# A Generalized Framework for Preserving Both Privacy and Utility in Data Outsourcing (Extended Abstract)

Shangyu Xie[1], Meisam Mohammady[2], Han Wang[1], Lingyu Wang[3], Jaideep Vaidya[4] and Yuan Hong[1]
[1]Illinois Institute of Technology [2]CSIRO Data61 [3]Concordia University [4]Rutgers University

2022 IEEE 38th International Conference on Data Engineering (ICDE) | 978-1-6654-0883-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDE53745.2022.00151

***Abstract*—In this paper, we propose a prefix-preserving encryption based data outsourcing framework which is applicable to multiple different types of data, such as geo-locations, market basket data, DNA sequences, numerical data and timestamps. It enables accurate data analyses on the encrypted data while ensuring strong privacy against inference attacks. The basic idea is to generates multiple *indistinguishable* data views in which one view fully preserves the utility for data analysis, and its accurate analysis result can be obliviously retrieved. We empirically evaluate the performance of our outsourcing framework against two common inference attacks on two different real datasets: the check-in location dataset and network traffic dataset, respectively. The experimental results demonstrate that our proposed framework preserves both privacy (with bounded leakage and indistinguishability of data views) and utility.**

***Index Terms*—Privacy, Prefix Preserving, Utility, Outsourcing**

## I. Introduction

With the significant development of cloud computing, an increasing number of data-related services have been prevalently outsourced to the cloud. This may result in immense privacy concerns with severe consequences for the enterprises. Many encryption algorithms have been proposed to protect the outsourced data. Different from homomorphic encryptions [1] which are computationally expensive and impractical for complicated analyses, property preserving encryptions (PPE) [2], [3] have enabled service providers to directly perform efficient and accurate data analyses on the encrypted data. For instance, order preserving encryption (OPE) [3] retains the order of the plaintexts in the ciphertexts; in prefix preserving encryption [2], if any two plaintexts share a prefix, the ciphertexts will share the same length of prefixes.

However, most existing PPE schemes [2], [3] have two major limitations. First, PPE is typically limited to specific data or applications. For instance, OPE is limited in range queries on numerical data while prefix-preserving encryption (i.e., CryptoPAn [2]) is only applicable to IP addresses. Second, most PPEs are vulnerable to various forms of inference attacks [4], which attempt to link the encrypted data to the original data with background knowledge or auxiliary data.

In this paper, we first propose a novel *prefix-aware encoding* scheme to encode a variety of data types into bit strings (viz. bit strings with prefix-based utility) for the prefix preserving encryption. Essentially, if any data is naturally hierarchical (e.g., IP addresses as bit strings) or can be indexed by a prefix-aware tree (e.g., location data, DNA sequences, and market basket data), the prefixes in the encoded bit strings could be fully preserved to ensure utility when directly analyzing the outsourced data. For instance, the distance between any two locations can be fully preserved in the outsourced data. Second, we design a general multi-view outsourcing framework which can significantly reduce the information leakage in two folds: 1) generating multiple *indistinguishable* data views; 2) bounding the overall information leakage. We empirically evaluate the performance of the generalized outsourcing framework against various inference attacks [4] on real datasets (the check-in location dataset and the network traffic dataset). The experimental results demonstrate that our framework preserves both privacy (with bounded leakage and indistinguishability of data views) and full analysis utility.

This work is partially supported by the National Science Foundation (NSF) under the Grants No. CNS-2046335 and CNS-2034870.

## II. Model Formulation

### A. Generalized CryptoPAn and Prefix-aware Encoding

CryptoPAn [2] was originally designed to generate deterministic ciphertexts for IP addresses (32-bit IPv4 or 128-bit IPv6). The utility of preserving prefixes in the encrypted IP can be realized since the ciphertexts can preserve all the original subnet structure (sharing a prefix in the original IP addresses also results in the same length of shared prefix in the encrypted IP addresses). In our framework, we generalize CryptoPAn to encrypt any length of bits (indexed for different data types).

**Prefix-aware Encoding.** Motivated by such prefix properties, we can encode other data types into prefix-aware bits, e.g., geo-location data, DNA sequences, items in market baskets, and timestamps using a ***prefix-aware tree*** (fully preserving the utility for performing analyses on the encrypted data).

### B. Threat Model and Privacy Property

We assume a *honest-but-curious* service provider, which has possessed background knowledge (e.g., the set of attributes in the outsourced data, and the domain for the attributes) to implement inference attacks [4]. In addition, all the communications are in secure channel. We also define two privacy properties in the data outsourcing framework;

- Indistinguishability: ensures that all the data views are *indistinguishable* (inspired from differential privacy [5]).

2375-026X/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDE53745.2022.00151

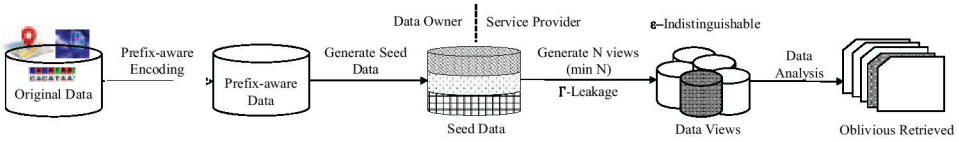



Fig. 1. Generalized Multi-view Outsourcing Framework.

