Published in final edited form as:
Proc IEEE Int Conf Big Data. 2020 December ; 2020: 3119–3122. doi:10.1109/bigdata50022.2020.9378390.

Leveraging Differential Privacy in Geospatial Analyses of Standardized Healthcare Data

Daniel R. Harris
Center for Clinical and Translational Sciences, Institute for Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Kentucky, Lexington, KY USA

Abstract

We present a collection of geodatabase functions which expedite utilizing differential privacy for privacy-aware geospatial analysis of healthcare data. The healthcare domain has a long history of standardization and research communities have developed open-source common data models to support the larger goals of interoperability, reproducibility, and data sharing; these models also standardize geospatial patient data. However, patient privacy laws and institutional regulations complicate geospatial analyses and dissemination of research findings due to protective restrictions in how data and results are shared. This results in infrastructures with great abilities to organize and store healthcare data, yet which lack the innate ability to produce shareable results that preserve privacy and conform to regulatory requirements. Differential privacy is a model for performing privacy-preserving analytics. We detail our process and findings in inserting an open-source library for differential privacy into a workflow for leveraging a geodatabase for geocoding and analyzing geospatial data stored as part of the Observational Medical Outcomes Partnership (OMOP) common data model. We pilot this process using an open big data repository of addresses.

Index Terms—

geographic information systems; data privacy; big data applications

I. Introduction

Geospatial analyses involving population health and healthcare data play an important role in public health and environmental health research, which includes supporting epidemiological pursuits of analyzing, visualizing, and tracking population health conditions and understanding environmental factors related to health [1]–[3]. The volume and diversity of healthcare data being generated and stored in clinical data warehouses is rapidly increasing and creates practical challenges in secondary use for research purposes [4]. Data standardization plays a pivotal role in current strategies for dealing with the volume and variety of healthcare data [4]. Common data models such as the Observational Medical Outcomes Partnership (OMOP) common data model and the United States Patient Centered

daniel.harris@uky.edu .

Outcomes Research Network (PCORNet) common data model were designed to support interoperability, expedite research, and enable easier data sharing [5]–[7].

Privacy is a key consideration when working with data in the healthcare domain; regulations protecting patient confidentially span from institutional requirements to government-mandated legal requirements. In the US, the Health Insurance Portability and Accountability Act (HIPAA) outlines privacy expectations and identifies data elements which include geospatial patient information. HIPAA outlines the minimum guidelines for protecting patient privacy and in practice, additional efforts outside of routine censoring should be done to truly protect patient privacy and to avoid unintended information leakage [8]–[10]. Specifically, releasing exact patient counts and other aggregations may be unsafe in certain circumstances and may accidentally reveal private health data [10]–[12]. Techniques exist for computationally controlling privacy in the geospatial domain [10] but require significant effort to integrate with healthcare standards.

Differential privacy is a conceptual model for privacy-preserving data analysis that attempts to minimize the analytic impact of adding or removing a single record so that in theory there is little risk to patients when their information is included into the data set [13], [14]. The formal definition is given in Definition 1 where the probability is determined by coin tosses of $\mathscr{K}$ and ϵ.

Definition 1: A randomized function $\mathscr{K}$ gives ϵ-differential privacy if for all data sets $\mathscr{D}_1$ and $\mathscr{D}_2$ differing on at most one element, and all $\mathscr{S} \subseteq Range(\mathscr{K})$,

$$Pr[\mathscr{K}(\mathscr{D}_1) \in \mathscr{S}] \leq exp(\epsilon) \times Pr[\mathscr{K}(\mathscr{D}_2) \in \mathscr{S}]$$

A review of differential privacy in the healthcare domain gave a primary example of counting patient cohorts for feasibility estimates in clinical trials [15]; other work includes supporting ranged queries [16]. Our work differs in that we focus on privacy concerns during the geospatial analyses of standardized healthcare data.

We will demonstrate that differential privacy can be inserted into workflows for geospatial analyses of standardized healthcare data. In particular, we leverage differential privacy to produce anonymized aggregated data from patient data stored in the OMOP common data model. It is worth noting that Definition 1 does not guarantee complete privacy but rather minimizes risk by limiting the significance of any particular record so that any analytical output would not substantially change. Risk within geospatial analyses can be subjective depending on the context of the data and the aggregated analytical data being produced; we also provide functions to benchmark the trade off between privacy and noise.

