Qrlew: Differentially Privacte SQL Query Rewriting

Anonymous submission

Abstract

AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following instructions.

Introduction

In recent years, the importance of safeguarding privacy when dealing with personal data has continuously increased. Traditional anonymization techniques have proven vulnerable to re-identification, as demonstrated by numerous works (Archie et al. 2018; Dwork et al. 2017; Narayanan and Shmatikov 2008; Sweeney, Abu, and Winn 2013). The total cost of data breaches has also significantly increased (IBM 2023) and governments have introduced stricter data protection laws. Yet, the collection, sharing, and utilization of data hold the potential to generate significant value across various industries, including healthcare, finance, transportation, and energy distribution.

To realize these benefits while managing privacy risks, researchers have turned to *differential privacy (DP)* (Wood et al. 2018; Dwork, Roth et al. 2014), which has become the gold standard in academia since its introduction by Dwork et al. in 2006 (Dwork et al. 2006) due to its provable and automatic privacy guarantees.

Despite the availability of open-source tools, DP adoption remains limited. One of the reasons for this lack of adoption is the relative complexity of the existing tools considered the utility of the results. *Qrlew* has been designed to solve these problems by providing the following features:

***Qrlew* provides automatic output privacy guarantees** With *Qrlew* a *data owner* can let an analyst (*data practitionner*) with no expertise in privacy protection run arbitrary SQL queries with strong privacy garantees on the output.

***Qrlew* leverages existing infrastructures** *Qrlew* rewrites a SQL query into a *differentially private* SQL query that can be run on any data-store with a SQL interface from lightweight DB to big-data stores. This removes the need for a custom execution engine and enables *differentially private analytics with virtually no technical integration*.

***Qrlew* leverages synthetic data** . Synthetic data are an increasingly popular way of *privatizing* a dataset. Using jointly *differentially private* mechanisms and *differentially private* synthetic data can be a simple, yet powerful, way of managing a privacy budget and reaching better utility-privacy tradeoffs.

Assumptions and Design Goals

In this work, we assume the *central model of differential privacy* (Near 2020), where a trusted central organization: Hospital, Insurance Company, Utility Provider, called the *data owner*, collects and stores personal data in a secured database. and whishes to let untrusted *data-practitionners* run SQL queries on its data. Furthermore, the *Qrlew* was designed to ease the

General architecture

In this work, we assume the *central model of differential privacy* (Near 2020), where a trusted central organization: Hospital, Insurance Company, Utility Provider, called the *data owner*, collects and stores personal data in a secured database. and whishes to let untrusted *data-practitionners* run SQL queries on its data. Furthermore, the *Qrlew* was designed to ease the

Paul on compilation

Victoria on DP mech and DP test

Comparison to other systems

Known limitations

Qrlew relies on the random number generator of the SQL engine used. It is usually not a cryptographic noise.

Qrlew uses the floating-point numbers of the host SQL engine, therefore our system is liable to the vulnerabilities described in

Useful links

PPAI

Last year papers: https://aaai-ppai23.github.io/#sp2 This year program: https://ppai-workshop.github.io/

Comparable open-source projects

- Paszke et al. 2017 - Automatic differentiation in PyTorch https://openreview.net/pdf?id=BJJsrmfCZ
- Frostig et al. 2018 - Compiling machine learning programs via high-level tracing https://mlsys.org/Conferences/2019/doc/2018/146.pdf

Comparable DP SQL papers

- Lessons Learned: Surveying the Practicality of Differential Privacy in the Industry (Garrido et al. 2022)
- Tumult Analytics: a robust, easy-to-use, scalable, and expressive framework for differential privacy (Berghel et al. 2022)
- Differentially Private SQL with Bounded User Contribution (Wilson et al. 2019)
- CHORUS: a Programming Framework for Building Scalable Differential Privacy Mechanisms (Johnson et al. 2020)
- Towards Practical Differential Privacy for SQL Queries (Johnson, Near, and Song 2018)

Thank you for reading these instructions carefully. We look forward to receiving your electronic files!

References

Archie, M.; Gershon, S.; Katcoff, A.; and Zeng, A. 2018. Who's watching? de-anonymization of netflix reviews using amazon reviews.

Berghel, S.; Bohannon, P.; Desfontaines, D.; Estes, C.; Haney, S.; Hartman, L.; Hay, M.; Machanavajjhala, A.; Magerlein, T.; Miklau, G.; et al. 2022. Tumult Analytics: a robust, easy-to-use, scalable, and expressive framework for differential privacy. *arXiv preprint arXiv:2212.04133*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, 265–284. Springer.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Dwork, C.; Smith, A.; Steinke, T.; and Ullman, J. 2017. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4: 61–84.

Garrido, G. M.; Liu, X.; Matthes, F.; and Song, D. 2022. Lessons learned: Surveying the practicality of differential privacy in the industry. *arXiv preprint arXiv:2211.03898*.

IBM. 2023. Cost of a Data Breach Report 2023.

Johnson, N.; Near, J. P.; Hellerstein, J. M.; and Song, D. 2020. Chorus: a programming framework for building scalable differential privacy mechanisms. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 535–551. IEEE.

Johnson, N.; Near, J. P.; and Song, D. 2018. Towards Practical Differential Privacy for SQL Queries. *Proc. VLDB Endow.*, 11(5): 526–539.

Narayanan, A.; and Shmatikov, V. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111–125. IEEE.

Near, J. 2020. Threat Models for Differential Privacy.

Sweeney, L.; Abu, A.; and Winn, J. 2013. Identifying participants in the personal genome project by name (a re-identification experiment). *arXiv preprint arXiv:1304.7605*.

Wilson, R. J.; Zhang, C. Y.; Lam, W.; Desfontaines, D.; Simmons-Marengo, D.; and Gipson, B. 2019. Differentially private SQL with bounded user contribution. *arXiv preprint arXiv:1909.01917*.

Wood, A.; Altman, M.; Bembenek, A.; Bun, M.; Gaboardi, M.; Honaker, J.; Nissim, K.; O'Brien, D. R.; Steinke, T.; and Vadhan, S. 2018. Differential privacy: A primer for a non-technical audience. *Vand. J. Ent. & Tech. L.*, 21: 209.