Authenticating Aggregate Queries over Set-Valued Data with Confidentiality (Extended Abstract)

Cheng Xu*, Qian Chen*, Haibo Hu‡, Jianliang Xu*, Xiaojun Hei§

* *Department of Computer Science, Hong Kong Baptist University, Hong Kong*
‡ *Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong*
§ *School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China*
{chengxu, qchen, xujl}@comp.hkbu.edu.hk, haibo.hu@polyu.edu.hk, heixj@hust.edu.cn

Abstract—**With recent advances in data-as-a-service (DaaS) and cloud computing, aggregate query services over set-valued data are becoming widely available for business intelligence that drives decision making. However, as the service provider is often a third-party delegate of the data owner, the integrity of the query results cannot be guaranteed and is thus imperative to be authenticated. Unfortunately, existing query authentication techniques either do not work for set-valued data or they lack data confidentiality. In this paper, we propose authenticated aggregate queries over set-valued data that not only ensure the integrity of query results but also preserve the confidentiality of source data.**

I. Introduction

With recent advances in data-as-a-service (DaaS) and cloud computing, aggregate query services over set-valued data are becoming widely available for business intelligence, scientific research, and government policy study. For example, personal genomics analysis (e.g., 23andMe and the Personal Genome Project (PGP) at Harvard Medical School [1]) is based on aggregate queries on large genome datasets, the integrity of whose results is vital. Below is one example:

PID	**ZIP**	**Mut-Genes**
P1	95014	A-C130R, P-I696M
P2	20482	H-C282Y, P-P12A, R-G1886S
P3	95014	A-C130R, U-G71R, W-R611H
P4	01720	A-V2050L, H-C282Y, M-R52C, U-G71R
P5	20134	A-C130R, P-P12A, R-G1886S, S-E366K
P6	17868	C-R102G, R-G1886S
P7	55410	C-R102G, C-Q1334H, S-E288V
P8	20852	C-R102G, P-P12A, R-G1886S, K-T220M

TABLE I: Set-Valued Genome Dataset

Example 1: **Aggregate Queries on PGP Data.** Table I shows a sample genome dataset, where *PID* is participant ID, *ZIP* is ZIP code, and *Mut-Genes* is a sensitive set-valued attribute that records the mutation genes of each participant. Users (e.g., medical doctors) may be interested in the following aggregate queries:

- ***Q1****: Find the most common gene in the district of Cupertino, CA (ZIP: 95014).*
- ***Q2****: Count the number of participants who carry the gene 'R-G1886S'.*
- ***Q3****: Find the most frequent genes with supports $\geq$ 3 in ZIPs 20***.*

The corresponding query results are: {'A-C130R'}, 4, and {'P-P12A', 'R-G1886S'}, respectively.

However, as the service provider is often a third-party delegate of the data owner, the integrity of the query results cannot be guaranteed and is thus imperative to be authenticated. A large body of research on authenticated query processing has been carried out to verify the integrity of query results [2], [3], [4]. Nevertheless, only a few query studies target set-valued data. Papamanthou *et al.* proposed efficient verification protocols for primitive set operations [5]. A verifiable frequent itemset mining algorithm is developed by Dong *et al.* [6]. Unfortunately, none of the previous studies has considered the data confidentiality requirement.

In this paper, we study authenticated aggregate query services over set-valued data with confidentiality preservation. We make the following contributions. First, to the best of our knowledge, this is the first work that addresses both integrity and confidentiality for aggregate queries over set-valued data. Second, the proposed privacy-preserving authentication framework supports various aggregate queries by introducing privacy-preserving verification protocols on multiset operations. Finally, we propose several optimizations to enhance the performance and conduct both theoretical and empirical analysis.

II. Problem Definition

A *data owner* (DO) owns a dataset $\mathbb{D} = \{o_1, o_2, \cdots, o_n\}$. Each object o_i is represented by $<A_i, X_i>$, where A_i is a set of non-sensitive attributes and X_i is a sensitive multiset of *features*. The DO outsources $\mathbb{D}$ to a third-party *service provider* (SP), together with an *authenticated data structure* (ADS) signed with the DO's private key. Based on this, the SP provides aggregate query services to clients (e.g., **Q1** and **Q2**, and **Q3** in Example 1). Our study mainly focuses on the following aggregate queries: (i) *count/sum*, which sums or counts the multiplicities of the queried feature in all selected objects; (ii) *max/min*, which selects the feature with the maximum or minimum summed multiplicity in all selected objects; (iii) *top-k*, which selects top-k features with the largest summed multiplicities in all selected objects; and (iv) *frequent feature query (FFQ)*, which selects the features whose summed multiplicities in all selected objects are no less than a threshold.

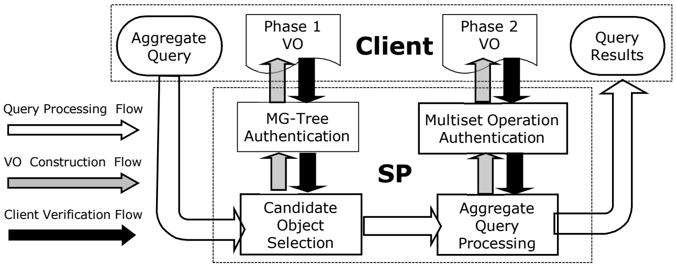


Fig. 1: Authentication Framework Overview

We consider two potential security threats: (i) the SP could provide unfaithful query execution, thereby returning incorrect or incomplete query results; and (ii) data privacy could be breached if sensitive source data are disclosed to the query client.