II. Methods

We depend on several open-source libraries and in turn, we make our demonstration code available as open-source [17]. In particular, we depend upon a common data model popular in healthcare informatics, a geodatabase we leverage for geocoding and analysis of patient data, and an open-source implementation of differential privacy. Figure 1 gives a high-level overview of the different components of which we will discuss in turn.

A. Common Data Models in Healthcare

The Observational Medical Outcomes Partnership (OMOP) common data model was designed to facilitate universal analysis of individual healthcare databases, where data from their original systems could be extracted, transformed, and loaded into a common representation [6]. With this common representation, the promise of a common analytical library and open-source community of contributors is enabled. The model itself is divided into three large parts: standardized clinical data tables, tables needed to support standardized vocabularies, and other ancillary tables. The key healthcare entities are patients, visits, and conditions. The patient table is linked to the location table which includes fine grain address details that are captured by most electronic health record systems (address lines, city, state, zip code, country) and some derived fields (latitude and longitude) which can be externally computed if not supplied.

B. Geospatial Workflow

Our electronic health record system did not supply geospatial coordinates; to maintain compliance with institutional policy, we calculated these using geocoding functions from an “in-house” HIPAA-compliant geodatabase, PostGIS [18]. PostGIS is an open-source extension adding geospatial objects and analytic functionality to the PostgreSQL database [19].

As part of our solution, we package a PostgreSQL procedure *omop_geocode*() designed to update OMOP’s location tables with calculated geospatial coordinates in the event they are missing. Specifically, this utilizes the TIGER geocoder of PostGIS extension which uses reference data from the US Census Bureau [18].

Additionally, we provide a PostgreSQL procedure *omop_census_link* that maps locations to census geographies (blocks, block-groups, and tracts). This is achieved by doing a spatial join to census block data on a polygon point intersection using the patient’s address. Neighborhood-level groupings are incredibly important for public health research; they also have complicated privacy concerns. If a patient’s location becomes another facet of their medical history, researchers and care providers may also utilize knowledge of their geographical neighborhood, including social-determinants of health [20]. Because both medical conditions and social determinants of health are potentially sensitive, the reportable overlap between the two must be treated in a way as which to protect patient privacy, motivating the inclusion of differential privacy in our workflow.

C. Differential Privacy

Our workflow uses Google’s open-source library for differential privacy [21]. We deployed this to the same Post-greSQL/PostGIS server housing the OMOP geospatial data and compiled the PostgreSQL extension; this extension supports anonymized aggregate functions parameterized by ϵ to control the noise-to-privacy ratio. Smaller values of ϵ result in more noise and thus more privacy; the ideal value for ϵ is context sensitive according to the geospatial query being developed. A less common condition that carries social stigma, such as HIV, requires more neighborhood-level privacy than a very common condition, such as high blood pressure, that carries very little social stigma. We include this benchmarking of noise-to-privacy ratio as a database function in our demonstration code; this is helpful for evaluating ϵ in different contexts. We discuss our findings on tuning differential privacy in greater detail in the next section. Figure 1 shows how our solution is organized: Postgresql hosts OMOP formatted data originally from clinical source systems; PostgreSQL extensions for PostGIS and differential privacy help analyze and interact with the data. To summarize, we expedite the following workflow: geocoding patient data stored in the OMOP common data model, mapping patients to different geographic boundaries, generating derived data using differential privacy, and benchmarking how the choice of *epsilon* impacts results.

III. Discussion

We piloted our procedures on a publicly available and "open" big data set before deploying our code on data from our clinical data warehouse. This helped us prototype functions to help tune ϵ and test our workflows on a data set that is not truly sensitive or regulated by privacy policies and laws.

A. Open Big Data

OpenAddresses is an online repository containing over half a billion geocoded addresses, which are annotated with longitude and latitude coordinates [22]. OpenAddresses has been used to benchmark geocoding solutions when privacy concerns prevent external validation of results due to data sharing limitations [23]. Approximately 200 million of these addresses are from the US, which is divided into 4 regions; we selected a random subset of one million addresses from our region to test our procedures. We populated a minimum set of OMOP tables by assigning a random address from OpenAddresses to a simulated patient record. We deployed our procedure to geocode the simulated patient addresses in order to prepare the data for privacy-preserving querying.

