Master's research internship

Stability and private algorithms for δ-covers of set systems

Level	Master 1 / Master 2 or equivalent
Duration	4-6 months, between March 2027 and August 2027
Team and place	INRIA center at the University of Lille, MAGNET Team
Supervision	Alexandre Louvet - alexandre.louvet@inria.fr
Keywords	differential privacy, approximation, set systems, VC-dimension, covers

1. Context

Differential privacy

Differential privacy (DP) aims to ensure that individuals whose data are used in data analysis processes cannot be identified from the process's result. Informally, DP requires that the probability to obtain the same result from a differentially private algorithm on two datasets that only differ by one individual's records is high.

Adjacent databases. Let $\mathcal{D}$ be the universe of possible databases, we say that $D, D' \in \mathcal{D}$ are neighbouring databases iff they differ by a single record. That is, there exists $y, y' \in D, D'$ such that $D \setminus \{y\} = D' \setminus \{y'\}$.

With this notation, we use two parameters ε and δ to bound the probability difference over all possible outcomes of a differentially private algorithm.

Differential privacy. Let $\varepsilon > 0, \delta \geq 0$, a random algorithm $\mathcal{A} : \mathcal{D} \mapsto \mathcal{S}$ is (ε, δ)-DP if and only if for all neighbouring D, D' and $S \subset \mathcal{S}$:

$$P(\mathcal{A}(D) \in S) \leq e^{\varepsilon} P(\mathcal{A}(D') \in S) + \delta$$

DP has become the gold standard for the design and analysis of private data processing [DR14].

η-covers

The objects we will study in this project are combinatorial objects to represent data called **set systems**. Their definition follows.

Set systems. Let X be a set of n points (for instance $[\![1, n]\!]$) and let $\mathcal{F}$ be a collection of m subsets of X called ranges. We call $(X, \mathcal{F})$ a set system.

These objects are ways to represent data with very few assumptions. In particular, one can consider set systems with constraints of various natures on $\mathcal{F}$ (combinatorial constraint like VC-dimension or geometric constraints like sets determined by geometric shapes [Mus22]).

To measure the distance between two sets, we use the symmetric difference defined as follows.

Symmetric difference. Let F, F' be two sets, the symmetric difference is defined as:

$$\Delta(F, F') = (F \cup F') \setminus (F \cap F')$$

In this project we will work on computing private η-covers. These objects are approximations of a set system such that all ranges of the set system are close (w.r.t. symmetric difference) to one range in the cover.

η-cover. Given $(X, \mathcal{F})$ and $\eta \in [0, n]$, a η-cover of $(X, \mathcal{F})$ is a collection $\mathcal{C}$ of subsets of X such that:

$$\forall F \in \mathcal{F}, \exists C \in \mathcal{C} \text{ s.t. } |\Delta(F, C)| \leq \eta$$

The construction typically proceeds in three steps [Lou25; MWW93]:

- Draw a uniform random sample $\mathcal{N}$ of X with suitable size.
- Group the ranges by their intersection with $\mathcal{N}$.
- Keep one representative per group.

With high probability, the sets that share the same intersection are close w.r.t. symmetric difference. Otherwise, the probability to randomly pick in the sample an element from their symmetric difference would be large.

η-covers are a basic tool for approximation of set systems. They find applications in range querying problems as they can be used to answer linear queries (such as counting queries) on the ranges of set systems. More generally, they enable the approximation of the solution to any problem as long as the error is a function of the symmetric difference.

Goal

The goal of this internship is to develop a differentially private algorithm to compute an η-cover of a set system under the assumption that X is public and $\mathcal{F}$ private (known only to the algorithm). The classical algorithm, described above, is highly non-private as it returns elements of $\mathcal{F}$.

Previous work on private approximation often relies on the assumption that there exists a public family from which elements of the approximation can be drawn and express results depending on the properties of that public family [Bha+10]. Such an approach can suffer from a misalignment between the public family and the actual data; in particular, using a sufficiently general family to ensure coverage can result in wasteful, oversized approximations.

We aim to remove this assumption by building private η-covers whose size depends only on combinatorial or geometric properties of the data. This would also allow us to understand the actual cost of the public family assumption in a variety of setups and the properties of data that can affect this cost.

2. Research environment

The internship will take place in the INRIA center at University of Lille. The INRIA center is part of the computer science laboratory of the University CRIStAL. The intern will join the MAGNET team. The project fits within the team's broader research agenda, which explores

the synergies between machine learning and privacy, specifically within the contexts of decentralized systems, algorithmic fairness, and natural language processing.

3. Main activities

The main activities during this internship are:

- Reading and understanding the relevant literature on differential privacy and combinatorial approximations.
- Designing a DP η-cover algorithm.
- Studying the stability conditions of η-covers in general set systems.
- Participating in team activities (mainly seminars).

4. Profile sought

The ideal candidate should have:

- A strong background in discrete mathematics or theoretical computer science.
- Proficiency in English (both written and oral).
- A strong interest in theoretical research and mathematical proofs.
- Knowledge of differential privacy, randomized algorithms, or approximation algorithms would be a plus.

To apply send a CV and a cover letter explaining your interest for this internship to Alexandre Louvet (alexandre.louvet@inria.fr).

Bibliography

[Bha+10] Bhaskar, Raghav ; Laxman, Srivatsan ; Smith, Adam ; Thakurta, Abhradeep: Discovering Frequent Patterns in Sensitive Data. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010

[DR14] Dwork, Cynthia ; Roth, Aaron: The Algorithmic Foundations of Differential Privacy. In: *Foundations and Trends® in Theoretical Computer Science* vol. 9, Now Publishers, Inc. (2014)

[Lou25] Louvet, Alexandre: *Faster Algorithms for Approximating Combinatorial and Geometric Data*, Doctoral dissertation, 2025

[Mus22] Mustafa, Nabil H.: *Sampling in Combinatorial and Geometric Set Systems*. vol. 265 : American Mathematical Society, 2022

[MWW93] Matoušek, Jiří ; Welzl, Emo ; Wernisch, Lorenz: Discrepancy and Approximations for Bounded VC-Dimension. In: *Combinatorica* vol. 13 (1993)