ScholarWorks@UMass Amherst

Title

Privacy-preserving Sanitization in Data Sharing

Author

Wentian Lu Follow

Document Type

Open Access Dissertation

Degree Name

Doctor of Philosophy (PhD)

Degree Program

Computer Science

Year Degree Awarded

Summer 2014

First Advisor

Gerome Miklau

Second Advisor

Neil Immerman

Third Advisor

David Jensen

Fourth Advisor

Krista Gile

Subject Categories

Computer Sciences | Databases and Information Systems | Information Security | Theory and Algorithms

Abstract

In the era of big data, the prospect of analyzing, monitoring and investigating all sources of data starts to stand out in every aspect of our life. The benefit of such practices becomes concrete only when analysts or investigators have the information shared from data owners. However, privacy is one of the main barriers that disrupt the sharing behavior, due to the fear of disclosing sensitive information. This dissertation describes data sanitization methods that disguise the sensitive information before sharing a dataset and our criteria are always protecting privacy while preserving utility as much as possible.

In particular, we provide solutions for tasks that require different types of shared data. In the case of sharing partial content of a dataset, we consider the problem of releasing a database under retention restrictions such that the auditing job can still be carried out. While obeying a retention policy often results in the wholesale destruction of the audit log in existing solutions, our framework allows to expire data at a fine granularity and supports audit queries on a database with incompleteness. Secondly, in the case of sharing the entire dataset, we solve the problem of untrusted system evaluation using released database synthesis under differential privacy. Our synthetic database accurately preserves the core performance measures of a given query workload, and satisfies differential privacy with crucial extensions to multi-relation databases. Lastly, in the case of sharing derived information from the data source, we focus on distributing results of network modeling under differential privacy. Our mechanism can safely output estimated parameters of the exponential random graph model, by employing a decomposition of the estimation problem into two steps: getting private sufficient statistics first and then estimating the model parameters. We show that our privacy mechanism provides provably less error than common baselines and our redesigned estimation algorithm offers better accuracy.

Recommended Citation

Lu, Wentian, "Privacy-preserving Sanitization in Data Sharing" (2014). *Doctoral Dissertations*. 235.

https://scholarworks.umass.edu/dissertations_2/235