Karlsruhe Institute of Technology

A Formal Study of Differential Privacy in Complex and Correlated Data

Zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Patricia Guerra Balboa

Tag der mündlichen Prüfung:	20.05.2026
1. Referent:	Prof. Dr. Thorsten Strufe
2. Referent:	Prof. Dr. Hannes Federrath

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

Abstract

Differential privacy (DP) has emerged as the standard for privacy-preserving data analyses, offering formal protection that can be monitored over time thanks to its composability properties. DP has proven to be highly effective for classical tabular data and simple queries; however, its deployment in modern data types—such as trajectories, time series, graphs, and other complex high-dimensional data—is not yet fully understood. These complex data types are characterized by rich semantics, strong correlations, and non-trivial structures, all of which deviate from the underlying assumptions of classical DP analyses and complicate the interpretation of privacy guarantees in practice.

In this thesis, we investigate the limitations of DP in complex data and develop new theoretical and practical tools to improve the adoption of DP in these new scenarios. We begin by systematizing the challenges that arise when DP is applied beyond its original tabular context, using trajectory data as a representative and practically relevant case study. This analysis reveals four core issues, which we then study throughout this dissertation: (i) the widespread formal errors in the design and implementation of DP mechanisms for complex domains, (ii) the issues in extending key DP properties like composability, (iii) the limited interpretability of DP parameters under realistic attack models, and (iv) the failure of classical DP guarantees in the presence of correlated data.

First, we address the challenge of composition in complex data domains. To provide general composition results applicable to the growing landscape of DP variations, we adopt the general metric privacy framework. Within this framework, we develop a unified composition theory that yields tighter privacy loss bounds by explicitly accounting for amplification and attenuation effects induced by preprocessing. As a result, we obtain tighter bounds that apply consistently across any domain and DP variation. Additionally, we extend our composition analysis to Gaussian DP, enabling a more interpretable characterization of the composition privacy loss over general data domains.

Second, we investigate the parameter interpretation and attack resilience of DP through a formal and empirical study on the risk incurred by individuals whose data are processed by DP mechanisms. Understanding this risk is essential for principled noise calibration: If the privacy parameters are set too loosely, sensitive information may be exposed; while if they are set too conservatively, utility is unnecessarily degraded. We analyze reconstruction attacks and demonstrate that existing adversarial metrics—such as reconstruction robustness (ReRo)—systematically overestimate the privacy risk by conflating statistical imputation and auxiliary knowledge with genuine participation leakage. To address this issue, we introduce *reconstruction advantage* (RAD), a novel advantage-based metric that explicitly incorporates auxiliary information. We establish worst-case and auxiliary-dependent tight bounds that link DP guarantees to RAD,

providing a sharper and more interpretable characterization than previous approaches. Our theoretical and empirical evaluation demonstrates both the robustness of RAD as a risk measure and the practical impact of our analysis. In particular, our bounds enable improved noise calibration, yielding better utility without sacrificing meaningful privacy guarantees. Moreover, we develop a RAD-based auditing framework that improves both efficiency and accuracy, especially for high-dimensional categorical data, and broadens the scope with respect to existing auditing tools. This framework enables earlier detection of formal errors and implementation flaws, ensuring that individuals' privacy guarantees are effectively enforced.

Finally, we study the vulnerability of DP to correlation-based inference attacks. Correlations—such as temporal dependencies in trajectories or social dependencies in networks—can substantially amplify an attacker's information gain, effectively shrinking DP privacy guarantees. To address this challenge, we investigate whether accurate inference guarantees can be achieved under *Bayesian differential privacy* (BDP), an enhanced DP notion that explicitly protects against correlation-based attacks. Particularly, we derive novel leakage bounds for BDP under arbitrary correlations, as well as tighter, correlation-specific bounds for Gaussian and Markov distributions. These results provide a systematic methodology for constructing accurate BDP mechanisms tailored to realistic correlation structures, a principled first step towards understanding when BDP guarantees and improved utility are simultaneously attainable.

Overall, this thesis demonstrates that privacy risk in complex data analysis depends critically on the data structure, correlations, and mechanism design—not solely on the DP parameters. By providing new theoretical foundations, tighter bounds, and practical auditing tools, our work advances the field towards reliable and interpretable deployments of DP in complex and correlated data.

Acknowledgments

There is no better way to begin these acknowledgments than by expressing my deepest and most sincere gratitude to my supervisor, Thorsten Strufe. He has been an attentive mentor, from whom I gained profound insights. He always granted me the freedom to pursue my interests, nurturing my ideas at every step. His patience has been boundless, and he has consistently made me feel valued and inspired me to strive for excellence. He ensured that our working environment was stimulating and cultivated the growth of a strong, collaborative team. Without a doubt, he has set an exemplary high standard for all my future managers.

I would also like to thank my second reviewer Prof. Dr. Hannes Federrath for accepting to review my dissertation and taking the time to grade my work.

I would also like to express my gratitude to all my collaborators with whom I had the pleasure of co-authoring papers during my PhD. Thanks to Jordi Forné, who introduced me to the fascinating world of trajectory privacy. Thanks to Javier Parra-Arnau, who introduced me to differential privacy and provided guidance and support at the start of my PhD. I would also like to thank Héber H. Arzolezi, from whom I gained extensive knowledge, especially about DP auditing, and with whom I had the opportunity to exchange ideas, insights, and great conversations.

I would also like to thank Martin Lange and Annika Sauer, who were wonderful students, always willing to explore every idea I proposed. Their dedication and curiosity inspired my passion for mentoring and teaching.

I would like to express my sincere gratitude to Ana-Maria Cretu, who was always friendly and supportive, and generously provided feedback on many of my early research ideas and shared her knowledge and experience as a scholar with me, which was invaluable.

I would like to thank all my fellow PhD students who created a great environment to work, discuss and enjoy every day at the office. Ein besonderer Dank gilt Christoph und Simon, die mich herzlich in der Arbeitsgruppe willkommen hießen und mir halfen, mich einzuleben, als ich erstmals nach Deutschland kam. Ebenso Julian und Felix, deren unendliche Geduld und Bereitschaft, mir bei der Programmierung von Experimenten zu helfen, äußerst wertvoll war. Und Daniel, der immer da war, um mich zu unterstützen, zu motivieren und zum Lachen zu bringen. Ich werde immer dankbar sein für jeden Streich aus The Office, jede deutsche Kaffeepause, und jeden Nachmittag, den wir mit Forschungsgesprächen verbracht haben. Es gibt nur wenige Menschen, die mir so selbstlos geholfen haben. Außerdem möchte ich Frau Sauer, unserer Sekretärin, danken, die stets Wege fand, bürokratische Prozesse für uns reibungslos zu gestalten. Ganz gleich, was benötigt wurde, Frau Sauer hat es immer geschafft, alles zu ermöglichen.

I, com no, gràcies al meu català favorit, l'Àlex, el meu col·laborador, company d'oficina i amic. Amb ell he rigut i he cridat, però, sobretot, he trobat un company i un suport en aquest doctorat que se m'hauria fet molt llarg sense ell. Gràcies per fer de la nostra oficina se senti com a casa nostra.

Thanks to my friends in Karlsruhe who kept me sane throughout these years. To those who made the WG feel like home, to those who were there every Monday for calisthenics training in the freezing winter, every Wednesday to help me unwind, and every weekend to join me on my adventures—whether it was a simple volleyball game or climbing the highest mountain in Germany. Without you, these years wouldn't have been the same.

Grazas tamén aos de casa, á miña Ohana, os que levan ahi dende sempre e que, aínda que dende lonxe, sempre estiveron para escoitarme nas queixas e acompañarme nas celebracións.

Las gracias a mi familia, sin duda, se quedan cortas. A mi hermano, Diego, que siempre me ha valorado y apoyado, que ha sido mi cómplice y muchas veces mi motor para seguir adelante. A la abuela Pura, que con su coraje, su carácter y su espíritu de lucha, capaz de cruzar el océano las veces que haga falta para darnos una vida mejor, se convirtió en mi referente para luchar por mis metas. Y a mis padres, Adriana y Daniel, a quienes les debo cada palabra de esta tesis. Si no fuera por vosotros que desde pequeña os esforzasteis por estimular mi curiosidad y me empujasteis a reflexionar y a tener espíritu crítico; si no fuera por vosotros que me enseñasteis a no conformarme, a luchar y a seguir adelante frente a la adversidad, predicando con el ejemplo; si no fuera por vosotros que me empujasteis a ver el mundo, a no tener miedo a aventurarme a nuevos desafíos y a descubrir otros países; si no fuera por vosotros que apoyasteis económica y emocionalmente cada paso de mi carrera académica, si no fuera por vosotros yo no estaría aquí.

Y, para finalizar, gracias a ti: a ti que te embarcaste en este viaje a otro país conmigo; a ti que soportaste mis frustraciones y celebraste mis victorias más que yo misma; a ti que te aseguraste de que ninguna *deadline* me dejase desnutrida—lo cual no era tarea fácil; a ti que escuchaste cada una de mis presentaciones cinco veces; a ti que creíste en mí incluso cuando yo no lo hacía; a ti que lo hiciste posible; gracias Josemi. Te quiero.

Publication List

The following is a list of the publications done during my PhD. The superscript $\star$ indicates equal contributions to the paper.

