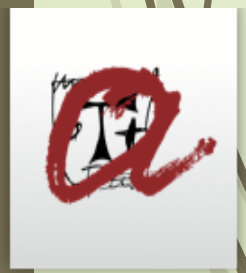
What is Reconstruction and Reidentification?

Illustrations from the 2010 US Census Tabular Data Release

Krish Muralidhar (University of Oklahoma)

Josep Domingo-Ferrer (Universitat Rovira i Virgili)



2010 US Decennial Census Data

- "Somewhere between 52 and 179 million person who responded to the 2010 Census can be correctly re-identified from the re-constructed microdata."
  - Abowd, J.M. 2021. Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN, U. S. District Court for the Middle District of Alabama. Apr. 13.
- Was the US Census Bureau (USCB) so negligent in their tabular data release that more than 50% of the respondents to the 2010 Census could be re-identified?

USCB tabular data release

- The original data are corrected for errors and the Census Edited File is created
- Disclosure limitation (primarily swapping) is applied to the microdata in the Census Edited File to create the Hundred-Percent Detail File (protected microdata)
- All tables are created from the Hundred-Percent Detail File
- Any reconstruction using tabular data can only recreate the protected microdata (and not the original data)

Tabular data release from 2010 US Decennial Census

- Cross tabulations at the individual level
  - Age, Sex, Race, Ethnicity, Geography
- Cross tabulations at the household level
  - Age, Household size, Sex, Race, Ethnicity, Relationship, Geography
- No cross tabulations between individual and household data
- USCB reconstructed (Geography, Sex, Age, Race, Ethnicity). This is a partial reconstruction. The reconstruction does not include one attribute (relationship to head of household)

USCB Reconstruction Methodology

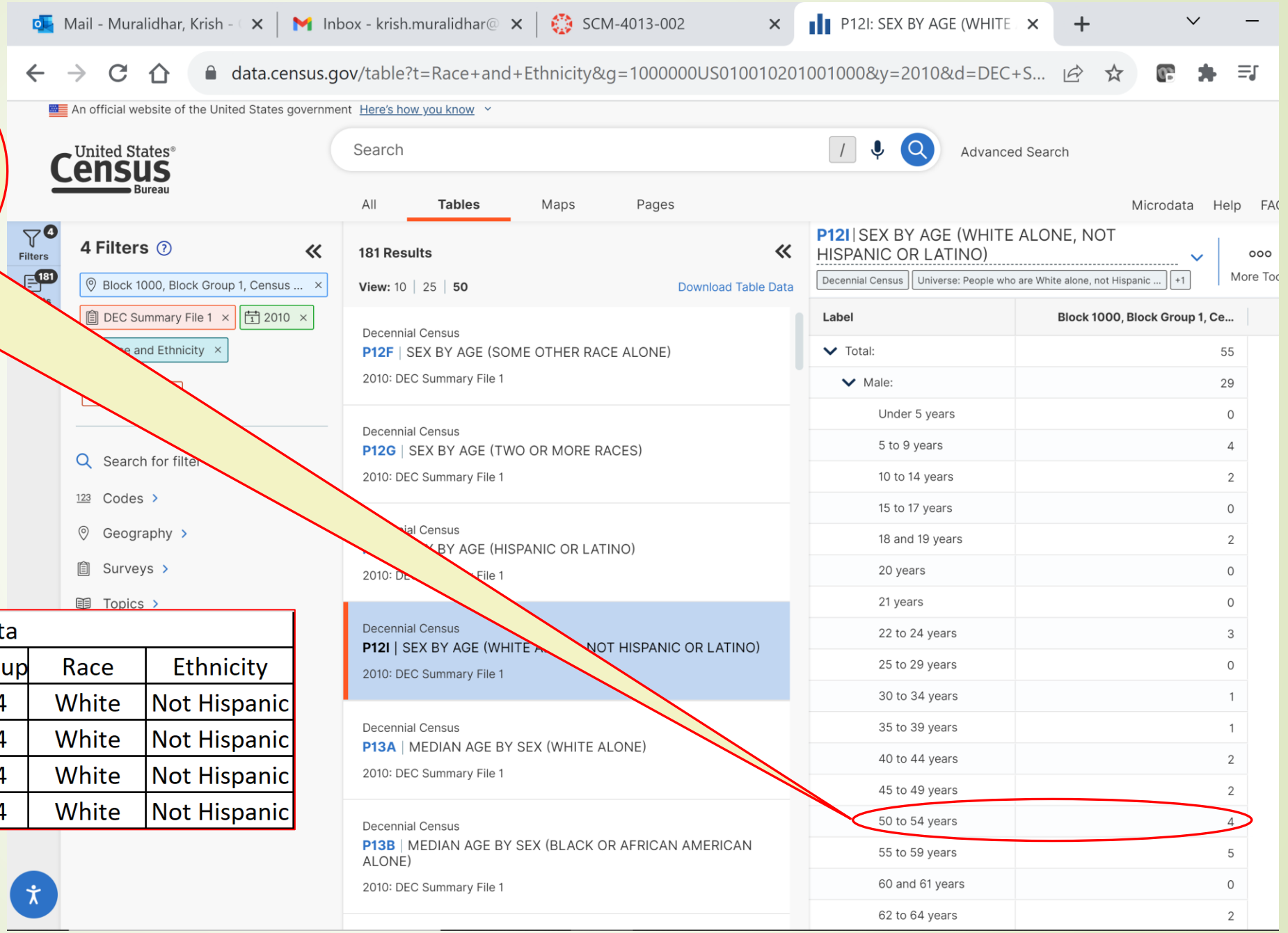
- "The reconstruction of the 2010 Census microdata for the sex, age, race, Hispanic/Latino ethnicity, and census block variables was carried out by constructing a system of equations consistent with the published tables listed above that, once solved, could then be converted into microdata. This system of equations was solved using commercial mixed-integer linear programming software (Gurobi)." (Abowd 2021a)
- Entirely unnecessary! See Muralidhar (2022)
  - Muralidhar, K., "A Re-examination of the Census Bureau Reconstruction and Reidentification Attack," In: Domingo-Ferrer, J., Laurent, M. (eds) Privacy in Statistical Databases. PSD 2022. Lecture Notes in Computer Science, vol 13463. Springer, Cham.

