IHME

The risk of linked census data to transgender children: a simulation study

Abraham D. Flaxman

June, 2022

(joint work with Os Keyes)

UNIVERSITY of WASHINGTON

Institute for Health Metrics and Evaluation

Acknowledgement and Disclosures

Land acknowledgement: This research was conducted on the traditional lands of the Coast Salish people.

My grant and research support: The Alfred P. Sloan Foundation (this work); Washington State Department of Health, Bill and Melinda Gates Foundation, US Census Bureau, NORC, NIH (other work).

My other consulting: Janssen; Sanofi; SwissRe; Merck for Mothers; Agathos, Ltd (startup).

Sharing data is a matter of ethics and the U.S. Census Bureau's data are a public good. So the need to provide greater access to Bureau data seems obvious to me. But I also see the possibility that some data can be misused, either by government officials or by others who access them. There is an ethical obligation not to aid and abet that abuse.

–Stephen Fienberg, 2006

Outline

1. Background: how linked census data might disclose sensitive gender identity information
2. Methods: computer simulation setting, alternative disclosure avoidance scenarios, and reconstruction-abetted linkage attacks
3. Results: the number of trans kids with gender identity disclosed
4. Discussion: limitations and directions for future work

Background: how linked census data might disclose sensitive gender identity information

- In recent years there has been heightened scrutiny of transgender people, with a particular focus on trans children.
- A prominent recent example: the governor of Texas directed the state Department of Family and Protective Services to investigate the parents of any trans child who receives gender-affirming medical care.
- In this work, we investigate the risk of disclosing a child's transgender status, through discordant reporting of binary gender in successive censuses.

Methods: computer simulation setting

We used computer simulation to investigate the risk of disclosing a child's transgender status, through discordant reporting of binary gender in successive censuses. Our simulation has 5 steps:

1. Reconstruct population with household and population structure from 2010 decennial census
2. Add gender using labels and rates matching CDC BRFSS (0.18% trans boys, 0.23% trans girls, and 0.12% gender nonconforming)
3. Filter population to include only 0-7 year olds
4. Simulate 10 years of aging and mobility (23% of simulants resided at the same address in 2010 and 2020 [American Community Survey])
5. Simulate reported sex in 2020 decennial census based on gender

Scenario 1: Extreme disclosure

- We counted all simulants with differing values recorded for sex in 2010 and 2020 to estimate the number of trans youth who would have their gender identity revealed if census microdata including names were to be released (or re-identified).
- We found disclosure of gender identify for over 6,000 simulants (38% of all trans kids in our simulated Texas; the remaining 62% were not identified because their reported sex was concordant in both censuses).

Synthetic data from 2010

unique_id	first_name	last_name	address	dob	race_eth	recorded sex
0	Mason	**Calvin**	**742 katie road austin tx**	6/9/2005	Black	**Male**
1	Kevin	**Francois**	1838 dragonia ennis tx	**1/13/2005**	Black	**Male**
2	Garrison	Ordonez Abril	1758 n 134th round rock tx	6/29/2002	Latino	Male
3	Damarion	Torres	5046 faulkner arlington tx	2/25/2003	Latino	Male
...	...	...	...	...	...	...

Synthetic Data (not released by Census!)