- **Patricia Guerra-Balboa**, Annika Sauer, Héber H. Arcolezi and Thorsten Strufe. "Understanding Disclosure Risk in Differential Privacy with Applications to Noise Calibration and Auditing". In Proceedings of the VLDB Endowment, 2026, DOI: 10.14778/3801059.3801069.
- Martin Lange*, **Patricia Guerra-Balboa***, Javier Parra-Arnau, and Thorsten Strufe. "Balancing Privacy and Utility in Correlated Data: A Study of Bayesian Differential Privacy". In: Proceedings of the VLDB Endowment, 2025, DOI: 10.14778/3749646.3749679.
- **Patricia Guerra-Balboa**, Annika Sauer, and Thorsten Strufe. "Analysis and Measurement of Attack Resilience of Differential Privacy". In: ACM Workshop on Privacy in the Electronic Society (WPES), 2024, DOI: 10.1145/3689943.3695046.
- Àlex Miranda-Pascual*, **Patricia Guerra-Balboa***, Javier Parra-Arnau, Jordi Forné, and Thorsten Strufe. "An overview of proposals towards the privacy-preserving publication of trajectory data". In: International Journal of Information Security (IJIS), 2024, DOI: 10.1007/s10207-024-00894-0.
- **Patricia Guerra-Balboa***, Àlex Miranda-Pascual*, Javier Parra-Arnau, and Thorsten Strufe. "Composition in Differential Privacy for General Granularity Notions". In: IEEE Computer Security Foundations Symposium (CSF), 2024, DOI: 10.1109/CSF61375.2024.00004.
- Àlex Miranda-Pascual*, **Patricia Guerra-Balboa***, Javier Parra-Arnau, Jordi Forné, and Thorsten Strufe. "SoK: Differentially Private Publication of Trajectory Data". In: Proceedings on Privacy Enhancing Technologies (PoPETS), 2023, DOI: 10.56553/popets-2023-0065.
- **Patricia Guerra-Balboa***, Àlex Miranda Pascual*, Javier Parra-Arnau, Jordi Forne and Thorsten Strufe. "Anonymizing trajectory data: limitations and opportunities". In: AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-22), 2023. URL: https://aaai-ppai22.github.io/files/25.pdf.

Contents

List of Figures

List of Tables

1. Introduction

Data analytics enables societal benefits across highly relevant domains [1], such as human mobility—for instance, large-scale trajectory data support accurate travel-time estimation and personalized routing [2]—and healthcare—for instance, medical record analysis enables early disease detection, improved diagnoses, and personalized treatments [3]. This impact is further amplified by the widespread availability of personal devices, such as smartphones and wearables, that facilitate the large-scale collection and processing of complex data [4], leading to unprecedented ease of data collection and analysis [5].

At the same time, this data-driven paradigm raises significant privacy concerns, leading to increased legal and ethical scrutiny [6], [7]. Location traces, health information, and social interactions are explicitly classified as special categories of personal data under regulatory frameworks, such as the European General Data Protection Regulation (GDPR), imposing strict constraints on their collection and sharing and promoting privacy enhancing technologies as a core compliance principle [8]. However, numerous real-world attacks on naïvely protected datasets have demonstrated how easily individuals can be re-identified [9], [10], [11], [12], [13], [14]. Consequently, ensuring privacy-preserving data analysis is not only an ethical and legal requirement but also a scientific challenge.

To address the tension between extracting reliable population-level insights and protecting individual privacy, differential privacy (DP) [15] has emerged as a principled framework providing rigorous guarantees under a well-defined threat model. Within its assumptions—most notably, an adversary who knows the entire dataset except for the target record—DP bounds the risk associated with an individual's participation [16]. The underlying principle is that nothing about a target should be learnable from the dataset that could not also be learned if that individual had not participated. Formally, DP compares the output distributions of a mechanism applied to two databases differing in a single record. If these distributions are close (measured by the privacy budget ε), the attacker's ability to infer information about the record is *almost* the same as if the record had not been included [15], hence participating in the dataset does not entail *significantly higher* privacy risk than not participating, as quantified by the privacy budget.

Unlike earlier privacy notions such as k-anonymity or ℓ-diversity [17], DP offers dataset-independent, quantifiable privacy guarantees and, crucially, enables formal tracking of privacy loss across multiple analyses, a property known as *composability* [18]. Particularly, two composition results exist: *sequential*, where the privacy loss increases linearly with the number of mechanisms [16], and *parallel* composition [19], which reduces the overall privacy loss to the maximum among the composed mechanisms, but only applies when they access mutually disjoint data.

These properties have made DP particularly effective for simple aggregate statistics and tabular datasets. In such setting, DP mechanisms enable useful outputs under tight privacy budgets (i.e., strong privacy). As a result, organizations such as Apple [20], LinkedIn [21] and the US Census Bureau [22], [23], [24] have widely adopted DP.

The success of DP in tabular data, where mechanisms are well understood and can achieve favorable utility–privacy trade-offs, has motivated efforts to extend DP to more complex data domains, i.e., data sets that are high-dimensional, exhibit strong dependencies, and/or cannot be naturally represented in tabular form, such as graphs, time series, and images [25]. However, such settings were not considered in the original design of DP and introduce fundamental challenges in both adapting DP mechanisms and interpreting privacy guarantees beyond the classical tabular model [26].

To better assess the feasibility of DP as a general-purpose tool for emerging complex data structures, we first study its application to trajectory data—a representative and practically relevant domain that exhibits correlations, dependencies, and structure beyond standard tabular data. This analysis enables a systematic understanding of how DP behaves in such settings and allows us to identify the main challenges in achieving meaningful and interpretable privacy–utility trade-offs. Based on this systematization, we identify several key challenges of DP, which form the focus of the remainder of this thesis. Specifically:

In complex data domains, the information encoded in the data is substantially richer and more difficult to isolate at the level of a single individual, which blurs the classical notion of a "record" or "individual contribution." For example, in social networks, an edge represents information shared by two individuals, rather than belonging to one in isolation. Consequently, the original DP definition has undergone numerous reformulations to clarify what information we aim to protect and, equivalently, what should be difficult for an attacker to infer—a concept referred to as the *granularity level* [16]. Consider, for instance, a binary sensitive query like "drug abuse." Here, we may aim to prevent an attacker from confidently determining a "yes" or "no" answer, corresponding to a *bounded* granularity level [18]. In contrast, in a social network, the goal might be to protect relationships between individuals, making it difficult to infer whether an edge exists between two nodes—this corresponds to an *edge* granularity level [18].

However, while these alternative granularities are essential for modeling complex data, they introduce new challenges for privacy accounting: standard composability results [18] do not uniformly generalize across data domains, composition protocols, or granularity notions. In particular, parallel composition does not extend straightforwardly to more complex granularities [27], complicating practical deployment and often leading to unnecessary utility loss. Addressing this issue is crucial, as composability is one of the primary advantages of DP. If it is not applicable or if privacy loss estimates are severely overestimated, the resulting noise injection can degrade utility to the point that DP becomes impractical for complex data.

Moreover, complex data typically contain more information or attributes that can be inferred about users, increasing the risk of revealing sensitive information beyond mere participation or the exact record value. For example, from a geospatial trajectory, one

can infer not only a user's location, but also whether their home is unoccupied, or even religious beliefs by identifying prayer stops and routines [28]. These emerging threats raise practical questions about the protection DP provides in real-world scenarios: What is the actual impact of a specific parameter choice on different types of attacks, and whether some mechanisms offer stronger protection against certain attacks than others—even when they satisfy DP with the same privacy parameters. Addressing these questions is highly relevant, not only to improve understanding and transparency, but also because it is crucial for correctly calibrating the noise in DP mechanisms. Overestimating risk results in unnecessary utility loss, while underestimating it can lead to severe privacy breaches for users.

New challenges also arise, such as the effect of correlations or dependencies in the data (e.g., social links in a network or spatio-temporal correlations in trajectories) on the information gain of an attacker who leverages these correlations. DP protection guarantees are limited to statistically independent data records, i.e., DP can underestimate private information leakage when the underlying data is correlated. The limitations of DP for protecting correlated data have been theoretically exposed [29], [30], [31], [32] and empirically confirmed with attacks on real databases [33]. This is a significant issue, as correlations among data records are common in real-world databases, such as those induced by friendships in social networks [34], genetic similarities among family members [35], or spatio-temporal correlations in human trajectories [32].

In particular, granularity definitions that were designed without accounting for the strong dependencies inherent in their respective domains are highly vulnerable to real-world attacks. For example, event-level privacy in streaming data [36] ignores spatio-temporal correlations, leading to significant information leakage, while Pixel-DP [18] in the image domain disregards the fact that knowledge of surrounding pixels can help reconstruct a missing one, making it vulnerable to attacks [14].

Hence, understanding how correlations impact privacy risk is essential for evaluating the guarantees of DP in complex domains. Failing to account for these correlations can provide users with a misleading sense of privacy.

To address this, the literature has proposed strengthened privacy notions, such as Bayesian DP (BDP) [37], which explicitly accounts for data dependencies. However, if maintaining acceptable utility is already challenging under classical DP, enforcing stronger guarantees to account for correlations can make the problem infeasible or result in solutions with sharply reduced utility.

Finally, the naïve application of existing DP mechanisms for simple query answering—such as the Laplace and exponential mechanisms [16]—without proper adaptation to complex data structures leads to fundamental formal errors and, consequently, to severe privacy failures. In our survey of trajectory privacy mechanisms, we find that nearly 60% of the surveyed mechanisms contain such flaws, resulting in privacy guarantees that do not hold in practice (see Section 3.4 and Table 3.2). These failures are not inherent limitations of DP as a framework, but rather stem from errors in its human implementation. Nonetheless, dismissing human error as a marginal concern would be unrealistic: When DP mechanisms are deployed incorrectly, individuals may be exposed

to unauthorized inference of sensitive attributes and violations of data protection rights. As a result, privacy mechanisms and analyses that rely exclusively on lengthy or complex formal proofs are inherently vulnerable to undetected flaws, underscoring the need for robust, systematic auditing frameworks for DP implementations.

1.1. Contributions

This thesis focuses on advancing the applicability of DP in complex systems by introducing novel formalizations of DP properties, insights, and approaches that address the aforementioned challenges arising in complex data structures. We elaborate on our specific contributions to each challenge in the following paragraphs.

