(ℓ, δ)-Diversity: Linkage-Robustness via a Composition Theorem

Arvind Rameshwar

Dept. of EE, IIT Madras

in collaboration with

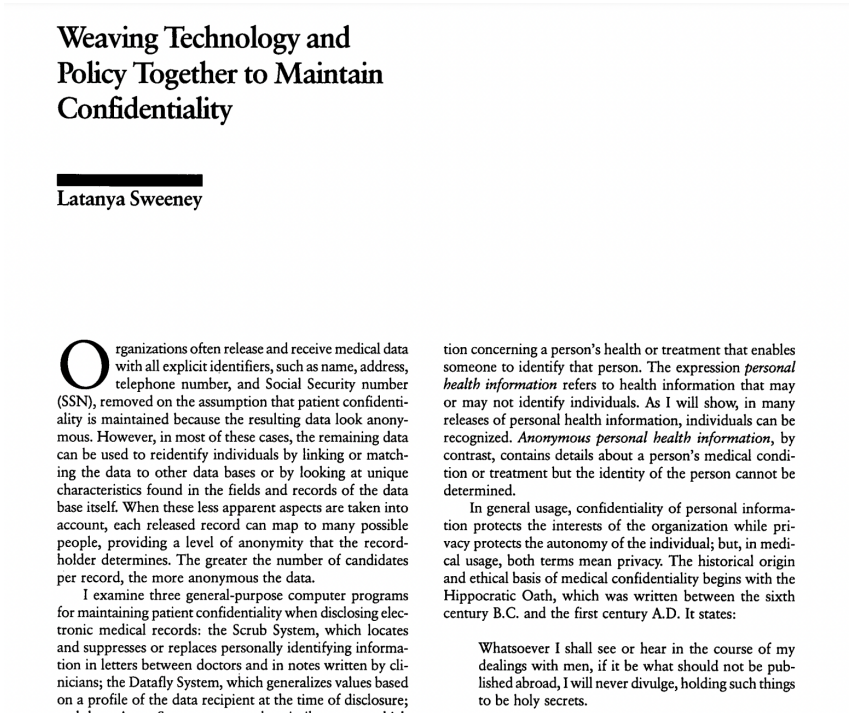
Anshoo Tandon

Centre of Data for Public Good, IISc

HyTS 2026

Some background: Privacy

- Privacy of users is a fundamental desideratum in the release of datasets for statistical analysis.



Weaving Technology and Policy Together to Maintain Confidentiality

Latanya Sweeney

Organizations often release and receive medical data with all explicit identifiers, such as name, address, telephone number, and Social Security number (SSN), removed on the assumption that patient confidentiality is maintained because the resulting data look anonymous. However, in most of these cases, the remaining data can be used to reidentify individuals by linking or matching the data to other data bases or by looking at unique characteristics found in the fields and records of the data base itself. When these less apparent aspects are taken into account, each released record can map to many possible people, providing a level of anonymity that the record-holder determines. The greater the number of candidates per record, the more anonymous the data.

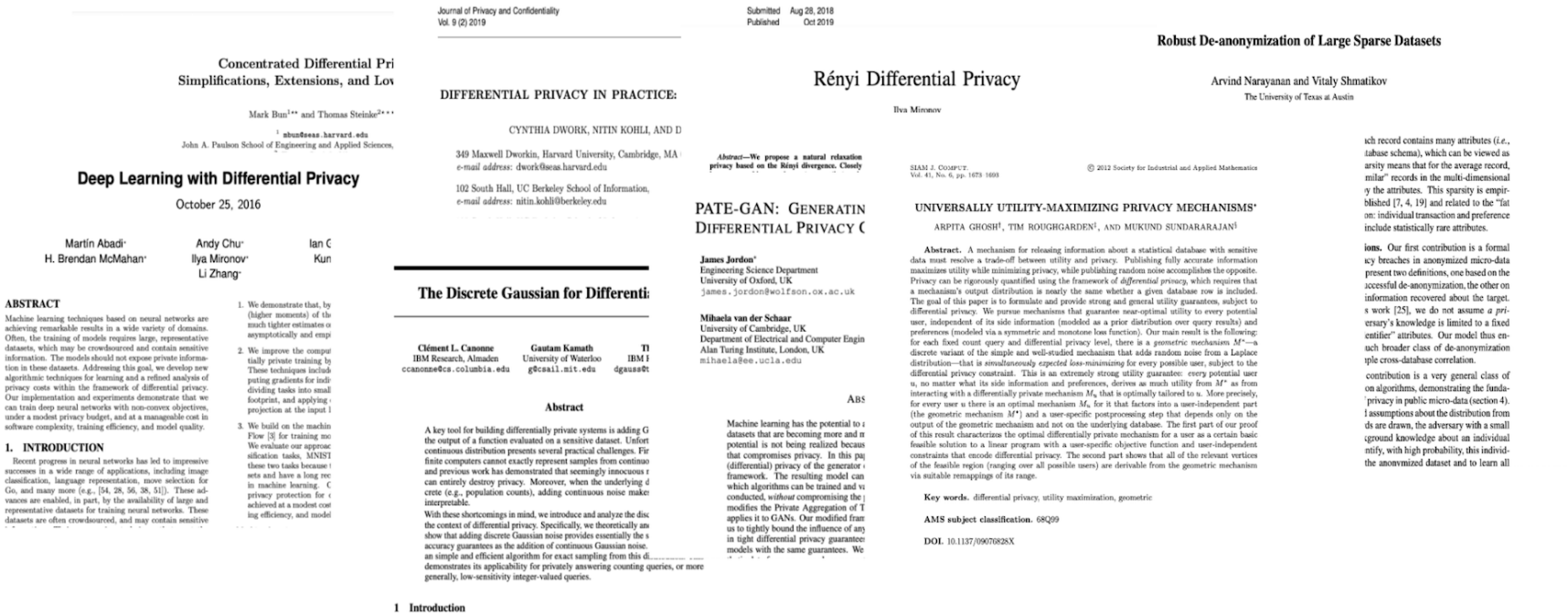
I examine three general-purpose computer programs for maintaining patient confidentiality when disclosing electronic medical records: the Scrub System, which locates and suppresses or replaces personally identifying information in letters between doctors and in notes written by clinicians; the Datafly System, which generalizes values based on a profile of the data recipient at the time of disclosure; tion concerning a person's health or treatment that enables someone to identify that person. The expression *personal health information* refers to health information that may or may not identify individuals. As I will show, in many releases of personal health information, individuals can be recognized. *Anonymous personal health information*, by contrast, contains details about a person's medical condition or treatment but the identity of the person cannot be determined.

In general usage, confidentiality of personal information protects the interests of the organization while privacy protects the autonomy of the individual; but, in medical usage, both terms mean privacy. The historical origin and ethical basis of medical confidentiality begins with the Hippocratic Oath, which was written between the sixth century B.C. and the first century A.D. It states:

> Whatsoever I shall see or hear in the course of my dealings with men, if it be what should not be published abroad, I will never divulge, holding such things to be holy secrets.

- Differential privacy (DP) [Dwork et al. (2006)] is now widely adopted as the "gold standard" of privacy, for aggregate statistics/learning outcomes release.

