Differential Privacy at the US Census Bureau: Status Report

Simson L. Garfinkel
Rolando Rodriguez
Phil Leclerc

U.S. Census Bureau

January 27, 2020

The views in this presentation are those of the author, and not those of the U.S. Census Bureau.

Abstract

The US Census Bureau will be using differential privacy as the primary privacy protection mechanism for the 2020 Census.

Following this transition, the Census Bureau will be working to use differential privacy in the American Community Survey, the Economic Census, and other data products.

In differential privacy, the parameter "epsilon" is used to control the tradeoff between privacy and accuracy. The Census Bureau has created an "epsilon registry" to track all uses of DP within the Census Bureau.

Acknowledgments

This presentation incorporates work by:

John Abowd (Chief Scientist)

Dan Kifer

William Sexton

Pavel Zhuravlev

Knexus Research Corporation

Briefing outline

- History of Differential Privacy and the US Census Bureau
- Differential Privacy and the 2020 Census
- Differential Privacy and the ACS
- Differential Privacy and the Economic Census

Differential Privacy and US Census Bureau: A Brief History

Shape
your future
START HERE >

United States®
Census
2020

Imagine a typical block in the US

The job of an official statistics agency is to collect data and publish useful statistics.

Punch Cards were invented for the 1890 Census

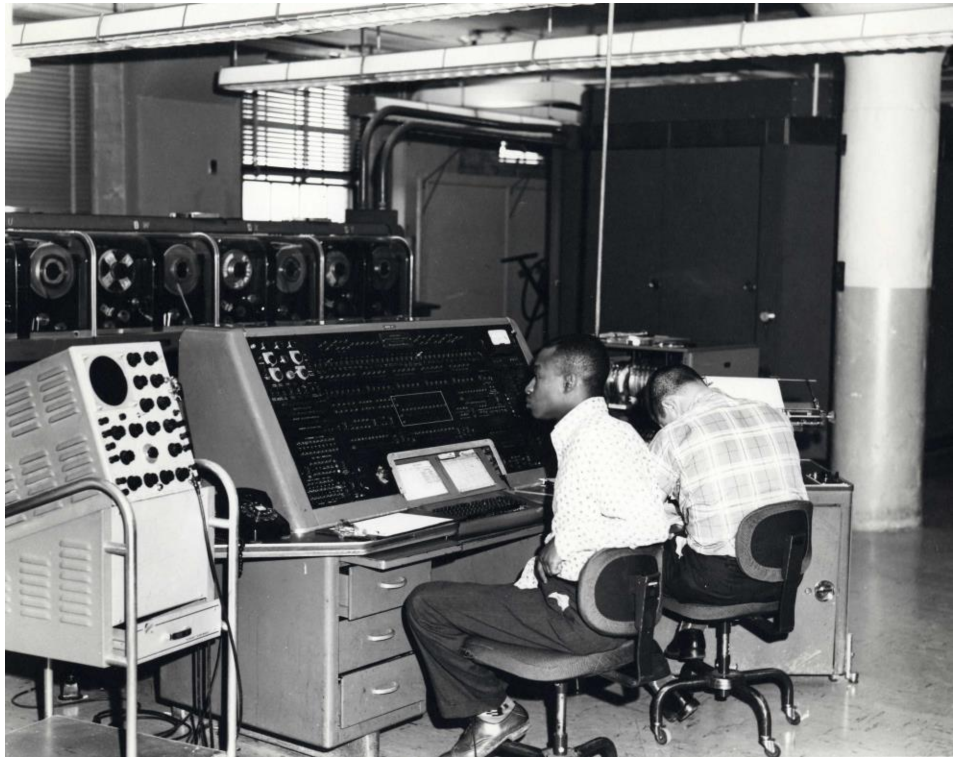
https://www.census.gov/history/www/innovations/technology/the_hollerith_tabulator.html