Contributions towards composition in complex data: We prove novel composition results for metric privacy—a generalization of DP that effectively addresses all possible granularity definitions simultaneously. Our theorems move beyond the traditional binary paradigm of sequential versus parallel composition, fully capturing the richness of the metric privacy framework and the influence of arbitrary preprocessing functions on composition privacy loss. This fine-grained analysis reduces the required noise injection, improving utility. Additionally, our results provide solutions to open questions in the literature, such as extending parallel composition to general metric spaces.

Furthermore, to enable direct interpretation of metric privacy in terms of attack mitigation—an interpretability previously limited to bounded DP—we introduce the first metric-based formulation of Gaussian differential privacy (GDP) [38].

Summarizing our contributions in this field are:

- We prove novel theorems that allow us to reduce the estimated privacy loss and design improved metric private mechanisms in general contexts. Moreover, our theorems make it possible to mix different granularity mechanisms while controlling the privacy guarantees offered.
- We show that sequential and parallel composition arise as special cases of our unifying framework, which enables their extension to arbitrary metrics and levels of granularity.
- We extend all our results to metric GDP, enabling the interpretability of composition directly in terms of attack mitigation.

Overall, our results allow to reduce risk overestimation in complex systems that perform several queries and in more general settings, hence improving the utility and applicability of DP in broader contexts.

Contributions towards parameter interpretability: In order to understand the real impact of the privacy budget in practical privacy, in this thesis, we mathematically and experimentally analyze the adversarial bounds of DP. We theoretically expose and empirically confirm that ReRo [39], the first metric for data reconstruction attacks (DRAs), has key theoretical limitations as a comprehensive adversarial metric. Particularly, ReRo fails to account for imputation-based success—leading to unnecessary utility loss when

used for noise calibration. Moreover, the existing bounds do not hold for attackers with target-specific auxiliary knowledge.

We address these issues by introducing *reconstruction advantage* (RAD), which extends advantage-based metrics to the DRA framework. RAD naturally incorporates auxiliary knowledge and avoids overestimating risk. We establish tight DP-to-RAD bounds enabling noise calibrated to true participant risk: (i) a worst-case bound independent of auxiliary knowledge, and (ii) an auxiliary-dependent bound that is universally tight. We further construct the optimal attack for any reconstruction goal, auxiliary knowledge, and mechanism—proving tightness and yielding a practical DP auditing tool. We also provide closed-form upper bounds without auxiliary knowledge and for perfect reconstruction, particularly relevant to categorical data. In summary, our contributions in this field are:

- We empirically show that the existing reconstruction attack metric, and its corresponding bounds, fail to account for imputation-based success and target-specific auxiliary knowledge, limiting applicability.
- We introduce reconstruction advantage (RAD) as a consistent, unifying risk metric that naturally incorporates auxiliary knowledge.
- We establish tight worst-case and auxiliary-dependent bounds for RAD, along with black-box bounds for attackers lacking auxiliary knowledge
- We construct the optimal attack strategy for any reconstruction goal, mechanism, and prior distribution, proving its optimality and demonstrating empirical utility for auditing.

Overall, our work demonstrates that privacy risk depends on the mechanism's structure, not just its nominal privacy parameters, and provides both fundamental insight and practical tools for privacy risk assessment and calibration—enabling notable utility gains without increasing the effective privacy risk in complex settings.

Contributions towards DP misuse: To prevent future flaws in DP mechanism design and implementation that can severely undermine the guarantees afforded to individuals, we develop formal impossibility results that enable early detection of DP misuses. Furthermore, building on our novel RAD bounds, we introduce a RAD-based auditing framework that generalizes beyond prior tools [40], [41], capturing the full spectrum of reconstruction risks and yielding more accurate and actionable privacy assessments. Our framework overcomes the fundamental scalability limitations of learning-based approaches [42], [43], enabling efficient auditing in high-dimensional settings. Our approach is strictly more general and produces *tighter empirical estimates* of the effective privacy budget than the state-of-the-art auditor [41]. In summary, our contributions in this field are:

- We expose formal mistakes in the literature and prove impossibility results for early detection and prevention of DP formalization flaws.
- We propose a RAD-based DP auditing framework that provides broader threat analyses and more accurate privacy-budget estimates than existing DP auditing techniques for distributed systems.

Overall, our results provide practical tools towards a safer use of DP technologies in practice.

Contributions towards Correlation-based Inference Attacks: We present theoretical bounds on the accuracy of Bayesian DP mechanisms and derive specific utility guarantees when certain correlation models are assumed. For each correlation model studied, we prove novel theorems that bound the Bayesian DP leakage of a DP mechanism. Finally, we provide insight into how our theoretical results apply in practice to real-world data containing Gaussian and Markov correlations. This allows us to confirm that our results enhance the utility of Bayesian DP mechanisms in actual applications.In summary, our contributions in this field are:

- We prove a bound on the BDP leakage of a DP mechanism with a fixed number of arbitrarily correlated records, showing it is tight. We call this the general bound.
- We derive a tighter BDP leakage bound for DP mechanisms under multivariate Gaussian correlations, improving on the general bound and prior work. This provides a systematic method for constructing more accurate BDP mechanisms tailored to Gaussian dependencies.
- We derive a BDP leakage bound for DP mechanisms under Markovian correlations, improving on the general bound when transition probabilities are similar. This enables the design of more accurate mechanisms than prior approaches in Markov settings.

Overall, our results for arbitrary Gaussian and Markov correlation models (Theorems 6.3, 6.14 and 6.20) advance the theoretical and practical understanding of BDP, enabling the reuse of DP mechanisms in correlated settings. This opens future directions for deriving correlation-specific bounds to design more accurate BDP mechanisms protecting against real-world correlation-based attacks.

1.2. Collaborations

During my thesis, I had the opportunity to collaborate with several co-authors across the papers that form the basis of this work. In the thesis, I use the academic "we" to honor these collaborations, since the research presented would not have been possible without them. In the following, I clarify the contributions of each collaborator to the relevant parts of the work.

Thorsten Strufe, my supervisor, collaborated with me on all papers. He provided extensive feedback on research ideas, helped refine the technical direction of the work, and contributed substantially to writing, editing, and presentation.

Javier Parra-Arnau and Jordi Forné contributed to the editing and writing of the papers they co-authored. Javier Parra-Arnau additionally contributed the initial ideas on the parallel composition problem.

Àlex Miranda-Pascual was a close collaborator during the early stages of my PhD. We worked hand in hand on both the systematization on trajectory privacy and the

composition works, which led to new results and insights. In particular, in the SoK and its extensions, he contributed to the utility analysis, the survey of similarity metrics, the syntactic notions review, and the systematization of syntactic, clustering and sampling mechanisms; while I focused on attacks, semantic privacy notions (DP granularities), and the systematization of additive noise mechanisms, synthetic DP generation and LDP. In the composition work, Àlex and I jointly developed and refined the extensions of composition results from pure DP to metric DP, including all novel theorems and their proofs. I subsequently extended these results to metric Gaussian DP, while he focused on approximate and zero-concentrated DP.

Annika Sauer, as my master thesis student, assisted with code and experiments for the attack-resilience papers. Héber H. Arcolezi contributed to the development of our new auditing framework with his analysis of the literature and experiment design. They both contributed to the writing and editing of the paper.

Martin Lange, also a master student of mine, supported the development of code and experiments during his master's thesis, leading to preliminary results that formed the foundation of the final work on BDP. I further refined these initial ideas into the final version, which would not have been possible without his early contributions.

2. Background

In this chapter, we introduce the relevant concepts for understanding this work and present the notation used throughout the thesis as summarized in Table 2.1. Particularly, we provide general background on DP, from the original definition to the most recent generalization to metric privacy, and a concise overview of composition properties. Additionally, we introduce the connection between DP and attack mitigation, along with extensions of DP designed to better capture this relationship: f-DP for the independent setting, and Bayesian DP for the case of correlated data. Finally, we introduce the probability and measure theory concepts required for the main results of this thesis.

2.1. Differential Privacy and Metric Privacy

We assume the database to consist of a finite number n of rows, $D = (x_1, \ldots, x_n) \in \mathcal{X}^n$, drawn from the joint distribution of the random vector $\mathbf{X} = (X_1, \ldots, X_n)$, where each row represents data associated with an individual, sampled from a universe of records $\mathcal{X}$. We denote by Π the joint distribution of $\mathbf{X}$ and by π the marginal distribution of individual records. Note that in the case where the entries of $\mathbf{X}$ are independent and identically distributed (i.i.d.), we have $\Pi = \pi^n$.

We use $[n] := \{1, \ldots, n\}$ to denote the set of indices. For a subset $K = \{i_1, \ldots, i_k\} \subseteq [n]$, we define the subvector $\mathbf{X}_K \in \mathcal{X}^k$ as $\mathbf{X}_K := (X_{i_1}, \ldots, X_{i_k})$. In particular, $\mathbf{X}_{-i}$ denotes $\mathbf{X}_K$ with $K = [n] \setminus \{i\}$, i.e., the whole dataset but one record, and we denote $\mathbf{x}_{-i} \equiv D_-$ when the position i is not relevant.

Let $\mathcal{D}(\Theta)$ denote the space of probability distributions over the output space Θ. We consider a (randomized) mechanism $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ that, given an input database $D \in \mathcal{X}^n$, produces a global output $\theta \in \Theta$ (e.g., an aggregate statistic or a trained model) with probability/density function $p_{\mathcal{M}}(\theta \mid D)$.

