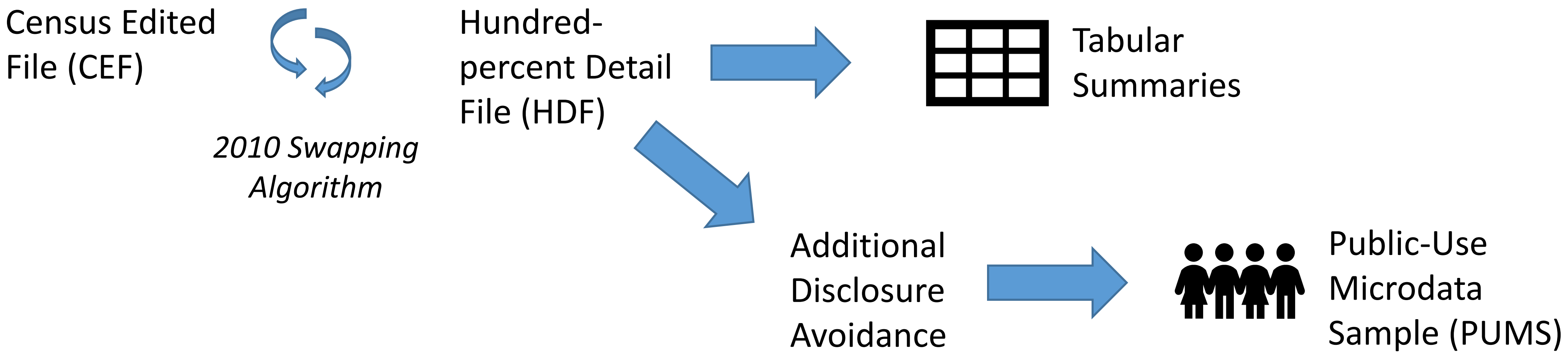
Reconstruction and Re-identification of the Demographic and Housing Characteristics File (DHC)

Michael Hawes

Research and Methodology Directorate

United States® Census Bureau

Disclosure avoidance for the 2010 Census



Census Edited File (CEF)

2010 Swapping Algorithm

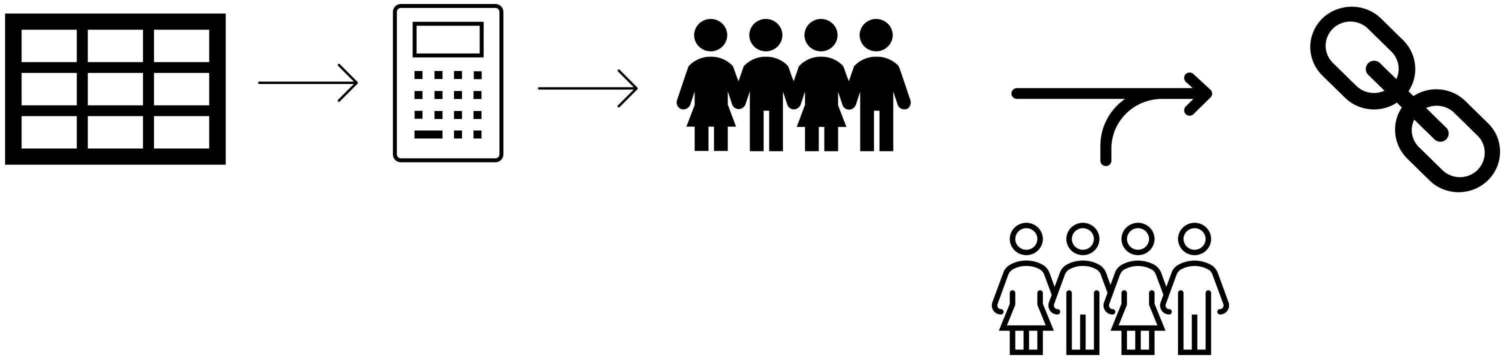
Hundred-percent Detail File (HDF)

Tabular Summaries

Additional Disclosure Avoidance

Public-Use Microdata Sample (PUMS)

What is reconstruction-abetted re-identification?



Published Tabular Summaries

System of Equations

Reconstructed Confidential Microdata (without names)

External data source (with names)

Linked microdata with names and confidential information

What 2010 Census data can an attacker use?

Anything published from the 2010 Census!

Our simulated attack used only a small subset:

- P1 (Total Population by Block)
- P6 (Total Races Tallied by Block)
- P7 (Hispanic or Latino Origin by Race by Block)
- P9 (Hispanic or Latino, and Not Hispanic or Latino by Race by Block)
- P11 (Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over by Block)
- P12 (Sex by Age by Block)
- P12A-I (Sex by Age by Block, iterated by Race)
- P14 (Sex by Single-year-of-age for the Population under 20 Years by Block)
- PCT12 (Sex by Single-year-of-age by Tract)
- PCT12A-O (Sex by Single-year-of-age by Tract, iterated by Race)

What external files can an attacker use?