Shape your future START HERE >

United States® Census 2020

The US Census Bureau Bought the First commercial computer

https://www.census.gov/history/www/innovations/technology/univac_i.html



Shape your future START HERE >

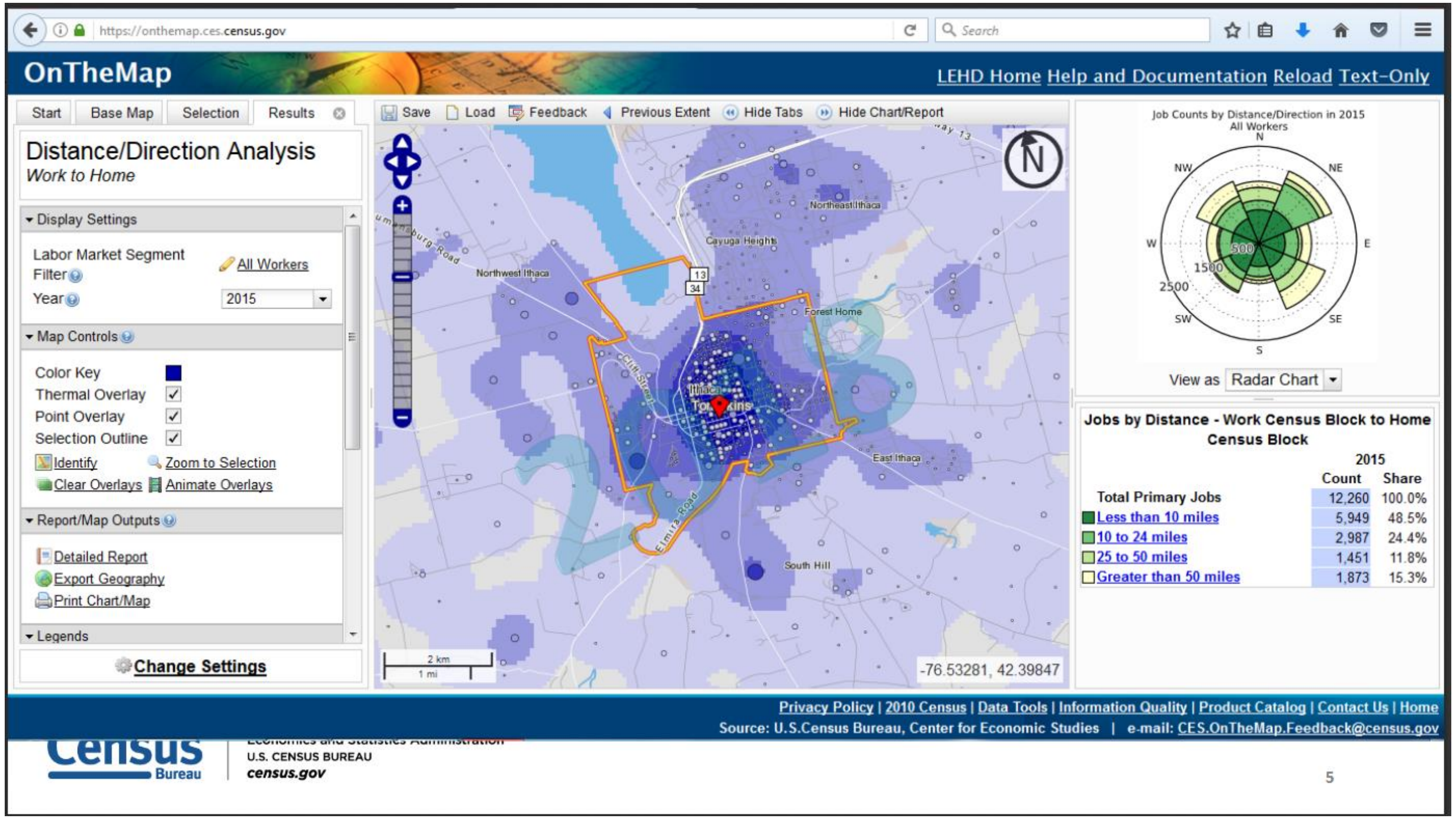
United States® Census 2020

Differential Privacy was invented for the US Census

October 25, 2019

The Census Bureau deployed differential privacy “on the map” in 2008!

https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf



Distance/Direction Analysis
Work to Home

Display Settings

Labor Market Segment Filter: All Workers
Year: 2015

Map Controls

Color Key
Thermal Overlay ☑
Point Overlay ☑
Selection Outline ☑

Report/Map Outputs

Jobs by Distance - Work Census Block to Home Census Block

	2015	
	Count	Share
Total Primary Jobs	12,260	100.0%
🟩 Less than 10 miles	5,949	48.5%
🟩 10 to 24 miles	2,987	24.4%
🟩 25 to 50 miles	1,451	11.8%
□ Greater than 50 miles	1,873	15.3%

Source: U.S.Census Bureau, Center for Economic Studies | e-mail: CES.OnTheMap.Feedback@census.gov

Shape your future START HERE >

United States® Census 2020

Differential Privacy, DRB, and DSEP

The DRB (Disclosure Review Board) has delegated authority from DSEP (the Data Stewardship Executive Policy committee) to release statistics based on Title 13-protected data.