The attacker is assumed to know all records except for a target index $i \in [n]$, for which all possible values x_i and x_i' must be indistinguishable—corresponding to bounded DP [19]. This model allows us to formalize DP in the following definition:

Definition 2.1 (Differential Privacy [16])**.** A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{D}(\Theta)$ is called (ε, δ)-*differentially private*, if for all measurable sets $S \subseteq \Theta$, any target index $i \in [n]$, any target values $x_i, x_i' \in \mathcal{X}$, and any remaining values $\mathbf{x}_{-i} \in \mathcal{X}^{n-1}$, we have

$$\Pr_{\mathcal{M}}[Y \in S \mid \mathbf{X}_{-i} = \mathbf{x}_{-i}, X_i = x_i] \leq \mathrm{e}^{\varepsilon} \Pr_{\mathcal{M}}[Y \in S \mid \mathbf{X}_{-i} = \mathbf{x}_{-i}, X_i = x_i'] + \delta.$$

The output of $\mathcal{M}$ is represented by the random variable Y, which depends on the input data. If $\delta = 0$, we speak of *pure* DP (ε-DP). If $\delta > 0$, we speak of *approximate* DP.

Notation	Description
$\mathbb{D}$	Arbitrary database class
$\mathcal{X}$	Domain of a single record $x \in \mathcal{X}$.
$\mathbb{D}_{\mathcal{X}}$	Universe of all databases drawn from $\mathcal{X}$.
$\mathcal{X}^n$	Universe of datasets with n elements drawn from $\mathcal{X}$.
D, D'	Pair of databases.
$\lvert D\rvert$	Size of D (number of records).
$m_D(x)$	Multiplicity of x in the multiset D.
Θ	Domain of outputs of a mechanism.
θ	Element of Θ.
S	Measurable subset of Θ.
$\mathcal{M} : \mathbb{D} \to \mathcal{D}(\Theta)$	Randomized mechanism with input from domain $\mathbb{D}$ and output in codomain Θ.
$\mathbf{X} = (X_1, \ldots, X_n)$	Random vector representing the input of $\mathcal{M}$ when its domain is $\mathcal{X}^n$.
Π	Probability distribution on $\mathbf{X}$.
π	Probability distribution of a record X.
κ_π	Probability of re-sampling from π.
Y	Random variable representing output of $\mathcal{M}$.
$[n]$	Set $\{1, \ldots, n\}$ for $n \in \mathbb{N}$.
$\mathbf{X}_K = (X_{i_1}, \ldots, X_{i_k})$	Random vector of a subset $K = \{i_1, \ldots, i_k\} \subseteq [n]$ of the random variables $X_1, \ldots, X_n$.
$\mathbf{x}_K = (x_{i_1}, \ldots, x_{i_k})$	Database with k records belonging to $\mathcal{X}^k$.
$\mathcal{G}$	Granularity notion/neighborhood definition.
$D \sim_{\mathcal{G}} D'$	D and D' are $\mathcal{G}$-neighboring.
$d_{\mathbb{D}}$ (or d)	Metric over $\mathbb{D}$.
$d_{\mathbb{D}}^{\mathcal{G}}$	Canonical metric of $\mathcal{G}$ over $\mathbb{D}$.
$\mathcal{U}$, $\mathcal{B}$	Unbounded and bounded granularity (resp.)
$D \triangle D'$	Symmetric difference $((D \cup D') \backslash (D \cap D'))$.
$d_H(D, D')$	Hamming distance between two databases.
$I_f(D, D')$	For $f = \{f_i\}_{i \in [k]}$, $\lvert\{i \in [k] \mid f_i(D) \neq f_i(D')\}\rvert$.
$\Phi(\cdot)$	Cumulative distribution function (CDF) of the standard normal distribution.

Table 2.1.: Notation Summary

The *privacy leakage* ε, also known as the *privacy budget*, determines how closely the probabilities of observing the same output must align for two databases D and D' such that $d_H(D, D') = 1$, where d_H denotes the Hamming distance, and D and D' are called *bounded-neighboring* databases.

Intuitively, if the output distribution of D is close (within a multiplicative factor of e^ε) to the output distribution of D', then an attacker cannot reliably determine whether D or D' was the input to the mechanism. Consequently, the two scenarios—one in which the dataset contains x_i and one in which it contains x_i'—are essentially *indistinguishable* (up to ε) [15]. A smaller ε provides stronger privacy guarantees—it is harder for an attacker to distinguish if the mechanism has been executed on D or on D'—but typically comes at the cost of utility [16] (cf. Proposition 2.10). The parameter δ permits certain violations of ε-DP while quantifying their probability and severity [44].

The original, *central* DP notion assumes the presence of a trusted party (data curator) who executes the mechanisms protecting the sensitive data. If no party with shared trust exists, it is necessary to distribute the curation to all participants. The corresponding *local differential privacy* (LDP) [16], assumes every individual holds their own data which is randomized on the client side before being transmitted to a data collector. They hence contribute partial answers to queries on the whole data, enforcing DP locally. Formally, in LDP mechanisms: $n = 1$, i.e., $\mathcal{M}$ takes as input a single data record $x \in \mathcal{X}$. LDP is a rigorous and increasingly relevant privacy model [16], especially suitable for privacy-sensitive applications such as telemetry and location-based services where no trusted data curator is considered [45].

Note that DP satisfies important properties, such as invariance by post-processing, i.e., just by operating in the output of a DP query, it is impossible to violate DP guarantees unless additional information about the input is provided.

Proposition 2.2 (Post-processing [16])**.** *Given $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ satisfying (ε, δ)-DP and $f\colon \Theta \to \Theta'$ be an arbitrary map, then $f \circ \mathcal{M}$ is (ε, δ)-DP.*

Here, $f \circ \mathcal{M}$ is an abuse of notation intended to represent the process of first sampling an output $\theta \sim \mathcal{M}(D)$ and then applying $f(\theta)$.

Another interesting property relates the protection of several records simultaneously. This is known as the *group privacy* property and shows that the privacy degrades linearly with respect to the group size for pure DP mechanisms:

Proposition 2.3 (Group privacy [16])**.** *Given $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ satisfying (bounded) ε-DP, then, for any $D, D' \in \mathcal{X}^n$, and for all measurable sets $S \subseteq \Theta$:*

$$\Pr_{\mathcal{M}}[Y \in S \mid D] \leq \mathrm{e}^{\varepsilon d_H(D,D')} \Pr_{\mathcal{M}}[Y \in S \mid D'].$$

This result shows that if the indistinguishability of one individual record is bounded by ε, changing m records is indistinguishable up to $m\varepsilon$.

This thesis primarily focuses on bounded DP (Definition 2.1), given its wide applicability and its close connection to attack resilience and DP extensions such as Gaussian DP (GDP) and Bayesian differential privacy (BDP). Nevertheless, many other granularities

of privacy[1] exist [18]—i.e., alternative definitions of what constitutes a sensitive "entry" in the database whose alteration should leave output probabilities nearly unchanged (up to a factor of ε) [16].

For example, consider a dataset represented as a graph, where nodes correspond to individuals, and edges encode social relationships. In this setting, it is impossible to change all information about one individual without simultaneously affecting others, since each edge represents information shared between at least two participants. In such cases, it may be more appropriate to define neighboring datasets by the addition or removal of a single edge (*edge-DP* [46]), thereby protecting connections between individuals as sensitive information.

Similarly, in streaming data applications, *event-level* DP [16] is frequently adopted. Although each stream belongs to a single individual, two datasets are considered neighboring if they differ in only one time-step value. This granularity is relevant for time series applications like trajectory data (see Section 3.3). Moreover, we present a practical use case for electricity consumption and activity time series in Section 6.5.

Even within the original domain of tabular data, different granularities of privacy can be defined. For instance, in *unbounded* DP, (i) the data domain comprises all datasets with rows drawn from $\mathcal{X}$, denoted $\mathbb{D}_{\mathcal{X}}$, rather than fixing the dataset size to n; and (ii) indistinguishability is enforced with respect to the addition or removal of a single row, i.e., $|D \triangle D'| = |(D \cup D') \backslash (D \cap D')| \leq 1$, rather than the modification of an existing row, as in bounded DP. This way, unbounded DP represents a pure membership inference, in which the attacker knows the exact record of the target and just tries to know if it participated or not without even knowing the actual database size.

Adjusting the granularity of privacy enables the modeling of protection against distinct types of privacy threats [18], [47] and supports the adaptation of the original definition to scenarios where what constitutes an individual's information is not trivially discretized, such as in social networks.

We generalize the definition of granularity notion $\mathcal{G}$ as follows:

Definition 2.4 ($\mathcal{G}$-neighborhood)**.** Given a database class $\mathbb{D}$, we define the *$\mathcal{G}$-neighborhood relation* as a binary symmetric relation $\sim_{\mathcal{G}}$ between elements in $\mathbb{D}$. We say that $D, D' \in \mathbb{D}$ are *$\mathcal{G}$-neighboring* if $D \sim_{\mathcal{G}} D'$.

We will use calligraphic letters to denote certain granularity notions (e.g., $\mathcal{U}$ for unbounded, $\mathcal{B}$ for bounded). Definition 2.4 generalizes (pure) bounded DP: A mechanism $\mathcal{M} \colon \mathbb{D} \to \mathcal{D}(\Theta)$ is $\mathcal{G}$ *ε-DP* ($\varepsilon \geq 0$) if for all $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$ and all measurable $S \subseteq \Theta$,

$$\Pr_{\mathcal{M}}[Y \in S \mid D] \leq \mathrm{e}^{\varepsilon} \Pr_{\mathcal{M}}[Y \in S \mid D'].$$

The group property of DP (Proposition 2.3) enables a formulation of DP in terms of the Hamming distance, thereby motivating the use of metrics to quantify privacy guarantees. This idea first appeared in [19] with the symmetric distance, $|D \triangle D'|$, for

[1] The granularity of privacy is also commonly referred to as the *neighborhood* or *adjacency definition*. We adhere to the terminology used in [16, Chapter 2, pp. 23–24].

unbounded DP reformulation. Later, Chatzikokolakis et al. [47] introduce generalization to arbitrary (not necessarily discrete) metrics, called *metric privacy* or $d_{\mathbb{D}}$-privacy, when the metric $d_{\mathbb{D}}$ is specified.

