer size)	−.372	−.414	48.4	47.3
ln(Employer payroll)	.397	.440	43.7	42.8
SIC Division = A	−.136	−.164	92.3	97.1
SIC Division = B	−.059	−.058	97.6	99.6
SIC Division = C	.031	.008	95.5	97.1
SIC Division = E	−.005	−.010	98.2	99.7
SIC Division = F	.041	.020	98.3	97.4
SIC Division = G	−.208	−.192	88.3	91.5
SIC Division = H	.089	.061	95.7	88.3
SIC Division = I	−.211	−.244	92.7	75.1
SIC Division = J	−.218	−.237	94.3	95.3
Year = 1991	−.003	.000	95.7	97.8
Year = 1992	.025	.023	97.3	99.3
Year = 1993	.040	.044	96.4	98.5
Year = 1994	.068	.069	95.2	98.3
Year = 1995	.081	.080	96.7	98.8
Year = 1996	.104	.102	95.9	98.8
Year = 1997	.126	.124	95.5	98.5
Year = 1998	.152	.148	95.1	98.4
Quarter = 2	.053	.051	93.9	97.1
Quarter = 3	.047	.048	93.5	97.9
Quarter = 4	.079	.089	94.3	96.4
Intercept	4.20	4.00	71.7	74.6
RMSE	.762	.864		
Number of Observations	7,145,344	11,910		

Current Work

We are working on a more substantive application based on the LEHD data.

These are administrative data, constructed from quarterly Unemployment Insurance (UI) system wage reports.

The Bureau of Labor Statistics (1997) claims that UI coverage is "broad and basically comparable from state to state" and that "over 96 percent of total wage and salary civilian jobs" were covered in 1994.

With the UI wage records as its frame, the LEHD data comprise the universe of employers required to file UI system wage reports — that is, all employment potentially covered by the UI system in participating states.

Nearly all states now participate in the US Census Bureau's LEHD partnership. Our application is based on one state, whose identity is confidential.

Structure of the LEHD Data

Structure corresponds to the prototypical case described earlier.

The UI wage records associate each individual with an employing firm in each quarter that the individual was employed. Also includes a measure of employment earnings.

The LEHD project adds demographic characteristics of individuals (sex, race, date of birth, county of residence), and characteristics of firms (industry, county), to the UI wage records. These characteristics are based on internal Census Bureau sources.

Relations between sampling frames define some additional derived characteristics of firms (size, payroll).

The LEHD data: final details

Sample comprises approx. 1 million individuals employed in this state between 1993 and 2004, at approx. 85,000 firms. About 3.5 million employment relationships total.

Some missing data, but not much. These have been multiply-imputed by Census Bureau staff for other purposes.

Our application is based on four completed data implicates. For each completed data implicate, we generate four partially-synthetic implicates $\Longrightarrow$ total of 16 partially-synthetic implicates.

