Differentially Private Group-by Data Releasing Algorithm

Iago Chaves[1], **Javam Machado**[1]

[1]Database and Systems Laboratory – Federal University of Ceará – Brazil

Abstract. *Privacy concerns are growing fast because of data protection regulations around the world. Many works have built private algorithms avoiding sensitive information leakage through data publication. Differential privacy, based on formal definitions, is a strong guarantee for individual privacy and the cutting edge for designing private algorithms. This work proposes a differentially private group-by algorithm for data publication under the exponential mechanism. Our method publishes data groups according to a specified attribute while maintaining the desired privacy level and trustworthy utility results.*

1. Introduction

Nowadays privacy is a trending topic for consumers and companies. For companies, privacy concern is growing fast because of data protection regulations around the world. And for the consumers, since companies are leaking information about individuals, such as Netflix leakage [Harmanci and Gerstein 2016]. To tackle privacy issues, the differential privacy was proposed [Dwork 2011]. Differential privacy is a strong and measurable guarantee that some method or mechanism do not leak information, depicted by equation 1. Intuitively, it ensures that the presence or absence of someone in the dataset does not change the answer of a query at all. In equation 1, $\mathcal{A}$ represents a mechanism, $o \in \mathcal{O}$ some possible output, ε symbolizes the privacy budget and x, y any neighboring datasets. The definition of neighboring datasets is that two datasets differ only in one user, so $x, y : |x-y| \leq 1$. The differential privacy has been made to run interactively e.g. average, count, sum. Someone submits a query to a database, and it returns via anonymization algorithms an answer with noise [Mendonça et al. 2017]. Occasionally, it is necessary to publish the data instead of statistics about it [Chen et al. 2011].

$$Pr[\mathcal{A}(x) = o] \leq \exp^{\varepsilon} Pr[\mathcal{A}(y) = o] \tag{1}$$

$$Pr[\mathcal{A}(x) = o] \propto \exp(\frac{\varepsilon u(x, o)}{2\Delta u}) \tag{2}$$

$$\Delta u = \max_{o \in \mathcal{O}} \max_{x,y:|x-y|\leq 1} |u(x, o) - u(y, o)| \tag{3}$$

A widely known differential privacy mechanism is the exponential mechanism [McSherry and Talwar 2007]. It works well with categorical answers. The set of all possible query answers is $\mathcal{O}$. The utility function u maps each dataset element, i.e. a user, and a possible result to a score number. The probability of the exponential mechanism outputs o as the answer with utility function u applied in the dataset x is showed in equation 2. But it is necessary to calculate the global sensitivity of Δu, which represents the highest impact regarding the presence or absence of someone in the data.

This paper proposes two methods for grouped data releasing under differential privacy framework. The first algorithm tries to find the optimal representation of the actual group but suffers from computational complexity. The second one reduces the space

search of the former method by a heuristic and achieves reliable results. The PINQ platform [McSherry 2009] address a similar problem, however only to statistical aggregation queries. In section 2 we introduce the two proposed methods TRIODE and TRIODE-H, afterward the experimental results, and finally the conclusion.

2. Our Method

Our goal is to proposes a privacy-preserving method that group a dataset by an attribute and release the groups. We intend to release data through a privacy-preserving group-by algorithm. In other words, a group-by query q response over the dataset $\mathcal{D}$ must do not leak information about individuals. To achieve our goal, we propose two methods via differential privacy mechanisms.

Our first method, named as TRIODE[1], finds each attribute-group applying the exponential mechanism. It is necessary to evaluate the score of all possible subsets, commonly named as powerset of $\mathcal{D}$ and denoted by $\mathbb{P}(\mathcal{D})$. The $\mathbb{P}(\mathcal{D})$ has cardinality $2^{|\mathcal{D}|}$. The second proposed method is TRIODE-H, a TRIODE-based approach that implements a new heuristic to assess the scores, circumventing the powerset complexity problem.

2.1. TRIODE

The TRIODE method groups the dataset by attributes and their values. In order to perform it in a differentially private manner, our method applies the exponential mechanism. So, our dataset $\mathcal{D}$ is composed by $\{a_0, a_1, \ldots, a_k\}$ categorical attributes, and for each attribute a_i, such that $0 \leq i \leq k$, it has κ_i categorical values, represented by $C_i = \{c_i^0, \ldots, c_i^{\kappa_i}\}$. When the group-by query is over the a_i attribute, the method response must contain $|C_i|$ groups.