Table I shows a sample run performing an anonymized count aggregated by patient state. The ideal ϵ is sensitive to the context of the data being analyzed. Our random sample from OpenAddresses was not evenly distributed by state. It is plausible that scenarios exist where state-level data is not particularly sensitive due to the state's large geographic coverage; it is also plausible that scenarios exist where state-level data is incredibly sensitive, such as reporting frequency of rare medical conditions. This heavy bias toward one location is not unlike hospital data, where the majority of patients live in the same neighboring service area.

B. Tuning Privacy

Tuning differential privacy for the correct trade off between privacy and noise is not an exact science [24]. Because context is important with geospatial healthcare data, we developed a procedure to help understand the impact of selecting different values for ϵ with three parameters for shifting ϵ *dp_benchmark*(*start, stop, step*); this will run differential privacy on a data set multiple times by using the *step* value to gradually increase ϵ from *start* to *stop*. We present the results of benchmarking with *dp_benchmark* in Figure 2. As expected, low values of ϵ produce noisy data sets that are great for preserving privacy and conversely higher values of ϵ produce data sets that accurately reflect the original values. Because the ideal ϵ is context-sensitive for geospatial analysis with healthcare data, shifts in totals as large as 200 may be acceptable in some situations and inappropriate in others.

This function should help the analyst understand the impact of ϵ in their application. As future work, we wish to include functions providing statistical guidance on what threshold the slope of the projected data is significantly different and suggests noisy decay.

C. Census-level Reporting

The examples above selected state from the patient's location upon which to aggregate. There are far fewer unique states than other smaller geographies, such as US Census-designated blocks, block-groups, and census tracts. The US Census Bureau distributes neighborhood-level demographic data with hundreds of levels of minutia such as age, race, poverty status, and so on; the US Census Bureau also distributes frequencies of how these demographics overlap, such as the number of people in this geography reported below the poverty level in the past 12 months who are of a certain age and race group. This vast amount of demographic data also carries major privacy concerns; the US Census Bureau has deployed differential privacy as a technique for preserving citizen privacy without harming data utility [25]

We extend our example outlined in Figure 2 by drilling down into a specific state (Florida) and producing anonymized aggregates for census tracts. There are too many census tracts to present as a table; we display how benchmarking this level of geography performs in Figure 3. It is important to note two major differences: there are far more census tracts than states, and census tracts are intentionally designed to not exceed more than approximately 4,000 people. This means our random sample from OpenAddresses has an approximate ceiling which is reflected in our population totals. With respect to reporting aggregated healthcare data, the intersection of census tract data with disease data will greatly reduce the frequency within these bins and may even be subject to regional biases.

IV. Conclusion and Future Work

We presented our work on integrating differential privacy into geospatial analyses involving standardized healthcare data. In particular, we show that patient location data stored in the OMOP common data model can be queried using privacy preserving aggregate functions. We further demonstrated the need for differential privacy benchmarking to find a suitable value for ϵ due to the context-sensitive nature of geospatial healthcare applications. Smaller geographic boundaries may be leveraged securely and we provide functions for mapping addresses to census-level geographies including blocks, block-groups, and tracts. As future work, we are formalizing our various workflows with the intent that anyone wishing to do privacy-preserving geospatial analysis of healthcare data may benefit from our work. The overhead cost would be greatly reduced for those already working with location data in the OMOP common data model format.

Acknowledgment

The project described was supported by the NIH National Center for Advancing Translational Sciences through grant number UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1]. Scholten HJ and De Lepper M, "The benefits of the application of geographical information systems in public and environmental health." World Health Statistics quarterly. Rapport Trimestriel de Statistiques Sanitaires Mondiales, vol. 44, no. 3, pp. 160–170, 1991. [PubMed: 1949884]

[2]. Kistemann T, Dangendorf F, and Schweikart J, "New perspectives on the use of geographical information systems (gis) in environmental health sciences," International journal of hygiene and environmental health, vol. 205, no. 3, pp. 169–181, 2002. [PubMed: 12040915]

[3]. Nykiforuk CI and Flaman LM, "Geographic information systems (gis) for health promotion and public health: a review," Health promotion practice, vol. 12, no. 1, pp. 63–73, 2011. [PubMed: 19546198]

[4]. Olaronke I and Oluwaseun O, "Big data in healthcare: Prospects, challenges and resolutions," in 2016 Future Technologies Conference (FTC). IEEE, 2016, pp. 1152–1157.