Currently, decisions regarding differential privacy drawn from a globally managed privacy loss budget are being made by DSEP.

The DRB has also approved noise infusion methodologies "inspired" by differential privacy.

- Additive noise drawn from a Laplace distribution.
- Attempting to compute some kind of sensitivity metric.

Decisions currently made by reviewing accuracy metrics at different levels of epsilon.

Epsilon Registry

Tracking:

ID#
Project Name
Project Sponsor
Dataset(s)
Formally Private (y/n)
Original DRB Date
Original DRB ID Number
Referred to Chief Scientist
Referred to DSEP

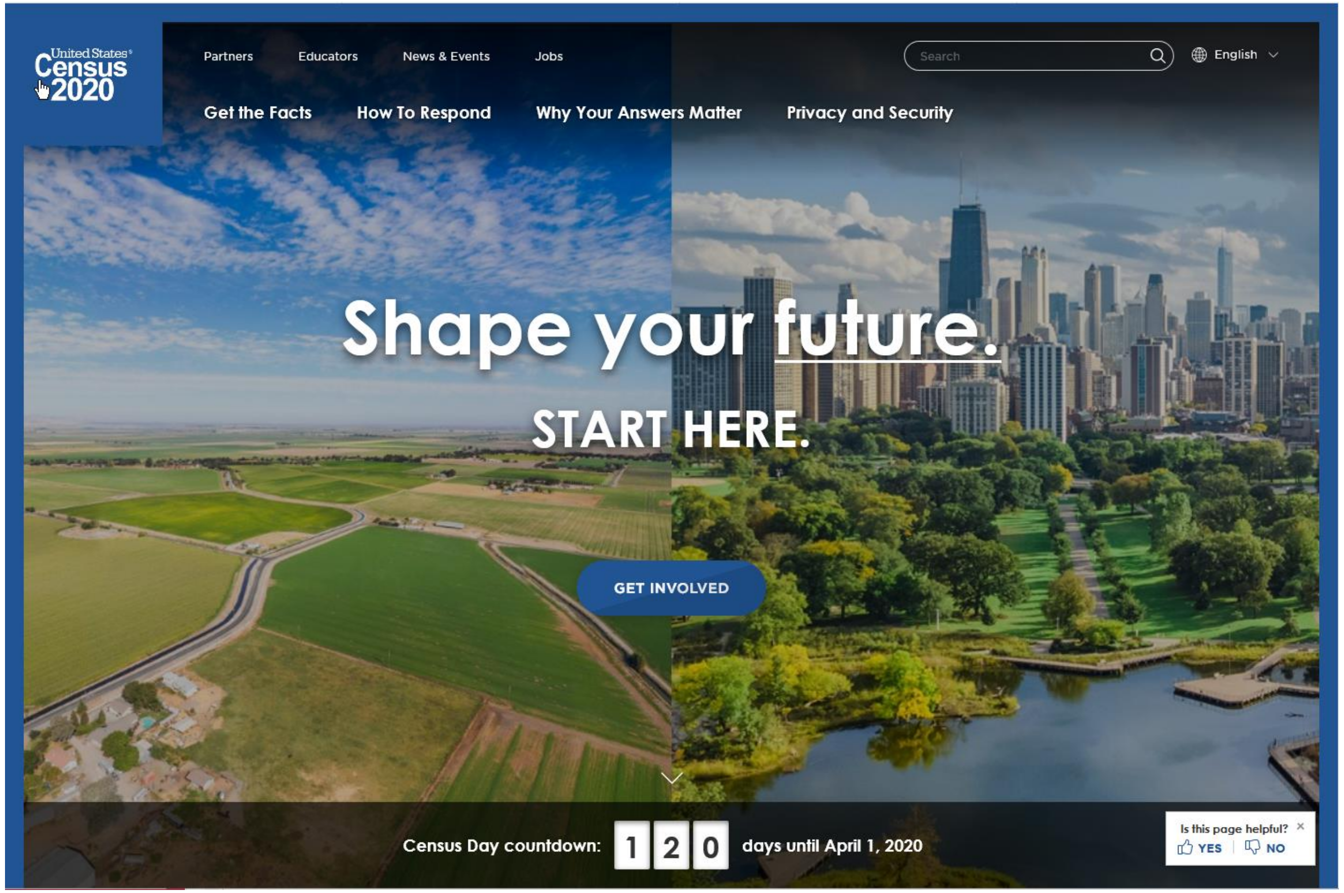
Original DRB Decision Number
Original Epsilon PLB Referred
DSEP Date of Review
DSEP Decision
Location of DSEP Minutes
Additional DRB Review Date
Additional DRB Agenda #
DRB Decision Number
DMS Project Number

