VLDB Journal manuscript No.
(will be inserted by the editor)

Appendix (Data use-oriented evaluation) to article "Enhancing Data Utility in Differential Privacy via Microaggregation-based k-Anonymity"

Jordi Soria-Comas · Josep Domingo-Ferrer · David Sánchez · Sergio Martínez

Received: date / Accepted: date

1 Evaluation on counting queries

In addition to the general-purpose SSE-based utility evaluation conducted and discussed in the body of the article, in this appendix we provide evaluation results for a specific data use, namely *counting queries*. The reason of focusing on this data use is that many related works on differentially-private data publishing aim at preserving the utility for counting queries over protected data [12–14, 1, 4, 7, 2].

We want to investigate how well our general-purpose method does for counting queries compared to methods that have been designed with this type of queries in mind.

1.1 Benchmark methods

As explained in Section 2 of the article, many related works aim at preserving the accuracy of query responses over the protected data by applying the noise required to fulfill differential privacy over counts or frequencies of records. To do so, the data domain is partitioned and the input record set is transformed into a contingency table (*i.e.* a frequency histogram), which accumulates record counts in each partition set. The partition does not violate differential privacy because it depends on the attribute domains rather than on the input data set. In the sequel we provide the implementation details of this family of methods for numerical and categorical data.

J. Domingo-Ferrer, J. Soria-Comas, D. Sánchez, S. Martínez
UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia.
E-mail: {josep.domingo, jordi.soria, david.sanchez, sergio.martinezl}@urv.cat

For numerical data (like our Census data set), whose attributes take continuous values and for which most records appear only once, we first generalize the attribute domains from continuous to interval-based. For the Census data set, we uniformly partitioned the joint domain of attribute values (defined in Section 5.1 of the article) into 100 intervals. This discretizes the continuous range of values and increases the count of the bin associated to each interval in the corresponding histogram. Differential privacy is not violated by this discretization, because the generalization does not depend on the input data set, but on the attribute domains. After that, the attribute values of records in each bin are replaced by the arithmetical mean of the corresponding interval. Then, geometric noise [5] (that is, the discrete version of Laplace noise) is added to the record count of each generalization in order to fulfill differential privacy. This approach covers in essence methods based on releasing differentially privacy contingency tables [2], histograms [12–14] or decomposed spatial data [10, 3, 9]; all of these methods rely on adding noise to record counts, rather than to the record values as our method does.

Given that a change of one record in the input data set could result in a change of the count of any partition set of records by at most one, the sensitivity that must be considered when adding noise for differential privacy is just 1. Since the added noise may be positive or negative, the record count for a partition might fall below zero after noise addition, in which case we recode it to zero. Doing so does not violate differential privacy, because recoding depends on the noise-added count alone. Even though this method focuses on retaining the utility on counting queries, it does not make any assumptions on the specific queries that will be performed (unlike [8, 6]).

For categorical data (like our Adult-Categorical data set), a similar procedure is applied but, since categorical data are already discrete and their attribute domains are typically small, we do not need the partition/generalization step. For the Adult-Categorical data set a histogram is directly created from the joint domain of the two attributes and geometric noise is added to the count of each bin of the histogram (again, the sensitivity of counts is one). As above, since the binning is defined on the attribute domains, rather than on the specific input data set, it does not violate differential privacy.

1.2 Evaluation measures

To evaluate the accuracy (*i.e.* utility) of counting queries in a general way, we used a well-known measure to compare the distributions of two histograms: the Earth Mover's Distance [11]. This measure quantifies the total cost of transforming a histogram (in our case, the histogram corresponding to the differentially private data set) into another histogram (*i.e.* the histogram of the original data set), so that their frequency distributions (*i.e.* record counts) match. This is done by optimally 'moving' mismatching records between histogram bins according to a distance between records located at different histogram bins. To compute such a record distance, we used the Euclidean distance for numerical attributes and the semantic distance for categorical ones, as done in Section 5 of the article.

To evaluate the practical privacy of the differentially private outputs, the Record Linkage measure introduced in Section 5.2 of the article was employed.