[5]. Rosenbloom S, Carroll R, Warner J, Matheny M, and Denny J, "Representing knowledge consistently across health systems," Yearbook of medical informatics, vol. 26, no. 1, p. 139, 2017. [PubMed: 29063555]

[6]. Overhage JM, Ryan PB, Reich CG, Hartzema AG, and Stang PE, "Validation of a common data model for active safety surveillance research," Journal of the American Medical Informatics Association, vol. 19, no. 1, pp. 54–60, 2011. [PubMed: 22037893]

[7]. Qualls LG, Phillips TA, Hammill BG, Topping J, Louzao DM, Brown JS, Curtis LH, and Marsolo K, "Evaluating foundational data quality in the national patient-centered clinical research network (pcornet®)," eGEMs, vol. 6, no. 1, 2018.

[8]. Blatt AJ, "Geospatial data mining and knowledge discovery," in Health, Science, and Place. Springer, 2015, pp. 77–87.

[9]. Benitez K and Malin B, "Evaluating re-identification risks with respect to the hipaa privacy rule," Journal of the American Medical Informatics Association, vol. 17, no. 2, pp. 169–177, 2010. [PubMed: 20190059]

[10]. Zandbergen PA, "Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data," Advances in medicine, vol. 2014, 2014.

[11]. Curtis A, Mills JW, Agustin L, and Cockburn M, "Confidentiality risks in fine scale aggregations of health data," Computers, Environment and Urban Systems, vol. 35, no. 1, pp. 57–64, 2011.

[12]. Kounadi O and Leitner M, "Why does geoprivacy matter? the scientific publication of confidential data presented on maps," Journal of Empirical Research on Human Research Ethics, vol. 9, no. 4, pp. 34–45, 2014. [PubMed: 25747295]

[13]. Dwork C, "Differential privacy: A survey of results," in International conference on theory and applications of models of computation. Springer, 2008, pp. 1–19.

[14]. Dwork C, McSherry F, Nissim K, and Smith A, "Calibrating noise to sensitivity in private data analysis," in Theory of cryptography conference. Springer, 2006, pp. 265–284.

[15]. Dankar FK and El Emam K, "Practicing differential privacy in health care: A review." Trans. Data Priv, vol. 6, no. 1, pp. 35–67, 2013.

[16]. Alnemari A, Romanowski CJ, and Raj RK, "An adaptive differential privacy algorithm for range queries over healthcare data," in 2017 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2017, pp. 397–402.

[17]. Harris/dp-omop: using differential privacy with omop. Accessed Oct. 1, 2020. [Online]. Available: https://bitbucket.org/_harris/dp-OMOP

[18]. Postgis: Spatial and geographic objects for postgresql. Accessed Oct. 1, 2020. [Online]. Available: https://postgis.net

[19]. Postgresql: the world's most advanced open-source database. Accessed Oct. 1, 2020. [Online]. Available: https://www.postgresql.org/

[20]. Sunderland N, Bristed H, Gudes O, Boddy J, and Da Silva M, "What does it feel like to live here? exploring sensory ethnography as a collaborative methodology for investigating social determinants of health in place," Health & Place, vol. 18, no. 5, pp. 1056–1067, 2012. [PubMed: 22722015]

[21]. google/differential-privacy: Google's differential privacy library. Accessed Oct. 1, 2020. [Online]. Available: https://github.com/google/differential-privacy

[22]. Openaddresses: the free and open global address collection. Accessed Oct. 1, 2020. [Online]. Available: http://openaddresses.io

[23]. Harris DR and Delcher C, "bench4gis: Benchmarking privacy-aware geocoding with open big data," in 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019, pp. 4067–4070.

[24]. Lee J and Clifton C, "How much is enough? choosing ε for differential privacy," in International Conference on Information Security. Springer, 2011, pp. 325–340.

[25]. Garfinkel SL, Abowd JM, and Powazek S, "Issues encountered deploying differential privacy," in Proceedings of the 2018 Workshop on Privacy in the Electronic Society, 2018, pp. 133–137.

