Master ISL, projet tueuré, de recherche
Répondre à des requêtes SQL en respectant la vie privée

J.-F. Couchot[1], V. Sonigo[1], and N. Bennani[2]

[1]Institut FEMTO-ST, UMR 6174 CNRS, Université de Franche-Comté, Besançon, France.
vsonigo,couchot@femto-st.fr
[2] Laboratoire d'InfoRmatique en Image et Systèmes d'information, UMR 5205 CNRS , INSA Lyon, France. , nadia.bennani@insa-lyon.fr

1 Contexte

L'Union Européenne souhaite responsabiliser les acteurs traitant des données en imposant que certaines décisions algorithmiques critiques ne soient prises par IA que si celle-ci est explicable, transparente et sûre [Com21]. Cette explicabilité dépend directement de la possibilité d'évaluer l'IA intensément et notamment parfois sur des données stockées dans des bases de données véritables. Cependant, les Systèmes de Gestion de Bases de Données (SGBD) peuvent manipuler des données personnelles et sensibles, et celles-ci ne doivent être révélées qu'aux personnes destinataires de ladite tâche (RGPD).

A contrario, ces données ne doivent pas être mises à la connaissance des personnes qui développent des algorithmes (de traitement, d'évaluation d'IA) sur ces bases, qui les corrigent aussi par la suite. Chiffrer ces données est nécessaire, pseudonymiser les identifiants est indispensable, mais cet ensemble est nullement suffisant. Savoir anonymiser les données pour interdire toute ré-identification, toute acquisition de connaissances personnelles est un enjeu crucial à l'ère du BIG DATA.

Différentes classes d'approches existent pour atteindre l'objectif d'anonymisation des résultats des requêtes : des méthodes syntaxiques (k-anonymat et dérivées), par ajout de bruit,...La confidentialité différentielle (Differential Privacy en anglais, DP) qui fait partie de cette seconde catégorie est acceptée comme un standard puisqu'elle permet de mesurer en termes probabilistes la quantité d'information fuitée à chaque réponse de requête. Ce confort a deux prix : d'une part, l'expression probabiliste est moins compréhensible par la personne en charge qu'un k-anonymat. D'autre part, le bruit introduit se fait souvent au détriment de l'utilité si l'on n'y prend pas garde.

Des extensions d'anonymisation aux SGBD usuels existent aujourd'hui. Dynamic DataMasking [1] (MS SQL Server), Masking with maxscale [2] (MariaDB), PostgreSQL Anonymizer [3] proposent des méthodes à base de pseudonymisation, de k-anonymat et d'ajout de bruit pour nettoyer plusieurs tables d'un SGBD. [KTH+19, WZL+19] proposent quant-à eux des méthodes permettant de nettoyer par confidentialité différentielle globale les réponses aux requêtes de type Select-Project-Join-Aggregation (SPJA) en autorisant plus ou moins de contributions par entité sensible (utilisateur, par exemple). [WZL+19] est complètement outillée [4] dans le langage ZETA-SQL de Google (nommée DP-ZETA-SQL ensuite).

2 Objectifs

Les objectifs de ce projet de recherche sont multiples.

1. Bibliographique d'abord. Il s'agira de comprendre les différentes méthodes d'anonymisation de données.
2. Pratique ensuite. Il s'agira ensuite d'évaluer a minima PostGreSQL Anonymizer pour l'approche syntaxique essentiellement et DP-ZETA-SQL sur des benchmarks reconnus.
3. Théorique enfin. Il s'agira d'évaluer comment des mécanismes récents à base de DP locale [DJW13], de d-privacy [CABP13], de Renyi-DP [Mir17], pourraient être exploitées particulièrement lorsqu'il s'agit de publier des releases. Le nettoyage par DP locale de publications de données catégorielles pour

1. https://docs.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking
2. https://mariadb.com/kb/en/mariadb-enterprise/mariadb-maxscale-21-masking/
3. https://postgresql-anonymizer.readthedocs.io/en/stable/
4. https://github.com/google/differential-privacy/tree/main/examples/zetasql

Master internship
Privacy-preserving SQL answers

J.-F. Couchot[1], V. Sonigo[1], and N. Bennani[2]

[1]Institut FEMTO-ST, UMR 6174 CNRS, Université de Franche-Comté, Besançon, France.
vsonigo,couchot@femto-st.fr
[2] Laboratoire d'InfoRmatique en Image et Systèmes d'information, UMR 5205 CNRS , INSA Lyon, France. , nadia.bennani@insa-lyon.fr

1 Context

The European Union aims to incite data processing stakeholders to be more responsible, by requiring certain critical algorithmic decisions to be taken only by explainable, transparent, and safe AI [Com21]. This explanability is obtained thanks to intensive AI evaluation, sometimes on data stored in real databases.