Note that while we generally refer to d as *metric* to simplify terminology, d can be any *extended pseudometric* $d_{\mathbb{D}} \colon \mathbb{D}^2 \to [0, \infty]$, i.e., a metric in which the distance between two different databases can also be 0 and ∞ [47]. Equipping $\mathbb{D}$ with a (pseudo)metric $d_{\mathbb{D}}$ induces a (pseudo)metric space, $(\mathbb{D}, d_{\mathbb{D}})$, which we call *privacy space*. With this notation in place, we formally define metric privacy.

Definition 2.5 ($d_{\mathbb{D}}$-privacy [47])**.** Let $(\mathbb{D}, d_{\mathbb{D}})$ be a privacy space. Then, a randomized mechanism $\mathcal{M} \colon \mathbb{D} \to \mathcal{D}(\Theta)$ is $d_{\mathbb{D}}$*-private* if for all $D, D' \in \mathbb{D}$ and all measurable $S \subseteq \Theta$,

$$\Pr_{\mathcal{M}}[Y \in S \mid D] \leq \mathrm{e}^{d_{\mathbb{D}}(D,D')} \Pr_{\mathcal{M}}[Y \in S \mid D'].$$

Observe that the metric absorbs the privacy budget ε, i.e., $d_{\mathbb{D}}$ can be written as $d_{\mathbb{D}} = \varepsilon d'_{\mathbb{D}}$ where $d'_{\mathbb{D}}$ is also a metric. This definition makes it challenging for an adversary to distinguish between databases D and D' that are "close" according to the metric d. However, if the two databases are significantly different, the output distributions can differ more, making it easier for the adversary to distinguish them.

Note that, due to the group property of DP, metric privacy is equivalent to DP when considering the Hamming distance scaled by ε. More generally, given a data domain $\mathbb{D}$ (e.g., social networks) we can construct a *canonical metric* $d^{\mathcal{G}}_{\mathbb{D}}$ for each granularity $\mathcal{G}$ over $\mathbb{D}$ (e.g., edge DP) [47]. It suffices to define the distance between two databases $d^{\mathcal{G}}_{\mathbb{D}}(D, D')$ as the minimum number of neighboring databases in $\mathbb{D}$ one needs to cross to obtain D' from D (e.g., number of edges one needs to add or delete to transform one network into the other), defining $d^{\mathcal{G}}_{\mathbb{D}}(D, D') = \infty$ if it is not possible to transform one dataset into the other through a chain of neighboring transformations.

As mentioned, the canonical metric for bounded DP is the Hamming distance, $d^{\mathcal{B}}_{\mathcal{X}^n}(D, D') = d_H(D, D')$ [47], that counts the number of record changes. As another example, in unbounded DP, we build neighbors by adding/deleting one record, hence the canonical metric is the symmetric difference between two sets:

$$d^{\mathcal{U}}_{\mathbb{D}_{\mathcal{X}}}(D, D') = |D \triangle D'| = |(D \cup D') \backslash (D \cap D')|,$$

which measures exactly the number of deletions and additions needed to go from D to D' [19].

Note that for every granularity, its canonical metric, $d^{\mathcal{G}}_{\mathbb{D}}(D, D')$, is always well defined (see Proposition A.1) and satisfies $d^{\mathcal{G}}_{\mathbb{D}}(D, D') = 1$ if and only if $D \sim_{\mathcal{G}} D'$. Moreover, we can extend the group privacy property of bounded DP for any granularity (see Proposition A.2 for proof details) obtaining the following relation.

Proposition 2.6. *Let $\mathcal{G}$ be a granularity notion over the database class $\mathbb{D}$. Then, a mechanism $\mathcal{M}$ with domain $\mathbb{D}$ is $\varepsilon d^{\mathcal{G}}_{\mathbb{D}}$-private if and only if it is $\mathcal{G}$ ε-DP.*

Given any granularity notion we can obtain a metric, but not all metrics are the canonical metric for a granularity notion. Therefore, the notion of $d_{\mathbb{D}}$-privacy is more general than $\mathcal{G}$ ε-DP.

Utility of private mechanisms. While protecting privacy is a primary goal of DP, privacy without utility is meaningless. From a technical standpoint, the most private mechanism is one that releases no information at all. However, DP is employed precisely because we seek to extract population-level insights—such as population averages or data distributions. Consequently, although the DP definition quantifies how effectively individual contributions are obscured, it must be complemented by metrics that capture how well global information is preserved, a notion commonly referred to as utility.

A well-established metric for quantifying the utility of a private mechanism is the (α, β)-accuracy [30], [48]. It captures how well the mechanism approximates a true statistic or function while considering the inherent randomness introduced by the mechanism:

Definition 2.7 ((α, β)-Accuracy [48])**.** A mechanism $\mathcal{M}$ is (α, β)-*accurate, with respect to a function* f and an error function Err, if for all databases $D \in \mathcal{X}^n$ we have

$$\Pr[\mathrm{Err}(\mathcal{M}(D), f(D)) \geq \alpha] \leq \beta.$$

A randomized mechanism $\mathcal{M}$ is (α, β)-accurate if an error of magnitude α has a probability of at most β. Thus, the smaller α and/or β, the better the accuracy of mechanism $\mathcal{M}$. Here, α quantifies the error tolerance, and β the failure probability. More precisely, it refers to the utility guarantee that with probability at least $1 - \beta$, the mechanism's output is within an interval of radius α centered on the true value.

For numerical queries, the absolute error (ℓ_1) is typically used as the error function Err [16]. For example, if an $(1, 0.05)$-accurate mechanism estimating the average age of a population outputs $\theta = 32$, then, with probability 0.95, the true average age lies between 31 and 32 years. In conclusion, the (α, β)-accuracy provides an interpretable theoretical measure of utility that accounts for the stochastic nature of DP mechanisms. Moreover, it can be empirically estimated using confidence intervals.

Although the privacy parameters influence the privacy–utility trade-off, different strategies and mechanisms may satisfy the same (ε, δ)-DP guarantee while offering substantially different utility and, in some cases, different levels of effective protection, as we illustrate in Chapter 5. One of the earliest and most widely used mechanisms proven to satisfy ε-DP is the Laplace mechanism [16]:

Definition 2.8 (Laplace Mechanism [16])**.** Let $f : \mathcal{X}^n \to \mathbb{R}^k$ be a function and its *sensitivity* defined as

$$\Delta f := \sup_{d_H(D, D')=1} ||f(D) - f(D')||_1.$$

Given that $\Delta f < \infty$ and $\varepsilon > 0$, the Laplace mechanism is defined for all $D \in \mathcal{X}^n$ as $\mathcal{M}_{\varepsilon, f}(D) = f(D) + (Z_1, \ldots, Z_k)$ where Z_i are i.i.d. random variables that follow the Laplace distribution centered at 0 and with scale $\frac{\Delta f}{\varepsilon}$.

Note that the sensitivity can be extended to metric spaces as:

Definition 2.9 (Sensitivity [47])**.** Let $(\mathbb{D}_1, d_1)$ and $(\mathbb{D}_2, d_2)$ be two privacy spaces and let $f \colon \mathbb{D}_1 \to \mathbb{D}_2$ be a deterministic map. We define the *sensitivity of* f with respect to d_1 and d_2 as the smallest value $\Delta f \in [0, \infty]$ such that $d_2(f(D), f(D')) \leq \Delta f \, d_1(D, D')$ holds for all $D, D' \in \mathbb{D}_1$ with $d_1(D, D') < \infty$.

Hence, we can obtain the metric version of the Laplace mechanism considering the sensitivity of f with respect to $(\mathbb{D}, d_{\mathbb{D}})$ and $(\mathbb{R}^n, \ell_1)$.

The Laplace mechanism privacy-utility trade-off offers a simple characterization that relates ε directly with its accuracy:

Proposition 2.10 ([16])**.** *Let $\mathcal{M}_{\varepsilon,f}$ be the Laplace mechanism. Let $\beta \in (0, 1]$ be a probability. Then $\mathcal{M}_{\varepsilon,f}$ is (α, β)-accurate with respect to f with $\alpha = \ln(\beta^{-1}) \frac{\Delta f}{\varepsilon}$.*

This accuracy result for the Laplace mechanism is tight [16] and reflects the effect of ε in utility: A larger ε (corresponding to higher privacy leakage) results in a smaller error α at a fixed confidence level, thereby improving accuracy.

Among the mechanisms designed to achieve approximate DP, (ε, δ)-DP, the *Gaussian mechanism* is one of the most widely used [26], [38]. Especially in machine learning applications, as the core mechanism for DP-SGD [49]. Its favorable composition properties and analytical tractability make it a cornerstone of both theoretical developments and practical implementations of DP.

Definition 2.11 (Gaussian Mechanism [50])**.** Let $f : \mathcal{X}^n \to \mathbb{R}^k$ be a function and its *ℓ_2-sensitivity* defined as

$$\Delta f := \sup_{d_H(D,D')=1} ||f(D) - f(D')||_2.$$

Given that $\Delta f < \infty$, the Gaussian mechanism is defined for all $D \in \mathcal{X}^n$ as $\mathcal{M}_\sigma(D) = f(D) + (Z_1, \ldots, Z_k)$ where Z_i are i.i.d. random variables that follow $\mathcal{N}(0, \sigma^2 I_k)$; a multivariate Gaussian distribution with mean zero and covariance matrix $\sigma^2 I_k$. For every σ satisfying:

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) \leq \delta, \tag{2.1}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Balle and Wang [50] proved that any $\mathcal{M}_\sigma$ Gaussian mechanism with σ verifying Eq. 2.1 satisfies (ε, δ)-DP.