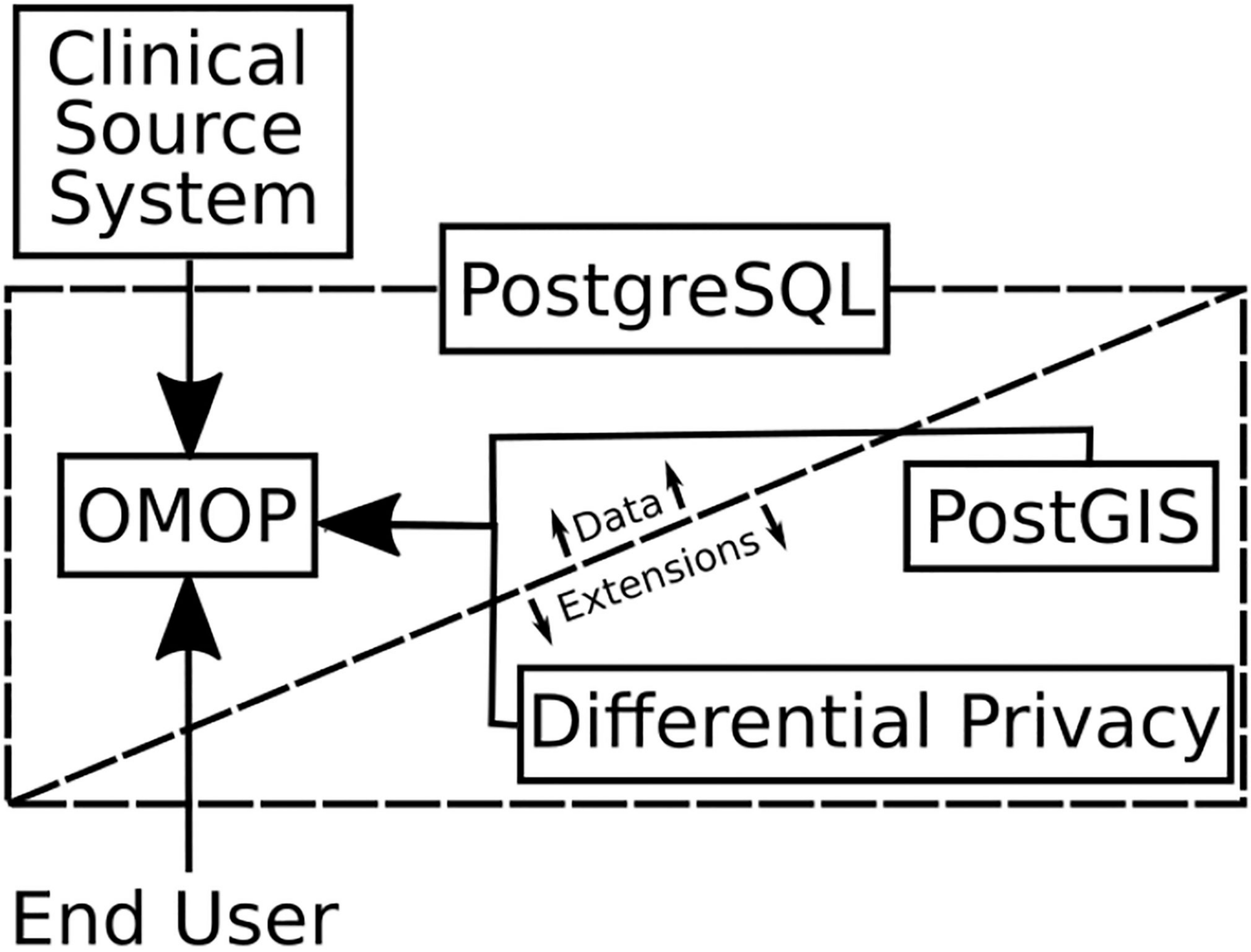

Fig. 1.

End users interact with an OMOP data source derived from a clinical source system and augmented with PostgreSQL extensions.