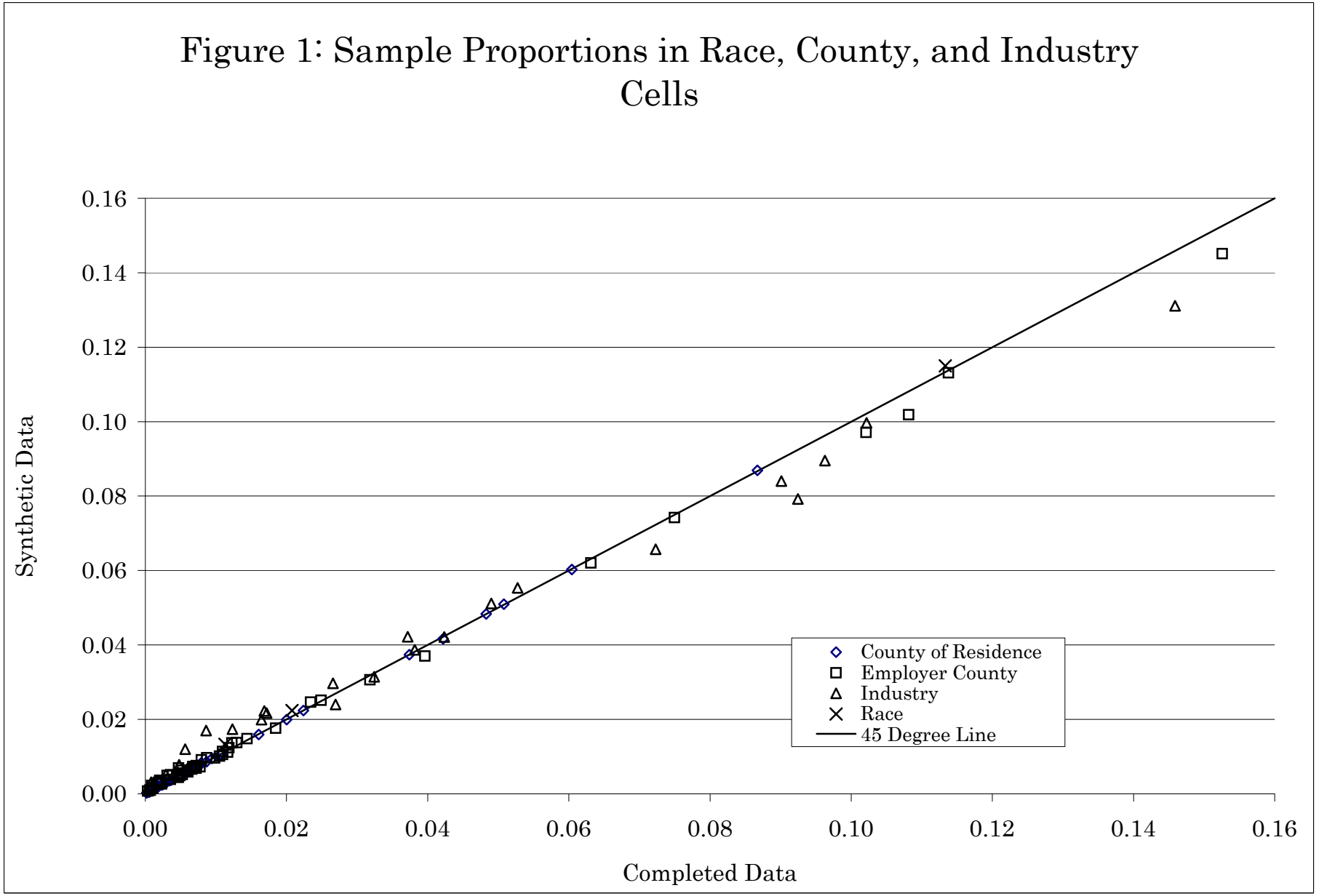
Synthesis order and details

1. Y_1 is all discrete individual characteristics: sex, race, and county of residence.
   - Multinomial likelihood, mixture of uninformative & Dirichlet priors.
2. Y_2 is all discrete firm characteristics: industry (NAICS sector) and county.
   - Multinomial likelihood, mixture of uninformative & Dirichlet priors.
3. Y_3 is date of birth (daily).
   - Linear regression + density-based transformation, uninformative prior.
4. Y_4 is the employment history.
   - Use a series of logit models to sequentially impute the quarters in which the job was active, uninformative priors throughout.
5. Y_5 is the earnings history.
   - Sequentially impute earnings in each quarter that the job is active, using linear regression + density-based transformation, uninformative priors throughout.

Table 1
Univariate Moments of Continuous Variables

Variable	Statistic	Value in Completed Data	Value in Synthetic Data
	Person- and Job-Level Variables		
Birthdate	Mean	1,213	1,214
	Standard deviation	5,743	5,738
	Skewness	-0.516	-0.519
	Kurtosis	-0.166	-0.177
Quarterly Earnings	Mean	4,653	4,649
	Standard deviation	9,563	7,286
	Skewness	357	281
	Kurtosis	301,809	249,558
In-sample Job Duration (Quarters)	Mean	5.34	5.43
	Standard deviation	7.87	7.84
	Skewness	3.01	2.97
	Kurtosis	9.74	9.44
	Derived Firm-Level Variables		
Number of Quarters with Positive Employment	Mean	17.2	13.7
	Standard deviation	14.8	14.7
	Skewness	0.728	1.079
	Kurtosis	-0.851	-0.246
Quarterly Employment	Mean	15.6	11.5
	Standard deviation	75.5	62.2
	Skewness	23.8	26.4
	Kurtosis	839	1025
Quarterly Payroll	Mean	72,519	53,562
	Standard deviation	490,288	381,557
	Skewness	31.3	34.9
	Kurtosis	1,420	1,832