Moreover, using the standard Gaussian tail bound we can express the accuracy of the Gaussian mechanism directly in terms of σ—also termed as noise scale:

Proposition 2.12. *Let $\mathcal{M}_{\sigma,f}$ be the Gaussian mechanism. Let $\beta \in (0, 1]$ be a probability. Then $\mathcal{M}_{\sigma,f}$ is (α, β)-accurate with respect to f with $\alpha \leq \sigma\, \Phi^{-1}(1 - \frac{\beta}{2}) \leq \sigma\sqrt{2\ln\left(\frac{2}{\beta}\right)}$.*

Importantly, while these mechanisms provide DP guarantees for a single query $f(D)$, most real-world data analysis tasks involve answering multiple queries. To address this more realistic scenario, the next section discusses the composition properties of DP, i.e., how the privacy loss evolves when publishing the output of several DP mechanisms.

2.2. Composition

One of the most useful properties of DP mechanisms relates to composition theorems. Sequential and parallel composition are considered key components of DP and are regularly used in the field [26].

The composition theorems share a common foundation. Simply put, these theorems state that given k ε_i-DP mechanisms $\mathcal{M}_i$, the composed mechanism $\mathcal{M}$ that applies each $\mathcal{M}_i$ to the dataset and releases the corresponding outputs satisfies ε-DP, where ε depends on $\varepsilon_1, \ldots, \varepsilon_k$. In other words, these theorems estimate the privacy loss (i.e., the final privacy budget) of the mechanism $\mathcal{M}$ that can be discretized on several mechanisms $\mathcal{M}_i$. To be precise, we formalize the adaptive-composed mechanism as follows:

Definition 2.13 (Adaptive-composed mechanism)**.** For $i \in [k]$, let $\overline{\Theta}_i \coloneqq \Theta_1 \times \cdots \times \Theta_{i-1}$ (where $\overline{\Theta}_0 = \{\varnothing\}$), and let $\mathcal{M}_i \colon \overline{\Theta}_i \times \mathbb{D} \to \mathcal{D}(\Theta_i)$ be randomized mechanisms. We define the *adaptive-composed mechanism* $\mathcal{M} \coloneqq (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ as the mechanism with domain $\mathbb{D}$ such that $\mathcal{M}(D) = (\mathcal{N}_1(D), \ldots, \mathcal{N}_k(D))$ for all $D \in \mathbb{D}$, where $\mathcal{N}_i(D)$ are defined recursively as $\mathcal{N}_i(D) = \mathcal{M}_i(\mathcal{N}_1(D), \ldots, \mathcal{N}_{i-1}(D), D)$ for $i \in [k]$ (where $\mathcal{N}_1 = \mathcal{M}_1$).

In other words, given $D \in \mathbb{D}$, $\mathcal{M}$ first draws θ_1 following the distribution of $\mathcal{M}_1(D)$; then, $\mathcal{M}_i$ draws θ_i following the distribution of $\mathcal{M}_i(\theta_1, \ldots, \theta_{i-1}, D)$ for each $i = 2, \ldots, k$ in order. At the end, $\mathcal{M}$ outputs $(\theta_1, \ldots, \theta_k)$. For example, we might first ask which road in a city is the most crowded. Once we obtain the noisy answer $\mathcal{M}_1(D) = \theta$, we may then ask how many cars are on that road, using $\mathcal{M}_2(D, \theta)$. If instead the first mechanism had returned θ', $\mathcal{M}_2$ would output the noisy count for θ'. In this sense, $\mathcal{M}_2$ *adapts* based on previous outputs.

Particularly, composition is *independent* in the particular case in which previous outputs are ignored by subsequent mechanisms, hence the outputs of each $\mathcal{M}_i$ are not affected by each other, i.e., $\mathcal{M}_1(D), \ldots, \mathcal{M}_k(D)$ are mutually independent random elements for all $D \in \mathbb{D}$. For instance, when we answer simultaneously many independent queries, such as the average salary and the average age—for each of them, we use a DP mechanism, but the output of one is not used to compute the other.

Note that adaptive-composed mechanisms are more general than independent-composed mechanisms. Consequently, all the results on adaptive composition include the independent composition as a particular case.

Sequential vs. parallel: Orthogonally, if every $\mathcal{M}_i$ takes the whole database D as input, the composition is called *sequential*. Alternatively, the composition is *parallel* if each $\mathcal{M}_i$ uses only data from a subset $D_i \subseteq D$ disjoint of the subset D_j used by any other $\mathcal{M}_j$ with $j \neq i$.

The first composition result of DP appeared in [51] for unbounded DP, however, it has already been generalized to both bounded and unbounded definitions:

Theorem 2.14 (Adaptive sequential composition (ASC) [27])**.** *Let $\mathcal{M}_1, \ldots, \mathcal{M}_k$ be k mechanisms such that for all $\overline{\theta}_i \in \overline{\Theta}_i$, $\mathcal{M}_i(\overline{\theta}_i, \cdot)$ satisfies [unbounded/bounded] ε_i-DP. The adaptive-composed mechanism $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ as in Def. 2.13, satisfies [unbounded/bounded] $(\sum_{i=1}^{k} \varepsilon_i)$-DP.*

This theorem corresponds to the adaptive definition and includes independent composition as a particular instance.

Sequential composition provides an explicit formula to compute the privacy leakage of a composed mechanism. However, the total privacy leakage increases linearly with the number of mechanisms. As a result, either the cumulative privacy leakage becomes significant when many queries are answered, or one must add substantial noise to each query to maintain the same privacy budget, leading to a considerable loss of utility.

At the same time, sequential composition assumes that each mechanism accesses the entire dataset. However, most queries only require information from a specific subset of the data. For example, to count the number of cars on a street, we only need the subdataset containing cars on that street, without accessing the locations of other cars. Taking this into account allows us to obtain a tighter bound on the overall privacy loss, which is formalized by the *parallel composition* theorem.

Theorem 2.15 (Parallel composition [19])**.** *Let $\mathcal{M}_i$ each provide [unbounded] ε-DP. Let $\mathcal{X}_i$ be arbitrary disjoint subsets of the universe of records $\mathcal{X}$. The sequence of $\mathcal{M}_i(D_i)$ provides [unbounded] ε-DP, where $D_i \subseteq D$ is the multiset such that element $x \in D$ has multiplicity $m_{D_i}(x) = \mathbf{1}_{\mathcal{X}_i}(x)\, m_D(x)$.*

By abuse of notation, D_i is also often denoted as $D \cap \mathcal{X}_i$. Alternatively, Li et al. [27] use a partitioning function p to define the disjoint subsets in the previous statement, i.e., $p_i(D) = D_i$ for all i and $D \in \mathbb{D}_{\mathcal{X}}$.

Even though sequential composition (Theorem 2.14) was initially stated for the unbounded granularity notion, it can easily be translated for other granularities, such as bounded DP. However, in Theorem 2.15, if instead of unbounded, we impose $\mathcal{M}_i$ to be bounded ε-DP, then it is not generally true that the sequence of $\mathcal{M}_i(D_i)$ provides bounded ε-DP for any $\varepsilon > 0$. Li et al. [27] show why the proof is not applicable: Even if $\mathcal{M}_i$ are bounded ε-DP, $\mathcal{M}'_i$ such that $\mathcal{M}'_i(D) = \mathcal{M}_i(D_i) = \mathcal{M}_i(D \cap \mathcal{X}_i)$ is not necessarily bounded ε-DP. This fact is clear in the following counterexample, which we provide to complete Li et al.'s claim [27]:

Example 2.16 (Parallel composition does not hold for bounded DP)**.** Let $\mathbb{D}_{\mathcal{X}}$ be a database universe and $\mathcal{X}_i$ arbitrary disjoint subsets of $\mathcal{X}$ (e.g., tuples of age and blood pressure divided by age ranges). We show that given $k > 1$ mutually independent bounded ε_i-DP mechanisms $\mathcal{M}_i \colon \mathbb{D}_{\mathcal{X}} \to \mathcal{D}(\Theta)$, it is not necessarily true that the composed mechanism $\mathcal{M} \colon \mathbb{D}_{\mathcal{X}} \to \mathcal{D}(\Theta)$ such that $\mathcal{M}(D) = (\mathcal{M}_1(D_1), \ldots, \mathcal{M}_k(D_k))$ is bounded DP, where $D_i \subseteq D$ is the multiset such that element $x \in D$ has multiplicity $m_{D_i}(x) = \mathbf{1}_{\mathcal{X}_i}(x)\, m_D(x)$, i.e., the subset of elements belonging to $\mathcal{X}_i$.

To do so, we prove that we can select $k > 1$ mutually independent bounded ε_i-DP mechanisms $\mathcal{M}_i \colon \mathbb{D}_{\mathcal{X}} \to \mathcal{D}(\Theta)$ such that mechanism $\mathcal{M} \colon \mathbb{D}_{\mathcal{X}} \to \mathcal{D}(\Theta)$ with $\mathcal{M}(D) = (\mathcal{M}_1(D_1), \ldots, \mathcal{M}_k(D_k))$ is not bounded ε-DP for any $\varepsilon \geq 0$.

For all $i \in [k]$, we choose $\mathcal{M}_i \colon \mathbb{D}_{\mathcal{X}} \to \mathcal{D}(\Theta)$ such that they output the number of elements of the input database, i.e., $\mathcal{M}_i(D) = \mathcal{M}^*(D) = |D|$ for all $D \in \mathbb{D}_{\mathcal{X}}$. It can easily be checked that this mechanism is bounded 0-DP: Since all bounded neighbors

have the same size by definition, revealing the size does not allow to distinguish them. Observe that in this case, the composed mechanism

$$\mathcal{M}(D) = (\mathcal{M}^*(D_1), \dots, \mathcal{M}^*(D_k)) = (|D_1|, \dots, |D_k|),$$

outputs the number of participants in each class (e.g. the number of participants in each age range).