Some background: Privacy research



Concentrated Differential Pri
Simplifications, Extensions, and Lov

Mark Bun[1]** and Thomas Steinke[2]***

[1] mbun@seas.harvard.edu
John A. Paulson School of Engineering and Applied Sciences,

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi* Andy Chu* Ian G
H. Brendan McMahan* Ilya Mironov* Kun
Li Zhang*

ABSTRACT

Machine learning techniques based on neural networks are achieving remarkable results in a wide variety of domains. Often, the training of models requires large, representative datasets, which may be crowdsourced and contain sensitive information. The models should not expose private information in these datasets. Addressing this goal, we develop new algorithmic techniques for learning and a refined analysis of privacy costs within the framework of differential privacy. Our implementation and experiments demonstrate that we can train deep neural networks with non-convex objectives, under a modest privacy budget, and at a manageable cost in software complexity, training efficiency, and model quality.

1. INTRODUCTION

Recent progress in neural networks has led to impressive successes in a wide range of applications, including image classification, language representation, move selection for Go, and many more (e.g., [54, 28, 56, 38, 51]). These advances are enabled, in part, by the availability of large and representative datasets for training neural networks. These datasets are often crowdsourced, and may contain sensitive

1. We demonstrate that, by (higher moments) of th much tighter estimates o asymptotically and empi
2. We improve the compu tially private training by These techniques includ puting gradients for indi dividing tasks into smal footprint, and applying projection at the input
3. We build on the machin Flow [3] for training mo We evaluate our approac sification tasks, MNIST these two tasks because neta and have a long re in machine learning. privacy protection for achieved at a modest cos ing efficiency, and model

Journal of Privacy and Confidentiality
Vol. 9 (2) 2019

Submitted Aug 28, 2018
Published Oct 2019

DIFFERENTIAL PRIVACY IN PRACTICE:

CYNTHIA DWORK, NITIN KOHLI, AND D

349 Maxwell Dworkin, Harvard University, Cambridge, MA
e-mail address: dwork@seas.harvard.edu

102 South Hall, UC Berkeley School of Information,
e-mail address: nitin.kohli@berkeley.edu

The Discrete Gaussian for Differenti

Clément L. Canonne
IBM Research, Almaden
ccanonne@cs.columbia.edu

Gautam Kamath
University of Waterloo
g@csail.mit.edu

Abstract

A key tool for building differentially private systems is adding G the output of a function evaluated on a sensitive dataset. Unfort continuous distribution presents several practical challenges. Fir finite computers cannot exactly represent samples from continuo and previous work has demonstrated that seemingly innocuous r can entirely destroy privacy. Moreover, when the underlying d crete (e.g., population counts), adding continuous noise make interpretable.
With these shortcomings in mind, we introduce and analyze the disc the context of differential privacy. Specifically, we theoretically an show that adding discrete Gaussian noise provides essentially the s accuracy guarantees as the addition of continuous Gaussian noise. an simple and efficient algorithm for exact sampling from this di demonstrates its applicability for privately answering counting queries, or more generally, low-sensitivity integer-valued queries.

1 Introduction

Rényi Differential Privacy

Ilya Mironov

Abstract—We propose a natural relaxation privacy based on the Rényi divergence. Closely

PATE-GAN: GENERATIN
DIFFERENTIAL PRIVACY (

James Jordon*
Engineering Science Department
University of Oxford, UK
james.jordon@wolfson.ox.ac.uk

Mihaela van der Schaar
University of Cambridge, UK
Department of Electrical and Computer Engin
Alan Turing Institute, London, UK
mihaela@ee.ucla.edu

ABS

Machine learning has the potential to a datasets that are becoming more and m potential is not being realized becaus that compromises privacy. In this pa (differential) privacy of the generator framework. The resulting model can which algorithms can be trained and v conducted, *without* compromising the modifies the Private Aggregation of T applies it to GANs. Our modified fram us to tightly bound the influence of any in tight differential privacy guarantee models with the same guarantees. We

SIAM J. COMPUT.
Vol. 41, No. 6, pp. 1673–1693

© 2012 Society for Industrial and Applied Mathematics

UNIVERSALLY UTILITY-MAXIMIZING PRIVACY MECHANISMS*

ARPITA GHOSH†, TIM ROUGHGARDEN‡, AND MUKUND SUNDARARAJAN§

Abstract. A mechanism for releasing information about a statistical database with sensitive data must resolve a trade-off between utility and privacy. Publishing fully accurate information maximizes utility while minimizing privacy, while publishing random noise accomplishes the opposite. Privacy can be rigorously quantified using the framework of *differential privacy*, which requires that a mechanism's output distribution is nearly the same whether a given database row is included. The goal of this paper is to formulate and provide strong and general utility guarantees, subject to differential privacy. We pursue mechanisms that guarantee near-optimal utility to every potential user, independent of its side information (modeled as a prior distribution over query results) and preferences (modeled via a symmetric and monotone loss function). Our main result is the following: for each fixed count query and differential privacy level, there is a *geometric mechanism* M^*—a discrete variant of the simple and well-studied mechanism that adds random noise from a Laplace distribution—that is *simultaneously expected loss-minimizing* for every possible user, subject to the differential privacy constraint. This is an extremely strong utility guarantee: every potential user u, no matter what its side information and preferences, derives as much utility from M^* as from interacting with a differentially private mechanism M_u that is optimally tailored to u. More precisely, for every user u there is an optimal mechanism M_u for it that factors into a user-independent part (the geometric mechanism M^*) and a user-specific postprocessing step that depends only on the output of the geometric mechanism and not on the underlying database. The first part of our proof of this result characterizes the optimal differentially private mechanism for a user as a certain basic feasible solution to a linear program with a user-specific objective function and user-independent constraints that encode differential privacy. The second part shows that all of the relevant vertices of the feasible region (ranging over all possible users) are derivable from the geometric mechanism via suitable remappings of its range.

Key words. differential privacy, utility maximization, geometric

AMS subject classification. 68Q99

DOI. 10.1137/09076828X

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

ich record contains many attributes (*i.e.*, tabase schema), which can be viewed as arsity means that for the average record, milar" records in the multi-dimensional y the attributes. This sparsity is empir- blished [7, 4, 19] and related to the "fat on: individual transaction and preference include statistically rare attributes.

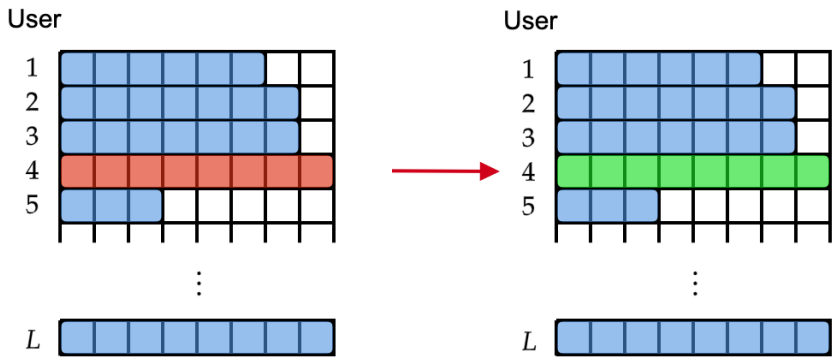
ions. Our first contribution is a formal acy breaches in anonymized micro-data present two definitions, one based on the accessful de-anonymization, the other on information recovered about the target. s work [25], we do not assume a *pri-* ersary's knowledge is limited to a fixed entifier" attributes. Our model thus en- uch broader class of de-anonymization uple cross-database correlation.

contribution is a very general class of on algorithms, demonstrating the funda- privacy in public micro-data (section 4). l assumptions about the distribution from ds are drawn, the adversary with a small ground knowledge about an individual ntify, with high probability, this individ- the anonymized dataset and to learn all

. . . and continuing to multiply

DP: A mathematically grounded framework for privacy

- We say that two datasets $\mathcal{D}_1, \mathcal{D}_2$ are neighbours if they differ in the sample values of a single user.