- Bounded leakage: the total leakage of all the encrypted data (generated based on all the data views) against the adversary's inference attacks is bounded.

## III. Generalized Multi-view Framework

### A. Overview

The framework (see Figure 1) involves the data owner and the service provider, e.g., a company who provides analysis services on the cloud. The data owner first encodes its private data (e.g., locations and IP addresses) into prefix-aware data. Then, the data owner encrypts and generates a seed data with *prefix-based* partitions, which will be outsourced to the service provider along with other related parameters. Next, the service provider will generate $N$ data views on the seed data, and perform the analysis on all the views, *only one* out of which is the real view (with the true analysis result). Finally, they will leverage oblivious random access memory (ORAM) protocol [6] to privately retrieve the analysis results.

### B. Prefix-based Partition

The initially encrypted data is partitioned by assigning all the values sharing at least $x$-bit prefix into the same partition. However, the adversary can potentially identify the real data view from such partitioning due to the collision property of CryptoPAn [2], [7]. We generalize such attack as *Subprefix Collision Attack*, which is caused by similar prefixes or subprefixes ("close prefix"). We define $\beta$-closeness for quantifying such close prefix [7]. To address such attack, the proposed scheme aims to create more similar collisions among $\beta$-close partitions while generating multiple data views.

### C. Multiple Data Views Generation

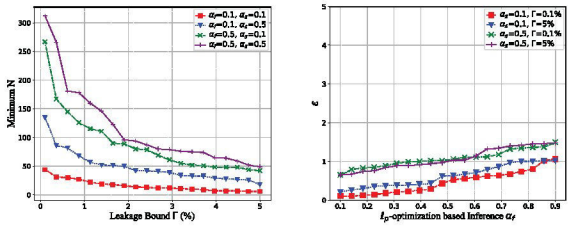
To generate the seed data, the framework applies CryptoPAn to obfuscate different partitions of prefix-aware data with random number of iterations. Then, the service provider will generate multiple data views on the seed data, which could reduce the leakage against inference attacks since the probabilities of matching the encrypted data to the original data can be greatly reduced with more data views. To further mitigate the collision attacks, we apply CryptoPAn for the same times (pseudorandom) on the $\beta$-close partitions in the fake data views, and apply CryptoPAn for different times (also pseudorandom) for partitions with different subprefixes (collisions may naturally occur in this case).

*1) Minimum $N$ with Bounded $\Gamma$-Leakage:* Our multi-view framework aims to seek a minimum $N$. Specifically, before partitioning the data, the data owner can empirically find an $x$ such that the required number of data views $N$ (for bounding the total leakage with $\Gamma$) is minimized – searching the $x$ and minimum $N$ takes $O(n\log(n))$ since the leakage derived from the fixed inference attacks is anti-monotonic on $N$ [7].

*2) Privacy Analysis:* In practice, an adversary will exploit any related information (received data, background knowledge, etc.) to identify if a data view is the real or fake one. We have derived that the data views satisfy $\epsilon$-indistinguishability [7].

## IV. Experimental Evaluations

We conducted experiments on two real datasets: check-in location from a social network and network traffic data collected from DoS attacks (see details in [7]). We have evaluated the bounded leakage of our framework against the inference attacks. Figure 2(b) presents the required minimum number of data views $N$ on the encrypted location data. If the leakage bound $\Gamma$ increases (from $0.1\%$ to $5\%$), the required minimum number $N$ declines from $\sim 300$ to $\sim 50$ (against strong attackers who know more percent of background knowledge with higher $\alpha_s$ and $\alpha_f$ [7]), and declines from $\sim 50$ to $\sim 5$ (against weak attackers). We also demonstrate the indistinguishablity bound $\epsilon$ w.r.t. various attack parameters. In Figure 2(a), $\epsilon$ increases as $\alpha_s$ or $\alpha_f$ grows. This indicates that a stronger attacker (with more knowledge) would be more likely to identify the real data view. However, $\epsilon$ is relatively small ($\leq 1.5$) even if the adversary is strong (with more knowledge).




(a) $N$ vs. $\Gamma$ (b) $\epsilon$ vs. Attack Parameter

Fig. 2. Minimum $N$ (a) and Indistinguishability (b) on Location Data

## References


[1] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *STOC*, 2009.

[2] J. Xu, J. Fan, M. H. Ammar, and S. B. Moon, "Prefix-preserving ip address anonymization: Measurement-based security evaluation and a new cryptography-based scheme," in *ICNP*, IEEE, 2002.

[3] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in *SIGMOD*, ACM, 2004.

[4] M. Naveed, S. Kamara, and C. V. Wright, "Inference attacks on property-preserving encrypted databases," in *CCS*, ACM, 2015.

[5] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12.

[6] X. S. Wang, Y. Huang, T.-H. H. Chan, A. Shelat, and E. Shi, "Scoram: Oblivious ram for secure computation," in *CCS*, ACM, 2014.

[7] S. Xie, M. Mohammady, H. Wang, L. Wang, J. Vaidya, and Y. Hong, "A generalized framework for preserving both privacy and utility in data outsourcing," *TKDE*, IEEE, 2021.