Any external files that contain name and address (or other unique identifiers) and pseudo-identifiers contained in the census data (e.g., sex and age)

Our simulated attack used a combination of 4 commercially available datasets. But there are higher quality data out there. This is a lower-bound analysis.

The impact of higher quality name and address data can be estimated by using the CEF as the external file. This is an upper-bound analysis.

Exact age vs. binned age

The subset of tables we used for our simulated attack do not always provide precise age reconstruction. Thus, we present re-id statistics for Exact Age matches, and for Binned Age matches using the following age bins from the block-level 2010 Summary File 1 tables:

Single year of age from 0 - 21
22-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-61
62-64
65-66
67-69
70-74
75-79
80-84
85+

Agreement rates for reconstructed microdata

Percentage of reconstructed records that exactly agree with the CEF on location, sex, age (exact/binned), race, and ethnicity

Agreement Rates	Exact Age	Exact and Binned Age
Published 2010 Tables (swapping)	46.5	91.8
High Swapping Experiment	26.5	52.1
DDP1 2022-03-16 (ρ=3.325)	15.7	33.1
DDP2 2022-08-25 (ρ=3.65)	18.1	35.1

Agreement rates by block size

Percentage of reconstructed records that exactly agree with the CEF on location, sex, age (exact/binned), race, and ethnicity

Block Population	Published 2010 Tables (swapping)		High Swapping Experiment		DDP1 2022-03-16 (ρ=3.325)		DDP2 2022-08-25 (ρ=3.65)	
	Exact Age	Exact & Binned Age	Exact Age	Exact & Binned Age	Exact Age	Exact & Binned Age	Exact Age	Exact & Binned Age
1-9	30.0	74.0	21.2	50.2	2.6	7.3	2.8	7.7
10-49	43.6	93.0	22.4	48.3	6.4	16.1	7.1	17.6
50-99	45.2	93.1	22.8	48.4	10.2	24.1	11.6	26.2
100-249	45.9	92.1	24.9	51.2	15.8	34.7	17.9	36.9
250-499	48.2	91.3	29.7	56.2	23.8	48.3	27.2	50.2
500-999	52.4	90.6	35.6	60.1	30.9	58.6	36.4	60.6
1,000+	62.7	91.5	49.0	67.5	40.1	70.2	51.3	73.0

United States® Census Bureau

How well accurately can an attacker re-identify the characteristics of specific individuals from the reconstructed records?

Defining the universe for analysis

Successfully re-identifying specific individuals requires more than just a match on location, sex, age, race, and ethnicity.

It also requires being able to link a name to that record.

Not all records in the CEF have unique Protected Identification Key (PIK) identifier within the block.*

To evaluate the success of our simulated attack, we define the universe (denominator) as the data-defined population (individuals with unique PIKs within the block).

*A PIK is the Census Bureau's individual record linkage identifier produced by the Person Identification Validation System (PVS), the production name and address linkage system. The vintage is the same as for the 2010 Census.

Definitions

Putative Re-identification Rate:

$$\frac{\#\ of\ records\ that\ agree\ on\ Block, Sex, Age\ (exact, binned)}{\#\ of\ records\ with\ unique\ PIK\ in\ block}$$

Confirmed Re-identification Rate:

$$\frac{\#\ of\ records\ that\ agree\ on\ PIK, Block, Sex, Age\ (exact, binned)\ Race\ and\ Ethnicitiy}{\#\ of\ records\ with\ unique\ PIK\ in\ block}$$

Re-identification Precision Rate:

$$\frac{\#\ of\ records\ that\ agree\ on\ PIK, Block, Sex, Age\ (exact, binned)\ Race\ and\ Ethnicitiy}{\#\ of\ records\ that\ agree\ on\ Block, Sex, Age\ (exact, binned)}$$

Re-identification statistics

(Exact and Binned Age)	Putative Rate	Confirmation Rate	Precision Rate
Published 2010 Tables (swapping) to Commercial	60.2	24.8	41.2
High Swapping Experiment to Commercial	56.2	17.2	30.7
DDP1 2022-03-16 ($\rho=3.325$) to Commercial	38.5	11.1	28.7
DDP2 2022-08-25 ($\rho=3.65$) to Commercial	39.7	11.4	28.7
Published 2010 Tables (swapping) to CEF	97.0	75.5	77.8
High Swapping Experiment to CEF	75.4	46.6	61.8
DDP1 2022-03-16 ($\rho=3.325$) to CEF	44.4	27.4	61.7
DDP2 2022-08-25 ($\rho=3.65$) to CEF	45.8	28.5	62.2

Re-identification of population uniques

Re-identification statistics for “population uniques” of the linking pseudo-identifiers (those who are unique within their block on sex, and either exact age [SAB] or binned age [SAbB])