Let $D, D' \in \mathbb{D}_{\mathcal{X}}$ be two bounded-neighboring databases such that $D \triangle D' = \{x, x'\}$ with $x \in D_j$ and $x' \in D'_l$, $j \neq l$ (e.g. we change one young healthy person by an elderly with hypertension). Then it is clear that $\mathcal{M}^*(D_j) = |D_j| \neq |D'_j| = \mathcal{M}^*(D'_j)$ (analogously to l), so

$$\Pr[\mathcal{M}^*(D_j) = |D_j|] = 1 \not\leq 0 = \Pr[\mathcal{M}^*(D'_j) = |D_j|].$$

Note that this is not a contradiction with $\mathcal{M}^*$ being bounded DP, since D_j and D'_j are not bounded-neighboring databases.

Consequently, taking $\theta = (|D_1|, \dots, |D_n|) \in \Theta$ we obtain $\Pr[\mathcal{M}(D) = \theta] = 1$, but $\Pr[\mathcal{M}^*(D'_j) = |D_j|] = 0$. Therefore $\Pr[\mathcal{M}(D) = \theta] = 1 \not\leq 0 = \mathrm{e}^{\varepsilon} \Pr[\mathcal{M}(D') = \theta]$ for all $\varepsilon \geq 0$, so the mechanism $\mathcal{M}$ is not bounded DP.

Importantly, this example illustrates that the composition results proved for one granularity (in this case unbounded DP in $\mathbb{D}_{\mathcal{X}}$) cannot be trivially generalized to other data domains or neighborhood definitions. The failure of bounded DP on satisfying $(\max_{i \in [k]} \varepsilon_i)$-DP when composed in parallel opens a new question about how to measure the privacy of composed mechanisms for general data domains and granularities.

We answer this question by generalizing these composition theorems to more general scenarios, in which the domain of the mechanism is not necessary $\mathbb{D}_{\mathcal{X}}$, and the given granularity notion is not necessarily unbounded in Chapter 2.

2.3. Differential Privacy and Attack Resilience

Inference attacks aim to extract sensitive information from released data. Real-world case studies have shown that such attacks can successfully re-identify individuals in pseudonymized datasets—i.e., data where only name and direct identifiers are removed—exposing severe privacy risks [9], [10], [11], [13], [14]. These concerns motivated the development of DP, which enables individuals to participate in data analysis while limiting the additional risk incurred by their inclusion—capturing the principle that nothing about an individual should be learnable from a dataset that could not be learned had they not participated [15].

Adversaries are commonly distinguished according to their level of access to the mechanism. This is typically captured by a binary distinction: *white-box* versus *black-box* access. In the original DP setting, the adversary is assumed to have *white-box* access, meaning that they know the mechanism exactly—including its internal structure and parameters—and can choose the queries to be executed. With the emergence of more complex tasks such as machine learning systems, a different threat model has been considered: the *black-box* adversary, who only observes input–output pairs obtained

through querying the system [39]. For example, when a statistic is computed adaptively by combining multiple queries under composition (see Section 2.2), the adversary may either observe all intermediate query results (white-box access) or only the final released statistic (black-box access). Similarly, in a learning setting, the adversary may have access to the trained model parameters (white-box) or may only be able to query the model and observe its outputs (black-box).

Orthogonal to this classification, we identify two main lines of research concerning the resilience of DP to attacks: those assuming independence among data records and those considering correlated data. The scope and maturity of these two lines of work are markedly unbalanced. Independence among data records has been extensively studied and remains a central and active topic in the literature, whereas correlation-aware analyses have only recently emerged and are still in their early stages. In the following, we review and contextualize the existing work in both directions.

2.3.1. Independence Assumption

Most of the literature considers the original DP assumption, in which each record is independent from the others, hence changing one record does not affect the rest. Consequently, previous work on attack resilience considers an *informed adversary* [39], since, under the assumption that records are independently drawn from π, bounding the performance of such an attacker also bounds the performance of any attacker with less information [39]. Formally, for any target record x an *informed adversary* [39] has access to: the fixed dataset $D_- = D \setminus \{x\}$, the distribution of data records π, the output θ of the model trained on $D_x = D_- \cup \{x\}$, the mechanism $\mathcal{M}$, and optional target-specific auxiliary knowledge $a(x)$ about target record x.

Among the various adversarial objectives, membership inference attacks (MIAs)[2] have been the most extensively studied threat in the context of DP [33], [54], [55], [56]. In an MIA, the attacker knows the full target record, $a(x) = x$, and seeks only to determine whether this record was included in the dataset. This attack model aligns closely with the core intuition of DP, which aims to ensure that the participation of x is indistinguishable from that of any alternative record (see the discussion following Definition 2.1). It is therefore natural that MIAs have received significant attention within the DP community.

However, MIAs capture only one possible threat—sometimes not even the most relevant one (see Section 3.2). In many settings, an attacker may aim to infer information beyond mere membership. For instance, in an attribute inference attack (AIA), each record is structured as $x = (x_1, x_2)$, where $a(x) = x_1$ represents the public attributes, and the attacker's goal is to reconstruct the sensitive attribute x_2. This threat model is therefore more general, as it captures a broader range of privacy risks, including different types of sensitive information encoded in the attributes of a record.

Recently, data reconstruction attacks (DRAs) have been proposed as unifying framework that generalizes a wide range of attack models [39]. This perspective enables reasoning about different privacy threats within a single, coherent formal framework. In a DRA

[2] MIAs generalize table attacks [52], [53] for non-tabular data.

the adversary's goal is to correctly reconstruct completely or partially the target record x, potentially given auxiliary knowledge $a(x) \in aux$ about the target. For example, an attacker that tries to reconstruct a license plate number from a target's car image, may already know the color of the car.

Formally, a DRA, denoted by $A\colon \Theta \times aux \to \mathcal{X}$ uses the output of a DP mechanism $\theta \sim \mathcal{M}(D)$ and the target auxiliary information $a(x)$ to produce a candidate $\tilde{x} = A(\theta, a(x))$. Note that, in case of composing several mechanisms, we consider the final output after the whole process.

The attack is considered successful if the output is similar enough (according to a success threshold η) to the real record x: $\ell(\tilde{x}, x) \leq \eta$ [39]. The error function ℓ depends on the context and allows one to model the fact that, in complex data domains, it may be sufficient to partially reconstruct the target. For instance, in the image domain, even if not all pixels are correct, we may gather sensitive information such as the action performed in the image. Consequently, ℓ may be chosen as an image-specific metric, such as the learned perceptual image patch similarity (LPIPS) [39]. Given the error function ℓ and the threshold η, we define the *success set* of a target x as $S_\eta(x) = \{x' \in \mathcal{X}\colon \ell(x, x') \leq \eta\}$.

DRAs cover AIAs and MIAs as particular cases [39]: In a classic AIA, given $x = (x_1, x_2)$ we define $a(x) = x_1$, $\phi(x) = x_2$ and $\ell(\tilde{x}, x) = 0$ if $\phi(\tilde{x}) = \phi(x)$ and one otherwise. Similarly, in a MIA, $a(x) = x$ and ℓ is the characteristic function such that $\ell(\tilde{x}, x) = 0$ when $\tilde{x} = x$ and one otherwise. Thus, DRAs cover a broad range of commonly studied privacy risks, including MIAs and AIAs as particular instances [39]. Accordingly, in this thesis we focus on analyzing DRAs.

Now the questions arises about (i) how to evaluate the performance of a DRA and (ii) how much DP mitigates that performance. For the particular cases of AIA and MIAs, the current literature [54], [57] agrees on the following metric:

Definition 2.17 (Adapted from [54])**.** Given π the distribution of data records and $\mathcal{M}, \phi(x), a(x), A$ as defined above, the *attribute advantage*, Adv_{AIA}, is defined as

$$\Pr_{\substack{x_0 \sim \pi \\ \theta \sim \mathcal{M}(D_{x_0})}} [A(\theta, a(x_0)) = \phi(x_0)] - \Pr_{\substack{x_0, x_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{x_1})}} [A(\theta, a(x_0)) = \phi(x_0)] \,.$$

The attribute advantage quantifies the adversary's gain in correctly inferring a sensitive attribute $\phi(x)$ when a record $x \in D$ is included in the dataset, compared to when it is drawn from the underlying distribution π. The second term in Definition 2.17 accounts for cases where the attribute could be predicted even without the record in the dataset (e.g., via imputation [58]). In the context of MIAs, the advantage reduces to the true positive rate (TPR) minus the false positive rate (FPR), effectively discounting trivial attacks, such as always predicting "member", which achieve high success probability (the probability to correctly identify a member is one) without revealing any meaningful private information.

Values $\mathrm{Adv} \leq 0$ indicate no privacy risk, as the attacker performs equally well (or even better) on non-members than on members. The maximum achievable advantage is $1 - \kappa_\pi$, where $\kappa_\pi = \Pr_{X, X' \sim \pi}[X = X']$, i.e., the probability of resampling from the distribution π [54]. This value corresponds to total disclosure: the attacker perfectly infers the

attribute or membership of dataset members while always failing on non-members. The advantage can be normalized by $1 - \kappa_\pi$ to obtain an upper bound of one.

Many works have formalized direct connections between privacy parameters and membership advantage (e.g. [54], [59], [60], [61]), leading to a tight bound for the *strong attacker* presented by Humphries et al. [33]. The strong attacker is a particular case of informed attacker that only hesitates among two possible members, i.e. $\mathcal{X} = \{x_0, x_1\}$ and π is uniform among those. Hence, $\kappa_\pi = 0.5$ and the normalized advantage can be bounded for every DP mechanism for i.i.d. data:

$$\mathrm{Adv}_{MIA} \leq \frac{\mathrm{e}^\varepsilon - 1 + 2\delta}{\mathrm{e}^\varepsilon + 1}. \tag{2.2}$$

The authors also prove that this bound holds for any other informed or even weaker attacker models [33] under independence assumption.

AIAs have been less studied: Existing works that provide theoretical bounds for AIAs either analyze specific attack strategies [54] or adopt more general DRA frameworks [39].

