The modernization of statistical disclosure limitation at the U.S. Census Bureau

Aref N. Dajani, Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vihesh Karwa, Hang Kim, Philip Leclerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd

Acknowledgements: Katherine J. Thompson and Michael Freiman

Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality, Skopje
September 21, 2017

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Here is what you need to know from today's presentation. (Slide 1 of 2)

Maintaining data privacy and confidentiality is critical to the release of all public-use statistical data products at the U.S. Census Bureau.

The Census Bureau is currently conducting research that may lead to implementing formally private releases for all of our public-use statistical data products.

The Census Bureau data product OnTheMap was the first real world production deployment of formal privacy (2008).

The 2020 Census of Population and Housing Disclosure Avoidance System (DAS) will use differential privacy, without modification of the original definition, to develop the inputs to the public-use tabulation system (redistricting data and summary files).

Here is what you need to know from today's presentation. (Slide 2 of 2)

If current experiments are successful, the American Community Survey may release a synthetic public-use microdata sample file and some components of this file may be formally private.

The 2017 Economic Census is conducting "proof of concept" research on synthesizing product data for tabulations based on its new North American Product Classification System.

Differentially private synthetic micro-data will allow the Census Bureau to produce public-use statistical data products at finer granularity with provable, strong confidentiality protection that does not depend upon hypotheses about the attackers' future information sets.

The Census Bureau is legally mandated to collect and protect personal information.

Title 13, U.S. Code, Section 9

- May not “use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied”.
- May not “make any publication whereby the data furnished by any particular establishment or individual under this title can be identified”.
- May not “permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports”.

Title 26 places similar requirement on tax data from the U.S. Internal Revenue Service when used by the Census Bureau for statistical purposes.

The Census Bureau applies disclosure avoidance procedures to all its public-use data products.

The Census Bureau releases reports, tables, infographs, and public-use microdata files.

The Census Bureau applies disclosure avoidance procedures to satisfy the statutory requirements of its legal mandate.

Current methods include suppression, aggregation/coarsening, perturbation with random input or output noise, and synthetic data, which is a special case of perturbation.

There are considerable problems with current statistical disclosure avoidance methods.

They are *ad hoc*: not based on a principled measure of global disclosure risk.

They cannot characterize the set of feasible attacks for which they are effective or the set of future attacks for which they will remain effective.

Some parameters; e.g., swapping rate, amount of noise, are not revealed. This prevents researchers from compensating properly for disclosure protection methods when doing estimation and inference.

All data publications lead to some privacy loss.

Privacy loss may be incremental because data releases must be considered in context of data already released.

Prior work in the mathematical landscape validate the concerns expressed above.

Database Reconstruction Theorem (Dinur and Nissim 2003)

- Any finite database can be reconstructed with arbitrary accuracy using finitely many accurately-answered queries.

Dwork et al. (2006) and Dwork and Naor (2010)

- Inferential disclosure cannot be fully eliminated unless published data are useless (fully encrypted). [Exact identity and attribute disclosures are special cases of inferential disclosures.]

Formal Privacy establishes definitions and mechanisms that are proven to implement it.

Differential Privacy ensures that the addition or removal of a single row in a database has a bounded impact on the probability distribution of outputs.

This bounded impact is e^{ε}.

$$\Pr[A(D_1) \in S] \leq e^{\varepsilon} Pr[A(D_2) \in S] \; \forall S, \|D_1 - D_2\|_1 = 1$$

Implies that the amount of information disclosed about a single row is limited relative to what is already known and must hold true regardless of other knowledge.

Can be applied to data inputs or outputs, typically based on noise infusion.

Suppose an agency has the choice to release disclosure-protected output based on the full dataset or based on the dataset with a single record omitted.

User's probabilistic belief about any individual record must be approximately the same regardless of which of these two datasets is used; approximation is bounded by e^{ε}.

ε is the privacy-loss budget that provides proven protection into the indefinite future.

This budget often called an R-U curve in SDL, or Production Possibilities Frontier (PPF), in economics, is also equivalent to Receiver Operating Characteristics (ROC) curve in statistics.

It shows how the SDL technology constrains the aggregate risk of partial database reconstruction given all published statistics.

All released statistics can *never* permit a database reconstruction more accurate than the budget.

For differential privacy, the guarantee is over all future attackers and any database with the same schema.