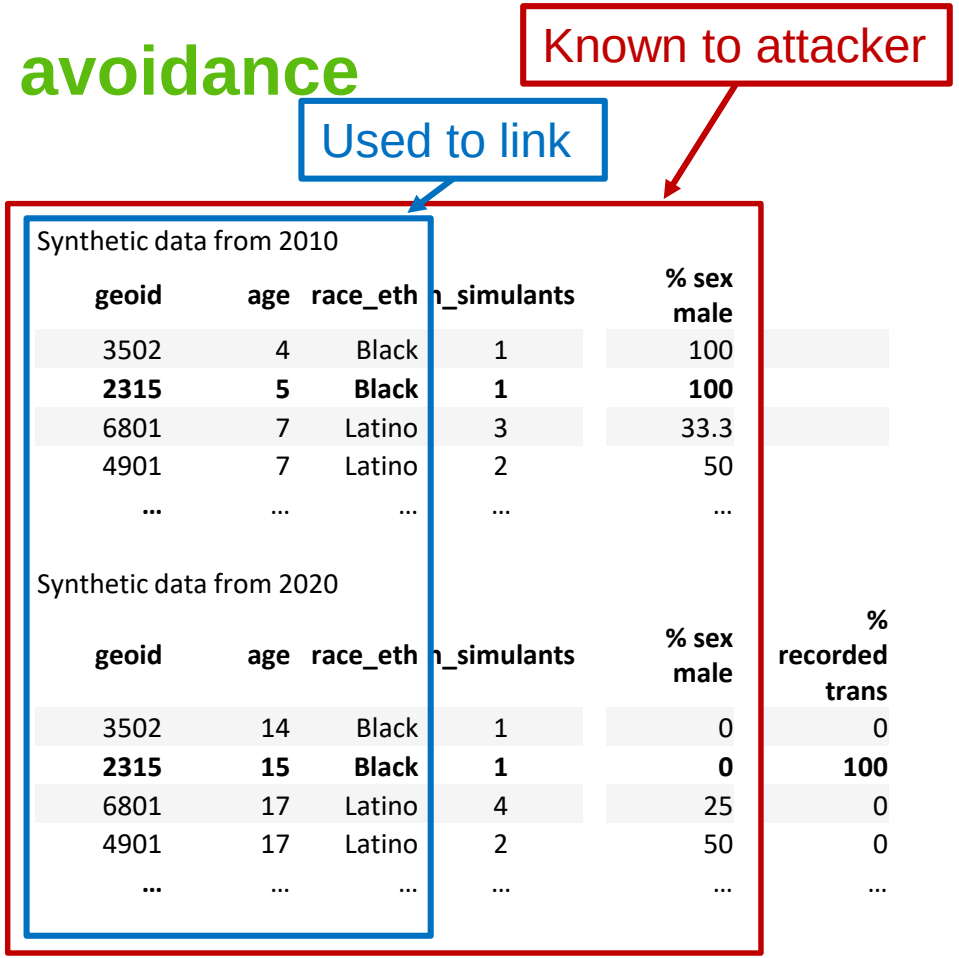
Synthetic data from 2020

unique_id	first_name	last_name	address	dob	race_eth	recorded sex
0	Mason	**Calvin**	**742 katie road austin tx**	6/9/2005	Black	**Female**
1	Kelly	**Francois**	263 noel st tomball tx	**1/13/2005**	Black	**Female**
2	Garrison	Ordonez Abril	12297 budlong lake avenue cypress tx	6/29/2002	Latino	Male
3	Damarion	Torres	7 haddington houston tx	2/25/2003	Latino	Male
...	...	...	...	...	...	...

Synthetic Data (not released by Census!)

Scenario 2: No disclosure avoidance

- Reconstructed-abetted linkage attack without a reidentification step
- Targets simulants age seven and younger in 2010 who had a unique combination of age, race and ethnicity in their census block.
- Trans kids who moved between the 2010 and 2020 censuses likely not revealed.
- Trans kids who did not move *might* not have their transgender status revealed, if in-migration resulted in them no longer having a unique combination of attributes in 2020.



Scenario 3: Swapping for disclosure avoidance

- Instead of using each simulant's geography directly in the reconstruction-abetted linkage attack, we first chose a random subset of households to have their reported location swapped to somewhere other than their true location.
- We selected some households to swap independently at random, with probability $p_{\text{swap}} = 5\%$.
- We chose a reported location to swap to by selecting uniformly from all synthetic households in Texas.

Synthetic data from 2010

geoid	age	race_eth	n_simulants	% sex male	
3502	4	Black	1	100	
~~2315~~ → 3781	5	**Black**	**1**	**100**	
6801	7	Latino	3	33.3	
4901	7	Latino	2	50	
…	…	…	…	…	

Synthetic data from 2020

geoid	age	race_eth	n_simulants	% sex male	% recorded trans
3502	14	Black	1	0	0
2315	**15**	**Black**	**1**	**0**	**100**
6801	17	Latino	4	25	0
4901	17	Latino	2	50	0
…	…	…	…	…	…

Scenario 4: TDA for disclosure avoidance

- Instead of simulating forward from 2010 to 2020, we initialized simulants in 2020 and simulated time backwards to 2010.
- This allowed using the DHC Demonstration Product instead of swapping to quantify the impact of TDA on the reconstruction-abetted linkage attack.
- Central question: how many fewer trans kids are identified by the reconstruction-abetted linkage attack against TDA compared to swapping?

Synthetic data from 2010

geoid	age	race_eth	n_simulants	% sex male	
3502	4	Black	1	100	
2315	**5**	**Black**	**1**	**100**	
6801	7	Latino	3	33.3	
4901	7	Latino	2	50	
…	…	…	…	…	

Synthetic data from 2020

geoid	age	race_eth	n_simulants	% sex male	% recorded trans
3502	14	Black	1 + DP Noise	0	0
2315	**15**	**Black**	**1 + DP Noise**	**0**	**100**
6801	17	Latino	4 + DP Noise	33.3	0
4901	17	Latino	2 + DP Noise	33.3	0
…	…	…	…	…	…

Results: the number of trans kids identified

	Trans kids disclosed	False Positives	Positive Predictive Value
Scenario 1: Extreme Disclosure	6,200	0	100.00%
Scenario 2: No disclosure avoidance	657	69,527	0.94%
Scenario 3: Swapping for disclosure avoidance	605	77,426	0.78%
Scenario 4: TDA for disclosure avoidance	170	36,267	0.47%

Limitations and directions for future work

- Components of model that are perhaps overly simplistic:
  - Migration lack heterogeneity
  - Mechanism of how gender maps to reported sex
  - Assumption that race/ethnicity is reported identically in 2010 and 2020
  - Census block boundaries don’t stay the same from 2010 to 2020

- Inside Census Bureau, it would be possible to investigate how results of simulation compare to real (but restricted) data.

Conclusion

1. Linked data from decennial censuses contains sensitive gender identity information.

2. TDA improves on swapping in protecting this sensitive information.

Replication archive and draft report:
https://github.com/aflaxman/linked_census_disclosure