1.3 Results and discussion

Evaluation results (Earth mover's distance and RL) for the benchmark approach detailed in the previous section and for our proposal with the same ε and k values used in Section 5 of the article are depicted in Figures 1 and 2 for the Census and Adult-Categorical data sets, respectively. Since the benchmark methods described in the previous section do not depend on the parameter k, they are represented with horizontal lines whose styles match the curves obtained with our proposal for the corresponding ε values.

Regarding the counting accuracy (*i.e.* Earth mover's distance), we notice that the behavior of our method for both data sets is similar to what was observed in Section 5 for the SSE-based evaluations. For low ε values (0.01 and 0.1), the baseline noise is so high that even with the noise reduction provided by the prior microaggregation step, the accuracy of the output is hardly improved as k grows. For high ε values (1.0 and 10.0) there is an improvement of the counting accuracy as the microaggregation parameter k increases.

Comparing these results to those of the benchmark methods focused on counting queries, the following can be concluded:

- For $\varepsilon = 0.01$, the benchmark methods produce significantly less accurate results than those yielded by our method for both data sets (notice the $\log_{10}$ scale for the Y-axis). Since to fulfill differential privacy, noisy record counts may be produced for *any* possible value in the joint attribute domains, a potentially large number of low density histogram bins will be obtained. Thus, the low or zero record count corresponding to these partitions combined with the high noise required by 0.01-differential privacy produces highly inaccurate counts with regard to the original data set.
- For $\varepsilon = 0.1$, the benchmark methods outperform our proposal for both data sets, regardless of the value of the microaggregation parameter k. This suggests that, for counting queries, if the ε parameter is relaxed enough, it may be preferable to modify record counts (as benchmark methods do) than record values (as done by our proposal).
- For $\varepsilon = 1.0$ and $\varepsilon = 10.0$, we observe significant differences between the two data sets:
  - For the Census numerical data set, the benchmark method yields about the same accuracy for $\varepsilon = 1.0$ and $\varepsilon = 10.0$. This is due to the accuracy penalty incurred by the discretization of continuous numerical values needed to create bins. In contrast, our proposal, which uses a k-anonymous microaggregation, is able to produce a distribution of values that is more similar to the original data: with $\varepsilon = 1.0$, it improves on the benchmark method for $k > 72$; with $\varepsilon = 10.0$, it improves on the benchmark for $k > 24$.
  - On the other hand, for the Adult-Categorical data set, the benchmark method produces near-perfect results, especially for $\varepsilon = 10.0$. Since the input attributes are already discrete and have domains with just a few possible values, no further discretization is needed and hence there is no utility loss associated to binning. Furthermore, the noise being added to counts is discrete. As a result, the output distribution looks very much like the original distribution. This behavior, however, has a negative impact on the practical privacy achieved, as discussed below.