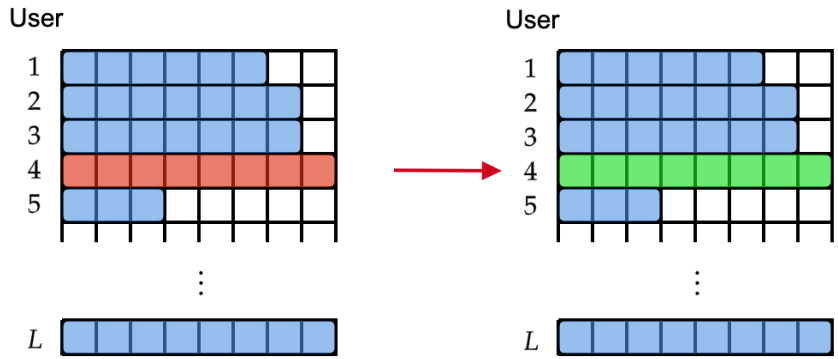


- A mechanism M is ε-DP if for every pair of datasets $\mathcal{D}_1, \mathcal{D}_2$ that are neighbours, and for every (measurable) Y,

$$e^{-\varepsilon} \Pr[M(\mathcal{D}_2) \in Y] \leq \Pr[M(\mathcal{D}_1) \in Y] \leq e^{\varepsilon} \Pr[M(\mathcal{D}_2) \in Y].$$

DP: A mathematically grounded framework for privacy

- We say that two datasets $\mathcal{D}_1, \mathcal{D}_2$ are neighbours if they differ in the sample values of a single user.



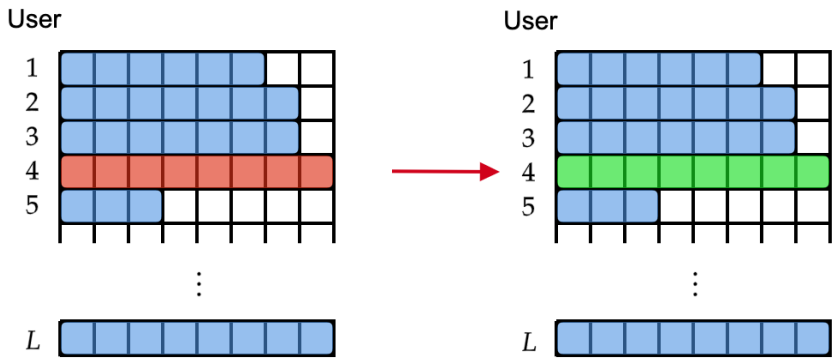
- A mechanism M is ε-DP if for every pair of datasets $\mathcal{D}_1, \mathcal{D}_2$ that are neighbours, and for every (measurable) Y,

$$e^{-\varepsilon} \Pr[M(\mathcal{D}_2) \in Y] \leq \Pr[M(\mathcal{D}_1) \in Y] \leq e^{\varepsilon} \Pr[M(\mathcal{D}_2) \in Y].$$

Think of $e^{\varepsilon} \approx 1 + \varepsilon$, for $\varepsilon > 0$ small

DP: A mathematically grounded framework for privacy

- We say that two datasets $\mathcal{D}_1, \mathcal{D}_2$ are neighbours if they differ in the sample values of a single user.



- A mechanism M is (ε, δ)-DP if for every pair of datasets $\mathcal{D}_1, \mathcal{D}_2$ that are neighbours, and for every (measurable) Y,

$$e^{-\varepsilon}\left(\Pr[M(\mathcal{D}_2) \in Y] - \delta\right) \leq \Pr[M(\mathcal{D}_1) \in Y] \leq e^{\varepsilon} \Pr[M(\mathcal{D}_2) \in Y] + \delta.$$

We must have $\delta \in (0, 1)$; think of $\delta \approx$ prob. of DP "failure"

A case for moving away from DP

- Data owners may often be incentivized to release "microdata," via the release of **entire datasets**.

One of the following alternatives can then be adopted (in theory):

- Add noise to every entry of the dataset and release
  [**Large noise addition $\equiv$ Poor accuracy!**]
- Select a "random" dataset that ensures high utility, via a DP selection mechanism
  [**High computational complexity!**]
- Generate and release DP synthetic datasets
  [**Could run into regulatory issues in practice!**]

A case for moving away from DP

- Data owners may often be incentivized to release "microdata," via the release of **entire datasets**.

One of the following alternatives can then be adopted (in theory):

- Add noise to every entry of the dataset and release
  [**Large noise addition $\equiv$ Poor accuracy!**]
- Select a "random" dataset that ensures high utility, via a DP selection mechanism
  [**High computational complexity!**]
- Generate and release DP synthetic datasets
  [**Could run into regulatory issues in practice!**]

In this work, we embrace early, heuristic notions of "**anonymity**", instead, and seek to formalize them.

Some background: Anonymization

The following heuristic notions of dataset anonymity are well-known:

- k-anonymity [Samarati and Sweeney (1998), Samarati (2001)]: Associated with every equivalence class (EC), there are $\geq k$ sensitive attributes
- ℓ-diversity [Machanavajjhala et al. (2007)]: Associated with every EC, there are $\geq \ell$ distinct sensitive attributes
- t-closeness [Li, Li, and Venkatasubramanian (2007)]: Distribution of sensitive attributes in any equivalence class is "close" to that in the overall dataset

We focus on the well-known problem of **dataset linkage**:

How do anonymity parameters degrade upon the linkage/composition of multiple datasets?

Glossary: Equivalence class (EC) $\equiv$ Binned/Anonymized quasi-identifiers

Dataset linkage: An example

Consider medical records in a hospital.

Gender	Postal Code	Disease
Male	600020	Heart disease
Male	600036	Lung infection
Female	600021	Osteoporosis
Female	600021	Cervical cancer
Female	600021	Osteoporosis

(a) Raw dataset

(Gender, Postal Code)	Disease
(Male, 600020) or (Male, 600036)	Heart disease Lung infection
(Female, 600021)	Osteoporosis Osteoporosis Cervical cancer

(b) Anonymized dataset that respects 2-anonymity and 2-diversity

Dataset linkage: An example

An adversary now has access to two, independently anonymized datasets, which he/she knows that $\underbrace{\text{(Female, 600021)}}_{\text{public quasi-identifier}}$ participates in.

(Gender, Postal Code)	Disease
(Male, 600020) or (Male, 600036)	Heart disease Lung infection
(Female, 600021)	Osteoporosis Cervical cancer

Anonymized dataset $\overline{\mathcal{D}}_1$

(Gender, Postal Code)	Disease
(Female, 600025)	Irritable bowel syndrome Lung infection
(Female, 600021)	Cervical cancer Amoebic dysentry

Anonymized dataset $\overline{\mathcal{D}}_2$

Dataset linkage: An example

An adversary now has access to two, independently anonymized datasets, which he/she knows that $\underbrace{\text{(Female, 600021)}}_{\text{public quasi-identifier}}$ participates in.