Most recent reconstruction

- Initial USCB reconstruction attempted to reconstruct individual year of age. Subsequent reconstruction only reconstructs "binned" (grouped) age
  - Hawes, M. 2022. Reconstruction and Reidentification of the Demographic and Housing Characteristics File (DHC). *Presentation to the Census Scientific Advisory Committee* (Washington DC, USA, 29-30 September 2022). https://www2.census.gov/about/partners/cac/sac/meetings/2022-09/presentation-reconstruction-and-re-dentification-of-dhc-file.pdf (accessed 14 March 2023).

Reconstruction is simply a matter of creating a list from the count data.

Reconstructed Data					
Respondent	Block	Sex	Age Group	Race	Ethnicity
1	Block 1000	Male	50 - 54	White	Not Hispanic
2	Block 1000	Male	51 - 54	White	Not Hispanic
3	Block 1000	Male	52 - 54	White	Not Hispanic
4	Block 1000	Male	53 - 54	White	Not Hispanic



Reconstruction

- "While the Census Bureau's confidentiality methodologies for the 2000 and 2010 censuses were considered sufficient at the time, advances in technology in the years since have reduced the confidentiality protection provided by data swapping." (Abowd 2021)

- Our analysis shows that this claim is absurd!

Reidentification

- Match the reconstructed data to an external source of data that contains identity information (Name and Address)
- USCB reidentification approach
  - Match the reconstructed data (Block, Sex, Age, Race, Ethnicity) to external data source (Name, Address, Sex, Age)
  - Matching variables (quasi-identifiers) are Sex and Age
    - See Abowd (2021) for complete details
- For reidentification to be meaningful, the matching must be one-to-one. USCB does not require one-to-one match.
  - One-to-many, many-to-one, or many-to-many implies that the identity of the match *cannot be confirmed*

Illustration 1

Hypothetical Data from one block, Males, age group (40 – 44)

- USCB procedure results in 100% confirmed reidentification.

External Data Source					Reconstructed Data			
Name	Address	Age	Sex		Age	Sex	Race	Ethnicity
ABCD	1001 Main Street	42	Male		40-44	Male	White	Not Hispanic
EFGH	1002 Main Street	44	Male		40-44	Male	White	Not Hispanic
IJKL	1003 Main Street	43	Male		40-44	Male	White	Not Hispanic
MNOP	1004 Main Street	41	Male		40-44	Male	White	Not Hispanic
QRST	1005 Main Street	41	Male		40-44	Male	White	Not Hispanic
UVWX	1006 Main Street	42	Male		40-44	Male	White	Not Hispanic
YZAB	1007 Main Street	41	Male		40-44	Male	White	Not Hispanic
CDEF	1008 Main Street	44	Male		40-44	Male	White	Not Hispanic
GHIJ	1009 Main Street	43	Male		40-44	Male	White	Not Hispanic
KLMN	1010 Main Street	43	Male		40-44	Male	White	Not Hispanic

- All respondents are identical and are all protected
- Any respondent can be assigned any identity (based on Sex and Age) (10-anonymity)

Illustration 2

Hypothetical Data from one block, Males, age group (40 – 44)

- USCB procedure results in minimum 80% confirmed reidentification.

