DPCube: Releasing Differentially Private Data Cubes for Health Information

Yonghui Xiao [#1], James Gardner [*#2], Li Xiong [#3]

[#]*Department of Mathematics and Computer Science, Emory University*
Atlanta, GA, USA
[1]yonghui.xiao@emory.edu
[3]lxiong@emory.edu

[*]*Digital Reasoning Systems, Inc.*
Franklin, TN, USA
[2]james.gardner@digitalreasoning.com

***Abstract*—We demonstrate DPCube, a component in our Health Information DE-identification (HIDE) framework, for releasing differentially private data cubes (or multi-dimensional histograms) for sensitive data. HIDE is a framework we developed for integrating heterogenous structured and unstructured health information and provides methods for privacy preserving data publishing. The DPCube component uses differentially private access mechanisms and an innovative 2-phase multidimensional partitioning strategy to publish a multi-dimensional data cube or histogram that achieves good utility while satisfying differential privacy. We demonstrate that the released data cubes can serve as a sanitized synopsis of the raw database and, together with an optional synthesized dataset based on the data cubes, can support various Online Analytical Processing (OLAP) queries and learning tasks.**

I. Introduction

Recent studies and advisory reports [16], [14] have pointed out that information sharing with appropriate privacy protection is one of the most critical challenges of our time to help revolutionizing health care. The current HIPAA Privacy Rule does not protect privacy as well as it should and an entirely new approach to protecting privacy in health research is needed.

Privacy preserving data analysis and data publishing [3], [6], [4] has received considerable attention in recent years as a promising approach for sharing information while preserving data privacy. There are two models for privacy protection [3]: the interactive model and the non-interactive model. In the interactive model, a trusted *curator* (e.g. hospital) collects data from *record owners* (e.g. patients) and provides an access mechanism for *data users* (e.g. public health researchers) for querying or analysis purposes. The result returned from the access mechanism is perturbed by the mechanism to protect privacy. In the non-interactive model, the curator publishes a "sanitized" version of the data, simultaneously providing utility for data users and privacy protection for the individuals represented in the data.

Differential privacy [3], [4] is widely accepted as one of the strongest known unconditional privacy guarantees. It requires that the outcome of computations to be formally indistinguishable when run with and without any particular record in the dataset, as if it makes little difference whether an individual is being opted in or out of the database. Non-interactive data release with differential privacy has been recently studied with hardness results obtained and it remains an open problem to find efficient algorithms for many domains [5]. A few recent works [12], [17], [10] considered the problem of releasing data for general predicate counting queries. While promising, the query strategies being used are data-oblivious in that they are determined by the query workload, a wavelet matrix, or a hierarchical matrix without taking into consideration the underlying data. DPCube uses an adaptive query strategy that explicitly exploits the underlying data indirectly observed by a differentially private interface. Our viewpoint is that more efficient and effective solutions could be achieved by exploiting the characteristics of the underlying dataset.

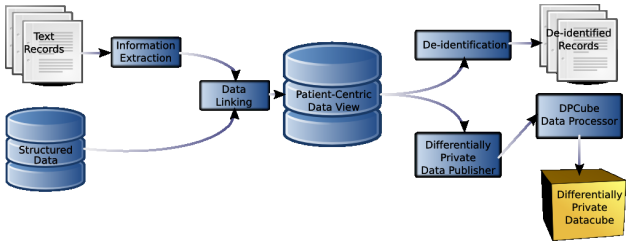

Fig. 1. HIDE Framework

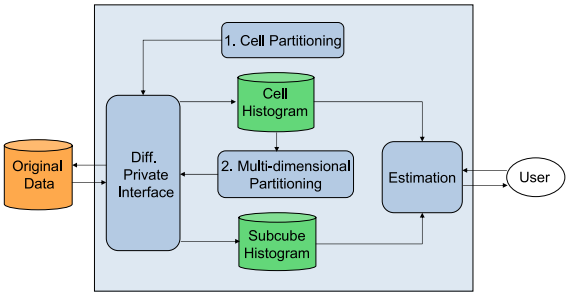

Fig. 2. Differentially Private Datacube Release

We demonstrate DPCube, as a component in our Health Information DE-identification (HIDE) framework, for non-interactive release of differentially private data cubes (or multi-dimensional histograms), and demonstrate its feasibility using real medical data and use cases. Health Information DE-identification (HIDE) [8], [7], [9] is an open-source software and framework we have developed for integrating heterogenous structured and unstructured data sources and provides methods for privacy preserving data publishing. An overview of the HIDE framework is shown in Figure 1. DPCube is the component for releasing statistical data cubes (in addition to the de-identified records) implementing and extending the multidimensional partitioning techniques [18]. An overview of DPCube is shown in Figure 2. An interactive differential privacy interface is used to provide a differentially private access to the raw database. The DPCube algorithm implements a 2-phase partitioning strategy which accesses the data through the interface and generates a differentially private equi-width cell histogram and a v-optimal subcube histogram of the raw database. Given a user-issued query, an estimation component uses the histograms and computes an answer using inference or estimation techniques [10], [18]. We demonstrate that the histograms can serve as a sanitized synopsis of the raw database and, together with an optional synthesized dataset based on the histograms, are useful to support count queries and other types of Online Analytical Processing (OLAP) queries and learning tasks.

