Multi-Analyst Differential Privacy with Fine-Grained Provenance for Databases

Xi He
University of Waterloo
xi.he@uwaterloo.ca

Motivation. Differential Privacy (DP) [8] has emerged as a promising standard for safeguarding the privacy of data contributors when their information is used in data-driven applications and research. Answering queries using DP has a *bounded* information disclosure of the data contributors to the data analyst. This information disclosure is quantified by a privacy parameter in DP, also known as the *privacy budget*. In practice, internal data analysts or applications with a higher privilege level for critical tasks like security alerts might be granted a bigger privacy budget and, hence, more extensive access to sensitive data, than an external application such as third-party advertisements.

Challenges. The current design of existing DP systems [2, 10, 11, 13, 15] does not distinguish different data analysts but regards them as a unified entity. This approach gives rise to potential problems. First, a low-privilege data analyst making queries earlier could deplete more of the privacy budget than an internal one if the system does not interfere with the sequence of queries. Second, if naïvely tracking and answering each analyst's queries independently, the system can waste the total privacy budget when two data analysts ask similar queries. These challenges stem primarily from the inherent "statelessness" of existing systems, meaning no one records the individual budget expenses per analyst and their historical queries. Specifically, the lack of metadata detailing the query's origin, computation method, and frequency of the results, which is related to the *provenance information* in database research [3, 4], creates a notable gap. Without the provenance of the metadata, answering queries in a multi-analyst use case can be unfair or wasteful in budget allocation.

Our Approach. To tackle these challenges, our research group at the University of Waterloo introduces `DProvDB` [19], a "stateful" DP query processing system. This system incorporates a novel privacy provenance framework designed for multi-analyst scenarios when data analysts are obliged not to share their answers, unlike prior work [16]. In this framework, we meticulously trace historical queries with their responses and privacy consumption on a per-analyst and per-view basis. This approach allows novel use of correlated Gaussian noise, ensuring more queries can be accurately answered while achieving two privacy objectives: (i) fair budget allocation according to the analysts' privilege levels and (ii) bounding the total privacy loss by the system-wise privacy budget even if all the analysts collude.

Open Questions. The granularity of provenance information is a crucial consideration in `DProvDB`. Rather than capturing the direct responses to all queries, `DProvDB` employs private synopses — materialized results for histogram views — to respond to incoming queries. In a recent work [14], we demonstrate that a more fine-grained tracking of historical information can further enhance the design of DP algorithms, albeit with an associated increase in performance costs. It is important to note that in extreme cases, adopting a fine-grained provenance at the tuple level may pose a privacy risk [6, 7]. Furthermore, stateful systems for multi-analyst DP, like `DProvDB`, have issues maintaining the states and difficulty providing fault-tolerance upon system crashes. The optimal balance between the system performance, query utility, and robust privacy protection with provenance remains an open question.

`DProvDB` demonstrates how incorporating provenance information about *the methods and results* can enhance DP systems. Beyond this, the provenance of *the input data source* can also be valuable. Often, privacy analyzers focus on a single DP algorithm [1]. For multiple DP algorithms, without tracing the access to the input data source, the system may overlook repeated access by different algorithms to the same data or data from the same contributors, leading to unexpected privacy loss (e.g., Apple's initial DP attempt [18]). Conversely, if two DP algorithms independently operate on data from different contributors, adding their privacy loss naively as the final privacy loss would be overkill [9]. This issue becomes more critical for dynamic databases where contributors join or opt to leave [5, 17], or when various privacy resolutions [13] or personalized preferences [12] come into play. How to leverage provenance information for input data to enhance privacy analyzer is another important open question in the field.

References

[1] Chiké Abuah, David Darais, and Joseph P. Near. Solo: A lightweight static analysis for differential privacy. *Proc. ACM Program. Lang.*, 6(OOPSLA2), oct 2022.

[2] Kareem Amin, Jennifer Gillenwater, Matthew Joseph, Alex Kulesza, and Sergei Vassilvitskii. Plume: Differential privacy at scale. *CoRR*, abs/2201.11603, 2022.

[3] Peter Buneman and Wang Chiew Tan. Provenance in databases. In Chee Yong Chan, Beng Chin Ooi, and Aoying Zhou, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, 2007*, pages 1171–1173. ACM, 2007.

[4] James Cheney, Laura Chiticariu, and Wang Chiew Tan. Provenance in databases: Why, how, and where. *Found. Trends Databases*, 1(4):379–474, 2009.

[5] Rachel Cummings, Sara Krehbiel, Kevin A. Lai, and Uthaipon Tantipongpipat. Differential privacy for growing databases. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8878–8887, Red Hook, NY, USA, 2018. Curran Associates Inc.

[6] Susan B. Davidson, Sanjeev Khanna, Sudeepa Roy, Julia Stoyanovich, Val Tannen, and Yi Chen. On provenance and privacy. In Tova Milo, editor, *Database Theory - ICDT 2011, 14th International Conference, Uppsala, Sweden, March 21-24, 2011, Proceedings*, pages 3–10. ACM, 2011.

[7] Daniel Deutch, Ariel Frankenthal, Amir Gilad, and Yuval Moskovitch. On optimizing the trade-off between privacy and utility in data provenance. In Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava, editors, *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pages 379–391. ACM, 2021.

[8] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[9] Sameera Ghayyur, Dhrubajyoti Ghosh, Xi He, and Sharad Mehrotra. MIDE: accuracy aware minimally invasive data exploration for decision support. *Proc. VLDB Endow.*, 15(11):2653–2665, 2022.

[10] Noah M. Johnson, Joseph P. Near, Joseph M. Hellerstein, and Dawn Song. Chorus: a programming framework for building scalable differential privacy mechanisms. In *IEEE European Symposium on Security and Privacy, EuroS&P 2020, Genoa, Italy, September 7-11, 2020*, pages 535–551. IEEE, 2020.

[11] Noah M. Johnson, Joseph P. Near, and Dawn Song. Towards practical differential privacy for SQL queries. *Proc. VLDB Endow.*, 11(5):526–539, 2018.

[12] Zach Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? personalized differential privacy. In Johannes Gehrke, Wolfgang Lehner, Kyuseok Shim, Sang Kyun Cha, and Guy M. Lohman, editors, *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 1023–1034. IEEE Computer Society, 2015.

[13] Ios Kotsogiannis, Yuchao Tao, Xi He, Maryam Fanaeepour, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau. Privatesql: A differentially private SQL query engine. *Proc. VLDB Endow.*, 12(11):1371–1384, 2019.

[14] Miti Mazmudar, Thomas Humphries, Jiaxiang Liu, Matthew Rafuse, and Xi He. Cache me if you can: Accuracy-aware inference engine for differentially private data exploration. *Proc. VLDB Endow.*, 16(4):574–586, 2022.

[15] Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In Ugur Çetintemel, Stanley B. Zdonik, Donald Kossmann, and Nesime Tatbul, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29 - July 2, 2009*, pages 19–30. ACM, 2009.

[16] David Pujol, Albert Sun, Brandon Fain, and Ashwin Machanavajjhala. Multi-analyst differential privacy for online query answering. *Proc. VLDB Endow.*, 16(4):816–828, 2022.

[17] Yuan Qiu and Ke Yi. Differential privacy on dynamic data. *CoRR*, abs/2209.01387, 2022.

[18] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12, 2017.

[19] Shufan Zhang and Xi He. DProvDB: Differentially private query processing with multi-analyst provenance. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2024*. ACM, 2024.