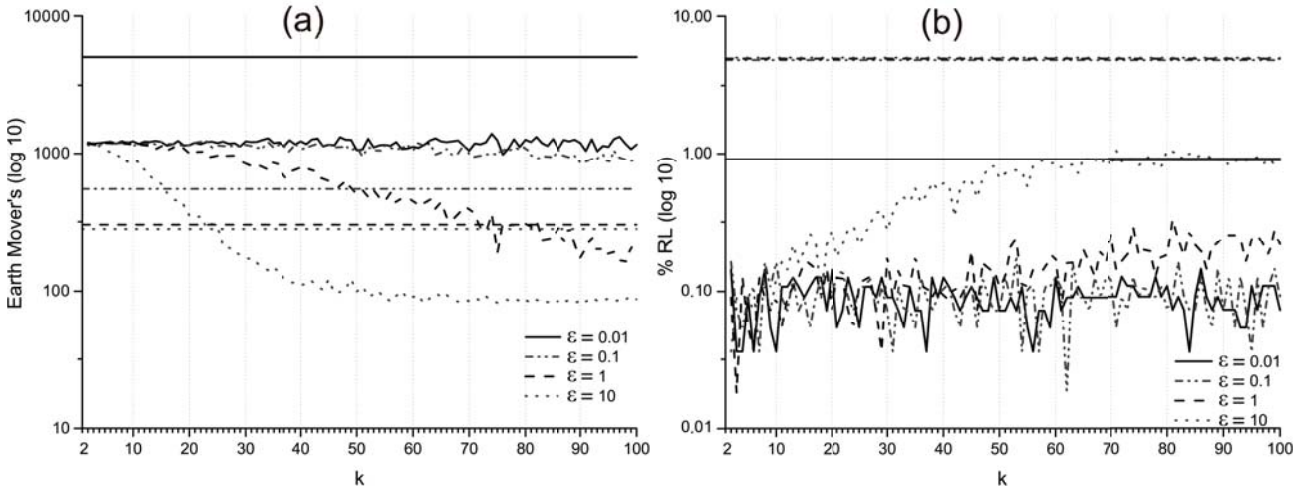

Fig. 1 Census data set. Earth mover's distance (a) and RL results (b) for our approach (non-horizontal lines, because our method depends on the microaggregation parameter k) vs. benchmark methods (horizontal lines, because these methods do not depend on k) for different ε values with a $\log_{10}$ scale and k varying with step 1.

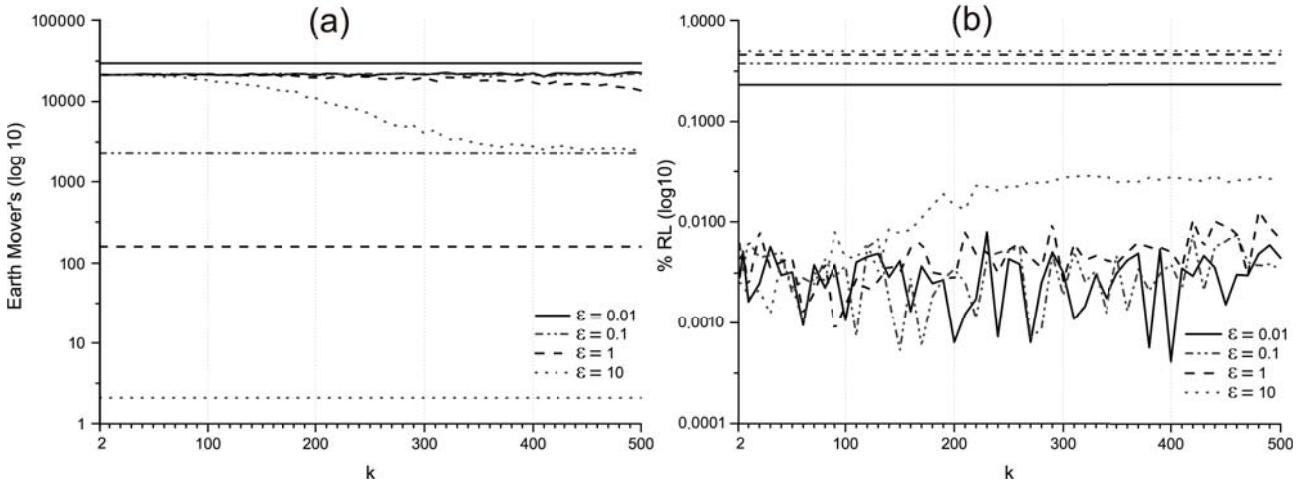

Fig. 2 Adult-Categorical data set. Earth mover's distance (a) and RL results (b) for our approach (non-horizontal lines, because our method depends on the microaggregation parameter k) vs. benchmark methods (horizontal lines, because these methods do not depend on k) for different ε values with a Log_{10} scale and k varying with step 10.

Regarding the practical privacy of the protected data set (*i.e.* RL), we notice that, for all ε values and for both data sets, our approach produces significantly less disclosure risk than the benchmark methods (again, notice the $\log_{10}$ scale for the Y-axis). Differences range from around 5 times lower RL for the Census data set and ε=10.0 (with the highest k) to 100 times lower RL for the Adult-Categorical and ε=0.01 (with the highest k). The reason is that, whereas benchmark methods modify record counts (leaving original values intact), our proposal adds noise to attribute values. As a result, our protected records are unlikely to unambiguously match any of the original ones, because all the attribute values have changed. In contrast, benchmark methods result in a practical privacy that is even weaker than the one obtained with the non-differentially private method (MDAV, see Figures 4(b) and 5(b) in Section 5 of the article).