Notes

Total projects in registry: 5

Currently this is a manual process…

Differential Privacy and the 2020 Census



Shape your future START HERE >

2020 DAS: Team

Federal Leadership

- John Abowd: Sponsor and Chief Scientist
- Rob Sienkiewicz: Portfolio Manager
- John Fattaleh: Project Manager
- Simson Garfinkel: Engineering Lead
- Phillip Leclerc: Science Lead

Contractors

- Knexus Research Corporation (KRC): Data modeling, programming, technical writing
- Econometrica: Project/Sprint management, programming, technical writing
- MITRE Corporation: Detailed tables, differential privacy SMEs, IT and AWS configuration support, application development, technical management (some of this work is being done by Tumult Labs under sub-contract to Mitre)

2020 DAS: Scope

Implement a disclosure avoidance method that ensures dissemination of high quality data while fully meeting legal and ethical obligations to protect the confidentiality of respondents and their information.

Development of applications that applies differential privacy in the generation of protected data products.

Development and operation of the applications to occur in TI GovCloud "ITE" environment.

Internal Use Only: Pre-decisional

2020 DAS: Data Products

Group I	Group II	Group III
• PL94 • Demographic Profiles • DHC-Persons • DHC-Households • CVAP (special tabulation)	• AIAN • Detailed Race • Person & Household joins • Averages	• PUMS • Special Tabulations

2020 DAS: Scope - Engineering

DAS Development to happen in TI GovCloud “ITE” environment

Develop methods to automate processing

Conduct research to determine best configuration that balances cost vs. performance

Coordinate code and data codification

2020 DAS: Scope - Challenges

Verify capability to implement the full DHC specification

Determine optimal methods (i.e., code and hardware) for implementation of disclosure avoidance to the Group I data products

Integration of the Tumult Labs solution

Define schedule for the Group II and Group III data products

Coordination on implementation of new IT system requirements

Coordination with external contractors for detailed tables

Need to determine when to work on HDMM integration and public-historical data improvements