III. Summary of the Proposed Framework

Fig. 1 illustrates both the query and authentication flow charts of the proposed privacy-preserving authentication framework for aggregate queries. The framework consists of two phases: candidate object selection and aggregate query processing. The key components are outlined below.

A. Privacy-Preserving Authentication on Candidate Object Selection

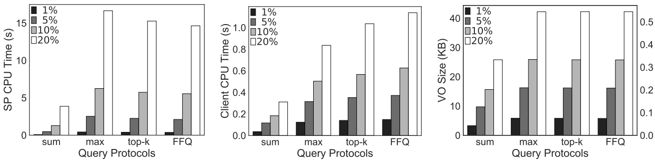
The candidate object selection is responsible to process the range selection criteria over the non-sensitive attributes. The results of this phase serve as the input for the following aggregate query processing. Here, we propose *Merkle Grid tree* (MG-tree), an ADS for the DO to construct and sign. The MG-tree partitions the space of non-sensitive attributes recursively into multiple levels of grid cells. Every tree node, which corresponds to a grid cell, stores the bounding box of the cell and a digest. The digest is computed from its child nodes and can be used to authenticate the candidate objects.

B. Privacy-Preserving Authentication Protocols on Multiset Operations

In our framework, we express multiset as a randomized bilinear-map accumulator. Specifically, given a multiset X, its accumulative value is $acc(X) = g^{r_X \cdot \prod_{x \in X}(x+s)}$. Here, g is a group generator of a cyclic multiplicative group, s is a secret known only by the DO, and r_X is a random value hidden from the query client but disclosed to the SP. A nice property of $acc(X)$ is that even without knowing the secret s, it can still be computed from X and the public seeds $g, g^s, g^{s^2}, \ldots$ through polynomial interpolation. By leveraging the properties of this accumulator and bilinear pairing [7], we propose five core privacy-preserving authentication protocols on multiset operations, namely, $subset$, sum, $empty$, $union$, and $times$.

C. Privacy-Preserving Authentication Algorithms on Aggregate Queries

The results of the *count* and *sum* queries can be verified by performing: (i) the inflation checking: $R \subseteq S$; and (ii) the deflation checking: $(S - R) \cap R = \emptyset$. Here, R denotes the query result multiset and S is the sum of the candidate objects. For *max*, *top-k* and *FFQ* queries, their results can be verified by the above checks and an additional completeness check $(S - R) \subseteq \tau \cdot (U - \widehat{R})$. Here, τ is the multiplicity threshold based on the query semantic, U denotes the union of the candidate objects, and $\widehat{R}$ is the set version of multiset R. The *min* query is similar except that we verify $(S - R) \supseteq \tau \cdot (U - \widehat{R})$ in the completeness checking.



(a) SP CPU Time (b) Client CPU Time (c) VO Size

Fig. 2: Query Performance vs. Selectivity on PGP dataset

IV. Performance and Security Analysis

Fig. 2 shows the query performance of our proposed framework on the PGP dataset, which contains personal genome data of 600 participants. It is observed that all the costs are generally sublinear to the query range. Another key observation is that both the client time cost and the VO size are determined only by the result set size and are independent of the number of sensitive features. We obtain similar result for other larger datasets [8].

With regard to the security, the proposed framework guarantees correctness and soundness of the query results. The confidentiality of the sensitive features is also preserved. The detailed security analysis is available in the full version of this paper [8].

Acknowledgement

This work was supported by Research Grants Council (RGC) of Hong Kong under GRF Projects 12244916, 15238116, 12202414, 12200114, 12200914, CRF Project C1008-16G, and NSFC Grant 61370231. Haibo Hu is also supported by PolyU Start-up Fund.

References

[1] "Personal genome project," http://www.personalgenomes.org, 2016.

[2] H. Hu, J. Xu, Q. Chen, and Z. Yang, "Authenticating location-based services without compromising location privacy," in *Proc. SIGMOD*, 2012, pp. 301–312.

[3] Q. Chen, H. Hu, and J. Xu, "Authenticated online data integration services," in *Proc. SIGMOD*, 2015, pp. 167–181.

[4] C. Xu, J. Xu, H. Hu, and M. H. Au, "When query authentication meets fine-grained access control: A zero-knowledge approach," in *Proc. SIGMOD*, 2018.

[5] C. Papamanthou, R. Tamassia, and N. Triandopoulos, "Optimal verification of operations on dynamic sets," in *Proc. CRYPTO*, 2011, pp. 91–110.

[6] B. Dong, R. Liu, and W. Wang, "Integrity verification of outsourced frequent itemset mining with deterministic guarantee," in *Proc. ICDM*, 2013, pp. 1025–1030.

[7] L. Nguyen, "Accumulators from bilinear pairings and applications," in *Proc. CT-RSA*, 2005, pp. 275–292.

[8] C. Xu, Q. Chen, H. Hu, J. Xu, and X. Hei, "Authenticating aggregate queries over set-valued data with confidentiality," *IEEE Trans. Knowl. Data Eng.*, 2017.