(Gender, Postal Code)	Disease
(Male, 600020) or (Male, 600036)	Heart disease Lung infection
(Female, 600021)	Osteoporosis Cervical cancer

Anonymized dataset $\overline{\mathcal{D}}_1$

(Gender, Postal Code)	Disease
(Female, 600025)	Irritable bowel syndrome Lung infection
(Female, 600021)	Cervical cancer Amoebic dysentry

Anonymized dataset $\overline{\mathcal{D}}_2$

Some unavoidable notation

Any dataset $\mathcal{D} = \{(\mathbf{q}_i, s_i) : i \in [N]\}$. We focus on ℓ-diversity as our notion of anonymity.

Some unavoidable notation

Any dataset $\mathcal{D} = \{(\mathbf{q}_i, s_i) : i \in [N]\}$. We focus on ℓ-diversity as our notion of anonymity.

Let $n^{(\mathbf{q})}(s) = \#$ samples with record $(\mathbf{q}, s)$.

Some unavoidable notation

Any dataset $\mathcal{D} = \{(\mathbf{q}_i, s_i) : i \in [N]\}$. We focus on ℓ-diversity as our notion of anonymity.

Let $n^{(\mathbf{q})}(s) = \#$ samples with record $(\mathbf{q}, s)$.

Definition

A dataset $\mathcal{D}$ satisfies ℓ-diversity, if $\ell \in [|\mathcal{S}|]$ is the largest integer such that for any vector $\mathbf{q} \in \mathcal{Q}$ with $\sum_{s \in \mathcal{S}} n^{(\mathbf{q})}(s) > 0$, we have $|\{s \in \mathcal{S} : \ n^{(\mathbf{q})}(s) > 0\}| \geq \ell$.

In order to satisfy ℓ-diversity, the data owner bins quasi-identifiers $\mathbf{q}$ into "equivalence classes" $\overline{\mathbf{q}}$, resulting in the anonymized dataset $\overline{\mathcal{D}}$.

Some unavoidable formalism

Any dataset $\mathcal{D} = \{(\mathbf{q}_i, s_i) : i \in [N]\}$. We focus on ℓ-diversity as our notion of anonymity.

- Datasets $\mathcal{D}_1, \ldots, \mathcal{D}_t$ (with a common user) are independently anonymized to yield $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$.

Some unavoidable formalism

Any dataset $\mathcal{D} = \{(\mathbf{q}_i, s_i) : i \in [N]\}$. We focus on ℓ-diversity as our notion of anonymity.

- Datasets $\mathcal{D}_1, \ldots, \mathcal{D}_t$ (with a common user) are independently anonymized to yield $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$.
- An adversary has access to $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$ and knows the quasi-identifier $\mathbf{q}$ of a particular user, who participates in **all** these datasets.

Some unavoidable formalism

Any dataset $\mathcal{D} = \{(\mathbf{q}_i, s_i) : i \in [N]\}$. We focus on ℓ-diversity as our notion of anonymity.

- Datasets $\mathcal{D}_1, \ldots, \mathcal{D}_t$ (with a common user) are independently anonymized to yield $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$.
- An adversary has access to $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$ and knows the quasi-identifier $\mathbf{q}$ of a particular user, who participates in **all** these datasets.
- The adversary identifies equivalence classes $\overline{\mathbf{q}}_i$, $i \in [t]$, which contain $\mathbf{q}$. Let

$$n^{(\mathbf{q})}_{[t]}(s) := \min\{ \underbrace{n^{(\overline{\mathbf{q}}_j)}(s)}_{\#s \text{ samples in } \overline{\mathbf{q}}_j} : j \in [t]\}.$$

Some unavoidable formalism

Any dataset $\mathcal{D} = \{(\mathbf{q}_i, s_i) : i \in [N]\}$. We focus on ℓ-diversity as our notion of anonymity.

- Datasets $\mathcal{D}_1, \ldots, \mathcal{D}_t$ (with a common user) are independently anonymized to yield $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$.
- An adversary has access to $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$ and knows the quasi-identifier $\mathbf{q}$ of a particular user, who participates in **all** these datasets.
- The adversary identifies equivalence classes $\overline{\mathbf{q}}_i$, $i \in [t]$, which contain $\mathbf{q}$. Let

$$n^{(\mathbf{q})}_{[t]}(s) := \min\{ \underbrace{n^{(\overline{\mathbf{q}}_j)}(s)}_{\#s \text{ samples in } \overline{\mathbf{q}}_j} : j \in [t]\}.$$

- The adversary constructs the linkage

$$\mathcal{L}^{(\mathbf{q})}_{[t]} := \{(s, n^{(\mathbf{q})}_{[t]}(s)) : n^{(\mathbf{q})}_{[t]}(s) > 0\}.$$

The linkage gives rise to a post-linkage dataset $\overline{\mathcal{D}}^{(\mathbf{q})}_{[t]}$.

Some unavoidable formalism

Any dataset $\mathcal{D} = \{(\mathbf{q}_i, s_i) : i \in [N]\}$. We focus on ℓ-diversity as our notion of anonymity.

- Datasets $\mathcal{D}_1, \ldots, \mathcal{D}_t$ (with a common user) are independently anonymized to yield $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$.
- An adversary has access to $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$ and knows the quasi-identifier $\mathbf{q}$ of a particular user, who participates in **all** these datasets.
- The adversary identifies equivalence classes $\overline{\mathbf{q}}_i$, $i \in [t]$, which contain $\mathbf{q}$. Let

$$n^{(\mathbf{q})}_{[t]}(s) := \min\{ \underbrace{n^{(\overline{\mathbf{q}}_j)}(s)}_{\#s \text{ samples in } \overline{\mathbf{q}}_j} : j \in [t]\}.$$

- The adversary constructs the linkage

$$\mathcal{L}^{(\mathbf{q})}_{[t]} := \{(s, n^{(\mathbf{q})}_{[t]}(s)) : n^{(\mathbf{q})}_{[t]}(s) > 0\}.$$

The linkage gives rise to a post-linkage dataset $\overline{\mathcal{D}}^{(\mathbf{q})}_{[t]}$. **Is $\overline{\mathcal{D}}^{(\mathbf{q})}_{[t]}$ ℓ'-diverse?**

Worst-case degradation of diversity

We are interested in computing the worst-case anonymity parameter of $\overline{\mathcal{D}}_{[t]}^{(\mathbf{q})}$, i.e.,

$$\ell_{[t]} := \min_{\overline{\mathcal{D}}_1,\ldots,\overline{\mathcal{D}}_t} \{\widehat{\ell} : \ \overline{\mathcal{D}}_{[t]}^{(\mathbf{q})} \text{ obeys } \widehat{\ell}\text{-diversity}\}.$$

Theorem

Under some very mild regularity conditions, for any $\ell \leq [|\mathcal{S}|]$, we have that

$$\ell_{[t]} = \begin{cases} 1, \ \textit{if } \ell \leq L \cdot \left(\frac{t-1}{t}\right) + 1, \\ L + 1 - (L - \ell + 1)t, \ \textit{otherwise.} \end{cases}$$

Worst-case degradation of diversity

We are interested in computing the worst-case anonymity parameter of $\overline{\mathcal{D}}^{(\mathbf{q})}_{[t]}$, i.e.,

$$\ell_{[t]} := \min_{\overline{\mathcal{D}}_1,\ldots,\overline{\mathcal{D}}_t} \{\widehat{\ell} : \ \overline{\mathcal{D}}^{(\mathbf{q})}_{[t]} \text{ obeys } \widehat{\ell}\text{-diversity}\}.$$

Theorem

Under some very mild regularity conditions, for any $\ell \leq \lceil |\mathcal{S}| \rceil$, we have that

$$\ell_{[t]} = \begin{cases} 1, \ \textit{if } \ell \leq L \cdot \left(\frac{t-1}{t}\right) + 1, \\ L + 1 - (L - \ell + 1)t, \ \textit{otherwise.} \end{cases}$$

Key take-away: The worst-case anonymity parameter is very poor (equals 1!) for most ℓ values.