Figure 1: Sample Proportions in Race, County, and Industry Cells



Assessing Attribute disclosure risk

We presume an intruder can link records across synthetic implicates.

In most applications, this would be conservative. Here, it is probably realistic.

Because we do not perturb the employment graph, some simple summaries of employment histories are replicated across partially-synthetic implicates

The number of distinct firms at which each individual was employed (R), coupled with the number of distinct employees (E) at each of those firms, the value of R for each of individual ever employed at one of their employers (their coworkers), and the value of E for each of their coworkers' employers, uniquely identifies about 80 percent of individuals.

Similar exercise will uniquely identify many firms.

Does this matter for risk of identity disclosure?

A measure of attribute disclosure risk

Assume an intruder estimates unit i's value of the k^{th} confidential variable, $y_{k,i}$, by averaging the unit's synthetic values across all partially synthetic implicates: $\bar{y}_{k,i} = \sum_{m=1}^{M} \tilde{y}_{k,i}^{m}$.

Our main measure of attribute disclosure risk is based on the *RRMSE* of this estimator of $y_{k,i}$ for each unit:

$$RRMSE_{k,i} = \left(\sqrt{\left(y_{k,i} - \bar{y}_{k,i}\right)^2 + M^{-1}\left(M-1\right)^{-1} \sum_{m=1}^{M} \left(\tilde{y}_{k,i}^{m} - \bar{y}_{k,i}\right)^2} \right) / y_{k,i}.$$

The distribution of *RRMSE* in the synthetic data provides a measure of variability in the imputations.

A second measure of attribute disclosure risk

Assume the intruder estimates $\bar{y}_{k,i}$ as before, and its variance based on the Reiter (2004) combining rules, and uses these to construct a 95 percent confidence interval for $y_{k,i}$.

We then calculate the proportion of the empirical density of y_k that lies within the interval.

Idea: predictions are more informative when the interval contains a small proportion of the empirical density (either the interval is narrow, or the prediction lies in a low-density region of the distribution).

Table 3
Attribute Disclosure Risk

	Percentiles of RRMSE of Prediction				
	1st	5th	10th	25th	50th
Avg Quarterly Earnings	0.035	0.064	0.087	0.151	0.309
In-sample Job Duration	0.014	0.088	0.122	0.187	0.347
	Percent of Empirical Distribution Covered by Synthetic 95% CI				
	≤ 10%	10-20%	20-30%	30-40%	> 40%
Avg Quarterly Earnings					
Synthetic 95% CI **Does Not** Contain Completed Value	5.22	3.54	2.15	1.18	0.85
Synthetic 95% CI **Does** Contain Completed Value	10.9	13.7	13.2	11.4	37.8
In-sample Job Duration					
Synthetic 95% CI **Does Not** Contain Completed Value	2.29	1.49	4.5	2.09	1.12
Synthetic 95% CI **Does** Contain Completed Value	7.02	5.32	5.74	8.29	62.1

Future directions: further reducing disclosure risk

Idea: reducing an intruder's ability to combine information across synthetic data implicates reduces risk (attribute and identity).

One possibility: release a sample of observations.

- Unique summaries of the employment graph in a sample do not guarantee uniqueness in the population, so intruder must assign probabilities that records with identical summaries correspond to the same unit.
- Most units will not appear in all samples, so an intruder has fewer implicates on which to base predictions about any unit's confidential values, and hence predictions are less precise.

Another possibility: partially synthesize relations between sampling frames.

- Expect that ia fairly small number of imputations of "who works where when" will introduce enough between-implicate variability to make summaries of the employment graph non-unique.