

Privacy-Preserving Computation and Verification of Aggregate Queries on Outsourced Databases*

Brian Thompson¹, Stuart Haber², William G. Horne²,
Tomas Sander², and Danfeng Yao¹

¹ Department of Computer Science
Rutgers University

Piscataway, NJ 08854, USA

{bthom,danfeng}@cs.rutgers.edu

² Hewlett-Packard Labs

5 Vaughn Drive, Suite 301

Princeton, NJ 08540, USA

{stuart.haber,william.horne,tomas.sander}@hp.com

Abstract. Outsourced databases provide a solution for data owners who want to delegate the task of answering database queries to third-party service providers. However, distrustful users may desire a means of verifying the integrity of responses to their database queries. Simultaneously, for privacy or security reasons, the data owner may want to keep the database hidden from service providers. This security property is particularly relevant for aggregate databases, where data is sensitive, and results should only be revealed for queries that are aggregate in nature. In such a scenario, using simple signature schemes for verification does not suffice. We present a solution in which service providers can collaboratively compute aggregate queries without gaining knowledge of intermediate results, and users can verify the results of their queries, relying only on their trust of the data owner. Our protocols are secure under reasonable cryptographic assumptions, and are robust to collusion among k dishonest service providers.

Keywords: Aggregate query, outsource, privacy, integrity, secret sharing, verification.

* This work has been supported in part by NSF grant CCF-0728937, CNS-0831186, and the Rutgers University Computing Coordination Council Pervasive Computing Initiative Grant. This material is also based upon work supported by the U.S. Department of Homeland Security under grant number 2008-ST-104-000016. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

1 Introduction

Privacy concerns are still a major obstacle that makes sensitive data inaccessible to data mining researchers and prevents collaborative data analysis and filtering among multiple organizations from becoming a reality. Many databases contain sensitive information, and the data owner may not want to share it in full with untrusted entities. Thus, the data owner may only want to allow queries of a statistical or aggregate nature. This privacy requirement has become a common issue for large collections of sensitive data, with applications to census data, medical research, and educational testing [19]. For example, aggregate medical information about a group of patients may be accessible for research purposes. However, medical records of individual patients are confidential and should be kept hidden from all parties except for the hospital maintaining them [17].

An increasing trend in today's organizational data management is data outsourcing and cloud computing. An owner may choose to outsource the data, that is, to allow the data to be hosted by third-party service providers. The data hosts would be given the ability to store full or partial information from the database, and the capability to answer queries of a certain type. Data outsourcing alleviates the workload of the data owner in answering queries by delegating the tasks to powerful third-party servers with large computational and network resources.

However, data outsourcing poses additional privacy risks to the sensitive contents. The outsourcing service providers may not be fully trusted by the data owner, or may be susceptible to attacks by malicious parties (both internal and external). Studies have shown that in an outsourced setting it is extremely easy for malicious employees at the service provider organization to access the passwords of business owners and thus their customer data [5]. Security breaches at providers caused by outside adversaries may expose sensitive hosted information.

However, existing database-as-a-service (DAS) models are unable to support sophisticated queries such as aggregation while simultaneously maintaining the secrecy of microdata (i.e., individual data entries). Existing approaches based on the encryption of outsourced contents [1,31] apply to models where the user who queries the encrypted outsourced data is the data owner herself. We consider a more general setting where the database can be queried by anyone. Thus, there is a gap between the security guarantees provided by existing data outsourcing systems and the privacy needs of the data owners. To protect sensitive data from these threats, it is desirable to outsource the data in such a way that *aggregate queries can be computed without revealing microdata to service providers*. This paper presents a solution that realizes this goal.

Cross-domain collaborative data analysis is another application that motivates our work. For example, multiple regional hospitals collaborate to discover the most frequently occurring flu strain of the season in that area. Existing solutions that support multi-party privacy-preserving data mining require either a trusted or semi-trusted third-party to moderate the computation [28] or the active online participation of players in order to complete the computation [6,33]. Neither approach provides a practical solution that can be deployed and operated in a completely decentralized fashion. As it will soon become clear, we aim to realize