Differential Privacy and the Overall Privacy of Decennial Data

Presenter: Michael Hawes – Providence, RI – 6/20/19 1:00 pm
Senior Advisor for Data Access and Privacy

PRESENTATION NOTES:

Disclaimer – This presentation is for a non-technical audience – simple, high-level conceptual understanding of differential privacy.

Why Differential Privacy? The Census Bureau is committed to Data Stewardship, it is required by Law (Title 13), keeping the Public's Trust, and the Quality on Census Statistics depends on our ability to keep the public's trust.

Privacy Protections Over time – History of Privacy Protections

Every disclosure avoidance method reduces the accuracy and usability of the data. Traditional methods for protecting privacy (suppression, coarsening, and perturbation) can have significant impacts on the usability of the resulting data products. Historically, data users have not been aware of the magnitude of the impacts of these techniques.

- Suppression – redacting sensitive values and redacting additional non-sensitive values to prevent recalculation of the sensitive values. Great precision, but loss of coverage.
- Coarsening – Introduces uncertainty by decreasing the precision of the reported data. Includes geographic aggregation, rounding, collapsing categories, reporting in ranges, etc. Better coverage – but lower precision
- Perturbation – Introducing noise or error to create uncertainty. One example is data swapping – taking pairs of records (ppl, households, etc) and swapping their values on sensitive attributes….it is used for a lot of federal statistics. Adding statistically unbiased noise is another example. Largely preserves usability of the data for statistical analyses.

Most disclosure avoidance methods rely on expert opinion rather than on quantitative measurement. But, if you cannot measure the risk, how can you measure whether you've take the appropriate measures to mitigate the risk?

Disclosure avoidance is increasingly important because there are more data to protect and more powerful computers/better algorithms for reconstructing databases and re-identifying individuals.

Compromised Data

Database re-construction – re-creation of individual-level data form tabular or aggregate data. If you release enough tables or statistics, eventually there will be a unique solution for what the underlying individual-level data work. Computer algorithms can do this easily.

Re-identification – taking individual records + External Database (e.g., credit records). Occurring with increasing frequency (e.g. Netflix prize, MA Health records, etc.).

There are 1.9 billion confidential data points and 7.7 billion statistics available from the 2010 Census. An internal Census Bureau Team conducted an experiment to reconstruct the confidential microdata from the publicly released tables, and to link the reconstructed individual-level data to a commercially available database. The staff was able to confirm re-identification for 52 million individuals from the 2010 Census (17% of the population) using only 'some' of the data products released publically by the Census Bureau. If this information got into the wrong hands, there could be potential harm to affected individuals (e.g. data miners could determine self-response for race and ethnicity and use that information in a malicious way).

The Census Bureau's Decision & the Future

The Census Bureau has committed to modernizing its approach to privacy protections.

This brings us to Differential Privacy – aka Formal Privacy – Quantifies the amount of re-identification risk for all calculations/tables/data products produced, no matter what external data are available now, or at any point in the future.

1) Measuring Sensitivity – How much would a calculation be affected by removing any particular individual or altering their responses? – Impacted by type of calculation, size of population (more pop/less sensitivity), and diversity (heterogeneity) of values.
2) Inject precise (targeted) amount of noise into the data (for every tabulation) based on the sensitivity of the calculation being performed – to mitigate the privacy risk.
3) Privacy vs Accuracy – Differential Privacy also allows policymakers to be very deliberate and specific to precisely calibrate where they should be on the privacy/accuracy tradeoff curve.
4) Privacy Budget – "Epsilon" – Where are you on the curve. The only way to absolutely eliminate all risk of re-identification would be to never release any usable data. What is the "acceptable" level of risk? Scale Epsilon = 0 = perfect privacy vs Epsilon = infinity = perfect accuracy.
5) Allocation of Privacy Budget. Each calculation, query, or tabulation of the data consumes a fraction of the privacy budget. Calculations/tables for which high accuracy is critical can receive a larger share of the overall privacy budget.
6) Accuracy – Impacted by the number of calculations performed or tables being generated, the type of calculation (count vs mean), the size of the underlying populations for each calculation or table, the uniformity/diversity of the population, the overall privacy budget (epsilon) and the allocation of the privacy budget across calculations/tables.
7) Accuracy – Differential privacy is not inherently better or worse than traditional methods…depends on the privacy budget.
8) Privacy – Differential Privacy is substantially better than traditional methods for protecting privacy, because it actually allows for measurement of the privacy risk.

What the Redistricting Community Can Do to Help

Senior Census Bureau policy makers will be making important decisions – and they need your input! What will the overall privacy budget be? What statistics will the Census Bureau release at which levels of geography? How will the overall privacy budget be allocated across different geographies, tables, and statistics?

