Broadband Usage Percentages Data Set: A Differentially Private Data Release

Mayana Pereira Joshua Allen Allen Kim Kevin White
Amit Misra Juan Lavista Ferres

1 Introduction

The Broadband Usage Percentages Data Set was developed as part of our efforts with Microsoft's Airband Initiative to help close the rural broadband gap. The initial data set released in April 2020 provided broadband usage percentages at a US county-level. In December 2020, we are adding a zip code-level view of this same information. The data can be used for the purpose of analyzing, understanding, improving, or addressing problems related to broadband access.

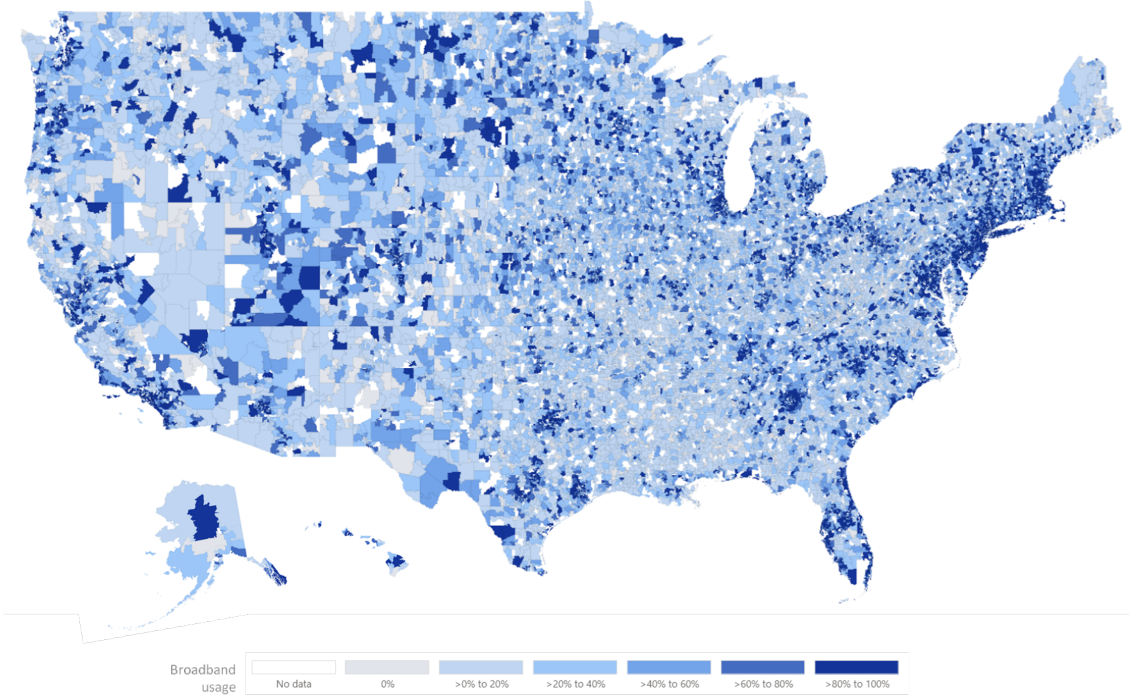

Figure 1: Map of the United States by postal codes with indicators of broadband usage.

2 Differentially Private Broadband Usage Percentages per Zip Code

The Broadband Usage Percentages Data Set is derived from aggregated and anonymized data Microsoft collects as part of our ongoing work to improve software and service performance and security. Given the zip code-level data set provides a more granular view of broadband usage percentages by households, we took the additional step to ensure data privacy guarantees by utilizing differential privacy. Differential privacy is a technique that adds noise to the data aggregations, preventing leakage about the presence of specific individuals in the data set.

We implemented differential privacy through the SmartNoise platform, a first-of-its-kind open-source platform for differential privacy co-developed by Microsoft and the OpenDP initiative led by Harvard. We estimate broadband usage by combining privatized data from multiple Microsoft services. The data privatization is done using differential privacy mechanisms on top of count queries, with a privacy parameter epsilon = 0.2.

3 Error Ranges Estimation

As differential privacy adds noise to protect privacy, the noise added to zip codes with a small number of households can impact utility. To ensure transparency into how zip codes with different population magnitudes are affected, we have included error range data.

For this data release, we empirically calculate the expected error range caused by differential privacy. Although we can estimate differential privacy effects in query counts, the impact of post-processing combining several differentially private data sources is better estimated through a simulation process.

The simulation process, illustrated in figure 2, generates several data privatization and broadband coverage calculation simulated processes. For all simulations, we compute the delta between the simulated coverage and the privatized coverage. We estimate the expected error range for zip codes based on their number of households. The error reported in our data set is the error range empirically calculated based on zip code households [5].

We report two distinct error metrics for each specific zip code: the mean absolute error and the 95^{th} percentile error. The non-private broadband coverage estimate will be, on average, within the mean absolute error (MAE) error range. Additionally, we provide the 95^{th} percentile error range. For 95% of the time, the non-private broadband coverage estimate for zip codes with a similar number of households will be within 95^{th} percentile error range.

We also provide the mean signed deviation (MSD). The mean signed deviation offers an estimate of bias introduced by the process.