However, Database Management Systems (DBMS) handle, among others, sensitive personal data whose access must be limited to people authorized to carry out some special tasks (RGPD). Conversely, these data must not be made available to the people who develop algorithms (for processing, AI evaluation) on these databases, who also correct them afterwards. Consequently, these data should be kept private. Encryption and identifiers' pseudo-anonymization represent the first mandatory processing, but they do not provide a sufficient guarantee of data security. Data anonymization remains the only mean to prevent any re-identification or acquisition of personal knowledge.

There exist several approaches to achieve the goal of anonymizing query results, among which syntactic methods (e.g., k-anonymity and derivatives) and noise addition methods(e.g., Differential Privacy (DP)). DP is accepted as a standard, since it enables to measure in probabilistic terms the amount of leaked information through each query response. However, DP efficiency is counterbalanced by a lack of data readability compared to syntactic methods and a drop in data utility.

Today, several common DBMSs have beed extended with data anonymization add-ons . For instance, Dynamic DataMasking[1] (MS SQL Server), Masking with maxscale[2] (MariaDB) and PostgreSQL Anonymizer[3]. These extensions offer methods based on pseudonymization, k-anonymization and noise addition to clean up several DBMS tables. Furthermore, [KTH+19, WZL+19] propose a global DP-based method to sanitize the answers to Select-Project-Join-Aggregation (SPJA) queries by authorizing more or less contributions per sensitive entity (user, for example). Last, a DP-based SQL has been fully implemented [WZL+19] [4] in Google's ZETA-SQL language (later named DP-ZETA-SQL).

2 Objectives

This research internship has multiple objectives.

1. Firstly, bibliographical. The aim is to understand the different methods of data anonymization.

2. Second, practical. The objectives will be to evaluate PostGreSQL Anonymizer for its syntactic approach and DP-ZETA-SQL on recognized benchmarks.

3. Finally theoretical. The aim is to evaluate how recent mechanisms based on local-DP [DJW13], such as d-privacy [CABP13] and Renyi-DP [Mir17], could be exploited particularly when it comes to publishing releases. In fact, local-DP cleaning of categorical data releases in order to build histograms (COUNT/GROUP BY) has made recently significant progress [ACABX21, ACABX22]. Thus, studying these advances for their integration is part of the theoretical goals of this work.

[1]https://docs.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking
[2]https://mariadb.com/kb/en/mariadb-enterprise/mariadb-maxscale-21-masking/
[3]https://postgresql-anonymizer.readthedocs.io/en/stable/
[4]https://github.com/google/differential-privacy/tree/main/examples/zetasql

construire des histogrammes (COUNT/GROUP BY) a récemment beaucoup progressé [ACABX21, ACABX22] et l'étude de ces progrès pour leur intégration fait partie des objectifs théoriques de ce travail.

3 Suites possibles

Ce projet pourra être suivi par un stage de recherche en laboratoire, financé et d'un doctorat dans le cadre d'un projet de recherche, pour lequel une bourse est déjà acquise.

Références

[ACABX21] Héber H Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. Random sampling plus fake data : Multidimensional frequency estimates with local differential privacy. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 47–57, 2021.

[ACABX22] Héber H Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates. *Digital Communications and Networks*, 2022.

[CABP13] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer, 2013.

[Com21] European Commission. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021.

[DJW13] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[KTH+19] Ios Kotsogiannis, Yuchao Tao, Xi He, Maryam Fanaeepour, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau. Privatesql : a differentially private sql query engine. *Proceedings of the VLDB Endowment*, 12(11) :1371–1384, 2019.

[Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[WZL+19] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private sql with bounded user contribution. *arXiv preprint arXiv :1909.01917*, 2019.