External Data Source					Reconstructed Data			
Name	Address	Age	Sex		Age	Sex	Race	Ethnicity
ABCD	1001 Main Street	42	Male		40-44	Male	Black	Not Hispanic
EFGH	1002 Main Street	44	Male		40-44	Male	White	Not Hispanic
IJKL	1003 Main Street	43	Male		40-44	Male	White	Not Hispanic
MNOP	1004 Main Street	41	Male		40-44	Male	White	Not Hispanic
QRST	1005 Main Street	41	Male		40-44	Male	White	Not Hispanic
UVWX	1006 Main Street	42	Male		40-44	Male	White	Not Hispanic
YZAB	1007 Main Street	41	Male		40-44	Male	White	Not Hispanic
CDEF	1008 Main Street	44	Male		40-44	Male	White	Not Hispanic
GHIJ	1009 Main Street	43	Male		40-44	Male	White	Not Hispanic
KLMN	1010 Main Street	43	Male		40-44	Male	White	Not Hispanic

- All respondents are not identical but are all protected
- Any respondent can be assigned any identity (based on Sex and Age) (10-anonymity)

Illustration 3

Hypothetical Data from one block, Males, age group (40 – 44)

- USCB procedure results in minimum 0% confirmed reidentification. Reidentification occurs only by chance.

External Data Source					Reconstructed Data			
Name	Address	Age	Sex		Age	Sex	Race	Ethnicity
ABCD	1001 Main Street	42	Male		40-44	Male	Black	Not Hispanic
EFGH	1002 Main Street	44	Male		40-44	Male	White	Not Hispanic
IJKL	1003 Main Street	43	Male		40-44	Male	Asian	Not Hispanic
MNOP	1004 Main Street	41	Male		40-44	Male	AIAN	Not Hispanic
QRST	1005 Main Street	41	Male		40-44	Male	NHPI	Not Hispanic
UVWX	1006 Main Street	42	Male		40-44	Male	Other	Not Hispanic
YZAB	1007 Main Street	41	Male		40-44	Male	2 or more	Not Hispanic
CDEF	1008 Main Street	44	Male		40-44	Male	White	Hispanic
GHIJ	1009 Main Street	43	Male		40-44	Male	Black	Hispanic
KLMN	1010 Main Street	43	Male		40-44	Male	Asian	Hispanic

- Every respondent is unique but are all protected
- Any respondent can be assigned any identity (based on Sex and Age) (10-anonymity)

Homogeneity and reidentification

- The USCB reidentification procedure is significantly affected by homogeneity of the block
  - Homogenous blocks are reconstructed and reidentified with greater accuracy (Illustration 1)
  - Heterogenous blocks are reconstructed and reidentified with lower accuracy (Illustration 3)
- Approximately 67% of the blocks are homogenous by (Race, Ethnicity)
  - Ruggles, S. and Van Riper, D. 2021. The role of chance in the Census Bureau database reconstruction experiment. *Population Research and Policy Review*, early access since Aug. 22. https://doi.org/10.1007/s11113-021-09674-3

Impact of homogeneity

(Block 3034, Block Group 3, Census Tract 63.01, Mobile County, Alabama)

- This block is over 90% (White, Not Hispanic)
- The percentage of confirmed reidentification is also very close to 90%

	Age Group							Total
	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	
White Not Hispanic Males	39	41	41	32	60	53	36	302
All Other Males	4	5	6	5	4	2	0	26
Minimum Confirmed Reidentification	35	36	35	27	56	51	36	276
Minimum Confirmed Reidentification (%)	90%	88%	85%	84%	93%	96%	100%	91%

- This defies common sense. (White, Not Hispanic Males) are protected by the fact that they are indistinguishable (basic concept of hiding in a crowd and k-anonymity)

General observations - Reconstruction

- USCB still continues to claim that "The technological advances that now permit database reconstruction at scale" (Keller and Abowd 2023) when it has been shown, repeatedly, that reconstruction can be performed even with century old punch card technology
- USCB claims that because of reconstruction "traditional approaches to SDL for aggregate statistics discussed above obsolete" (Keller and Abowd 2023). Such reconstruction can only reproduce the protected microdata and not the original data.
- Without subsequent reidentification, reconstruction alone may not necessarily be a problem.

General observations - Reidentification

- USCB continues to claim that "Somewhere between 52 and 179 million person … can be correctly re-identified from the re-constructed microdata" (Keller and Abowd 2023) even though, according to USCB: "To date, we are not aware of successful reidentifications by bad actors"
- Our analysis has shown that reidentification claims made by the USCB based on the reconstructed data is incorrect and vastly over-stated.

Reconstruction without Reidentification

- Reconstruction without subsequent reidentification, in and of itself, is not necessarily a problem in the context of the USCB
- Reconstruction without reidentification should be the objective of the USCB
  - Accurate reconstruction → Accurate Data
  - But no reidentification → Prevents identity disclosure

Conclusion

- Random assignment of identity (Name, Address) to a group of respondents who are identical does not constitute confirmed reidentification. When you have a large group of respondents with identical attributes, they are protected from reidentification due to group privacy
  - According to the USCB, the most common records are also the most reidentified!
  - The claims of the USCB contradict the basic definition of identity disclosure.
- We may have reconstruction, but not reidentification!

Vielen Dank und beste Wünsche!

Fragen?