THE OHIO STATE UNIVERSITY

Location Privacy in the Geographically Aggregated Data Protected by Differential Privacy

A Case of the United States Census

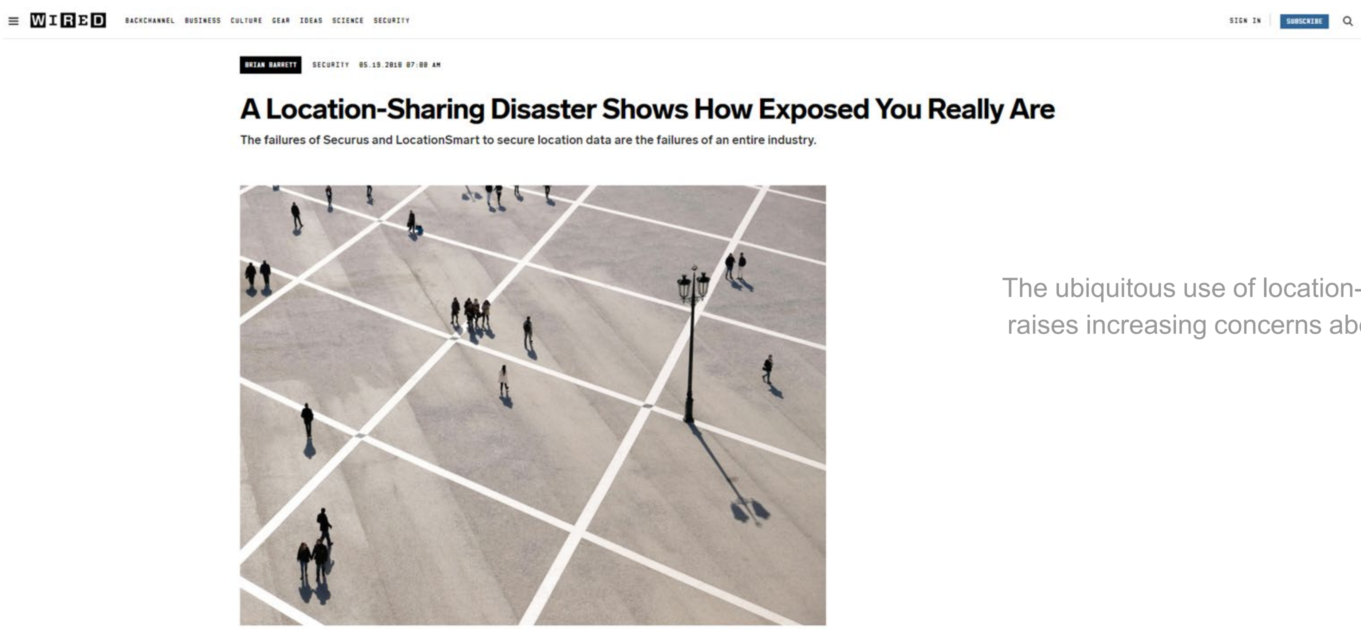
Yue Lin*, Ningchuan Xiao

Department of Geography | The Ohio State University

lin.3326@osu.edu

Location Privacy and Geographic Identity

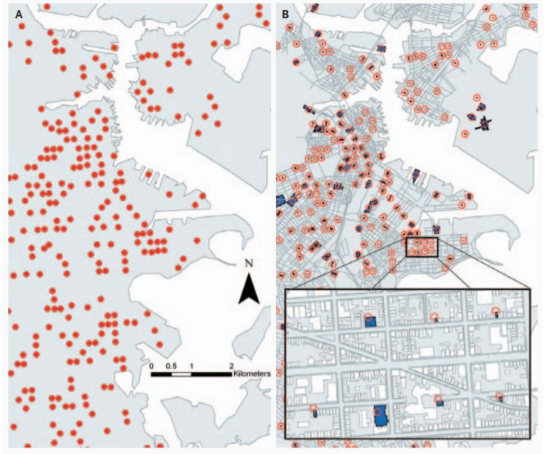
- **Location privacy** is the right of an individual to be free from unauthorized collection, disclosure, and use of his/her **personally identifiable location**
- The location identifiable to an individual, either alone or with other information, is referred to as a **geographic identity**
- Related research and acts:
    - Geoprivacy (Kwan et al., 2014; Kounadi & Leitner, 2014; Richardson et al., 2015)
    - Personal identifiable information (McCallister, 2010; Voigt & Von dem Bussche, 2017)



