User interface to support differentially private daily page views

Liudas Panavas
Northeastern University

Cody Dunne
Northeastern University

Abstract

The Wikimedia Foundation's release of differentially private pageview statistics offers enhanced privacy but presents a challenge: effectively communicating the inherent uncertainty in the data is a bottleneck in making this information accessible and useful. Our project aims to develop an interactive visualization tool to address this, enhancing data usability and clarifying differential privacy. Central to our work is the question: How can we best visualize and convey this uncertainty, facilitating understanding and informed decision-making among journalists, researchers, and the community?

Introduction

Wikimedia has embarked on a significant initiative, releasing detailed pageview data broken down by country and language with enhanced privacy through differential privacy. This method intelligently adds uncertainty to data, safeguarding individual privacy [1]. Currently, this rich dataset is constrained to TSV files, making effective use and interpretation challenging.

Addressing this, our project seeks to bridge the gap between data availability and user accessibility. One major barrier to this is how to represent the noise (uncertainty) inherent in differentially private data. Historical instances of similar private data releases, like the 2020 U.S. Census, have left organizations with decreased trust rather than increased trust [2]. To prevent this from occurring at WMF while still releasing the data in an accessible format, we need effective visualization strategies for communicating the uncertainty in the data [3].

Therefore, our investigation revolves around the primary question: How can we best display WMF's differentially private daily pageview data while communicating uncertainty effectively? This decomposes into two specific inquiries:

1. What visualization prototype most effectively represents the newly generated pageview data?
2. How can we best communicate the differentially private uncertainty of the data, especially when displayed cartographically, given its differential privacy nature?

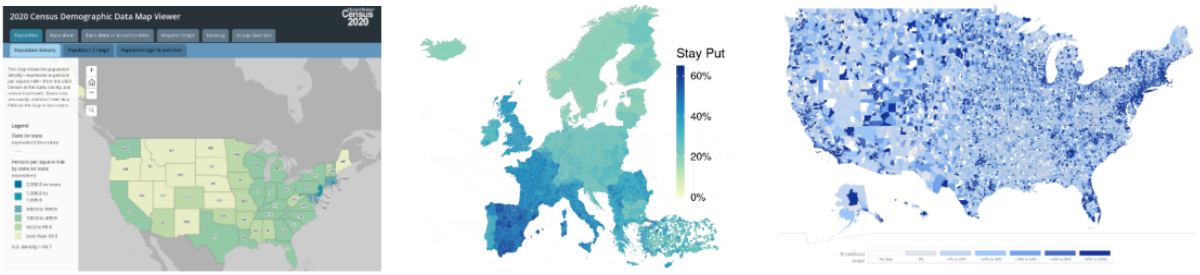

Figure 1: Major differential privacy data releases. From left to right: 2020 U.S. Census, Facebook Covid Mobility Reports, Microsoft Broadband Coverage.

By accurately displaying uncertainty, we enable a clear **identification of impactful topics**, ensuring decisions are based on solid statistical foundations. Our rigorously tested visualization method will **improve the user experience** by making the complex concept of statistical uncertainty both comprehensible and user-friendly. Furthermore, our initiative enhances **equity in decision-making**. By creating a user-centric visualization of the data, we extend its accessibility to those lacking technical expertise to analyze raw TSV files, thereby equipping a broader audience with the knowledge needed for meaningful participation in decision processes. Our findings aim to set practical guidelines for differential privacy implementations and contribute to the advancement of the field of uncertainty visualization.

Related work

Many of the current uses of differential privacy are related to releasing geographic data [4, 5] but there are no conclusive answers on how to show uncertainty in geographic data [6] (Figure 1). This research can help provide quantitative evidence to future differential privacy deployments.

The existing pageview data, valuable to Wikimedia developers and researchers, assists in assessing Wikimedia campaigns [7] and gauging public interest in conservation [8]. It's accessible via the page.view tool (figure 2). However, the new differentially private dataset at WMF, currently in static TSV format, lacks easy exploration. Developing an interactive interface for this dataset can enhance research on Wikimedia data.

Methods

First, we'll collect design inputs for our visualization prototype by discussing with the Wikimedia community and conducting user interviews, focusing on their preferences for visualizing the private pageview datasets.