1.4 Conclusion

In general, methods focused on counting queries have the advantage that, even for low ε values, the protected output is highly accurate, thanks to the discrete nature and low sensitivity of the counts to which the noise is added.

A first drawback of such methods is that the practical privacy may be significantly compromised if the

ε parameter is relaxed too much, which may produce results that are almost identical to the original data, especially when original record values have not changed (such as in the categorical data set). Another drawback is the accuracy loss caused by the prior domain partition required to discretize continuous numerical data. In contrast, our approach adds noise to the record values rather than to their count, a procedure that outputs values that are similar but never identical to the original values; this clearly increases the practical privacy, as unequivocal record linkages are less likely to occur than for benchmark methods that do not alter record values. As a result, our approach strikes a better balance between practical utility (*i.e.* count accuracy) and practical privacy (*i.e.* RL) than benchmark methods.

In summary, even though our method aims at producing differentially private data sets regardless of their data uses, it is still able to offer reasonably accurate results for counting queries compared to the typical strategy followed by methods specifically aimed at this type of queries, while significantly reducing the practical risk of disclosure.

Disclaimer and acknowledgments

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TIN2011-27076-C03-01 "CO-PRIVACY", TIN2012-32757 "ICWT", IPT2012-0603-430000 "BallotNext" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Commission under FP7 projects "DwB" and "Inter-Trust". The second author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. Blum, A., Ligett, K., Roth, A.: A learning theory approach to non-interactive database privacy. In: Proc. of the 40th Annual Symposium on the Theory of Computing-STOC 2008, pp. 609-618 (2008)
2. Chen, R., Mohammed, N., Fung, B.C.M., Desai B.C., Xiong, L.: Publishing set-valued data via differential privacy. In: 37th Intl. Conference on Very Large Data Bases-VLDB 2011/Proc. of the VLDB Endowment 4(11), 1087-1098 (2011)
3. Cormode, G., Procopiuc, C. M., Shen, E., Srivastava, D., Yu, T.: Differentially private spatial decompositions. In: IEEE International Conference on Data Engineering (ICDE 2012), pp. 20-31 (2012)
4. Dwork, C., Naor, M., Reingold, O., Rothblum G.N., Vadhan, S.: On the complexity of differentially private data release: efficient algorithms and hardness results. In: Proc. of the 41st Annual Symposium on the Theory of Computing-STOC 2009, pp. 381-390 (2009)
5. Ghosh, A., Roughgarden, T., Sundararajan, M.: Universally utility-maximizing privacy mechanisms. In: Proc. of 41st annual ACM symposium on Theory of computing (STOC 2009), pp. 351-360 (2009)
6. Hardt, M., Talwar, K.: On the geometry of differential privacy. In: Proc. of the 42nd ACM symposium on Theory of computing (STOC 2010), pp. 705-714 (2010)
7. Hardt, M., Ligett, K., McSherry, F.: A simple and practical algorithm for differentially private data release. Preprint arXiv:1012.4763v1 (2010)
8. Li, C., Hay, M., Rastogi, V., Miklau, G., McGregor, A.: Optimizing linear counting queries under differential privacy. In: Proc. of the 29th Symposium on Principles of database systems (PODS 2010), pp. 12-134 (2010)
9. Li, N., Yang, W., Qardaji, W.: Differentially private grids for geospatial data. In: IEEE International Conference on Data Engineering (ICDE 2013), pp. 757-768 (2013)
10. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: IEEE International Conference on Data Engineering (ICDE 2008), pp. 277-286 (2008)
11. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval. Int. J. of Comp. Vis. 40(2), pp. 99-121 (2000)
12. Xiao, X., Wang, G., Gehrke, J.: Differential Privacy via Wavelet Transforms. IEEE Trans. on Knowl. and Data Eng. 23(8), pp. 1200-1214 (2010)
13. Xiao, Y., Xiong, L., Yuan, C.: Differentially private data release through multidimensional partitioning. In: Proceedings of the 7th VLDB conference on Secure data management (SDM'10), pp. 150-168 (2010)
14. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G.: Differentially Private Histogram Publication. In: IEEE International Conference on Data Engineering (ICDE 2012), pp. 32-43 (2012)