The ubiquitous use of location-based technologies raises increasing concerns about location privacy

Geographic Identity Disclosure

- Often occurs in **fine-resolution individual data**
  - Through means such as simple tabulation (though rare) or dot mapping individual locations
  - Disclosing **point-based** location information such as residential addresses or geographic coordinates
- **Aggregation** has been considered as a safe measure to privacy protection
  - If only publishing aggregated data (e.g., population count) by geographic area, will we disclose geographic identities unexpectedly?
    - Yes! A single individual may be identified in an area where the combination of some attributes is unique



Unauthorized disclosure of patients' geographic identities (addresses) through reverse geocoding (Brownstein et al., 2006)

Block	White Alone
390490001101001	1
390490001101015	79
390490001103000	49
390490001104004	12

Block-level census table on race (2010 Census Summary File 1)

- One can be uniquely identified by their sex and block
- The block itself is a geographic identity that needs to be protected

Statistical Attacks on Location Privacy

- **Outlier attacks:** Identify individuals who contribute to unusual or outlying information in the **aggregated data directly**
    - For census tables, occur for cells with population uniques (count of one)
    - Risks of geographic identity disclosure affected by **types of query** and **aggregation levels**

Block	White Alone
390490001101001	1
390490001101015	79
390490001103000	49
390490001104004	12

Block-level census table on race (2010 Census Summary File 1)

- **Reconstruction attacks:** Identify individuals by **recovering individual data** of the entire population (not only the outlying ones)
    - For census data, this means to recover both areal locations and demographic attributes of the entire population from a combination of census tables

Block-level census tables on race and ethnicity (2010 Census Summary File 1)

Block	White Alone	Black or African American Alone
390490001101001	1	0
390490001101015	79	1
390490001103000	49	0
390490001104004	12	0

Block	Non-HispanicWhite	Non-Hispanic Black or African American
390490001101001	1	0
390490001101015	68	1
390490001103000	44	0
390490001104004	10	0

Person	Block	Race	Ethnicity
1	390490001101001	White	Non-Hispanic
2	390490001101015	Black	Non-Hispanic
3	390490001101015	White	Non-Hispanic
4	390490001101015	White	Non-Hispanic
5	390490001101015	White	Non-Hispanic
6	390490001101015	White	Non-Hispanic
7	390490001101015	White	Non-Hispanic
...	...	...	...

- ❑ Recovered individual data
- ❑ Can be linked to external databases for identification

Differential Privacy (DP)

- An emerging mechanism to safeguard aggregated data (including geographically aggregated data)
    - A recent use of this mechanism is in the **2020 United States Census**
- How differential privacy protects privacy in general?
    - Apply statistical noise during data production (Dwork & Roth, 2014)
        - Control trade-off between privacy and data utility using a parameter called **privacy loss budget (PLB)**
    - Resistant to **reconstruction attacks**
        - Individual records cannot be recovered using multiple aggregated data

Block	Non-HispanicWhite	Non-Hispanic Black or African American
390490001101001	1	0
390490001101015	68	1
390490001103000	44	0
390490001104004	10	0

2010 Census Summary File 1 (Original)

Block	Non-HispanicWhite	Non-Hispanic Black or African American
390490001101001	1	0
390490001101015	66	0
390490001103000	40	2
390490001104004	12	1

2010 Census Summary File 1 (Differentially private; from IPUMS NHGIS Privacy-Protected Demonstration Data vintage 2021-06-08)

Does Differential Privacy Guarantee Location Privacy?