. . . We hence need a different notion of anonymity that degrades gracefully upon linkage.

Proof idea

The proof makes use of a simple reduction to the analysis of arrays of binary strings.

- Given the equivalence classes $\overline{\mathbf{q}}_j$ in $\overline{\mathcal{D}}_j$, one can define

$$\iota_j := \left(\mathbb{1}\{ n^{(\overline{\mathbf{q}}_j)}(s) > 0 \} : s \in \mathcal{S} \right),$$

for $j \in [t]$. Clearly, we have that $w(\iota_j) \geq \ell$, for all $j \in [t]$.

Proof idea

The proof makes use of a simple reduction to the analysis of arrays of binary strings.

- Given the equivalence classes $\overline{\mathbf{q}}_j$ in $\overline{\mathcal{D}}_j$, one can define

$$\iota_j := \left(\mathbb{1}\{n^{(\overline{\mathbf{q}}_j)}(s) > 0\} : s \in \mathcal{S} \right),$$

  for $j \in [t]$. Clearly, we have that $w(\iota_j) \geq \ell$, for all $j \in [t]$.

- One can then define an array $M \in \{0,1\}^{t \times L}$ whose j^{th} row M_j is ι_j, with

$$\ell_{[t]} = w\left(\odot_{j=1}^{t} M_j\right).$$

- Observe that if $\ell > L \cdot \left(\frac{t-1}{t}\right) + 1$, then we cannot have at least one 0 in $L-1$ columns, resulting in $\ell_{[t]} > 1$.

Proof idea

The proof makes use of a simple reduction to the analysis of arrays of binary strings.

- Given the equivalence classes $\overline{\mathbf{q}}_j$ in $\overline{\mathcal{D}}_j$, one can define

$$\iota_j := \left(\mathbb{1}\{n^{(\overline{\mathbf{q}}_j)}(s) > 0\} : s \in \mathcal{S} \right),$$

  for $j \in [t]$. Clearly, we have that $w(\iota_j) \geq \ell$, for all $j \in [t]$.

- One can then define an array $M \in \{0,1\}^{t \times L}$ whose j^{th} row M_j is ι_j, with

$$\ell_{[t]} = w\left(\odot_{j=1}^{t} M_j\right).$$

- Observe that if $\ell > L \cdot \left(\frac{t-1}{t}\right) + 1$, then we cannot have at least one 0 in $L-1$ columns, resulting in $\ell_{[t]} > 1$.

The proof then proceeds by arguing a similar "converse" for other ℓ values, and constructing sequences that achieve the stated $\ell_{[t]}$.

(ℓ, δ)-Diversity: A definition

Consider the setting of datasets constructed from samples from volunteering users.

- A good assumption here is that the samples $(\mathbf{q}_i, s_i)$ are i.i.d. across $i \in [N]$, from a joint distribution $P_{\mathbf{Q},S} = P_{\mathbf{Q}} \cdot P_{S|\mathbf{Q}}$.
- WLOG, we assume that $P_{\mathbf{Q}} \succ 0$ and $P_S \succ 0$.

Definition

A dataset $\mathcal{D}$ obeys (ℓ, δ)-diversity, for some fixed $\delta \in (0, 1)$, if it respects ℓ-diversity w.p. $\geq 1 - \delta$.

(ℓ, δ)-Diversity: A definition

Consider the setting of datasets constructed from samples from volunteering users.

- A good assumption here is that the samples $(\mathbf{q}_i, s_i)$ are i.i.d. across $i \in [N]$, from a joint distribution $P_{\mathbf{Q},S} = P_{\mathbf{Q}} \cdot P_{S|\mathbf{Q}}$.
- WLOG, we assume that $P_{\mathbf{Q}} \succ 0$ and $P_S \succ 0$.

Definition

A dataset $\mathcal{D}$ obeys (ℓ, δ)-diversity, for some fixed $\delta \in (0, 1)$, if it respects ℓ-diversity w.p. $\geq 1 - \delta$.

1. How must a dataset owner ensure (ℓ, δ)-diversity? [Achievability]
2. How does (ℓ, δ)-diversity degrade upon linkage? [Linkage-resilience]
3. Can we achieve (ℓ, δ)-diversity while guaranteeing "closeness" to the true dataset? [Utility]

Achievability

We assume the existence of a Central Anonymizer (CA) who knows $P_{\mathbf{Q},S}$.

1: **procedure** MECH-(ℓ, δ)-DIVERSITY
2: Clients convey (ℓ, δ) to the CA.
3: CA chooses $p \in (0, p_\ell]$ and constructs eq. classes $\overline{\mathbf{q}}$ s.t. for each $\overline{\mathbf{q}}$, we have $|S_{\overline{\mathbf{q}}}| := |\{s : P(\overline{\mathbf{q}}, s) \geq p\}| \geq \ell$.
4: CA broadcasts $N = \frac{\ln\left(\frac{m\ell}{\delta}\right)}{\ln\left(\frac{1}{1-p}\right)}$ and $\{\overline{\mathbf{q}}\}$ to the data owners (DOs).
5: DOs obtain N samples and construct $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$ using $\{\overline{\mathbf{q}}\}$.

Here, $p_i = i^{\text{th}}$-largest $P_S(s)$; and $m := \min\left\{|\mathcal{Q}|, \frac{1}{\ell p}, \frac{\sum_{i=\ell}^{|\mathcal{S}|} p_i}{p}\right\}$.

Achievability

We assume the existence of a Central Anonymizer (CA) who knows $P_{\mathbf{Q},S}$.

1: **procedure** MECH-(ℓ,δ)-DIVERSITY
2: Clients convey (ℓ,δ) to the CA.
3: CA chooses $p \in (0, p_\ell]$ and constructs eq. classes $\overline{\mathbf{q}}$ s.t. for each $\overline{\mathbf{q}}$, we have $|S_{\overline{\mathbf{q}}}| := |\{s : P(\overline{\mathbf{q}}, s) \geq p\}| \geq \ell$.
4: CA broadcasts $N = \frac{\ln\left(\frac{m\ell}{\delta}\right)}{\ln\left(\frac{1}{1-p}\right)}$ and $\{\overline{\mathbf{q}}\}$ to the data owners (DOs).
5: DOs obtain N samples and construct $\overline{\mathcal{D}}_1, \ldots, \overline{\mathcal{D}}_t$ using $\{\overline{\mathbf{q}}\}$.

Theorem

Any dataset $\overline{\mathcal{D}}$ obtained via Mech-(ℓ,δ)-diversity satisfies (ℓ,δ)-diversity.

Here, $p_i = i^{\text{th}}$-largest $P_S(s)$; and $m := \min\left\{|\mathcal{Q}|, \frac{1}{\ell p}, \frac{\sum_{i=\ell}^{|\mathcal{S}|} p_i}{p}\right\}$.