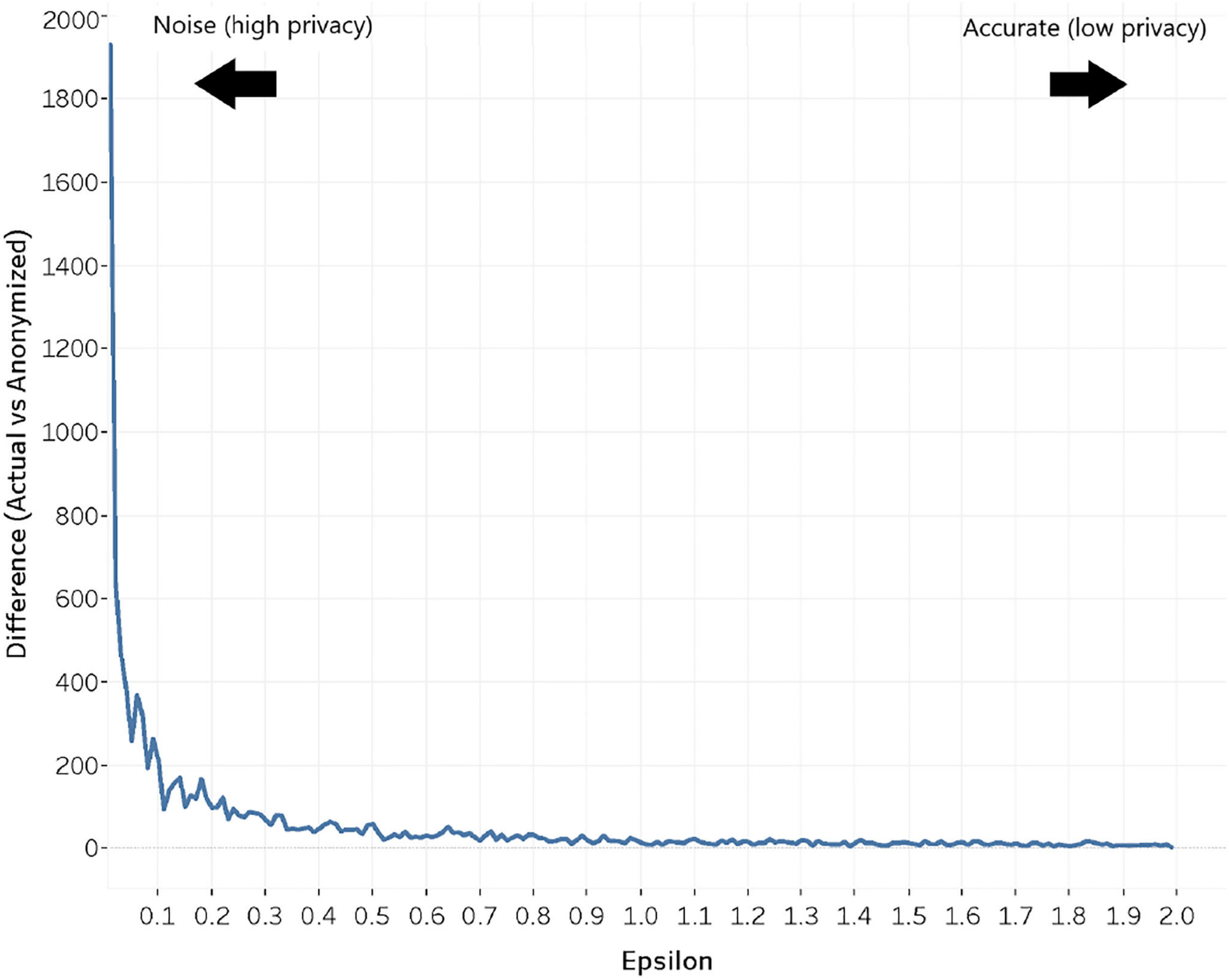

Fig. 2.

A plot of *dp_benchmark*(*start* = 0.1, *stop* = 2.0, *step* = 0.1) showing the quick decay into noise for low values of ϵ.