	Putative Rate		Confirmation Rate		Precision Rate	
	SAB	SAbB	SAB	SAbB	SAB	SAbB
Published 2010 Tables (swapping) to Commercial	32.9	23.1	28.3	21.8	86.1	94.6
High Swapping Experiment to Commercial	26.6	17.6	18.2	12.7	68.5	72.4
DDP1 2022-03-16 (ρ=3.325) to Commercial	13.4	7.0	8.9	4.7	66.6	66.5
DDP2 2022-08-25 (ρ=3.65) to Commercial	14.1	7.6	9.5	5.1	67.0	67.3
Published 2010 Tables (swapping) to CEF	95.0	93.1	84.2	87.2	88.6	93.6
High Swapping Experiment to CEF	69.2	64.0	46.7	44.5	67.5	69.6
DDP1 2022-03-16 (ρ=3.325) to CEF	30.3	22.1	19.6	13.9	64.6	62.9
DDP2 2022-08-25 (ρ=3.65) to CEF	32.4	24.0	21.1	15.3	65.1	63.9

Re-identification of population uniques for non-modal race/ethnicity

Re-identification statistics for “population uniques” of the linking pseudo-identifiers (those who are unique within their block on sex, and either exact age [SAB] or binned age [SAbB]) for individuals of the blocks’ non-modal race/ethnicity

Non-Modal Race/Ethnicity	Putative Rate		Confirmation Rate		Precision Rate	
	SAB	SAbB	SAB	SAbB	SAB	SAbB
Published 2010 Tables (swapping) to Commercial	24.0	13.7	14.3	12.2	59.4	89.2
High Swapping Experiment to Commercial	20.6	11.5	5.0	3.5	24.4	30.6
DDP1 2022-03-16 (ρ=3.325) to Commercial	11.4	5.3	2.4	1.2	20.8	23.2
DDP2 2022-08-25 (ρ=3.65) to Commercial	12.0	5.7	2.4	1.2	20.0	21.6
Published 2010 Tables (swapping) to CEF	90.6	86.2	60.4	70.2	66.7	81.5
High Swapping Experiment to CEF	71.6	65.5	20.0	21.9	27.9	33.4
DDP1 2022-03-16 (ρ=3.325) to CEF	34.7	25.9	7.8	6.2	22.3	24.0
DDP2 2022-08-25 (ρ=3.65) to CEF	36.8	27.8	7.9	6.4	21.6	23.2

Re-identification of population uniques for modal race/ethnicity

Re-identification statistics for “population uniques” of the linking pseudo-identifiers (those who are unique within their block on sex, and either exact age [SAB] or binned age [SAbB]) for individuals of the blocks’ modal race/ethnicity

Modal Race/Ethnicity	Putative Rate		Confirmation Rate		Precision Rate	
	SAB	SAbB	SAB	SAbB	SAB	SAbB
Published 2010 Tables (swapping) to Commercial	35.1	25.3	31.8	24.2	90.5	95.3
High Swapping Experiment to Commercial	28.0	19.0	21.4	14.9	76.3	78.5
DDP1 2022-03-16 (ρ=3.325) to Commercial	13.8	7.4	10.5	5.5	75.8	73.9
DDP2 2022-08-25 (ρ=3.65) to Commercial	14.6	8.0	11.2	6.0	76.4	75.0
Published 2010 Tables (swapping) to CEF	96.1	94.8	90.0	91.3	93.6	96.3
High Swapping Experiment to CEF	68.6	63.6	53.2	50.0	77.5	78.5
DDP1 2022-03-16 (ρ=3.325) to CEF	29.2	21.2	22.4	15.8	76.7	74.3
DDP2 2022-08-25 (ρ=3.65) to CEF	31.3	23.1	24.3	17.5	77.6	75.6

Summary of re-identification statistics

Difference in percentage points across experiments, relative to the published 2010 tables, in the putative and precision rates. Comparisons shown use the CEF panels from the previous slides.

	High Swap		DDP1 2022-03-16		DDP2 2022-08-25	
Metric	Putative Rate	Precision Rate	Putative Rate	Precision Rate	Putative Rate	Precision Rate
National	-21.6	-16.0	-52.6	-16.1	-51.2	-15.6
SAB Uniques	-25.8	-21.1	-64.7	-24.0	-62.6	-23.5
SAB Non-Modal	-19.0	-38.8	-55.9	-44.4	-53.8	-45.1
SAB Modal	-27.5	-16.1	-66.9	-16.9	-64.8	-16.0

Blocks with zero solution variability

- For certain blocks, the SF1 tables used for reconstruction imply a single possible reconstructed microdata set (using binned age[1])
- These blocks are said to have *zero solution variability* (0-solvar)
- At least 65% of blocks in 2010 were 0-solvar
- 935 tracts consisted entirely of 0-solvar blocks

[1] Binned age is exact age for persons younger than 22 years old

Re-identification of population in blocks with zero solution variability

Zero Solution Variability Blocks	Putative Rate		Confirmation Rate		Precision Rate	
	Overall	Non-modal & SAbB Unique	Overall	Non-modal & SAbB Unique	Overall	Non-modal & SAbB Unique
Published 2010 Tables (swapping) to Commercial	58.2	18.3	33.3	16.7	57.3	**91.2**
Published 2010 Tables (swapping) to CEF	94.6	71.9	90.8	68.4	96.0	**95.1**