DP-tree: Indexing Multi-Dimensional Data under Differential Privacy

Shangfu Peng, Shanghai Jiao Tong University
Yin Yang, Advanced Digital Sciences Center
Zhenjie Zhang, Advanced Digital Sciences Center
Marianne Winslett, University of Illinois at Urbana-Champaign, and Advanced Digital Sciences Center
Yong Yu, Shanghai Jiao Tong University
pengshangfu@sjtu.edu.cn

ε-differential privacy (ε-DP) is a strong and rigorous scheme for protecting individuals' privacy while releasing useful statistical information. The main idea is to inject random noise into the results of statistical queries, such that the existence of any single record has negligible impact on the distributions of query results. The accuracy of such randomized results depends heavily upon the query processing technique, which has been an active research topic in recent years. So far, most existing methods focus on 1-dimensional queries. The only work that handles multi-dimensional query processing under ε-DP is [1], which indexes the sensitive data using variants of the quad-tree and the *k-d*-tree. As we point out in this paper, these structures are inherently suboptimal for answering queries under ε-DP. Consequently, the solutions in [1] suffer from several serious drawbacks, including limited and unstable query accuracy, as well as bias towards certain types of queries.

Motivated by this, we propose the DP-tree, a novel index structure for multi-dimensional query processing under ε-DP that eliminates the problems encountered by the methods in [1]. Further, we show that the effectiveness of the DP-tree can be improved using statistical information about the query workload. Extensive experiments using real and synthetic datasets confirm that the DP-tree achieves significantly higher query accuracy than existing methods. Interestingly, an adaptation of the DP-tree also outperforms previous 1D solutions in their restricted scope, by large margins.

[1] G. Cormode, M. Procopiuc, E. Shen, D. Srivastava and T. Yu. Differentially Private Spatial Decompositions. *ICDE*, 2012.

ACM Categories & Descriptors: H.2.8 **[DATABASE MANAGEMENT]**: Database Applications—Statistical databases

Keywords: Differential Privacy, Multi-Dimensional Index, Consistency, Workload-Aware Query Processing