Contributions. We summarize the technical contributions of the demonstrated system below. First, the DPCube component provides a non-interactive mechanism for releasing differentially private multi-dimensional data cubes. A common interactive differential privacy mechanism is to add calibrated noise to a query result determined by the privacy parameter and the sensitivity of a query. The composability of differential privacy [13] ensures privacy guarantees for a sequence of differentially-private queries or computations with additive privacy depletions in the worst case. Given an overall privacy requirement or privacy budget, it has to be allocated to subroutines or individual queries to ensure the overall privacy. This limits the applicability of an interactive differential privacy interface, especially for health data sharing scenarios where multiple users have to share a common privacy budget for exploratory data analysis. In contrast, the non-interactive approach uses a carefully designed query strategy to allocate the privacy budget and access the raw database resulting in released data cubes which can be then used to answer an arbitrary large number of queries or for various data analysis tasks.

Second, DPCube implements and extends the novel multi-dimensional partitioning strategy [18] which is crucial to the utility of the resulting data cubes or the synthetic dataset. For relational data, DPCube uses a two-phase algorithm that generates a most fine-grained equi-width cell (unit) histogram and a v-optimal subcube histogram based on the cell histogram. The multi-dimensional partitioning component incorporates a uniformity measure that seeks to produce close to uniform partitions so that approximation errors within partitions are minimized. In data warehouse literature, a *data cube* consists of a set of *cuboids* which can be viewed as the projection of a fact table on a subset of dimensions, producing a set of cells with associated aggregate measures. Using this terminology, DPCube publishes a base cuboid and a *generalized* cuboid where cells are grouped into sub-cubes or partitions and aims to minimize the error for random predicate counting queries. This differs from the recent work [2] which attempts to publish a subset of cuboids to minimize the error for the complete set of derived cuboids. Once the subcube histogram is generated, the estimation component further boosts the accuracy by applying estimation or inference techniques to answer a query using the released histograms.

Finally, as DPCube is integrated into the HIDE software, it gives practitioners an entire toolkit of privacy preserving data publishing techniques on both structured and unstructured data. The DPCube component can be applied to both structured data tables as well as the extracted patient-centric data from the text.

The demonstration will show various components and new features of HIDE (including more advanced extraction components), with a particular focus on DPCube, and demonstrate the utility of the released data cubes to several classes of OLAP queries and learning tasks. The audience can freely issue random count queries with different parameter settings, observe the result, and compare it with the result from the original database, the interactive mechanism, as well as other alternative approaches. It will also show under-the-hood how the partitions are generated for various datasets and combinations of dimensions. The audience will also be able to see how various fields are automatically extracted from text reports and the data is then used to generate differentially private multi-dimensional histogram views of the data.

Software Availability. HIDE including DPCube is an open-source project and the software is available for download (http://code.google.com/p/hide). More information about the project is also available online (http://www.mathcs.emory.edu/hide).

II. System Description

HIDE consists of three major layers: information extraction, data linking, and privacy preserving data publishing. The privacy preserving data publishing layer can be further classified into record and aggregate publishers. The record publisher is used to publish de-identified (using HIPAA safe-harbor method) or anonymized individual records with a given privacy principle such as k-anonymity and l-diversity. The aggregate publisher provides a differentially private interface to the underlying data and is used by the DPCube algorithm as described below for releasing statistical data with differential privacy.

A. Differentially Private Interface

DPCube uses a differentially private interface in HIDE similar to PINQ [13], for any access to the original database such that the released data guarantees differential privacy. HIDE provides differentially private operators for database aggregate queries such as count (**NoisyCount**) and sum (**NoisySum**) which add Laplace noise to the original answer to enforce differential privacy.

B. DPCube Partitioning

DPCube publishes statistical data of the original datasets through multi-dimensional hypercubes or cuboids. An OLAP cuboid is a multi-dimensional representation of a *measure* for a set of *dimensions* that are a projection of a relational table. Each dimension corresponds to an attribute and each cell represents an aggregated measure (such as count and sum). For example, a 3-dimensional cuboid for the census data may have the dimensions Age, $Education$, and $Income$ and the measure $Population\ Count$. One cell of the cuboid may correspond to (Age = 30, Education = "Bachelors", Income = 80K) with a population count of 5000. Our key novelty is that we publish *generalized* cuboids where cells are grouped into sub-cubes or partitions exploiting the underlying distribution of the data. When the measure is population count or frequency (relative count), the cube can be considered as a multi-dimensional histogram.