Proof idea

- The proof first argues that $|\overline{\mathcal{Q}}| \leq m$.
- Next, note that for any equivalence class $\overline{\mathbf{q}}$,

$$\Pr\left[N(\overline{\mathbf{q}}, s) = 0\right] \leq (1-p)^N.$$

Proof idea

- The proof first argues that $|\overline{\mathcal{Q}}| \leq m$.
- Next, note that for any equivalence class $\overline{\mathbf{q}}$,

$$\Pr\left[N(\overline{\mathbf{q}}, s) = 0\right] \leq (1-p)^N.$$

- Let $M_{\overline{\mathbf{q}}}$ denote # distinct sensitive attributes in $\overline{\mathbf{q}}$. Then,

$$\Pr\left[M_{\overline{\mathbf{q}}} \geq \ell\right] \geq 1 - \ell \cdot (1-p)^N.$$

Proof idea

- The proof first argues that $|\overline{\mathcal{Q}}| \leq m$.
- Next, note that for any equivalence class $\overline{\mathbf{q}}$,

$$\Pr\left[N(\overline{\mathbf{q}}, s) = 0\right] \leq (1-p)^N.$$

- Let $M_{\overline{\mathbf{q}}}$ denote # distinct sensitive attributes in $\overline{\mathbf{q}}$. Then,

$$\Pr\left[M_{\overline{\mathbf{q}}} \geq \ell\right] \geq 1 - \ell \cdot (1-p)^N.$$

- Finally, via a union bound again, we have

$$\begin{aligned}\Pr\left[M_{\overline{\mathbf{q}}} \geq \ell, \text{ for all } \overline{\mathbf{q}} \in \overline{\mathcal{Q}}\right] &\geq 1 - \ell|\overline{\mathcal{Q}}| \cdot (1-p)^N \\ &\geq 1 - m\ell \cdot (1-p)^N = 1 - \delta,\end{aligned}$$

by the choice of N.

Linkage resilience

The following composition theorem (à la the Basic Composition Theorem of DP [Dwork-Roth (2014)]) provides the degradation of anonymity on linkage:

Theorem

The overall post-linkage dataset $\overline{\mathcal{D}}_{[t]}$ *obtained via anonymization using* $\textsc{Mech}$-(ℓ, δ)-$\textsc{Diversity}$, *satisfies* $(\ell, t\delta)$-*diversity.*

The proof is via a simple application of the union bound.

Linkage resilience

The following composition theorem (à la the Basic Composition Theorem of DP [Dwork-Roth (2014)]) provides the degradation of anonymity on linkage:

Theorem

The overall post-linkage dataset $\overline{\mathcal{D}}_{[t]}$ *obtained via anonymization using* $\textsc{Mech}$-(ℓ, δ)-$\textsc{Diversity}$, *satisfies* $(\ell, t\delta)$-*diversity.*

The proof is via a simple application of the union bound.

Observe that the Theorem asserts that the entire post-linkage dataset, in **all** equivalence classes, satisfies $(\ell, t\delta)$-diversity.

Linkage resilience

The following composition theorem (à la the Basic Composition Theorem of DP [Dwork-Roth (2014)]) provides the degradation of anonymity on linkage:

Theorem

The overall post-linkage dataset $\overline{\mathcal{D}}_{[t]}$ *obtained via anonymization using* $\textsc{Mech}$-(ℓ, δ)-$\textsc{Diversity}$, *satisfies* $(\ell, t\delta)$-*diversity.*

The proof is via a simple application of the union bound.

We shall next make explicit the "generalization" strategy for obtaining eq. classes to ensure "high utility," in a special setting of practical relevance.

Utility - I

Consider the setting where there exists a total order on quasi-identifiers:

$$\mathbf{q}_1 < \mathbf{q}_2 < \ldots < \mathbf{q}_{|\mathcal{Q}|}.$$

- We consider the natural class of "generalization" (or binning) strategies that give rise to eq. classes with contiguous quasi-identifiers.

  [e.g., contiguous PIN codes, contiguous ages, etc.]

- A natural notion of "utility" hence is the number of eq. classes: large $|\overline{\mathcal{Q}}| \equiv$ high utility.

Q: What is the structure of the optimal (highest utility) contiguous generalization strategy that guarantees ℓ-diversity?

Utility - I

Consider the setting where there exists a total order on quasi-identifiers:

$$\mathbf{q}_1 < \mathbf{q}_2 < \ldots < \mathbf{q}_{|\mathcal{Q}|}.$$

- We consider the natural class of "generalization" (or binning) strategies that give rise to eq. classes with contiguous quasi-identifiers.

  [e.g., contiguous PIN codes, contiguous ages, etc.]

- A natural notion of "utility" hence is the number of eq. classes: large $|\overline{\mathcal{Q}}| \equiv$ high utility.

Q: What is the structure of the optimal (highest utility) contiguous generalization strategy that guarantees ℓ-diversity?

A: The greedy procedure that iteratively adds quasi-identifiers to eq. classes until $|S_{\overline{\mathbf{q}}}| \geq \ell$ is optimal (proof via induction).

Utility - I

Consider the setting where there exists a total order on quasi-identifiers:

$$\mathbf{q}_1 < \mathbf{q}_2 < \ldots < \mathbf{q}_{|\mathcal{Q}|}.$$

- We consider the natural class of "generalization" (or binning) strategies that give rise to eq. classes with contiguous quasi-identifiers.

  [e.g., contiguous PIN codes, contiguous ages, etc.]

- A natural notion of "utility" hence is the number of eq. classes: large $|\overline{\mathcal{Q}}| \equiv$ high utility.

Q: What is the structure of the optimal (highest utility) contiguous generalization strategy that guarantees ℓ-diversity?

A: The greedy procedure that iteratively adds quasi-identifiers to eq. classes until $|S_{\overline{\mathbf{q}}}| \geq \ell$ is optimal (proof via induction).

An explicit variant of this algorithm can be derived when $P_{\mathbf{Q},S} = P_{\mathbf{Q}} \cdot P_S$.

Numerical results

We present numerical results on the # samples required by Mech-(ℓ, δ)-Diversity.

We consider the setting of hospital records, where each record is

$$\mathbf{r} \in \underbrace{\textsc{Postal-Code} \times \textsc{Gender} \times \textsc{Age}}_{\text{quasi-identifiers}} \times \underbrace{\textsc{Disease}}_{\text{sens. attr.}},$$

with some fixed cardinalities of the underlying sets.

Numerical results

We present numerical results on the # samples required by Mech-(ℓ, δ)-Diversity.