The intuition behind the privacy guarantees provided by differential privacy lies in the fact that the inclusion or exclusion of any individual in a data set should not affect the results. Following this intuition, to achieve differential privacy, some randomness should be added to an analysis's results. Additionally, an analysis that is differentially private is probabilistic in nature. Executing a differentially private analysis several times on the same data can result in different answers. This is because such analyses introduce some uncertainty into the computation in the form of random noise. The results of a differentially private data analysis are not exact but an approximation.

It is worth noting that the broadband coverage estimation has other sources of error, such as sampling error and bias.

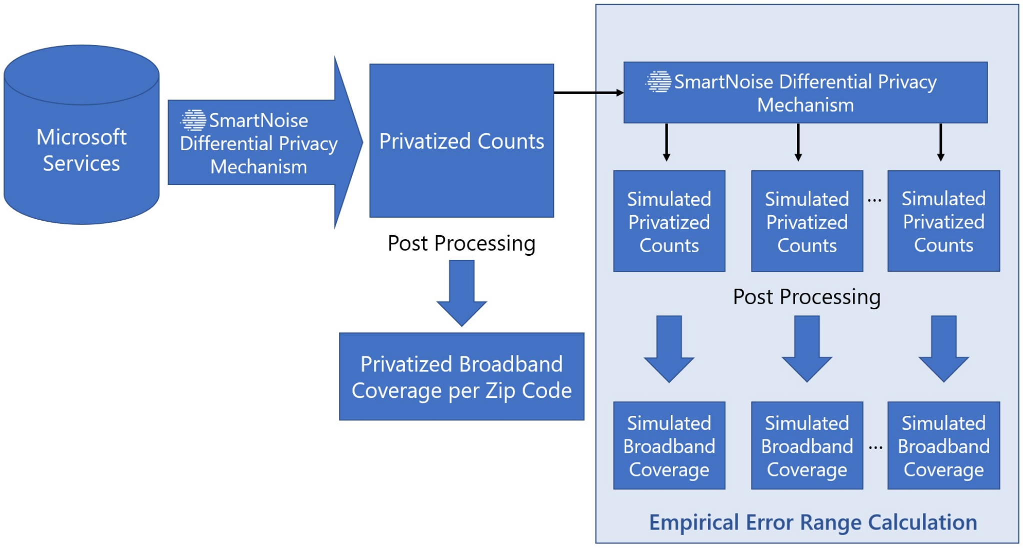
Microsoft Services
SmartNoise Differential Privacy Mechanism
Privatized Counts
Post Processing
Privatized Broadband Coverage per Zip Code
SmartNoise Differential Privacy Mechanism
Simulated Privatized Counts
Simulated Privatized Counts
...
Simulated Privatized Counts
Post Processing
Simulated Broadband Coverage
Simulated Broadband Coverage
...
Simulated Broadband Coverage
Empirical Error Range Calculation

Figure 2: Process for estimating the error introduced by differential privacy. The error ranges are estimated by generating several simulated private releases, and reporting the error introduced differential privacy.

4 Understanding Differential Privacy and SmartNoise

Differential Privacy is a standard for formalizing privacy-preserving data analysis. Privacy-preserving data analysis is defined as an analysis in which the output does not reveal information about any specific individual. We refer the reader to *Differential Privacy: A Primer for a Non-technical Audience* [6] for a more detailed explanation of differential privacy and an intuitive explanation of how an individual's privacy is preserved by applying differential privacy mechanisms in data analysis.

The SmartNoise library is an open source verified secure differentially private data curator application. It includes several differential privacy mechanisms in python and provides an easy-to-use and flexible framework for data set privatization. SmartNoise is powered by OpenDP, a community of collaborators in academia, industry, and government focused on developing open-source software tools for privacy-protective statistical analysis of sensitive personal data.

The SmartNoise GitHub contains several tutorials and examples, including how to release private statistics and how to compute accuracy guarantees of a differentially private data release with SmartNoise.

We encourage you to learn more, evaluate SmartNoise and OpenDP, and join the community.

SmartNoise:

https://smartnoise.org

https://github.com/opendifferentialprivacy/smartnoise-core-python

https://github.com/opendifferentialprivacy/smartnoise-sdk

OpenDP:

https://projects.iq.harvard.edu/opendp

https://github.com/opendifferentialprivacy

SmartNoise Gitter:

https://gitter.im/opendifferentialprivacy/SmartNoise

References

[1] Census. American fact finder. https://data.census.gov/.

[2] Federal Communications Commission. 2018 broadband deployment report — federal communications commission. https://www.fcc.gov/reports-research/reports/broadband-progress-reports/2018-broadband-deployment-report.

[3] OpenDP. Accuracy: Pitfalls and edge cases. https://github.com/opendifferentialprivacy/smartnoise-samples/blob/master/analysis/accuracy_pitfalls.ipynb.

[4] OpenDP. Privacy-preserving statistical release. https://github.com/opendifferentialprivacy/smartnoise-samples/blob/master/analysis/tutorial_mental_health_in_tech_survey.ipynb.

[5] U.S. Department of Housing and Urban Development. Office of policy development research. https://www.huduser.gov/portal/datasets/usps_crosswalk.html.

[6] Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R O'Brien, Thomas Steinke, and Salil Vadhan. Differential privacy: A primer for a non-technical audience. *Vand. J. Ent. & Tech. L.*, 21:209, 2018.