As aforementioned, a score function $u : \mathcal{D} \times R \to \mathbb{R}$ maps an individual from the dataset and a possible result to a score. The TRIODE answers set are all possibles subsets of $\mathcal{D}$, since a group can be empty or the entire dataset either. The answers set size grows exponentially with the dataset size, once $|R| = 2^{|\mathcal{D}|}$. It is necessary to evaluate the score of each possible category a_i separately. The score must measure a possible result $r \in R$ of a query q grouping over the attribute-value c_i^j concerning the ground truth grouped set $\mathcal{D}_{c_i^j} \subseteq \mathcal{D}$, where $0 \leq j \leq \kappa_i$. Consequently, an answer $r \in R$ with higher similarity with $\mathcal{D}_{c_i^j}$ must imply in a higher score.

$$u_{c_i^j}(r, \mathcal{D}) = 2 \times \frac{|r \cap \mathcal{D}_{c_i^j}|}{|r| + |\mathcal{D}_{c_i^j}|} \tag{4}$$

The score function is defined by equation 4. The fraction numerator is the number of matches between two sets, and the denominator represents the cardinality sum of the sets. Moreover, we need to calculate the global sensitivity of the utility function:

$$GS_u = \max_{\mathcal{D}, \mathcal{D}', r \in R : |\mathcal{D} - \mathcal{D}'| \leq 1} |u(r, \mathcal{D}) - u(r, \mathcal{D}')| \tag{5}$$

$$\leq \frac{2}{2|\mathcal{D}| - 1} \tag{6}$$

[1] Acronym for differen*T*ially p*RI*vate gr*O*up-by *D*ata r*E*leasing

From equation 5 to 6 we assume that $|r|$ is at most $|\mathcal{D}|$, since $r \in \mathbb{P}(\mathcal{D})$. Once the scores and global sensitivity were calculated, it is possible to get the group conforming to the exponential mechanism. The method TRIODE is ε-DP.

2.2. TRIODE-H

The measurement task of all scores in $\mathbb{P}(\mathcal{D})$ is a computationally tough task. For this purpose, we designed a TRIODE-based method, called TRIODE-H. The core concept behind TRIODE-H is to firstly reduce the search space using a heuristic to achieve results within larger datasets, timely. Furthermore, we find the group length through differential privacy to build the groups.

To prune our search space we randomly split our dataset $\mathcal{D}$ into m disjoint fragments, $\{F_1, \ldots, F_m\}$, which each fragment has a fixed length h. Evaluating it with the score function $\pi : F_v \times T_v \to \mathbb{R}$ for all subsets, where $0 \leq v \leq m$ and $T_v = \mathbb{P}(F_v)$ is the set of all possible results for F_v. It is defined as:

$$\pi_{c_i^j}(t, F_v) = -\left| \; |F_v^{c_i^j}| - |t| \; \right| \tag{7}$$

Where $F_v^{c_i^j}$ represents the group F_v grouped by attribute C_i with a value equal to c_i^j and $t \in T_v$. Now we are capable to calculate the score $s : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ for each individual $d \in \mathcal{D}$ separately. For this, we define the individual-based scoring function in equation 8, which calculates the sum of scores, provided by π, for all subsets $\{g \subseteq \mathbb{P}(F_v) \,|\, d \in g\}$.

$$s_{c_i^j}(d, F_v) = \sum_{g \in \mathbb{P}(F_v)} \psi(d, g, \pi_{c_i^j}(g, F_v)) \tag{8}$$

$$\psi(d, g, \pi) = \begin{cases} \pi & \text{if } d \in g \\ -|g| & \text{otherwise} \end{cases} \tag{9}$$

It is important to notice that the $s_{c_i^j}(d, F_v)$ index describes the score function for the group $F_v^{c_i^j}$. Now it is necessary to compute the global sensitivity of s:

$$\begin{aligned} GS_s &= \max_{F_v, F_v', d \in F_v : |F_v - F_v'| \leq 1} |s(d, F_v) - s(d, F_v')| && (10) \\ &\leq |F_v| \times 2^{|F_v|-1} && (11) \\ &= h \times 2^{h-1} && (12) \end{aligned}$$

From equation 10 to 11, we explore that $F_v' = F_v \setminus d$ where $d \in F_v$ is an individual, and $\mathbb{P}(F_v') = 2^{|F_v|-1}$. Now each individual from $\mathcal{D}$ has a score for c_i^j, so we need to publish its group. To do this, it is necessary to discover the size of $\mathcal{D}_{c_i^j}$ in a private manner. Therefore, we defined the set of possible answers for $|\mathcal{D}_{c_i^j}|$ as $L = \{0, \ldots, |\mathcal{D}|\}$ and a score function $\lambda : \mathcal{D} \times L \to \mathbb{R}$.

