

Portfolio Question 10.2.P - Privacy in the Census

Daniel Lawson University of Bristol

Portfolio 10.2.P (v4.0.0)

Portfolio Question 10.2.P - Privacy in the Census

- ▶ Should national statistics agencies deliberately make published statistics less accurate in order to protect privacy?
 - ▶ Contrast the **privacy** guarantee given to every respondent with the **accuracy** needed by the communities who rely on the statistics, and consider who bears the cost.

Why the census matters

- ▶ The census is **compulsory**: everyone must answer, in return for a legal promise of confidentiality.
- ▶ Its outputs are used to:
 - ▶ Allocate government funding to local areas,
 - ▶ Draw electoral boundaries,
 - ▶ Plan schools, hospitals and transport,
 - ▶ Provide the baseline for almost every other survey (as weights and sampling frames).
- ▶ Many of these uses need counts for **very small areas**: a single street block, or a small village.
- ▶ This is exactly where disclosure risk is highest. A table of “age by sex by ethnicity” for a block of 20 people is close to a list of individuals.

The reconstruction attack

- ▶ The US Census Bureau tested its own 2010 published tables using a **reconstruction attack**¹:
 - ▶ Billions of published table cells were treated as constraints, and a solver found individual records consistent with all of them.
 - ▶ This reconstructed block, sex, age, race and ethnicity **exactly** for about 46% of the population.
 - ▶ Linking the reconstructed records to commercial data **re-identified** about 17% of the population (roughly 52 million people).
- ▶ The 2010 tables had been protected using traditional methods, such as swapping some households between areas.
- ▶ This is a statistical disclosure attack (see the lecture) carried out at national scale: no single table revealed anything, but their **intersection** did.

¹Abowd et al. "A simulated reconstruction and reidentification attack on the 2010 U.S. Census." Harvard Data Science Review (2023).

The response: differential privacy

- ▶ For 2020, the US Census Bureau adopted **differential privacy** (the “TopDown Algorithm”):
 - ▶ Noise is added to counts at the national level, then state, county, tract and block.
 - ▶ Results are post-processed so that they are consistent (non-negative integers that add up).
 - ▶ A total **privacy-loss budget** ϵ is chosen, and shared out between the different tables.
- ▶ Choosing ϵ is a **policy decision**, not a technical one. Lower ϵ means more privacy and more noise.
- ▶ Total state population counts were kept exact. Almost everything below that was perturbed.
- ▶ For large areas the noise is negligible. For small areas it can be large relative to the true count: some blocks were published with people but no housing, or children with no adults.

The controversy

- ▶ **Data users** objected that small-area statistics became unreliable:
 - ▶ Small towns, rural areas and Native American reservations faced the largest relative errors, with consequences for funding.
 - ▶ Alabama sued the Census Bureau in 2021, partly over differential privacy (the challenge was not successful).
 - ▶ Researchers argued the noise could distort redistricting for racially mixed areas².
- ▶ **Privacy advocates** replied that the old methods only *seemed* safe, because nobody had tested them; differential privacy makes the risk **visible and measurable**.
- ▶ The UK took a different path: the ONS 2021 Census for England and Wales used **targeted record swapping** and **cell key perturbation** (adding small, consistent noise to table cells), without a formal ϵ guarantee.

²Kenny et al. “The use of differential privacy for census data and its impact on redistricting.” Science Advances (2021).

The trade-off

- ▶ Privacy and accuracy cannot both be maximised. Every published statistic “spends” some privacy.
- ▶ The costs fall unevenly:
 - ▶ Large groups lose almost no accuracy; small and minority groups lose the most.
 - ▶ But small groups are also the **easiest to re-identify**, so they have the most to gain from protection.
- ▶ Several questions are hidden inside the choice:
 - ▶ Is a **provable** guarantee worth more than an informal one that is more accurate?
 - ▶ Should the agency publish fewer tables (and keep them accurate), or many tables with more noise?
 - ▶ Who should decide ϵ : statisticians, politicians, or the public?

Discussion

- ▶ Should the census use **formal differential privacy**, **traditional disclosure control**, or simply publish **fewer small-area tables**?
- ▶ Does compulsory participation make the duty of confidentiality stronger than for voluntary data?
- ▶ If the privacy risk is real but rarely exploited, is it worth reducing the accuracy of statistics that decide funding?
- ▶ How would **you** explain a choice of ϵ to the community whose statistics it affects?