Anticipating Differential Privacy in the 2020 US Census

Abraham D. Flaxman (joint work with Samantha Petti)

July 2, 2019
“Our thriving agriculture sector is part of why Ethiopia’s overall economic growth has been so impressive, and why we are on track to become a middle-income country by 2025. But there is one thing that could derail us: climate change. Climate change is caused by the actions of richer countries, but the most vulnerable people in poorer countries are feeling it first.”

---Dr. Kaba Urgessa, State Minister for Natural Resources, Ethiopia
Global Climate Strike and Census Data

Climate-driven migration into the USA will be quantified by Census Bureau sources

• Decennial Census
• American Community Survey

How will differential privacy in Census Bureau products impact our understanding of climate-driven migration?
Outline

• What is differential privacy?
• How was it achieved in the 2018 end-to-end test?
• What effect did this have on accuracy?
• Why might differential privacy in 2020 Census be a good thing, and what is still to be decided?
What is Differential Privacy (DP)?

“Differential privacy is a definition, not an algorithm,” Dwork and Roth, 2014

Intuitive definition
Regardless of external knowledge, an adversary with access to the sanitized database will draw the same conclusions whether or not my data is included in the original database.
What is Differential Privacy (DP)? (continued)

“Differential privacy is a definition, not an algorithm,” Dwork and Roth, 2014

Mathematical definition

A randomized algorithm $\mathcal{A}$ is $\varepsilon$-DP if, for each possible output $\mathcal{P}$, for any pair of datasets $D$ and $D'$ that are the same everywhere except for on one person’s data,

$$\Pr[\mathcal{A}(D) = \mathcal{P}] \leq \exp(\varepsilon) \Pr[\mathcal{A}(D') = \mathcal{P}].$$
Outline

• What is differential privacy?
• How was it achieved in the 2018 end-to-end test?
• What effect did this have on accuracy?
• Why might differential privacy in 2020 Census be a good thing, and what is still to be decided?
Example: population counts by county

\[ \epsilon\text{-DP algorithm: } \text{tab}(\text{COUNTY}) \%\% \text{ add\_noise}(\text{epsilon}) \]

<table>
<thead>
<tr>
<th>county</th>
<th>exact</th>
<th>eps_0.25_rep_1</th>
<th>eps_0.50_rep_1</th>
<th>eps_0.75_rep_1</th>
<th>eps_1.00_rep_1</th>
<th>eps_2.00_rep_1</th>
<th>eps_4.00_rep_1</th>
<th>eps_6.00_rep_1</th>
<th>eps_8.00_rep_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6227</td>
<td>6203</td>
<td>6333</td>
<td>6278</td>
<td>6244</td>
<td>6232</td>
<td>6236</td>
<td>6230</td>
<td>6229</td>
</tr>
<tr>
<td>3</td>
<td>8466</td>
<td>8422</td>
<td>8485</td>
<td>8508</td>
<td>8539</td>
<td>8499</td>
<td>8465</td>
<td>8469</td>
<td>8473</td>
</tr>
<tr>
<td>5</td>
<td>12101</td>
<td>11854</td>
<td>12116</td>
<td>12070</td>
<td>12097</td>
<td>12109</td>
<td>12104</td>
<td>12106</td>
<td>12104</td>
</tr>
<tr>
<td>7</td>
<td>34510</td>
<td>34591</td>
<td>34555</td>
<td>34469</td>
<td>34491</td>
<td>34511</td>
<td>34513</td>
<td>34510</td>
<td>34515</td>
</tr>
<tr>
<td>9</td>
<td>22017</td>
<td>21857</td>
<td>21981</td>
<td>21938</td>
<td>22039</td>
<td>22005</td>
<td>22028</td>
<td>22013</td>
<td>22007</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>69</td>
<td>4295</td>
<td>4409</td>
<td>4383</td>
<td>4391</td>
<td>4298</td>
<td>4288</td>
<td>4310</td>
<td>4300</td>
<td>4298</td>
</tr>
<tr>
<td>71</td>
<td>30665</td>
<td>30651</td>
<td>30611</td>
<td>30650</td>
<td>30689</td>
<td>30660</td>
<td>30659</td>
<td>30658</td>
<td>30656</td>
</tr>
<tr>
<td>73</td>
<td>60444</td>
<td>60623</td>
<td>60429</td>
<td>60418</td>
<td>60449</td>
<td>60467</td>
<td>60456</td>
<td>60429</td>
<td>60442</td>
</tr>
<tr>
<td>75</td>
<td>27296</td>
<td>27351</td>
<td>27445</td>
<td>27392</td>
<td>27282</td>
<td>27274</td>
<td>27291</td>
<td>27305</td>
<td>27299</td>
</tr>
<tr>
<td>77</td>
<td>99424</td>
<td>99319</td>
<td>99222</td>
<td>99471</td>
<td>99362</td>
<td>99449</td>
<td>99413</td>
<td>99424</td>
<td>99431</td>
</tr>
</tbody>
</table>
Is it really that simple?
Distribution of Noise Added to Histogram Cells

Number of areas

\[ \text{Geometric Noise (people)} \]

(a)
Distribution of Noise Added to Histogram Cells

(a) Distribution of geometric noise added to histogram cells as a function of the number of areas.