$$\lambda(\mathcal{D}, \ell) = -| \; |\mathcal{D}| - \ell \; | \tag{13}$$

$$GS_\lambda = \max_{\mathcal{D}, \mathcal{D}', \ell \in L : |\mathcal{D} - \mathcal{D}'| \leq 1} |\lambda(\mathcal{D}, \ell) - \lambda(\mathcal{D}', \ell)| \leq 1 \tag{14}$$

Algorithm 1: 2ε-TRIODE-H

Input: $\mathcal{D}$, $c_i^j \in C_i$, 2ε
Result: group $\tilde{\mathcal{D}}_{c_i^j}$ published by the differentially private method
$\{F_1, \ldots, F_m\} \leftarrow$ split($\mathcal{D}$)
foreach $F_v \in \{F_1, \ldots, F_m\}$ **do**
  $T_v \leftarrow \mathbb{P}(F_v)$
  foreach $t \in T_v$ **do**
    foreach $d \in F_v$ **do**
      scores[d] $\leftarrow s(d, F_v)$
$L = \{0, \ldots, |\mathcal{D}|\}$
foreach $\ell \in L$ **do**
  scores$_L[\ell] \leftarrow \lambda(\mathcal{D}_{c_i^j}, \ell)$
$\ell_{c_i^j} \leftarrow$ ExponentialMechanism($\mathcal{D}$, scores$_L$, ε, $GS = 1$)
foreach $u \in 0, \ldots, \ell_{c_i^j}$ **do**
  $\tilde{\mathcal{D}}_{c_i^j}[u] \leftarrow$ ExponentialMechanism($\mathcal{D}$, scores, $\varepsilon/\ell_{c_i^j}$, $GS = h \times 2^{h-1}$)
return $\tilde{\mathcal{D}}_{c_i^j}$

Where $\ell \in L_{c_i^j}$ and to proof the $GS_\lambda \leq 1$ we used the triangle inequality. Once the scores and global sensitivity were calculated, it is possible to get the group size $\ell_{c_i^j}$ conforming to the exponential mechanism. After all, we are capable to discover the group of C_i with a value equal to c_i^j via private way, i.e. the set $\tilde{\mathcal{D}}_{c_i^j}$. To do this, it is necessary get one individual $\ell_{c_i^j}$ times from $\mathcal{D}$ via exponential mechanism with privacy budget $\frac{\varepsilon}{\ell_{c_i^j}}$ to populate $\tilde{\mathcal{D}}_{c_i^j}$. The algorithm 1 shows clearly the necessary flow for achieving a diferentially private group-by data release. Finally, we reached that the TRIODE-H is 2ε-differential private. The ε budget for querying the group length, and ε for create the group via sequential composition [McSherry 2009].

3. Results

Our experimented data are the Adult Dataset from UCI Machine Learning Repository [Dua and Graff 2017] with 48,842 real individuals and 14 attributes. In this data, for this work, we are interested only in the "sex" attribute, that has two possible value ''Female" and "Male". When the data are grouped, the size of the ''Female" group is 16,192 and so, the size of the "Male" group is 32,650.

The chosen metrics are Precision, Recall, and F1-score [Powers 2011]. The precision is defined by the number of true positives divided by the sum of true positives with false positives. Precision tells us what proportion of our private data are actually relevant. The recall is the number of true positives divided by the sum of true positives with false negatives, which expresses the proportion of actual positive cases that are in the private answer. Finally, the F1-scores is the harmonic mean of precision and recall: F1-score $= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$. The TRIODE method explained in section 2.1, has high computational complexity due to the calculation of $\mathbb{P}(\mathcal{D})$. Besides, the dataset size for this method was set as 20 individuals randomly chosen. Some experiments with 25 and 30 individuals lead to out of memory error, and as consequence, the dataset size is limited

to 20 persons.

All the experiments were made with 5 well-known privacy budgets: $\{0.01, 0.1, ln(2), 1, ln(3), 10\}$ [Dwork 2008]; and for each privacy budget, the experiment was run 10 times. The privacy budgets tries to represent environments from high to low privacy regime, where $\varepsilon = 0.01$ is a strong privacy parameter, and $\varepsilon = 10$ is a weak privacy guarantee. The ε controls the trade-off between utility and privacy. The higher ε more utility, and less privacy you have.