We consider the setting of hospital records, where each record is

$$\mathbf{r} \in \underbrace{\textsc{Postal-Code} \times \textsc{Gender} \times \textsc{Age}}_{\text{quasi-identifiers}} \times \underbrace{\textsc{Disease}}_{\text{sens. attr.}},$$

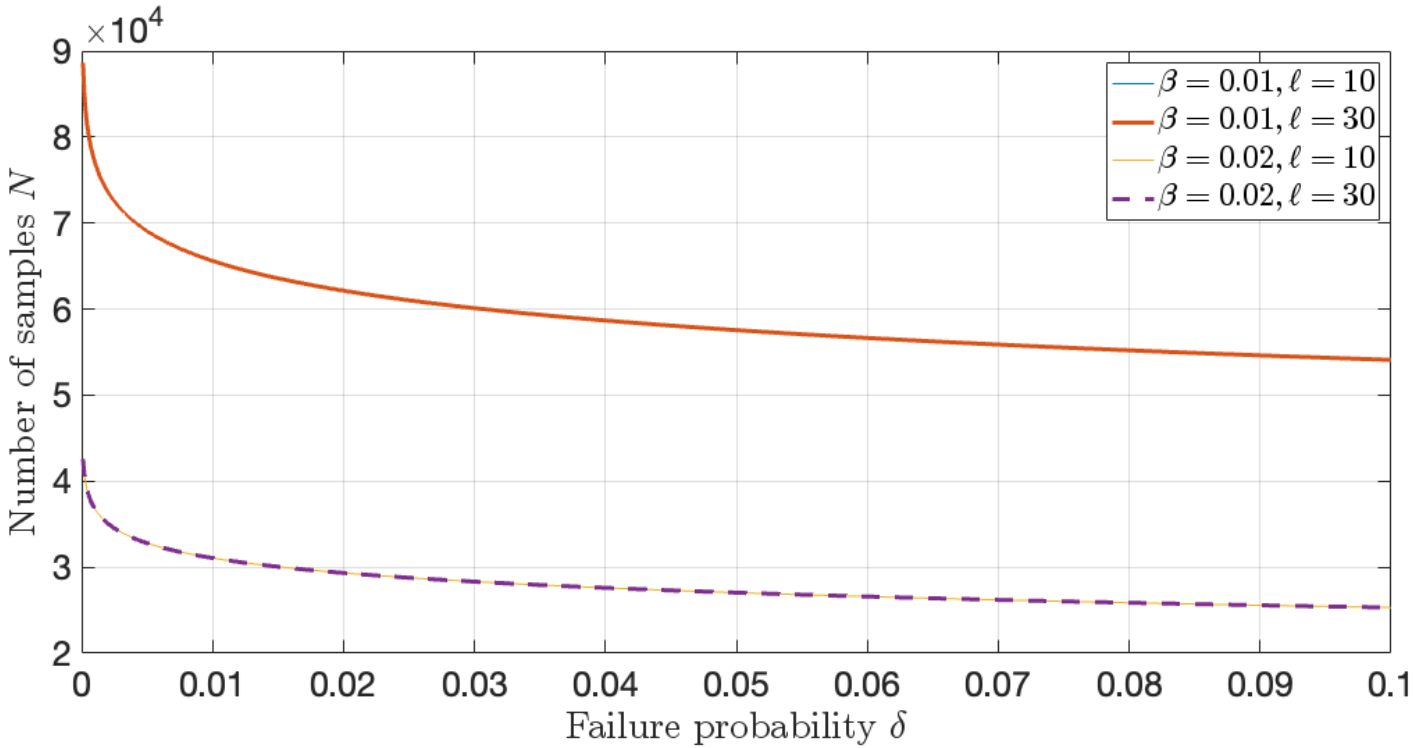
with some fixed cardinalities of the underlying sets.

We consider the following marginal distributions P_S:

1. **Uniform marginals**: Here, we let $P_S(s) = 1/|\mathcal{S}|$, for all $s \in [|\mathcal{S}|]$; observe that $p_i = 1/|\mathcal{S}|$, for all $i \in [|\mathcal{S}|]$.
2. **Geometric marginals**: Here, we let $P_S(s) = p_1 \rho^{s-1}$, for suitable $p_1, \rho \in (0, 1)$; observe that $p_i = P_S(i)$, for all $i \in [|\mathcal{S}|]$.

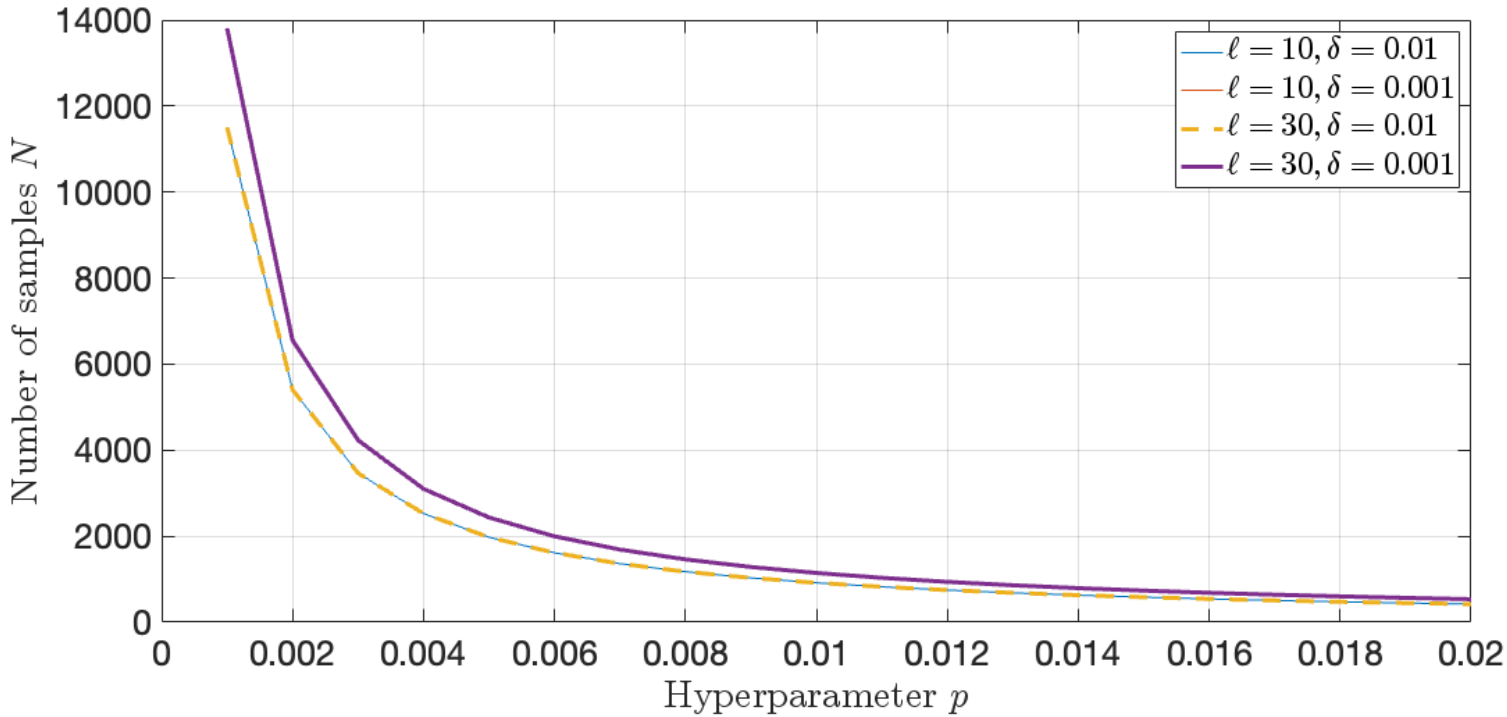
The exact parameters can be found in `https://arxiv.org/abs/2506.18405`.