(b) Empirical privacy loss as a function of the error in DP count.

Institute for Health Metrics and Evaluation
Is it really that simple?

DAS CORE APPLICATION FRAMEWORK

DAS FRAMEWORK CREATES HISTOGRAM

OPTIMIZE

DIFFERENTIALLY PRIVATE HISTOGRAM OF 340M PEOPLE IN US

OPTIMIZE

PEOPLE ALLOCATED TO STATES

OPTIMIZE

PEOPLE ALLOCATED TO COUNTIES

OPTIMIZE

PEOPLE ALLOCATED TO CENSUS TRACTS

OPTIMIZE

PEOPLE ALLOCATED BLOCK GROUPS

OPTIMIZE

PEOPLE ALLOCATED BLOCK & HOUSES

OPTIMIZE

DAS VALIDATOR

Amazon S3 uscb-decennial-ite-das

CEF FILE

Amazon S3 uscb-decennial-ite-das

MDF FILE
Outline

• What is (differential) privacy?
• How was it achieved in the 2018 end-to-end test?
• **What effect did this have on accuracy?**
• Why might differential privacy in 2020 Census be a good thing, and what is still to be decided?
We are sitting in enumeration district 35 of Tompkins County
How does TopDown add noise to this district?

Whites in Tompkins County, ED 35

Size of Population (people)

Exact value: 876

Privacy Budget (epsilon)
How does TopDown add noise to this district?

Blacks in Tompkins County, ED 35

Size of Population (people)

Privacy Budget (epsilon)

Exact value: 44
How does TopDown add noise to this district?

Chinese in Tompkins County, ED 35

Privacy Budget (epsilon)

Size of Population (people)

Exact value: 4
How does TopDown add noise to this district?

Japanese in Tompkins County, ED 35

Size of Population (people)

Privacy Budget (epsilon)

Exact value: 0
Distribution of noise after optimization

(for epsilon=1.0, for district-level stratified counts)
Empirical Privacy Loss decreases as a function of epsilon, but only for epsilon of at least 1.0
Empirical Privacy Loss decreases as a function of epsilon, but only for epsilon of at least 1.0

(for county-level stratified counts)
Outline

• What is differential privacy?
• How was it achieved in the 2018 end-to-end test?
• What effect did this have on accuracy?
• Why might differential privacy in 2020 Census be a good thing, and what is still to be decided?
Why might $\epsilon$-DP in 2020 Census be a good thing?

There is an opportunity to democratize tradeoff decision between privacy and accuracy.

This could be a major improvement for social research.

But only if we understand what they are doing well enough to be part of the decision.
Still to be decided

1. Overall privacy budget: “epsilon”
2. How to split this budget between national, state, county, tract, block group, and block
3. At each level, how to split level-budget between “detailed queries” and “DP queries”
4. What DP Queries to include
5. What invariants to include
6. What constraints to include
7. What to publish
Still to be decided

1. **Overall privacy budget: “epsilon”**
2. How to split this budget between national, state, county, tract, block group, and block
3. At each level, how to split level-budget between “detailed queries” and “DP queries”
4. What DP Queries to include
5. **What invariants to include**
6. What constraints to include
7. **What to publish**
My Recommendations

1. If TopDown performs on 2020 data like it does on 1940 data, we should chose a privacy loss budget of 1.0 or greater

2. Include total count at census block level as an invariant

3. Publish (a) “state-by-state synthetic microdata” files and (b) “replication archive noisy count” file
Thank you

Thanks to:

• Samantha Petti, CS PhDc at GATech, who did all the work understanding the code;

• Simson Garfinkel and Philip Leclerc at Census Bureau who graciously answered many questions from me and Sam

• Social researchers like you, who have been patient despite being anxious about these big changes happening for the 2020 US Census
For first time this decade, a dip in King County’s white population, census data shows

July 17, 2019 at 6:00 am | Updated July 17, 2019 at 11:25 am

Growing diversity in King County as white population dips
From 2017 to 2018, for the first time this decade, King County’s white population declined.

Source: U.S. Census Bureau

Reporting by GENE BALK, Graphic by MARK NOWLIN / THE SEATTLE TIMES

By Gene Balk / FYI Guy  
Seattle Times columnist
For first time this decade, a dip in King County’s white population, census data shows

July 17, 2019 at 6:00 am | Updated July 17, 2019 at 11:25 am