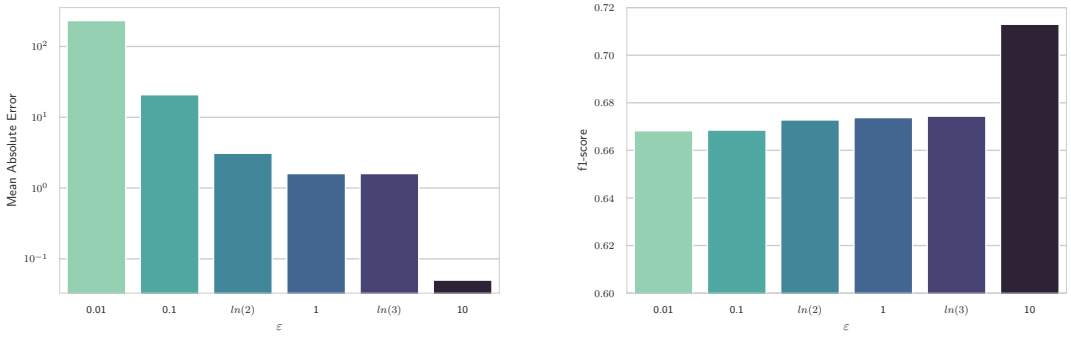
ε	Precision	Recall	F1-score
0.0100	0.7625	0.8133	0.7871
0.1000	0.7438	0.7933	0.7677
$ln(2)$	0.7500	0.8000	0.7742
1.0000	0.7812	0.8333	0.8065
$ln(3)$	0.7750	0.8267	0.8000
10.0000	0.9375	1.0000	0.9677

Table 1. TRIODE average results for the dataset with the length of 20 persons

ε	Precision	Recall	F1-score
0.0100	0.6683	0.6682	0.6682
0.1000	0.6686	0.6685	0.6685
$ln(2)$	0.6728	0.6727	0.6728
1.0000	0.6738	0.6737	0.6738
$ln(3)$	0.6744	0.6743	0.6744
10.0000	0.7131	0.7129	0.7130

Table 2. TRIODE-H average results applied to the entire Adult dataset

Table 1 shows the results for the query: $q =$ `FROM` $\overline{\mathcal{D}}$ `GROUP BY sex`; where $\overline{\mathcal{D}}$ is the dataset $\mathcal{D}$ with only 20 individuals, randomly chosen. Thus, we achieved good and trustworthy results. For the strict budget $\varepsilon = 0.01$ the mean of F1-scores is ≈ 0.79, and for $\varepsilon = 10$ the average F1-score is ≈ 0.97. This is a fine result, but the TRIODE bottleneck is the complexity (time and space), making the experiment with more data unfeasible.



(a) Mean Absolute Error from finding the group size

(b) The average F1-score for the group answer

Figure 1. TRIODE-H results applied to the entire dataset

For TRIODE-H method, the privacy budgets and the query q was the same used in TRIODE experiments. The query q is over the entire data $\mathcal{D}$. Firstly, was necessary to differentially private choose the group length. Figure 1a shows the mean absolute error for each privacy budget. It is worth mentioning that the mean absolute error axis is in logarithmic scale. Once the group length is known, we can settle the answer for query

q via the TRIODE-H technique. The dataset $\mathcal{D}$ has 48,842 rows, and it was split into disjoint fragments with size $m = 10$. The experiments with TRIODE-H was shown in Table 2 and Figure 1 shows the average F1-score for each ε.

4. Conclusion

In this work, we proposed two differentially private data releasing techniques for group-by queries. The first technique TRIODE achieved good utility results. However, the lack of scalability makes the method unfeasible for a large amount of data. The second method TRIODE-H addressed the scalability deficiency by splitting the entire dataset into small fragments. The TRIODE-H was performed with the complete dataset $\mathcal{D}$, and it attains quality results with small privacy budget. For future works we expect to measure the utility bounds for TRIODE and TRIODE-H and to find the optimal, or nearly optimal, space search for TRIODE; It might be valable to experiment with attributes with more than 2 possible classes and also to define an optimization problem to find the optimal set of parameters.

Acknowledgments

Thank you André Mendonça, Daniel Praciano, Eduardo Duarte and Israel Vidal. This research was partially supported by Lenovo Brasil, as part of its R&D investment under Brazil's Informatics Act, CAPES, LSBD/UFC.

References

Chen, R., Mohammed, N., Fung, B. C., Desai, B. C., and Xiong, L. (2011). Publishing set-valued data via differential privacy. *Proceedings of the VLDB Endowment*, 4(11):1087–1098.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C. (2008). Differential privacy: A survey of results. In Agrawal, M., Du, D., Duan, Z., and Li, A., editors, *Theory and Applications of Models of Computation*, pages 1–19, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dwork, C. (2011). Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340.

Harmanci, A. and Gerstein, M. (2016). Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature methods*, 13(3):251.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *FOCS*, volume 7, pages 94–103.

McSherry, F. D. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30. ACM.

Mendonça, A. L., Brito, F. T., Linhares, L. S., and Machado, J. C. (2017). Dipcoding: A differentially private approach for correlated data with clustering. In *Proceedings of the 21st International Database Engineering & Applications Symposium*, pages 291–297. ACM.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.