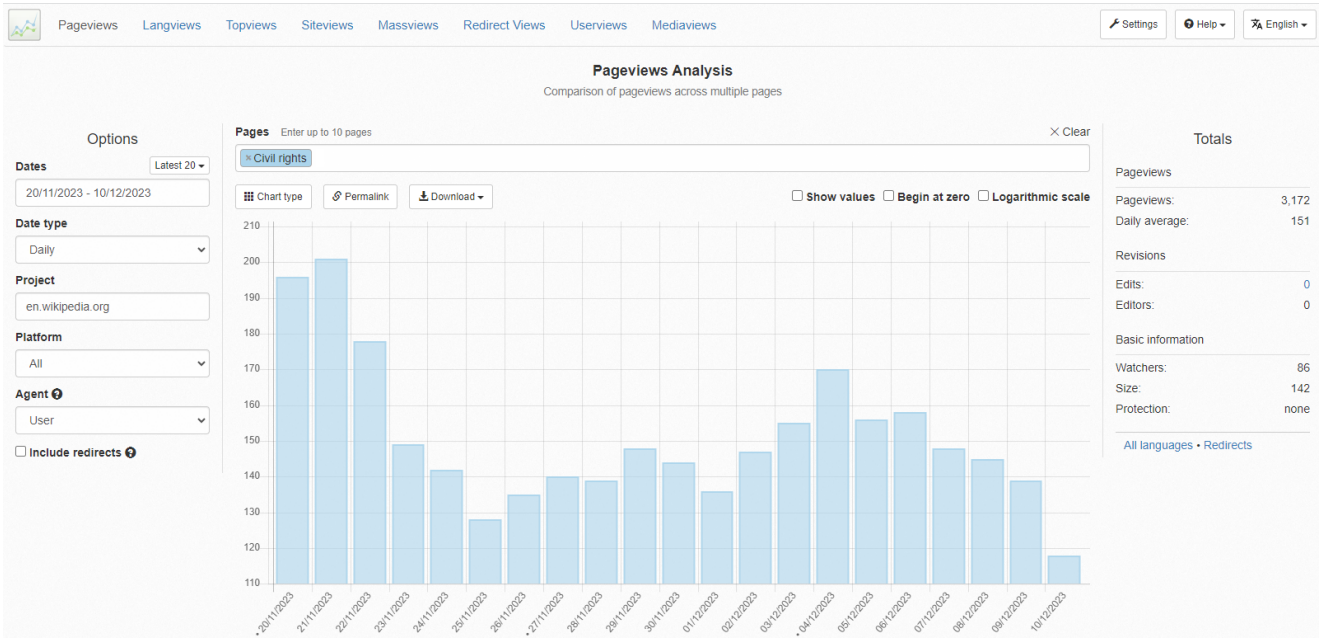

Figure 2: Existing Wikimedia pageview dashboard.

Secondly, the bulk of our project will research three distinct visualization approaches to represent uncertainty, drawing from the existing literature. User feedback will be solicited through surveys to determine the most effective and preferred visualization strategy. Throughout this process we will ensure we test on a variety of devices (mobile, desktop) and users (technical editors, newcomers) to ensure that our solution is universally accessible

Expected output

1. A prototype interactive visualization tool, tailored for integration with Wikimedia's pageview dashboards, aimed at improving understanding and exploration of user data for contributors and researchers.
2. An evaluated visualization method for representing differential privacy noise in geographic data, designed to address interface challenges and provide actionable guidance for similar data releases. Results to be shared at the IEEE VIS conference.

Risks

The project's primary challenge lies in the limited guidance available on differential privacy communication. The lack of definitive answers in both academic and non-academic literature poses a significant hurdle. Another challenge will be determining the most effective ways to evaluate our proposed solutions. With a variety of potential methods for assessing uncertainty visualization and differential privacy communication, selecting the most pertinent questions and criteria for the success of our visualizations is crucial.

Community impact plan

The existing differentially-private pageview pipeline releases ~40x more data than top-per-country, the non-DP baseline. This covers a significantly larger proportion of lower-traffic countries, and in the future (after the new Country Protection List policy is implemented) will include dozens of countries previously deemed too risky to release. However, WMF's inability to clearly communicate uncertainty in DP data holds the community back from taking advantage of this vast expansion.

This research will impact the community by removing the uncertainty visualization bottleneck and enable adoption of DP data into public understandings of WMF traffic patterns. To that end, we will be closely advised by Hal Triedman, the creator of the dataset, to ensure we correctly understand the privacy policies in place. We also plan to discuss with WMF engineers (Leon Ziemba and Dan Andreescu) on how to best integrate our findings into the pageviews dashboard.

Evaluation

Success for this project will be defined by the development of a set of best practices for communicating uncertainty in differential privacy for geographic data. A key measure of success will be the extent to which users of the tool engage with the new pageview datasets.

Budget

The project would fund one PhD student for 4 months at a cost of $24,219:

- Stipend: $16,667
- Benefits: $1,275
- Tuition: $ 3,586
- Institutional overhead of 15%: $2,691

We will post any publications as free preprints on OSF.

Prior contributions

Liudas Panavas specializes in differential privacy and visualization, designing private data interfaces and interviewing over 25 experts in the field.

Cody Dunne, with 16 years in visualization and HCI, will effectively oversee user studies and lead the publication process.

References

1. Hal Triedman, Isaac Johnson and Nuria Ruiz. "New Dataset Uncovers Wikipedia Browsing Habits While Protecting Users." Diff, 22 June 2023, diff.wikimedia.org/2023/06/21/new-dataset-uncovers-wikipedia-browsing-habits-while-protecting-users/.
2. Boyd, Danah, and Jayshree Sarathy. "Differential perspectives: Epistemic disconnects surrounding the US Census Bureau's use of differential privacy." *Harvard Data Science Review (Forthcoming)* (2022).
3. Hullman, Jessica, et al. "In pursuit of error: A survey of uncertainty visualization evaluation." *IEEE transactions on visualization and computer graphics* 25.1 (2018): 903-913.

4. Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. "Issues encountered deploying differential privacy." Proceedings of the 2018 Workshop on Privacy in the Electronic Society. 2018.
5. Herdağdelen, A., et al. "Protecting privacy in Facebook mobility data during the COVID-19 response." Facebook Research (2020).
6. Kinkeldey, Christoph, Alan M. MacEachren, and Jochen Schiewe. "How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies." The Cartographic Journal 51.4 (2014): 372-386.
7. Chelsy Xie, Xiaoxi, Isaac Johnson, and Anne Gomez. "Detecting and gauging impact on Wikipedia page views." Companion Proceedings of The 2019 World Wide Web Conference. 2019.
8. Guedes-Santos, Jhonatan, et al. "Evaluating public interest in protected areas using Wikipedia page views." Journal for Nature Conservation 63 (2021): 126040.