Resource management

What we’ve done to date

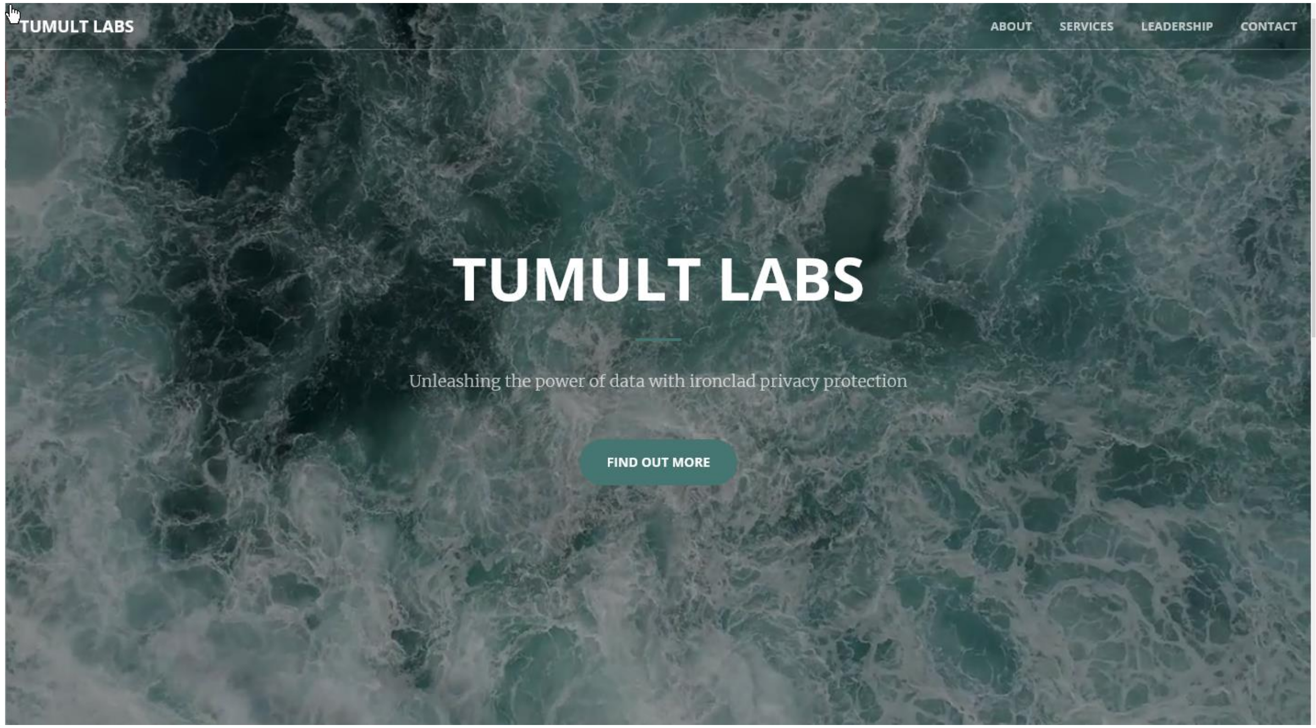
Scientific Development

Algorithmic Development

Education Efforts

Tumult Labs

"Tumult Labs builds state-of-the-art privacy technology to enable the effective use of data while respecting the privacy of contributing individuals. Our technology enables the safe release of de-identified data, statistics and machine learning models. All of our solutions satisfy differential privacy, an ironclad, mathematically-proven privacy guarantee."



Differential Privacy and the American Community Survey

Introducing the American Community Survey

The American Community Survey (ACS) is the U.S. Census Bureau's largest demographic survey

The main release has 1-year and 5-year microdata and tables - Over 10 billion estimates produced per year

ACS collects a wealth of information about households and people

Household characteristics (relationships, mortgage/rent, utilities); Person characteristics (age, sex, ancestry, schooling, o ccupation); >35 topic areas

The ACS determines the distribution of over $650 billion in federal funds annually

Status of formal privacy in the ACS

Public statements give 2025 as earliest DP ACS implementation

Our partners at Purdue and Brandeis are leading ACS DP research

Effect of ACS production processes on query sensitivity

- Missing data imputation; Weighting adjustments to control to census-based estimates

Smooth sensitivity for alternative production processes

Inroads for randomization to reduce noise needed for current processes

What is the best path forward for a formally private main release?

How do we satisfy the demand for detailed microdata?

The ACS is not just the main data release

Numerous organizations request special tabulations from the survey

Privacy practices for these currently grandfathered; But they provide a potential base for research and messaging

Results also come from:

The Census Bureau's Research and Methodology directorate; This include Federal Statistical Research Data Centers (FSRDCs)

Queries on sub-state populations must have noise injection prior to release; Common themes:

- Weighted generalized linear models
- Weighted quantiles

Formal privacy methods for these outputs would be immediately useful

Paraphrases from the data users

The ACS is a sample so it's already safe

The ACS already has disclosure avoidance procedures

How will differential privacy affect my particular analysis?

What if my city puts resources in the wrong place due to privacy?

We base our decisions on the estimates only

Differential Privacy and Economic Products

Roadmap

LEHD “On The Map” — The first product to use Differential Privacy

Our original plan was to make the Business Dynamics Statistics the second economic product.

What happened:

1. ECON 2017 was more complicated than expected. De-scoped to a single business sector.
2. Business Dynamics Statistics — Emergency deployment of DP using “Pufferfish” relaxation.
3. Current plan is for a demonstration product in ECON 2020. Ian Schmutte is the lead scientist for the formally private Econ Census team. Nick Orsini and John Abowd are working out the details.