The Census Bureau is developing a *framework* for a Disclosure Avoidance Systems (DAS).

Two modes

- Development & Test Mode
- Production Mode

Two Testing Systems

- DAS0 — 100% accuracy, no privacy
  (No disclosure avoidance)
- DAS1 — 100% privacy, no accuracy
- DAS2 — “bottom-up” engine

One Operational System

- DAS3 — “top-down” engine

The price of increasing data accuracy (public "good") in terms of increased privacy loss (public "bad") is the *slope* of the PPF.

Cost of Statistical Disclosure Limitation is the *Slope* of the Risk-Utility Curve (=PPF, =ROC curve) with privacy-loss on the x-axis and data accuracy on the y-axis.

The *slope* of this PPF is also called the marginal social cost.

Difficult to select an optimal point because that requires a model of preferences between data accuracy and privacy loss, which is outside the scope of the disclosure limitation technology.

The optimal choice equates Marginal Social Benefit (MSB, determined by the preference model) with Marginal Social Cost (MSC, determined by the PPF).

The privacy-loss budget must quantify the privacy-loss expenditure of each data release.

The collection of the algorithms taken altogether must satisfy the privacy-loss budget.

This means that the collection of algorithms used must have known sequential and parallel composition properties.

With the information environment changing much faster than before, *it may no longer be reasonable to assert that a product is empirically safe given best-practice traditional disclosure limitation prior to its release.*

Differentially private guarantees hold even if external sources are published or released later.

Formal privacy models replace empirical assessment with designed protection.

Resistance to all future attacks is a property of the design.

Differentially private data are robust to background knowledge of the data.

They allow for post-processing edits.

If a model is formally private, then synthetic data become provably safe microdata.

Synthetic process: Collected data ➔ Model ➔ Data sharing original data's properties

If the model is not formally private, then we use the term synthetic data without modification.

The synthetic data cannot reflect all properties of the original data.

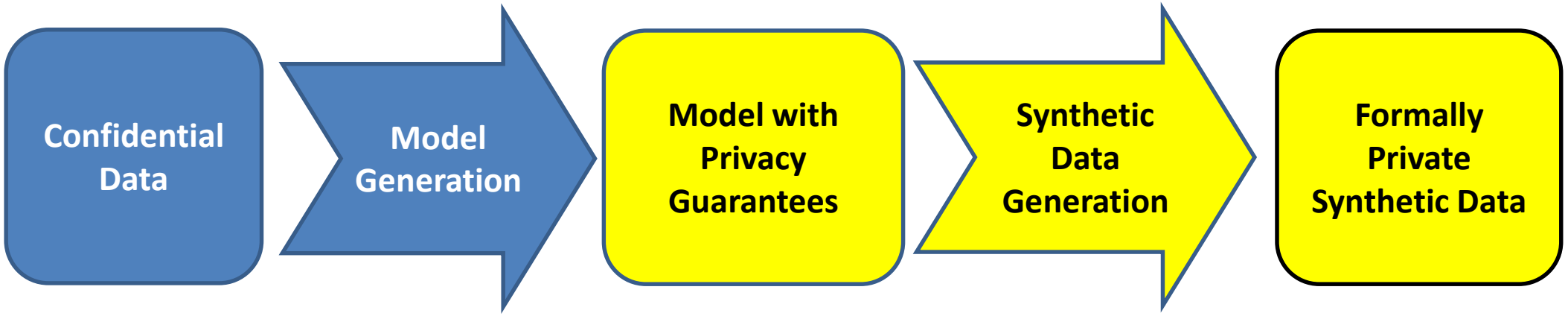
Synthetic data work well when paired with verification or validation servers.

Verification servers allow users to check whether synthetic data distorted results. If verification server calls into question validity of results, researcher may go to a research data center.

Validation servers allow models estimated on the synthetic data to be computed directly on the confidential data. If the validation server results are also protected, then they can be published directly.

Verification and validation servers also cause privacy loss.

Formally Private Synthetic Microdata have intentional errors introduced; i.e., noise infused.

