General Comments on TDA/DAS

V. Joseph Hotz

Duke University

CNSTAT Workshop on Disclosure Avoidance in DHC

June 21, 2022

My Remarks Today

- I haven’t had time to adequately assess Michael Hawes’s presentation slides, given their (necessary) late arrival to me.

- I will offer some ***general comments*** – less directly responsive to Michael’s presentation – about ***TDA for 2020 Census DHC*** & ***its DP underpinnings***.

- I will draw on the following:

  - JASON (2020), Formal Privacy Methods for the 2020 Census (April).

  - JASON (2022), JASON (2022), Formal Consistency of Data Products and Formal Privacy Methods for the 2020 Census (January).

  - boyd & Sarathy, “Differential Perspectives: Epistemic Disconnects Surrounding the US Census Bureau’s Use of Differential Privacy,” forthcoming in *Harvard Data Science Review.*

  - Hotz & Salvo, “A Chronicle of the Application of Differential Privacy to 2020 Census,” forthcoming in *Harvard Data Science Review.*

  - Hotz, Bollinger, Komarova, Manski, Moffitt, Nekipelov, Sojourner & Spencer, “Balancing Privacy and Usability in the Federal Statistical System,” forthcoming in *Proceedings of the National Academy of Sciences*.

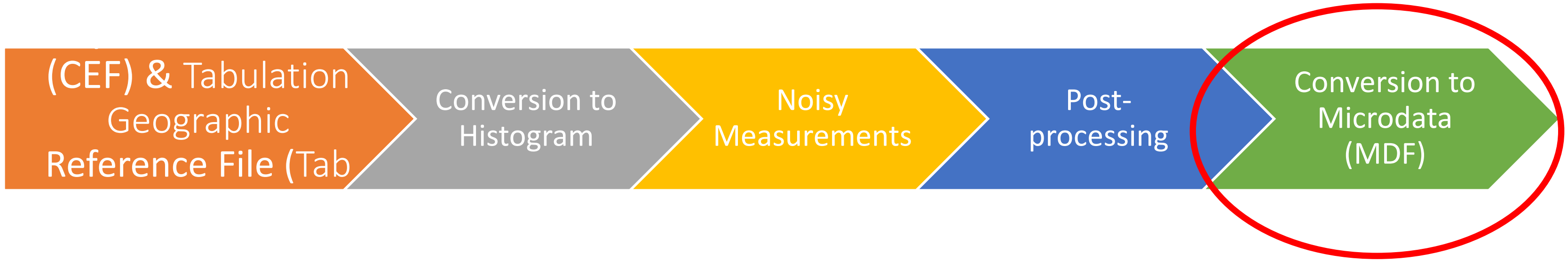
Giving Credit to Census Bureau

- There is ***clear evidence*** that ***TDA improved*** from Oct. 2019 to March 2022 Versions!
- Changes included:
  - Move from TDA based on "pure DP" with Laplace Mechanism to "Zero Concentrated DP" (zCDP) with Gaussian mechanism
  - Move to Optimized Geographic Spine
  - Better handling of Post-Processing constraints (more below)
- They improved ***accuracy*** of ***many statistics*** in both PL and DHC data – JASON Report (2022).
- Much of this, although not all, was in ***response to input from user communities***.

Some Remaining Issues

- **Synthetic Microdata Step** of TDA (see next slide).
  - Due to inherited ***production system*** for data products, TDA is required to produce ***synthetic microdata*** file in same form as Census Edited File (CEF)
  - JASON (2022) argues this step wasn't needed, may have introduced additional errors & certainly ***added to complexity*** of TDA
  - I understand this won't change for 2020 Census data, but…
    - It would be good to assess & report on its its consequences
    - More importantly, this step needs to be ***eliminated*** for 2030 Census & other data products.