Publications to date

The Most Technical 2020 Publications

Github: https://github.com/uscensusbureau/census2020-das-2010ddp

https://github.com/uscensusbureau/census2020-das-2010ddp/blob/master/doc/2010-Demonstration-Data-Products-Disclosure-Avoidance-System-Design-Specification%20FINAL.pdf

https://github.com/uscensusbureau/census2020-das-2010ddp/blob/master/doc/20191020_1843_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf

For more information…

THE WALL STREET JOURNAL.

English Edition ▾ | December 6, 2019 | Print Edition | Video

Politics Economy Business Tech Markets Opinion Life & Arts Real Estat

Census Overhaul Seeks to Avoid Outing Individual Respondent Data

Most Census 2020 results will be adjusted; measures would prevent targeting based on citizenship

By Paul Overberg

Nov. 10, 2019 7:00 am ET

practice

DOI:10.1145/3287287

Article development led by acmqueue
queue.acm.org

These attacks on statistical databases are no longer a theoretical danger.

BY SIMSON GARFINKEL, JOHN M. ABOWD, AND CHRISTIAN MARTINDALE

Understanding Database Reconstruction Attacks on Public Data

IN 2020, THE U.S. Census Bureau will conduct the Constitutionally mandated decennial Census of Population and Housing. Because a census involves collecting large amounts of private data under the promise of confidentiality, traditionally statistics are published only at high levels of aggregation. Published statistical tables are vulnerable to *database reconstruction attacks* (DRAs), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations. A DRA can be performed by using the tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. This article shows how such an attack can be addressed by adding noise to the published tabulations, so the reconstruction no longer results in the original data. This has implications for the 2020 census.

The goal of the census is to count every person once, and only once, and in the correct place. The results are used to fulfill the Constitutional requirement to apportion the seats in the U.S. House of Representatives among the states according to their respective numbers.

In addition to this primary purpose of the decennial census, the U.S. Congress has mandated many other uses for the data. For example, the U.S. Department of Justice uses block-by-block counts by race for enforcing the Voting Rights Act. More generally, the results of the decennial census, combined with other data, are used to help distribute more than $675 billion in federal funds to states and local organizations.

Beyond collecting and distributing data on U.S. citizens, the Census Bureau is also charged with protecting the privacy and confidentiality of survey responses. All census publications must uphold the confidentiality standard specified by Title 13, Section 9 of the U.S. Code, which states that Census Bureau publications are prohibited from identifying "the data furnished by any particular establishment or individual." This section prohibits the Census Bureau from publishing respondents' names, addresses, or any other information that might identify a specific person or establishment.

Upholding this confidentiality requirement frequently poses a challenge, because many statistics can inadvertently provide information in a way that can be attributed to a particular entity. For example, if a statistical agency *accurately* reports there are two persons living on a block and the average age of the block's residents is 35, that would constitute an improper disclosure of personal information, because one of the residents could look up the data, subtract their contribution, and infer the age of the other.

Communications of ACM March 2019
Garfinkel & Abowd

Can a set of equations keep U.S. census data private?

By Jeffrey Mervis

Science

Jan. 4, 2019 , 2:50 PM

http://bit.ly/Science2019C1

More Background on the 2020 Census Disclosure Avoidance System

September 14, 2017 CSAC (overall design) https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf?#

August, 2018 KDD'18 (top-down v. block-by-block) https://digitalcommons.ilr.cornell.edu/ldi/49/

October, 2018 WPES (implementation issues) https://arxiv.org/abs/1809.02201

October, 2018 *ACMQueue* (understanding database reconstruction) https://digitalcommons.ilr.cornell.edu/ldi/50/ **or** https://queue.acm.org/detail.cfm?id=3295691

December 6, 2018 CSAC (detailed discussion of algorithms and choices) https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#

April 15, 2019 Code base and documentation for the 2018 End-to-End Census Test (E2E) version of the 2020 Disclosure Avoidance System https://github.com/uscensusbureau/census2020-das-e2e

June 6, 2019 Blog explaining how to use the code base with the 1940 Census public data from IPUMS https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html

June 11, 2019 Keynote address "The U.S. Census Bureau Tries to Be a Good Data Steward for the 21st Century" ICML 2019 abstract, video

June 29-31, 2019 Joint Statistical Meetings Census Bureau electronic press kit
(See talks by Abowd, Ashmead, Garfinkel, Leclerc, Sexton, and others)