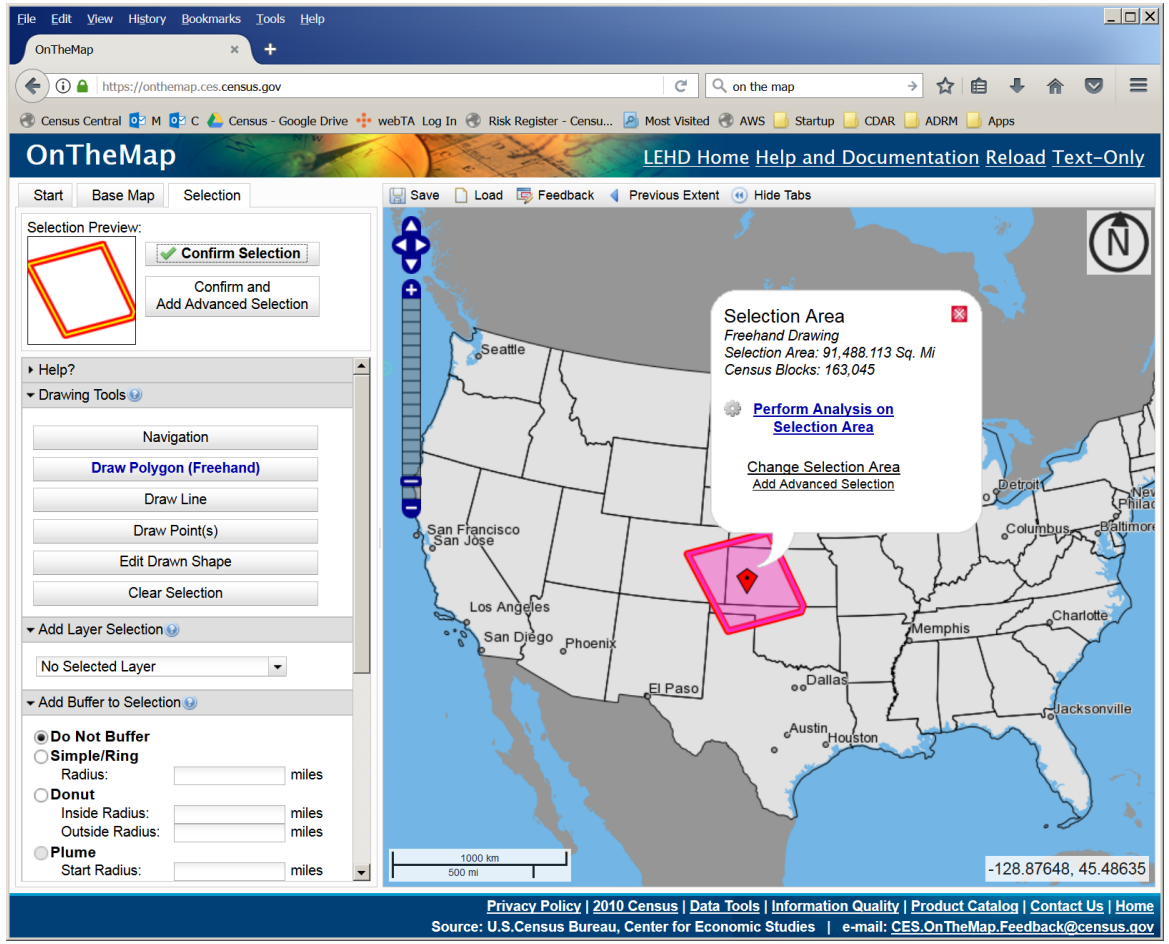


Quantifying privacy loss can be mathematically established and proven.

OnTheMap is the first real world production deployment of formal privacy.

OnTheMap is a web-based application that shows where workers are employed and where they live.

OnTheMap uses formally private synthetic data on the residence side (the system uses an early variant of differential privacy called probabilistic differential privacy).



To defend against a reconstruction attack, the 2018 End-to-End Census Test's disclosure avoidance system will use differential privacy.

Differential privacy provides provable bounds on the accuracy of the best possible database reconstruction given the released tabulations.

Sometimes, especially with complicated algorithms, the provable accuracy bounds are hard to calculate but some privacy-loss budget can be devoted to estimating them.

The 2020 Disclosure Avoidance System will rely on formally private output noise infusion.

Provides provable bounds on the accuracy of the best possible database reconstruction given the released tabulations.

Algorithms allow policy makers to decide the trade-off between accuracy and privacy loss.

Advantages of noise infusion with formal privacy: easy to understand, *tunable* privacy guarantees that do not depend on external data, protects against database reconstruction attacks, and privacy operations are *composable.*

Disadvantages: entire country must be processed at once for best accuracy and every use of private data must be tallied in the *privacy-loss budget.*

A differentially private microdata detail file is familiar to both internal and external stakeholders.