- Avoidance of reconstruction itself is **not a guarantee** of location privacy
  - Consider an algorithm that alters all individual data except the population uniques in tables
    - Low reconstruction rate and yet high risk under outlier attacks
- The privacy definition taken by differential privacy **differs** from the concept of location privacy
  - Differential privacy focuses on the **indistinguishability** of whether an individual’s data is used
  - Location privacy emphasizes **location-based identifiability**
- More research is still needed to understand the effectiveness of differential privacy for protecting location privacy in geographically aggregated data

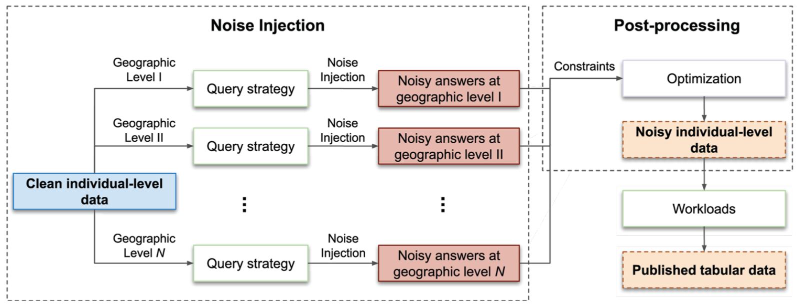
Research Objectives

- **Goal:** To investigate whether and how differential privacy protects location privacy in geographically aggregated data, **with a focus on census data**

- **Research questions:**
  - How to quantify risks of geographic identity disclosure under outlier attacks?
  - Is the differentially private mechanism effective at mitigating outlier attacks? What effect do different PLB (privacy loss budget) values have on the effectiveness of this mechanism?
  - Can PLB fully determine the risks? Are the risks consistent across different query types, aggregation levels, and geographical areas?

Data Preparation

- **U.S. Census Bureau's differentially private (DP) algorithm**
  - Noise injection and post-processing



- **Data:** Simulated individual-level population data
  - Based on the 2010 United States Census Summary File 1 (SF1)
  - Use linear programming to determine the individual data that minimize the difference between its summarized information and corresponding aggregated data from census tables

Assessing Disclosure Risks under Outlier Attacks

General idea:

- In an outlier attack, geographic identity disclosure occurs when **a published population unique is an actual unique (true unique)**

Block	Non-HispanicWhite	Non-Hispanic Black or African American
390490001101001	1	0
390490001101015	68	1
390490001103000	44	0
390490001104004	10	0

Actual table

2010 Census Summary File 1 (Original)

True uniques

Actual uniques

Block	Non-HispanicWhite	Non-Hispanic Black or African American
390490001101001	1	0
390490001101015	66	0
390490001103000	40	2
390490001104004	12	1

Published table

2010 Census Summary File 1 (Differentially private; from IPUMS NHGIS Privacy-Protected Demonstration Data vintage 2021-06-08)

Published uniques

Measures: PPV and TPR

- Positive predictive value (PPV): probability of finding a true unique among the published uniques
- True positive rate (TPR): probability of an actual unique being published

Actual table

Block	Non-HispanicWhite	Non-Hispanic Black or African American
390490001101001	**1**	0
390490001101015	68	1
390490001103000	44	0
390490001104004	10	0