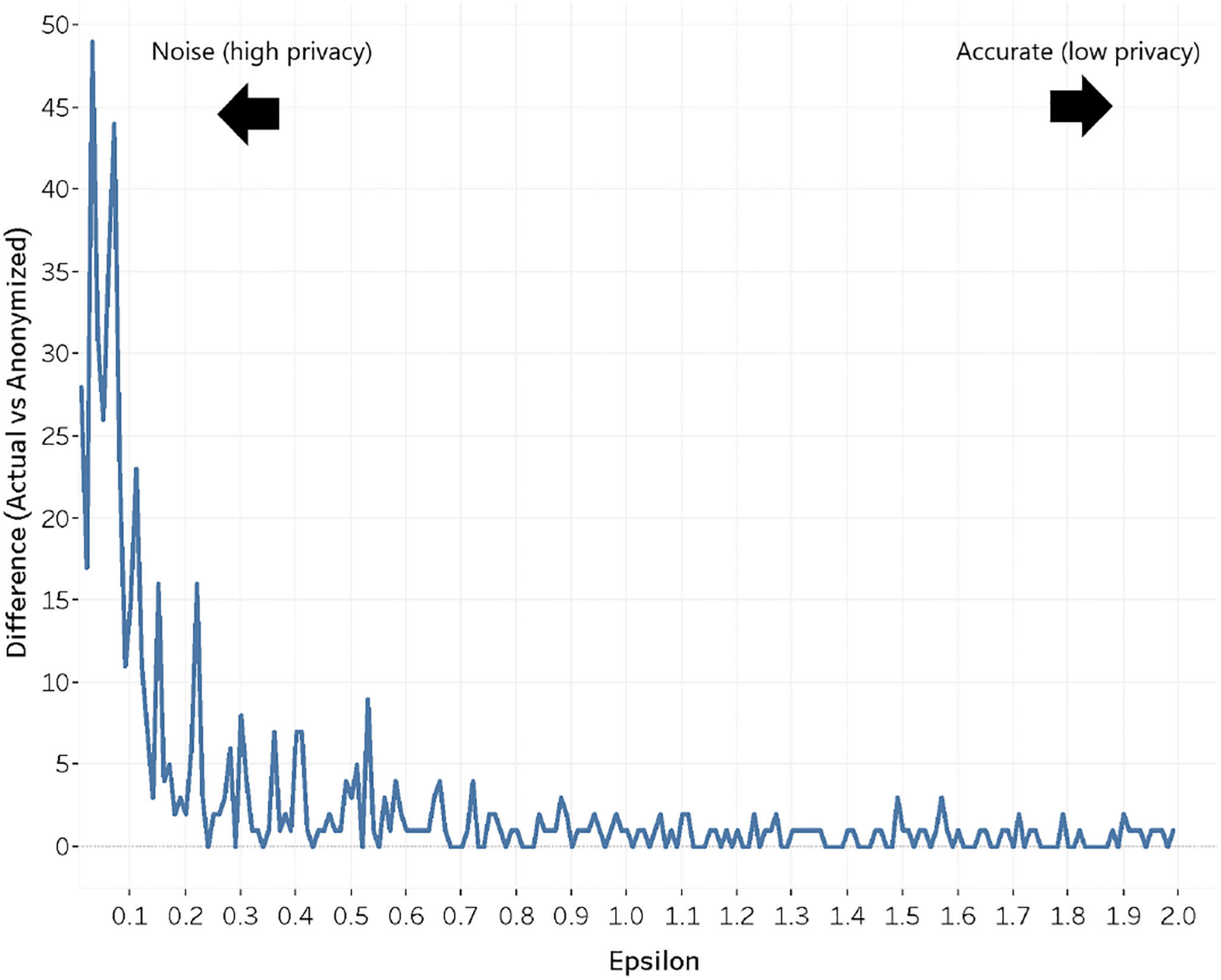

Fig. 3.

A plot of *dp_benchmark*(*start* = 0.1*, stop* = 2.0*, step* = 0.1) with census tract data shows slightly different behavior than state-level data due to distribution differences.

TABLE I

Anonymized counts aggregated at the state level

State	ϵ	Actual Count	Anon. Count	Diff.
MA	*ln*(3)/2	1	0	−1
OK	*ln*(3)/2	1	1	0
MI	*ln*(3)/2	2	0	−2
DE	*ln*(3)/2	3	2	−1
NY	*ln*(3)/2	3	3	0
OH	*ln*(3)/2	4	4	0
IL	*ln*(3)/2	5	3	−2
CA	*ln*(3)/2	12	20	8
PA	*ln*(3)/2	13	9	−4
MO	*ln*(3)/2	15	12	−3
AR	*ln*(3)/2	87	87	0
AL	*ln*(3)/2	826	822	−4
LA	*ln*(3)/2	7,521	7,512	−9
TN	*ln*(3)/2	1,1355	11,358	3
MD	*ln*(3)/2	11,610	11,611	1
SC	*ln*(3)/2	35,469	35,475	6
GA	*ln*(3)/2	40,866	40,865	−1
NC	*ln*(3)/2	57,898	57897	−1
WV	*ln*(3)/2	60,743	60,743	0
FL	*ln*(3)/2	135,147	135,146	−1
TX	*ln*(3)/2	247,239	247,230	−9
VA	*ln*(3)/2	391,180	391,182	2