Short Paper Track

Group-wise K-anonymity meets (ε, δ) Differentially Privacy Scheme

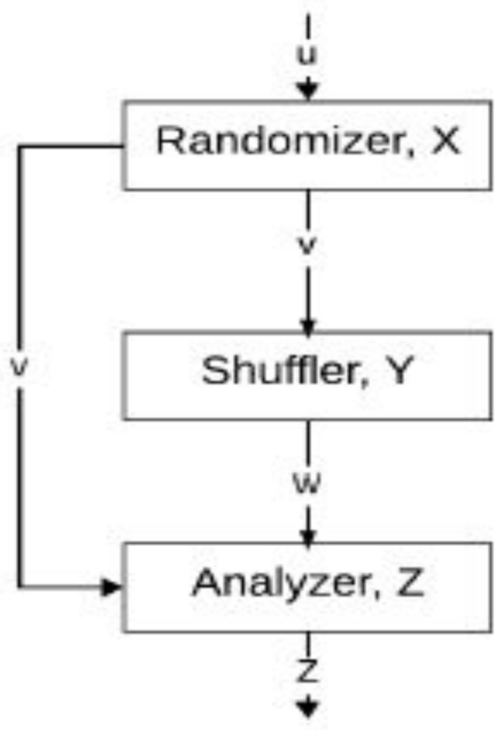
Kenneth Odoh

https://kenluck2001.github.io

Objectives

The work explores the intersection of **K-anonymity** and **differential privacy** to develop a novel method for estimating **noise levels** in privacy-preserving data schemes. We offer a fourfold contributions:

- Using the **birthday-bound paradox** for estimating noise levels in differential privacy schemes
- Proposing a **group-aware formulation** to enhance resilience against inference attacks
- Drawing connections to the attacker's advantage with the noise level in both **univariate and multivariate cases**.
- We present a case study demonstrating the applicability of our formulation in Laplacian, Gaussian, and Exponential mechanisms.

Architecture	DP Mechanism (Phases)
	• Randomizer, $X : u \to v$, where u is the original secret data, and v is the transformed output forwarded to the shuffler. Perturb the input data, u, by adding noise via the X routine. An example of a randomizer is $\mathcal{A}_q$ in Definition 1. • Shuffler, $Y : v \to w$, (optional) where v is transformed data from the randomizer, X, and w is the intermediary transformed output forwarded to the analyzer phase. Permute the data, v, utilizing the Y routine. • Analyzer, $Z : w \to z$, $Z : v \to z$ where v, w is transformed data from randomizer and shuffler respectively. z is the output of the privacy protocol, and we calculate aggregate statistics.

K-anonymity

K-anonymity provides a way to achieve data privacy where each record is similar to any corresponding set of at least k – 1 other records.

- K-anonymity is related to the **birthday-bound formulation,** which follows the pigeonhole principle.
- The birthday-bound paradox explains this example. In a group of 23 people, there is at least a 50% chance that at least two individuals have the same birthday.

$$\pi(k,N) = \frac{N!}{(N-K)!N^k}$$

$\pi(k,N)$ is the uniqueness probability where k individuals are unique from a population size, N,

Estimating noise level in (ε, δ) Differential Privacy Schemes

$$R := \max_{x \in X, x' \in X'} d(x, x') \qquad \epsilon = \frac{-\ln\left(\frac{p}{1-p}\cdot\left(\frac{1}{\delta+p}-1\right)\right)}{R}$$

Where X, X' are rows of data. We provide a noise estimation, ε with probability, p, set as $\pi(k,N)$, δ is guessing advantage, and data sensitivity, R, as a scaling factor for a **univariate case**.

For **multivariate case**, we discuss the following:

- AND-events
- OR-events.

Where k_i is the number of unique elements in a group, N_i is the number of elements in a group, n_{group} is the total number of groups, and $n = \sum_{i=1}^{n_{group}} N_i$ as number of records across every group.

Limitations / Future Work

- The paper assumes pairwise independence in its noise estimation method, which may limit its applicability in scenarios where dependencies exist among variables.
- Future work can incorporate adversarial uncertainty uses inherent variance as noise, making smaller noise values necessary to achieve suitable privacy guarantees.
- Increase data size for evaluation.

Multivariate case: AND-events

Adversary can reconstruct every field with the guessing advantage, δ,

$$\epsilon \le \frac{-\ln\left(\frac{\prod_{i=1}^{n_{group}} \pi(k_i,N_i)}{1-\prod_{i=1}^{n_{group}} \pi(k_i,N_i)}\cdot\left(\frac{1}{\delta+\prod_{i=1}^{n_{group}} \pi(k_i,N_i)}-1\right)\right)}{R}.$$

Where $R = \| R_1, \ldots, R_n \|_\infty$.

Conclusions

- The paper introduces a novel privacy scheme that merges group-wise K-anonymity with (ε, δ) differential privacy to enhance data protection in aggregation processes.
- We proposed a method for adjusting noise based on group characteristics to maintain utility while ensuring privacy. The approach utilizes the **birthday paradox** for estimating uniqueness within groups, offering a balance between data utility and privacy by dynamically adjusting the noise added to aggregated data.
- This scheme is particularly aimed at providing a theoretical foundation and potential for practical application in privacy-preserving data analysis.

Multivariate case: OR-events

Adversary can reconstruct at least one of the attributes with the guessing advantage, δ,

$$\epsilon_i \le \frac{-\ln\left(\frac{\pi(k_i,N_i)}{1-\pi(k_i,N_i)}\cdot\left(\frac{1}{\delta+\pi(k_i,N_i)}-1\right)\right)}{R}$$

The estimated noise, $\epsilon = \min i \in \{0, 1, \ldots, n_{group}\} (\epsilon_i)$

Where $R = \| R_1, \ldots, R_n \|_\infty$.