ISSUE BRIEF U.S. Policy

PETs Use Case: Measuring Digital Literacy with Telemetry Data Using Differential Privacy

APRIL 2026

FUTURE OF PRIVACY FORUM

ORGANIZATIONS

Microsoft

EDITORS

Andrew Gruen, FPF Senior Fellow

Shea Swauger, FPF

Laura Amortegui, FPF

ACKNOWLEDGEMENTS

The Research Coordination Network (RCN) for Privacy-Preserving Data Sharing and Analytics is supported by the U.S. National Science Foundation under Award #2413978 and the U.S. Department of Energy, Office of Science under Award #DE-SC0024884.

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

Table of Contents

Problem

Digital literacy, loosely defined as competency with a range of digital technologies, is a key skill for full social and democratic participation, but it is a difficult thing to measure across a large population[1]. Having accurate data about how people use technology can inform how educators, policymakers, and technology companies make decisions and are especially important for addressing the digital divide. To help address this gap, Microsoft used telemetry signals from a segment of Windows machines owned by U.S. based consumers to indirectly measure digital literacy. Building upon previous research using differential privacy to estimate broadband internet access, a research team comprised of Microsoft employees and three academic researchers designed two indirect measures (called indexes) for digital literacy, collected telemetry data, scored each machine, and aggregated scores by zip code. The first index, called "Media and Information Composite Index" (MCI) measures the usage of common desktop applications such as web browsers, Microsoft Office tools, or email. The second index, called "Content Creation and Computation Composite Index" (CCI) measures the usage of more advanced applications like Adobe Creative Suite or developer tools. The telemetry data only measures usage per machine, not per user, meaning it cannot differentiate between people if there are multiple users on one machine. The research team wrote a manuscript describing their project, which is undergoing internal compliance approval, an internal privacy review, and a peer-reviewed paper.

[1] Institute of Electrical and Electronics Engineers. (2020). IEEE Standard for Digital Intelligence (DQ)—Framework for Digital Literacy, Skills, and Readiness (IEEE 3527.1-2020). IEEE Standards Association. https://standards.ieee.org/ieee/3527.1/7589/

Proposed PET Details:

Technology Type:

The core Privacy Enhancing Technology (PET) in this use case is Differential Privacy.

What is Differential Privacy?

Differential Privacy is a controlled process of adding noise to data to protect privacy. Several factors inform how much noise is added to the data and how much analysis is allowed. These factors are combined into what is called a "privacy-loss budget" which is represented numerically. Differential Privacy may be applied at query time or to an entire dataset prior to sharing.

For Query-based Differential privacy, a value is calculator to determine how much privacy loss is created by viewing a single variable, which is then deducted from an overall "budget" of privacy loss deemed acceptable. Imagine a privacy budget to analyze it spends a portion of budget, in this example, 0.2 per variable and 0.5 per query. With a privacy budget of 4, an end user could see 5 variables (0.2 x 5 = 1) and make 6 queries (0.5 x 6 = 3) on the data. When Privacy budget is fully spent, any new variable shown or query made increases the risk of re-identification beyond the limit set at the beginning of the analysis. An advantage of this kind of Differential Privacy is that the individual researcher may choose which parts of a dataset are more important to them - and thus spends as much of their budget in those places to maximize the accuracy.

However, Differential Privacy may be applied to an entire dataset prior to release. In this case, the entire budget is spent to create a static dataset. Noise is added to individual records; the greater the noise, the lower the total privacy-loss budget. An advantage of this type of Differential Privacy is that it enables a one-time calculation of privacy loss and continual re-use of the same underlying data.

Dataset Development:

From October 2022 to March 2023, the research team collected non-personal data from anonymized Windows devices that share diagnostic data with Microsoft. This "telemetry data" includes logs of Windows devices' interactions with various applications and was collected during a machine's operating system update.

Functionality

The research team used OpenDP's Python library to conduct all data analysis. The total privacy-loss of the data release is $\epsilon = 4.1$. The privacy loss computation applied the parallel and sequential composition properties of differential privacy mechanisms. The privacy-loss results from the generate the indices results from weight calculations (PCAbased) and device index aggregations at ZIP code level.

The privacy-loss resulting from the principal component analysis (PCA) is of $\epsilon = 0.5$ for each index. The total privacy loss incurred from PCA analysis is of $\epsilon = 1.0$. The weights calculation were done over a baseline month (March 2023). The aggregation process for the index resulted in a privacy-loss of $\epsilon = 3.0$. Each histogram was privatized using a Laplace mechanism, with privacy-loss of $\epsilon = 0.05$. All Laplace mechanism and PCA implementations utilized in this project were developed by the OpenDP library.

Beneficiaries: Researchers and Policymakers

- Researchers who study digital literacy, technology access or adoption, secondary social or economic indicators, geographic distribution of technology use, and related academic fields.
- Policymakers interested in increasing technology equity or access.
- The data published alongside the manuscript has open licensing for secondary research.

Industry/Domain:

- Technology
- Education

Stakeholders:

Microsoft:

While Microsoft doesn't intend to develop commercial applications from this research, the company benefits in two key ways from creating datasets like these. First, like any large enterprise, the company benefits from finding socially useful ways to deploy resources that would otherwise be "wasted." Because telemetry data is a non-rivalrous good, finding ways to deploy it for business *and* social benefit produces goodwill. Second, by creating privacy-preserving datasets for social benefit, the company is also "training" itself in the complex process of producing privacy-preserving datasets that maintain significant utility. As the skills required to do so become more broadly available across the enterprise, it is more likely that privacy-preserved, high-utility datasets can be deployed in ever more places. This, in the long term, could enable a norm of working on and with privacy-protected data instead of individual-level, identifiable data.