Plots - I



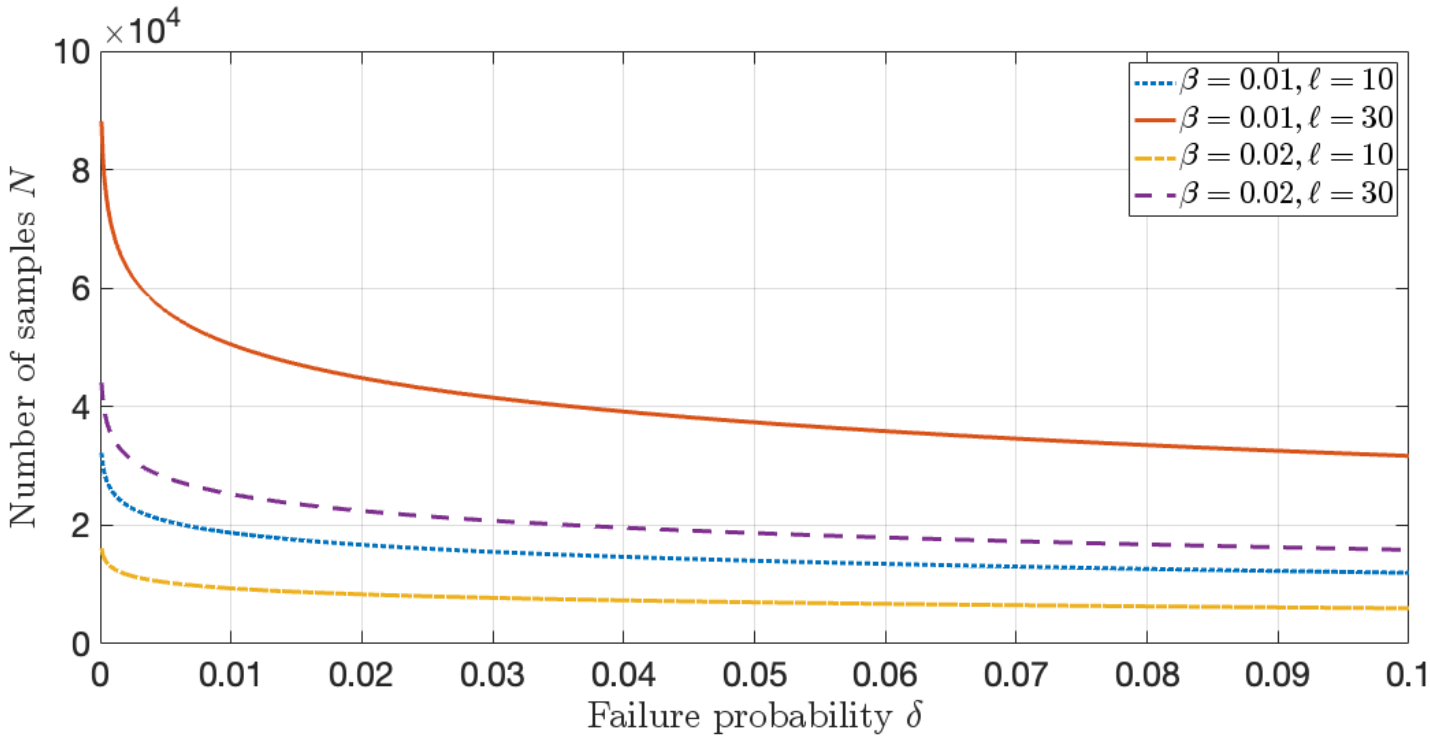
Plots showing the variation of the number of samples N with the parameter δ, under **uniform marginals** of sensitive attributes. Here, $\beta := p/p_\ell$.

Plots - II



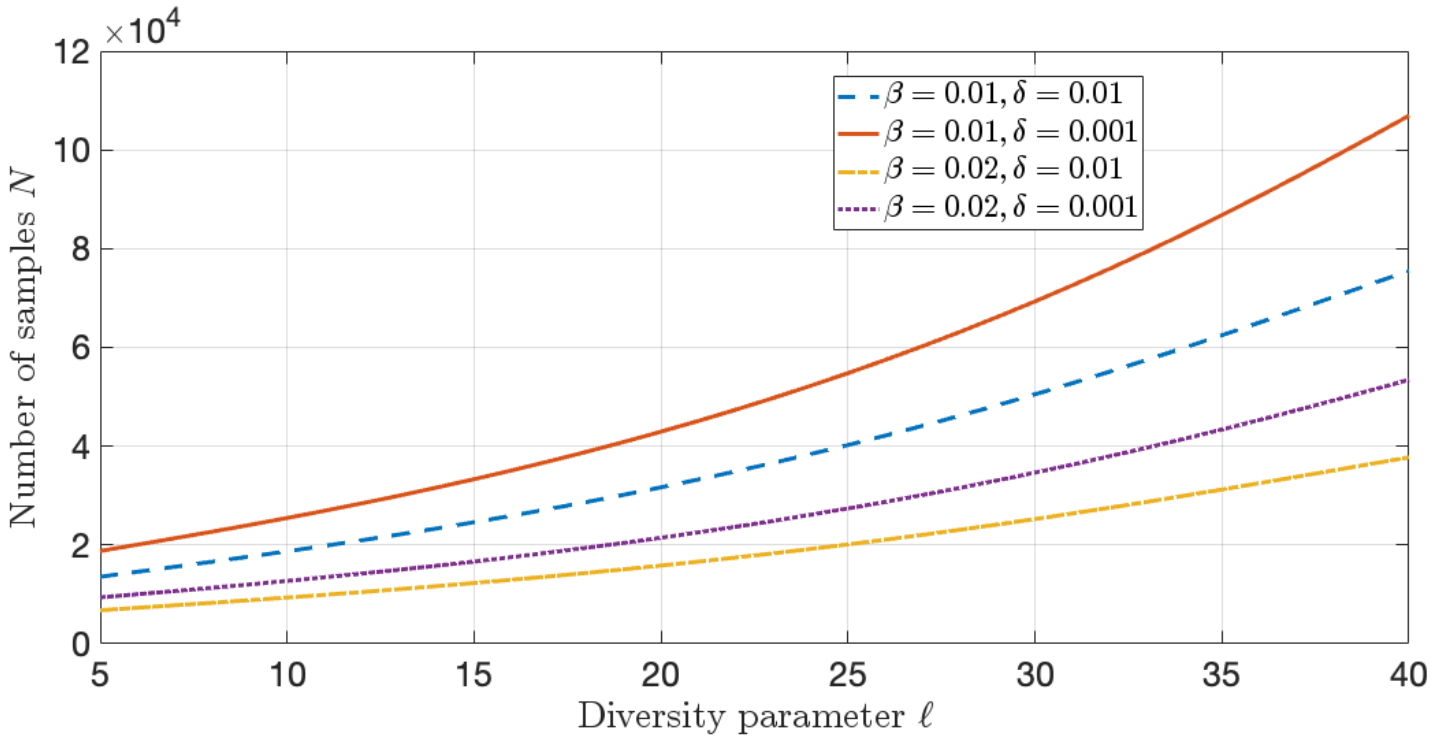
Plots showing the variation of the number of samples N with the parameter p, under **uniform marginals** of sensitive attributes.

Plots - III



Plots showing the variation of the number of samples N with the parameter δ, under **geometric marginals** of sensitive attributes. Here, $\beta := p/p_\ell$.

Plots - IV



Plots showing the variation of the number of samples N with the parameter ℓ, under **geometric marginals** of sensitive attributes. Here, $\beta := p/p_\ell$.

Ongoing and future research directions

- Extend our analysis for the setting of datasets with correlated (possibly Markovian) samples
- Design achievability schemes (generalization strategies) without the need for a Central Anonymizer (CA)
- Extend our analysis techniques to other notions of anonymity such as entropy ℓ-diversity

Thank You!