The TopDown Algorithm

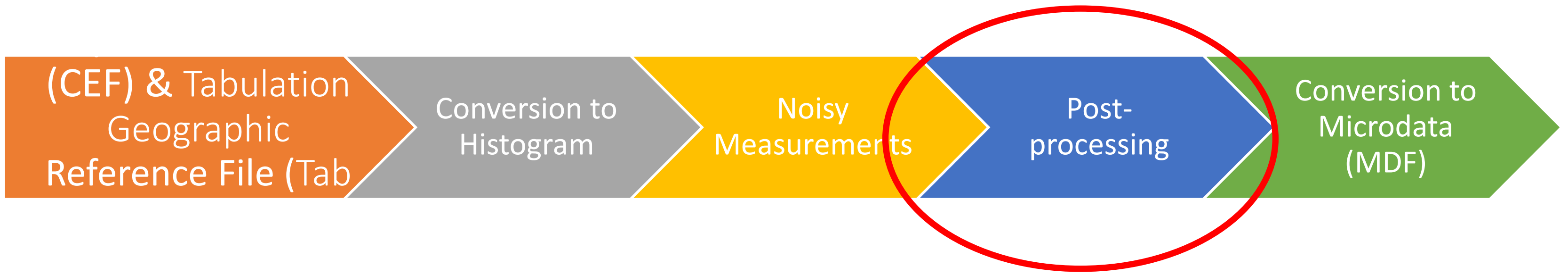


For complete details see: Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*. (June) https://doi.org/10.1162/99608f92.529e3cb9

Some Remaining Issues

- ***Imposing constraints*** in TDA
  - This the ***post processing*** step in the TDA (next slide).
  - Post-processing within TDA was ***improved***, ***but issues remain***.
    - Evidence suggested this imposed additional errors/inaccuracies in TDA processing. (JASON, 2022)
  - Important to ***explore other ways to impose constraints*** in advance of 2030 Census & other Census products. (Wang & Reiter, 2021)
  - Important to ***provide more information*** to users ***about uncertainties post processing imposed*** on data (more below).

The TopDown Algorithm



For complete details see: Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*. (June) https://doi.org/10.1162/99608f92.529e3cb9

Some Remaining Issues

- ***Understanding sources of uncertainty*** in data
  - Important to understand contributions of components of TDA to *uncertainty* (margins of error) in DHC & other data products.
    - ***Noise injection*** based on DP criterion is ***relatively straightforward*** & ***transparent***.
    - ***Uncertainties due to post-processing*** & ***synthetic microdata step*** are ***not***!
  - But also ***essential*** for Census to provide ***more full accounting*** of ***other non-sampling sources of error/uncertainty*** and **soon**! – JASON (2022); boyd & Sarathy (2022)
    - Users will need this to make informed & effective use of data.
    - Software & documentation will need to be developed & made known & available to users. – JASON (2022).
  - ***Related point***: as part of this, ***Census needs to either release noisy measurements*** file ***or close proximity*** to it. – JASON (2022) & others.

Communication, Communication, Communication!

- ***Essential for Census to improve communication about DAS/TDA for all levels of users***
  - This is ***really essential*** on multiple levels:
    - Essential for ***users & public*** – other than CS experts – ***to better understand Census's DAS***.
    - Deal with ***misconceptions*** that are replete & need to be addressed.
    - ***Clearer*** & ***more transparent explanations of DAS/TDA*** – including what it did & didn't do – will ***help users to make more effective use of DHC***, DDHC, other Census products.
  - See JASON (2022), boyd & Sarathy (2022), among others.