Researchers:

The individual academic research partners benefited from their participation in this work through the opportunity to generate peer-reviewed publications—the currency of career progression—and early access to and understanding of the dataset they helped to generate. Generally, academic researchers also have an interest in the work both for the methodology and the specific dataset. Methodologically, it suggests that significant quantities of business process data could be "unlocked" and made available for more research. Specifically, *this* dataset is of interest to a broad swath of social scientists and policy scholars who seek to improve the lives, economic opportunities, and education of people across the U.S.

Background and Motivation

At the highest level, this use case for differentially private flat data file release is about balancing between privacy and utility. The use case does so in a particularly politically salient space: the use of business

process data – data collected for a business purpose – for a social scientific purpose. In this specific instance, Microsoft sought to utilize telemetry it receives in the form of consumer device-level data, while maintaining a reasonable standard of privacy. While the dataset itself is subject to meaningful limitations (including, for example, that software application usage is an imperfect measure of digital literacy) the key issue here is the broadly applicable use case of repurposing business process data for other, more socially beneficial research activity.

Current Regulatory Environment

While there aren't any laws or regulations that specifically address Differential Privacy used in this way in the United States, there are laws such as the European Union's General Data Protection Requirements (GDPR), Singapore's Personal Data Protection Act (PDPA), or Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) that encourage using PETs to comply with privacy obligations. Despite that, there are company policies that regulate how Differential Privacy is conducted internally with a high value of ensuring customer data is secure and private.

The data collected was restricted to using IP addresses for US-based consumer machines only. While there may be edge cases of non-US machines being included if machines used a VPN, this number of these is estimated to be relatively small and statistically insignificant. Therefore, the research is intended to be viewed from a US regulatory perspective.

Current Adoption Status

Differential Privacy is used across industry, academic research, civil society, and governments around the world and is often an attractive PET because it can offer *technical* privacy guarantees regarding the likelihood of reidentification. can be considered one of the more developed PETs. It has robust research communities and high investments in its continued improvement.

Benefits of PET Implementation

Privacy Benefits

By applying differential privacy to the data, the research team could release their underlying dataset with extremely low risk that any individual in it would be identifiable. Differential Privacy allows the research team to add statistical noise, measured by epsilon (ε), to the dataset they want to make public. This allows them to set their preferred level of privacy *and* utility in the data. In addition, because the team produced a flat file, as opposed to a query engine, there is an opportunity to make the privacy guarantees composable—adding additional time series without impacting the overall privacy of individuals within the dataset.

Operational Benefits

Potential for Broader Use

Sharing what epsilon they chose and how they arrived at it can help other research teams decide on their own epsilon.

Societal Impact

The more Differential Privacy is used, including how epsilon is set, can help lead to better informed decision-making in future PETs implementation.

Risk and Ethical Analysis

Ethical Considerations

There was no informed consent process for participants in the research. While there is arguably consent given through the software update agreement or telemetry data-sharing agreement, this falls below what a traditional human subjects research protocol would involve.

Trade-offs

A smaller epsilon provides greater privacy by adding more noise, but decreases the data's analytical utility. A larger epsilon provides less privacy by adding less noise but increases the data's analytical utility. For comparison, The US Census Bureau set their global privacy budget at 6, 4 for population tables and 2 for household tables.[2]

Mitigation Strategies

Microsoft used the same Differential Privacy approach for this project as they did for estimating broadband internet access. In that project, they created multiple fake datasets and tested different epsilon levels to determine the appropriate number. They also had significant involvement from their legal team to ensure regulatory compliance at every stage. One difference between these projects is that the Broadband dataset's granularity was at the county level and the digital literacy dataset is at the zipcode level, which increases the utility but potentially decreases privacy.

Privacy Risks

1. Zip codes can be correlated with potentially sensitive data like education level or income, which can then be tied to the digital literacy scores. While no specific individual's data would be compromised, there is a small risk of this kind of secondary correlation.

[2] U.S. Census Bureau. (2019, October 31). *Design parameters and global privacy-loss budget* (2020 Census Program Memorandum Series No. 2019.25). U.S. Department of Commerce. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/memo-series/2020-memo-2019_25.html

2. While index scores are aggregated by zipcodes and the minimum cell size is 50 households, there are some cases where there are fewer than 50 households in a zip code.

Operational Risks

If the dataset is ever reidentified, Microsoft risks reputational damage that can negatively impact future revenue.

Known Regulatory Challenges and Barriers

Impact of Regulatory Barriers

The lack of laws or standards leaves determining a privacy budget (epsilon) up to the person or team's judgment with a given dataset. While some projects might appropriately protect sensitive data, others may set epsilon too high, risking privacy violations, or set epsilon too low, unnecessarily reducing the data's value for analysis and secondary use.

Specific Legal Questions to be Addressed (by Regulator Network)

Policy/legal issues

- What policy or legal concerns do you have about this use case?

Technical issues

- What technical questions or concerns do you have about this use case?

Does this use case comply with laws in my jurisdiction?

- If not, which ones and why?
- What would need to be done to bring it into compliance?

What in this use case are the most important or relevant features?

- Was there a central problem, concern, technology, or process that influenced your feedback the most?

Was there any additional information you needed but didn't have?

Additional Information and Resources

- The New Digital Divide, NBER Working Paper No. 32932
- U.S. Broadband Coverage Data Set: A Differentially Private Data Release

Washington, DC | Brussels | Singapore

FPF.org