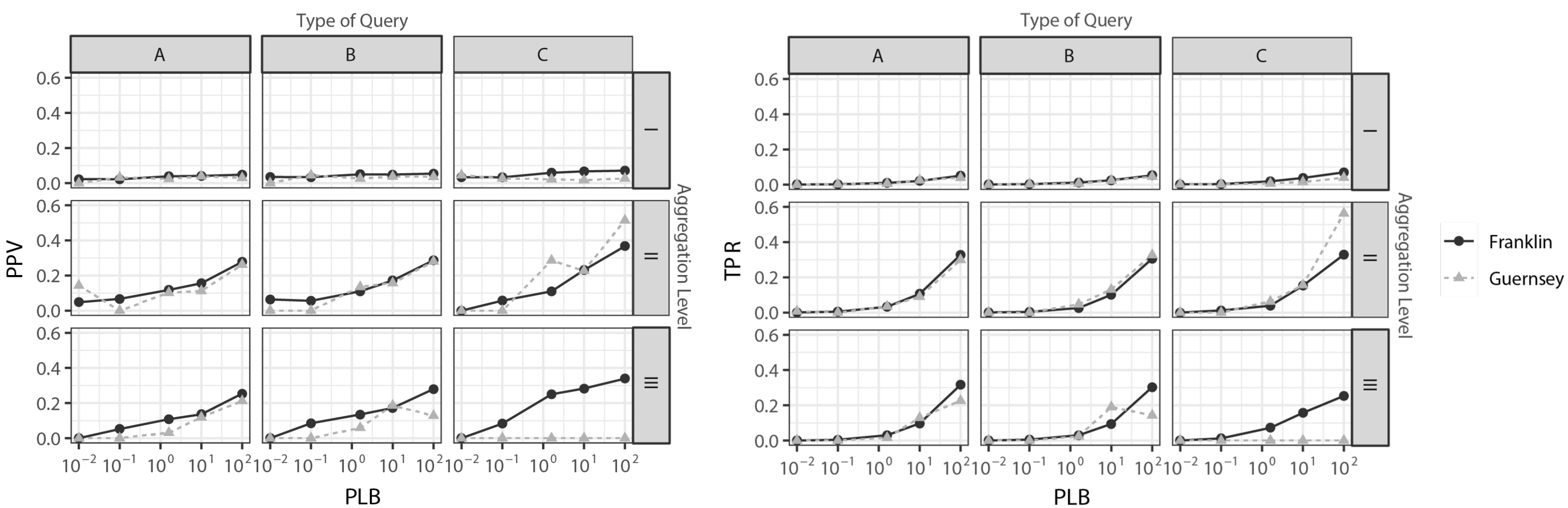
Published table

Block	Non-HispanicWhite	Non-Hispanic Black or African American
390490001101001	**1**	0
390490001101015	66	0
390490001103000	40	2
390490001104004	12	1

- A small value of PPV and TPR indicates a strong protection

Effectiveness of DP in Protecting Location Privacy

- **Findings:**
  - DP is generally effective to reduce both PPV and TPR when a small value of PLB (less than 1) is applied
  - PLB itself cannot determine the risks of geographic identity disclosure; effectiveness differs among tables and across geographic areas under outlier attacks
  - Effectiveness of DP is subject to substantial variability for geographic areas with small population sizes



Type of query:
- A: population count by housing type by voting age by ethnicity by race
- B: population count by voting age by race
- C: population count by race

Aggregation level:
- I: block
- II: block group
- III: tract

Summary

- **Examined the effectiveness of differential privacy for protecting location privacy in census data**
  - How to quantify risks of geographic identity disclosure under outlier attacks?
    - Developed measures of PPV and TPR to quantify the risks
  - Is the differentially private mechanism effective at mitigating outlier attacks? What effect do different PLB (privacy loss budget) values have on the effectiveness of this mechanism?
    - DP is generally effective when PLB is small (but not in all the cases)
  - Can PLB fully determine the risks? Are the risks consistent across different query types, aggregation levels, and geographical areas?
    - PLB cannot fully determine the risks. It is possible to have unexpectedly high risks with small PLB for areas with unusual demographic compositions and small population sizes

- **Ongoing and future work**
  - The accuracy side of differentially private census data
  - Protecting location privacy without much compromise of accuracy under differential privacy

References

- Brownstein, J. S., Cassa, C. A., & Mandl, K. D. (2006). No place to hide—reverse identification of patients from published maps. *New England Journal of Medicine*, 355(16), 1741-1742.
- Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4), 211-407.
- Kounadi, O., & Leitner, M. (2014). Why does geoprivacy matter? The scientific publication of confidential data presented on maps. *Journal of Empirical Research on Human Research Ethics*, 9(4), 34-45.
- Kwan, M. P., Casas, I., & Schmitz, B. (2004). Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks?. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39(2), 15-28.
- McCallister, E. (2010). *Guide to Protecting the Confidentiality of Personally Identifiable Information* (Vol. 800, No. 122). Diane Publishing.
- Richardson, D. B., Kwan, M. P., Alter, G., & McKendry, J. E. (2015). Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research. *Annals of GIS*, 21(2), 101-110.
- Voigt, P., & Von dem Bussche, A. (2017). The EU general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, *10*, 3152676.