Absolute vs Relative Disclosure Risk

- Following slides draw on:
  - Hotz, V.J., C. Bollinger, T. Komarova, C.F. Manski, R.A., Moffitt, D. Nekipelov, A. Sojourner & B.D. Spencer, "Balancing Privacy and Usability in the Federal Statistical System," forthcoming in *Proceedings of the National Academy of Sciences*.
  - Hotz, V.J. & J.J. Salvo, "A Chronicle of the Application of Differential Privacy to 2020 Census," forthcoming in *Harvard Data Science Review.*

Absolute vs Relative Disclosure Risk

- ***Absolute Disclosure Risk:***

$$\Pr\left(J = j, Y_j | D^*, A\right)$$

where J is ***target individual*** in an ***intruder's information set***, A;
j is ***individual*** in ***released data set***, D^*;
D is ***confidential data set***;
Y_j is ***true value*** of j's (***sensitive***) ***data*** in D;

- Using Bayes Rule:

$$\Pr\left(J = j, Y_j \middle| D^*, A\right) = \left[\Pr\left(D^* | J = j, Y_j, A\right) / \Pr(D^* | A)\right] \Pr\left(J = j, Y_j \middle| A\right)$$

where $\Pr\left(J = j, Y_j \middle| A\right)$ is ***intruder's prior*** about j being J & her data being Y_j, ***based on just intruder's info***, A, and,

$$\Pr\left(D^* | J = j, Y_j, A\right) / \Pr(D^* | A) = \Pr\left(J = j, Y_j \middle| D^*, A\right) / \Pr\left(J = j, Y_j \middle| A\right)$$

where RHS (blue) is ***Relative Disclosure Risk***, i.e., ***increase*** in (or ***incremental***) ***disclosure risk from releasing*** D^*.

Absolute vs Relative Disclosure Risk

- Consider ***analogy*** to ***mortality risks from diseases/health conditions***:
  - Epidemiological studies often focus on ***relative risks*** of ***dying*** from ***one disease/condition vs. another***.
  - These risks may be ***easier to estimate***, as they ***abstract from differences*** in individuals' ***underlying health conditions*** (their ***prior risks***), which are harder to adequately measure.
  - But ***most individuals*** – & their health care providers – ***care about absolute risk of their dying from disease/condition they have***.
    - Individuals **may not worry** about even a **large relative increase in risk of dying** from condition **if they are in "good health."**
    - But, individuals **may worry** a lot about even a **small relative increase in risk of dying** from condition **if they aren't in good health.**

Absolute vs Relative Disclosure Risk

- ***Case*** of ***privacy loss due to a data release***.
  - One can argue that ***individuals in confidential data set*** – & maybe stat agencies – ***may*** (***should***) ***care*** about ***absolute risk of disclosure*** from ***releasing their data***.
    - Individuals **may care** a lot about even **small relative increase in disclosure risk** due to data release **if prior prob of disclosure is high**, all else equal.
    - May **not be bothered** by even **large relative increase in disclosure risk** due to date release **if prior prob of disclosure risk is low.**

Two Observations concerning Disclosure Risks from Data Releases

1. One can show (Gong & Meng, 2020*, among others) that **DP criterion** – and **DP-based DASs** – **bound *relative disclosure risks,* not *absolute ones***.

   - ***Rationale for DP focus on relative risks***: Serious challenges to quantifying or knowing what information potential intruders may know.

   - DP-based mechanisms address ***worst-case scenarios***.

   - And, importantly, DP-mechanisms can quantify and **"guarantee"** control over the *relative increase in disclosure risks* from data releases in ***transparent way***.

*Gong, R. and X.-L. Meng. 2020. "Congenial Differential Privacy under Mandated Disclosure," *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*. Virtual Event, USA: Association for Computing Machinery, 59–70.

Two Observations concerning Disclosure Risks from Data Releases

2. But, there are **"risks"** to ***not paying more attention*** to ***absolute disclosure risks*** in designing DAS for data releases:

   - Will have ***harder time communicating disclosure risks individuals do face***.
   - May have ***harder time complying with privacy protection mandates*** (Title 13 & 26).
   - And, ***not paying attention to absolute disclosure risks will limit***:
     a) *assessing extent of these risks*, e.g., doing further reconstruction and re-identification assessments for 2020 Census data;
     b) *assessing how magnitudes of risks differ* across *groups* & *types* of information;
     c) *Improving* our *understanding which external data* (priors) *contribute* to *risks of disclosure*.

Thanks for Listening!