It guarantees that legally mandated population totals (e.g., voting age population) will be exact at all levels of geography.

It is consistent among query answers.

Conceptually, the new system is similar to swapping with these main differences:

- Every record in the population is at risk to be modified.
- Records in the tabulated data have no exact counterpart in the confidential data.
- Explicitly protected tabulations have either provable or estimable, public accuracy levels.

The American Community Survey (ACS) may release a synthetic public-use microdata sample file and is considering to make it formally private.

The ACS is the successor to the long form survey of the Census of Population and Housing.

The ACS is a monthly survey and receives approximately 2.5M responses each year.

The Census Bureau releases one-year and five-year ACS statistical data products.

The research team plans to build a chain of models, simulating each variable successively given the previous synthesized models.

Formally private models for the questions in common with the 2020 Census of Population and Housing will be integrated.

It may also be possible to build formally private versions of the models for other variables, then create microdata samples from these models, and finally create tables from these samples.

This work is experimental and has not cleared the internal Census reviews necessary for implementation.

The 2017 Economic Census is conducting "proof of concept" research on synthesizing product data in its new product classification system.

The Economic Census is conducted every five years to nearly four million U.S. business establishments.

Establishments are defined as a specific economic activity conducted at a specific location.

The goal is to release product and product-by-industry tabulations that satisfy predetermined privacy and reliability constraints and to release supplemental synthetic industry-level microdata files.

Differential privacy presents technical and policy challenges.

The commonly used, flattened histogram representation of the universe is calculated as the Cartesian product of all potential combinations of responses for all variables.

The number of potential combination of responses can be orders of magnitude larger than the total population being enumerated.

Policy makers (the Data Stewardship Executive Policy Committee at the Census Bureau) must have enough information about the privacy-loss/data accuracy tradeoff to make an informed recommendation to the Director about ε.

In some cases, the amount of noise infusion from differential privacy may limit the suitability for use of the published statistics to more narrowly defined domains than has historically been true.

The ACS and other Census Bureau surveys present specific challenges.

High dimensionality: There are roughly two hundred topical module variables with mixed continuous and categorical values.

Geography: Estimates are required at the block-group level.

Preserving associations among variables across people in the same household.

Outliers in economic variables are considerable.

Data are often collected with a complex sample design with considerable missing data and temporally distinct field periods.

Research is ongoing to assess the scientific validity of weighted, design-consistent analyses performed using differential privacy as compared to feasible alternatives.

There are algorithmic obstacles in generating high quality formally private synthetic microdata.

Integer counts

Non-negativity

Publicly known counts

Structural zeros

Small biases versus large aggregations

We have specific technical issues moving forward.

Currently, synthetic data methods must be customized for each confidential dataset.

- *We need generic methods that will work on a broader range of datasets.*

It may be difficult to find meaningful correlations not represented in the model.

- *The model must anticipate the analyses that will be done.*
- *We need better model-building tools.*
- *We also need generic tools for correlating arbitrary models with the ones used to build the synthetic data so that researchers can assess when to request an analysis on the original confidential data.*
- *Data custodians, like the Census Bureau, need tools for managing a dynamic privacy-loss budget so that reuse of the confidential data for new analyses is properly controlled.*

Reproducible-science methods will be required to use formally private synthetic data effectively.

Differentially private synthetic data will allow the Census Bureau to produce public-use statistical data products at finer granularity.

Demographic public-use statistical data products can be produced at lower geographies.

Economic public-use statistical data products can be produced at more specific industry, sector, and product levels.

Differentially private synthetic data protect against all future attacks and may eliminate the need to count “sample uniques” and conduct re-identification studies.

This transition involves retooling of methods for our career methodologists.

This transition will help the Census Bureau lead similar innovation across the U.S. Federal Government and beyond.

References

Dinur, Irit and Kobbi Nissim (2003). Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems(PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.

Dwork, C. (2006). Differential privacy. *Automata, languages and programming*, 1-12.

Freiman, M.H. (2017). Philosophy of disclosure avoidance for Census Bureau data. *Disclosure Avoidance Research Report Series* #2017-01. Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC.

Thanks!

Aref N. Dajani, Ph.D.

Center for Disclosure Avoidance Research

U.S. Census Bureau

Aref.N.Dajani@census.gov