Questions and Answers

1) Q: Who determines how much mitigation is necessary?
    a. A: This is determined by policymakers with input from data users – Including the redistricting community, and respondents (the public).
2) Q: Why is the definition of risk different for differential privacy? – Beveridge.
    a. A: In order to protect against future privacy threats (e.g., better computers/algorithms or more external data to use in a re-identification attack), Differential Privacy assumes a "worst case" scenario, and calculates the amount of noise needed to protect privacy against that.
3) Q: How was the 2010 Census Privacy Budget divvyed up?
    a. A: Michael – Differential privacy was not used in 2010, so this is really akin to comparing apples to oranges. In theory, you could measure the impact of the data swapping used for the 2010 census on the usability of the resulting data and find a corresponding point on the privacy/accuracy tradeoff curve, but the parameters for disclosure avoidance in the 2010 census are themselves confidential, so this isn't something that can be released publicly. One could, however, use the publicly available 1940 Census data, approximate the 2010 census disclosure avoidance by applying an arbitrary swap rate, and compare the resulting data to the output of the 2018 E2E Test code on the 1940 census data to get a sense of the relative impact of the two approaches on accuracy.
4) Q: What if a data user was interested in the number of Portuguese people in Kent County?
    a. Michael – All calculations/tabulations use a share of the overall budget. Census policymakers can allocate more/less of that budget to different tables or levels of geography, based on stakeholder feedback and use cases.
5) Q: What about margin of error
    a. Differential Privacy – you can be completely transparent. The precise noise/margin of error can be released, along with the epsilon privacy budget for all calculations. The Epsilon for the 2018 E2E Test was 0.25.
6) Q: What is the scale?
    a. LaPlace distribution – pointier at the top than a Normal/Gaussian distribution – width or narrowness – noise is determined by Epsilon – The higher the privacy budget, the narrower range of the LaPloss distribution. The info is in the source code. Because it is all calculable, it will likely be public, or can be easily calculated from the source code.
7) Q: Is the LaPloss distribution available for the E2E test?
    a. The privacy budget (epsilon) for the E2E test was 0.25. The noise distribution for each tabulation can be calculated by knowing the sensitivity of the calculation and the epsilon. This is all included in the 2018 E2E source code which the Census Bureau has released.
8) Q: The more questions we ask, the greater risk of privacy invasion?
    a. Generally speaking, this is the inherent problem with traditional disclosure avoidance methods. Releasing too many statistics, at too great a level of accuracy will violate privacy. With differential privacy, this is not the case anymore. However, under differential privacy, the more statistics released….the greater the impact on usability, for any fixed privacy budget (as each calculation uses up a share of that budget).
9) Q: What about the ACS?
    a. The Census Bureau has used a version of differential privacy since 2008 for "On the Map." We are looking expand the use of differential privacy to our other data products over the coming

years. The American Community Survey's confidentiality protections will be addressed with traditional methods until, or if, formal methods can be made to apply…We will have time to engage the user community thoroughly on the implications of privacy protections for the American Community Survey, and we intend to do so.

10) Q: PL 94-171 – Block level pop – aggregate to state-level, will I get the same number as the official state population number.
    a. The design is such that all child-level geographies will add up to their parent level; there will be additive consistency when data is aggregated up through the census geographic hierarchy.
11) Q: Deployed Military Overseas
    a. Deployed military are allocated to the block – used in apportionment and then removed. You will see deployed in the P.L. Redistricting Data.
12) Q: Summation for each field will be consistent? Are you going to publish data for these fields because this is not consistent for the redistricting community?
    a. Answer is the same as for Question 11 – the concept of additive consistency will be applied.
13) Q: Is noise inserted separately at each step of the geography?
    a. Yes, first, noise is inserted for all tabulations (except for those values that are being held invariant, e.g., state level population totals) at all levels of geography. However, the design of the system is such that all tabulations will aggregate consistently across geographic levels.
14) Q: In order to ensure privacy of individuals in one census block, noise is inserted. What if there is a cluster of homogenous blocks where pop are the same race? Does this mean there is potential for these blocks to not be homogenous when you add noise?
    a. This is possible, however, the top-down design of the disclosure avoidance system will often preserve homogeneity at lower levels of geography if it is reflected at higher levels.
15) How noisy will each level of geography be?
    a. That is determined by the privacy budget, but no decisions have been made/finalized regarding the privacy budget.
16) Question on accuracy? When you aggregate up to a higher-level of geography – Congressional District – SLD – doesn't the error propagate out? Shouldn't you have a very tight estimate at higher levels of geography – for congressional re-districting, you are going to be off by a few people with a margin of error not several percentage points.
    a. Since we are using a top down methodology for applying differential privacy, the larger the geography the higher the expected accuracy.
17) Question on accuracy? What about school districts – You can have pockets where synthetic white people are scattered in a majority minority community….thus causing issues with meeting the Voting Rights act of creating a 51% majority minority voting district.
    a. This depends on Privacy Budget, but the nature of the addition of noise is that some characteristics may be changed.
18) Redistricting Software – It is critical that the numbers summate correctly.
    a) Additive consistency will apply. The summation of lower levels of geography will add up to totals for larger geographies in which the lower geographies nest.