DPCube uses an innovative two-phase partitioning strategy as shown in Figure 2. First, a cell based partitioning based on the domains (not the data) is used to generate a fine-grained equi-width cell histogram. A differentially private data cube, D_c is created by adding Laplacian noise to the count of each cell. Second, a multi-dimensional partitioning is performed on D_c, the differentially private cell histogram which gives an approximation of the original data distribution.

The key step is the multi-dimensional partitioning. We want to find the partitioning that maximizes the utility of the released D_p data cube. DPCube uses an innovative kd-tree like partitioning strategy that seeks to produce close to uniform partitions, essentially resulting in a v-optimal histogram [15]. It starts from the root node which covers the entire space. At each step, a splitting dimension and a split value from the range of the current partition on that dimension are chosen heuristically to divide the space into subspaces. The algorithm repeats until a pre-defined requirement (such as number of data points in each partition) are met. In contrast to kd-tree construction which desires a balanced tree, our main goal is to generate uniform or close to uniform partitions so that the approximation error when answering a query with predicates smaller than the partitions is minimized. Thus DPCube uses uniformity based heuristics to make the decision whether and where to split the current partition. Concretely, we do not split a partition if it is close to uniform and split it otherwise. In addition to the variance-like metric defined in [18], DPCube also implements information gain to favor uniform or homogenous distributions in a similar way used in a decision tree construction to favor the class homogeneity.

For comparison purposes, DPCube also includes an implementation of the hierarchical strategy used in [10], the kd-tree strategy used in [11], and an implementation of the Wavelet method [17]. Finally, to cope with the high dimensionality and sparsity of the set-valued data such as symptoms, DPCube includes an implementation of the top-down partitioning strategy proposed in [1].

C. DPCube Estimation

DPCube also includes an estimation component that further boosts accuracy according to the released data cubes. In addition to the proportional estimation using only the subcube histogram assuming a uniform distribution within a partition [18], we also adapted the inference technique in [10] originally designed for its hierarchical strategy to our two-phase strategy. The basic idea is to apply probabilistic inference to integrate multiple differentially private views (histograms) of the original data to derive posterior distributions over the data sets.

III. Demonstration

We plan to demonstrate the functionalities and utilities of HIDE and DPCube and show the utility of the released data cubes to several classes of OLAP queries and learning tasks.

Datasets. We will use the Adult dataset from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) containing census data. In addition, we plan to use several medical datasets including the data from the Surveillance, Epidemiology and End Results (SEER) Program at NCI (http://seer.cancer.gov/data/) and the i2b2 datasets (https://www.i2b2.org/NLP/DataSets). The SEER dataset consists of a variety of cancer statistics. The i2b2 dataset consists of example pathology reports that have been re-synthesized with fake Protected Health Information (PHI). In addition, we will use a large set of synthesized pathology reports (1 million reports) generated at the Department of Pathology and Laboratory Medicine at the School of Medicine at UCLA to demonstrate the performance of HIDE and DPCube.

Demonstrating Basic Functionality. We will demonstrate loading structured and unstructured data into HIDE, de-identifying and anonymizing the data, and releasing differentially private data cubes using DPCube. We will use several dimension combinations to demonstrate the data cubes generated by DPCube. Figure 3 shows a snapshot of the DPCube interface for constructing the data cubes based on user input parameters and the resulting cell and partition histograms using a single dimension of Age. Using two dimensions, Age and Income, the original data cube is shown in Figure 4. Figure 5 shows an example cell data cube with differential privacy parameter $\alpha_1 = 0.05$. Figure 6 shows an example partition data cube with each horizontal plane being one partition. Finally, we will show the estimated cell data cube using the estimation techniques and compare it to the original cell data cube. Figure 7 shows an example estimated cell data cube. We will show the errors of the different data cubes during the demo.

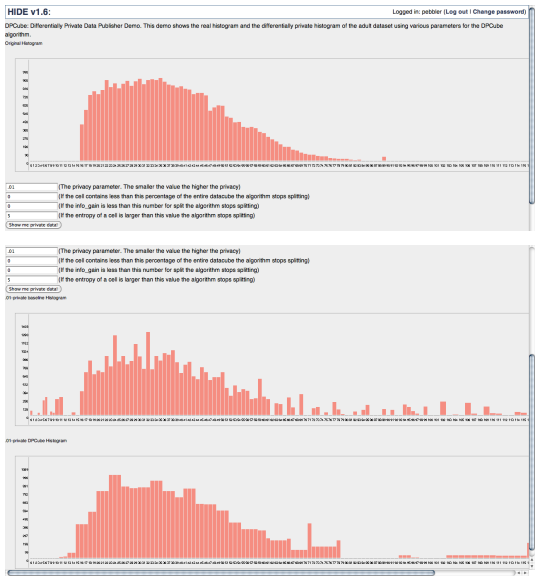

Fig. 3. DPCube Interface

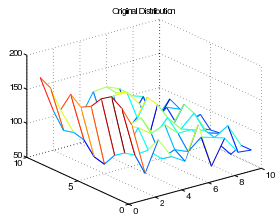

Fig. 4. Original Data

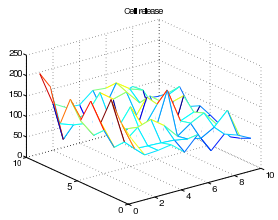

Fig. 5. Cell Data Cube

Applications and User Interactions. We will demonstrate the utility of the released data cubes for several applications with user interactions and inputs. We will first show the utility for several classes of OLAP queries. Count queries are supported directly by the released data. Sum queries sum(A) for an attribute or dimension A can be computed as $\sum_{i \in S_\varphi}(a_i * c_i)$. Average queries avg($A$) for an attribute or dimension A can be computed as $\frac{\sum_{i \in S_\varphi}(A_i * c_i)}{\sum_{i \in S_\varphi}(c_i)}$. Through the user interface, the conference audience can freely issue predicate queries using different parameter settings and observe the result. The interface will compare the result with those from the original database as well as other alternative approaches. In addition to showing the data cubes visually and the query results, we will also show the errors of the query results and compare them with several other state-of-the-art approaches.

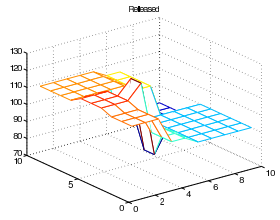

Fig. 6. Partition Data Cube

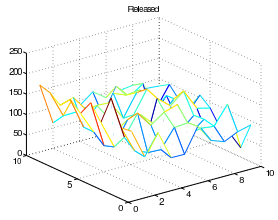

Fig. 7. Estimated Cell Data Cube

The released data cubes can be also used for learning tasks such as construction of decision tree and record linkage. We will use classification and record linkage [11] as examples to illustrate the utility of the data cubes.

Under-the-hood. For interested audiences, the demo will show under-the-hood how the partitions are generated for various combinations of dimensions and parameters, illustrating the effect of the data distributions and the impact of algorithmic parameters such as the partitioning threshold, different partitioning metrics and heuristics, and different allocation of the overall privacy budget among the two phases.

ACKNOWLEDGEMENT

This research was supported in part by NSF grant CNS-1117763, a Cisco Research Award, and an Emory URC grant. The authors would like to thank the anonymous reviewers for their comments which helped improve the final version of this demo paper.

REFERENCES

[1] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong. Publishing set-valued data via differential privacy. In *VLDB*, 2011.
[2] B. Ding, M. Winslett, J. Han, and Z. Li. Differentially private data cubes: optimizing noise sources and consistency. In *SIGMOD*, 2011.
[3] C. Dwork. Differential privacy: a survey of results. In *TAMC*, 2008.
[4] C. Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54(1), 2011.
[5] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *STOC*, 2009.
[6] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys*, 42(4), 2010.
[7] J. J. Gardner and L. Xiong. An integrated framework for de-identifying unstructured medical data. *Data Knowl. Eng.*, 68(12), 2009.
[8] J. J. Gardner, L. Xiong, K. Li, and J. J. Lu. Hide: heterogeneous information de-identification. In *EDBT*, 2009.
[9] J. J. Gardner, L. Xiong, F. Wang, A. Post, J. H. Saltz, and T. Grandison. An evaluation of feature sets and sampling techniques for de-identification of medical records. In *IHI*, 2010.
[10] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially-private queries through consistency. In *VLDB*, 2010.
[11] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino. Private record matching using differential privacy. In *EDBT*, 2010.
[12] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *PODS*, 2010.
[13] McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD*, 2009.
[14] S. J. Nass, L. A. Levit, and L. O. Gostin, editors. *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research*. Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule; Institute of Medicine; The National Academies Press, 2009.
[15] V. Poosala, P. J. Haas, Y. E. Ioannidis, and E. J. Shekita. Improved histograms for selectivity estimation of range predicates. *SIGMOD Rec.*, 25, 1996.
[16] W. W. Stead and H. S. Lin, editors. *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*. Committee on Engaging the Computer Science Research Community in Health Care Informatics; National Research Council, The National Academies Press, 2009.
[17] X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. In *ICDE*, 2010.
[18] Y. Xiao, L. Xiong, and C. Yuan. Differentially private data release through multidimensional partitioning. In *Secure Data Management*, 2010.