The current proposed performance metric for general DRAs [39] is reconstruction robustness (ReRo). ReRo does not define an advantage but instead only accounts for the success probability of an attack that has as input solely the output of the DP mechanism and the known dataset D_-, ignoring any possible target-specific auxiliary knowledge:

Definition 2.18 (ReRo [39])**.** Let π be a prior over $\mathcal{X}$ and $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ an error function. Mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{D}(\Theta)$ is (η, γ)-reconstruction robust with respect to π, ℓ if for any dataset $D_- \in \mathcal{X}^{n-1}$ and any reconstruction adversary $A : \Theta \to \mathcal{X}$,

$$\Pr_{\substack{X \sim \pi, \\ \theta \sim \mathcal{M}(D_X)}} [\ell(X, A(\theta)) \leq \eta] \leq \gamma.$$

The first bound for ReRo under ε-DP was given by [39]:

$$\gamma \leq \kappa^+_{\pi,\ell}(\eta)\mathrm{e}^\varepsilon, \tag{2.3}$$

where $\kappa^+_{\pi,\ell}(\eta) = \sup_{x_0 \in \mathcal{X}} \Pr_{X \sim \pi}[\ell(x_0, X) \leq \eta]$. Intuitively, $\kappa^+_{\pi,\ell}(\eta)$ represents the success probability of an oblivious attack that always selects the most likely reconstruction under the prior π.

Note that ReRo differs fundamentally from previous performance metrics, as it measures a success probability rather than an advantage. In Chapter 5, we discuss the formal impact of this change. Moreover, we empirically study the shortcomings of ReRo and existing bounds, as a general performance metric.

Recent work [62] refined ReRo bound in Eq. 2.3 using f-DP [38], a characterization of DP that captures the exact statistical indistinguishability between neighbors through the functional f. Intuitively, f-DP characterizes privacy directly through the ROC curve (i.e., the plot of the true positive rate (TPR) against the false positive rate (FPR)) of a MIA that seeks to distinguish between two possible members x and x'. By explicitly linking privacy guarantees to attack performance, it provides a natural and operational interpretation of privacy parameters. For this reason, f-DP serves as a fundamental tool throughout this thesis that we present in detail in the following section.

2.3.1.1. f-DP and Gaussian DP

Recently, f-DP was proposed [38] as a characterization of DP that captures the exact statistical indistinguishability between neighbors through the functional f using the hypothesis testing interpretation of DP. Among its advantages, it provides a more interpretable approach of DP allowing us to characterize the advantage of an MIA. Formally, we consider an attacker trying to solve a hypothesis testing problem for two neighboring databases $D = D_- \cup \{x\}$ and $D' = D_- \cup \{x'\}$ as

$$\begin{cases} H_0 \colon \text{The input database is } D, \\ H_1 \colon \text{The input database is } D'. \end{cases}$$

Specifically, given an output θ, an attacker will use a rejection rule ϕ to decide whether D or D' was the initial database. The difficulty in distinguishing between the two hypotheses is then described by the optimal trade-off between the *type I error* (i.e., rejecting H_0 when it is true, 1−TPR) and the *type II error* (i.e., failing to reject H_0 when it is false, FPR). If P and Q are the distribution functions of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ respectively, then the type I and type II errors are defined respectively as $\alpha_\phi \coloneqq \mathbb{E}_P[\phi]$ and $\beta_\phi \coloneqq 1 - \mathbb{E}_Q[\phi]$, given a rejection rule $0 \leq \phi \leq 1$. This motivates the definition of trade-off function [38].

Definition 2.19 (Trade-off function [38])**.** Let P and Q be two probability distributions on the same measurable space. A *trade-off function* is defined as $T(P, Q) \colon [0, 1] \to [0, 1]$ such that

$$T(P, Q)(\alpha) = \inf_\phi \{\beta_\phi \mid \alpha_\phi \leq \alpha\},$$

where the infimum is taken over all (measurable) rejection rules ϕ.

Note that a function $f \colon [0, 1] \to [0, 1]$ is a trade-off function if and only if it is continuous, convex, and non-increasing such that $f(x) \leq 1 - x$ [38].

A trade-off function $T(P, Q)(\alpha)$ represents the minimum achievable type II error β for a given level of type I error α. Note that the minimum β_ϕ can be achieved by the likelihood-ratio test, since it is the test with the highest *power* (i.e., lowest type II error for a prespecified type I error α) according to the Neyman–Pearson lemma [63]. The larger the trade-off function, the harder it is to distinguish between the two hypotheses. This idea of "hard to distinguish" leads us to the definition of f-DP:

Definition 2.20 (f-DP [38])**.** A mechanism $\mathcal{M}$ with domain $\mathcal{X}^n$ is said to be *f-differentially private* if, for all $D, D' \in \mathcal{X}^n$ such that $d_H(D, D') = 1$,

$$T(\mathcal{M}(D), \mathcal{M}(D'))(\alpha) \geq f(\alpha)$$

for all $\alpha \in [0, 1]$.

First, note that $T(\mathcal{M}(D), \mathcal{M}(D'))$ is the trade-off function of the distribution of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ (by abuse of notation). Specifically, consider the attack A as a test of H_0: the

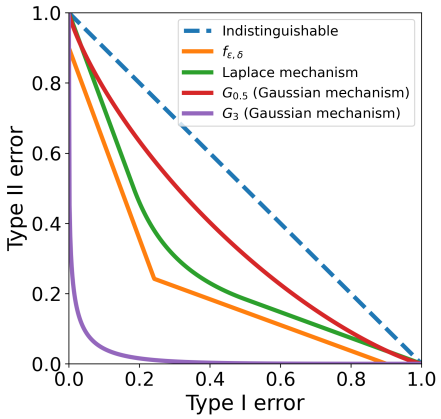


Figure 2.1.: Comparison of different DP mechanism trade-off functions for $\varepsilon = 1$, $\delta = 0.1$ and Gaussian mechanism with $\mu = 1/\sigma$.

input is D_0 vs. H_1: the input is D_1, applied to the output of $\mathcal{M}$. Then $\Pr(A(\mathcal{M}(D_0)) = 1)$ is the significance level and $\Pr(A(\mathcal{M}(D_1)) = 1)$ is the power of the test. Under this interpretation, for a given significance level, f bounds the maximum achievable power, which leads to the following equivalent definition:

Definition 2.21 ([23])**.** Let $f\colon [0,1] \to [0,1]$ be a continuous, convex, non-increasing function such that $f(x) \leq 1 - x$. A mechanism $\mathcal{M}$ satisfies f-DP if for all $D_0, D_1 \in \mathcal{X}^n$ such that $d_H(D_0, D_1) \leq 1$ and all post-processing algorithms $A\colon \mathrm{Range}(\mathcal{M}) \to \mathcal{D}(\{0,1\})$,

$$\Pr(A(\mathcal{M}(D_0)) = 1) \leq 1 - f\big(\Pr(A(\mathcal{M}(D_1)) = 1)\big).$$

Intuitively, f-DP characterizes the advantage of a hypothesis-testing-based MIA. Moreover, it generalizes DP since every mechanism $\mathcal{M}$ is (ε, δ)-DP if and only if $\mathcal{M}$ satisfies $f_{\varepsilon,\delta}$-DP, with

$$f(\alpha) = \max\{0, 1 - \delta - \mathrm{e}^{\varepsilon}\alpha, \mathrm{e}^{-\varepsilon}(1 - \delta - \alpha)\}. \tag{2.4}$$

We visualize the f-DP trade-off curves of different DP mechanisms in Figure 2.1. The closer a curve lies to the coordinate axes (such as G_3 in Figure 2.1), the weaker the privacy guarantee, as the attacker can simultaneously achieve small type I and type II errors. In other words, the test can reliably distinguish between H_0 and H_1. Conversely, the closer the curve is to the dashed diagonal line, the stronger the privacy guarantee. In this regime, any attempt to control the type I error necessarily results in a large type II error, and vice versa. In the extreme case, the optimal attack degenerates into a trivial strategy—for instance, always predicting membership—thereby offering no meaningful discriminative power beyond random guessing.

A particularly relevant case, is considering the trade-off function between two normal distributions with different means, the resulting notion is known as *Gaussian DP* (μ-GDP). GDP establishes that distinguishing between $\mathcal{M}(D)$ and $\mathcal{M}(D')$ is at least as hard as distinguishing between the normal distributions $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$, formally:

Definition 2.22 (Gaussian DP [38])**.** Given $\mu \geq 0$, a mechanism $\mathcal{M}$ with domain $\mathcal{X}^n$ is said to be μ-*Gaussian differentially private* (GDP) if, for all $D, D' \in \mathcal{X}^n$, such that $d_H(D, D') = 1$,

$$T(\mathcal{M}(D), \mathcal{M}(D'))(\alpha) \geq T(\mathcal{N}(0,1), \mathcal{N}(\mu,1))(\alpha)$$

for all $\alpha \in [0,1]$. We denote $G_\mu \coloneqq T(\mathcal{N}(0,1), \mathcal{N}(\mu,1))$.

By the Neyman–Pearson lemma, we can explicitly express G_μ as

$$G_\mu(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \mu),$$

for all $\alpha \in [0,1]$, where Φ is the distribution function of a standard normal distribution $\mathcal{N}(0,1)$. Note that this trade-off function decreases with respect to μ, i.e., $G_\mu \leq G_{\mu'}$ if $\mu \geq \mu'$ as we can see in Figure 2.1.

GDP satisfies a group privacy property that establishes that privacy degrades linearly with respect to the number of changes between the two databases [38]. Moreover, the Gaussian $\mathcal{M}_\sigma$ mechanism (see Definition 2.11) satisfies μ-GDP for $\mu = \Delta f/\sigma$ [64].

Remark 2.23 (f-DP composition [65])**.** Beyond advancing DP interpretability in terms of attacks, the f-DP framework satisfies improved composability properties: Specifically, the T-adaptive sequential composition of an f-DP mechanism satisfies $f^{\otimes T}$-DP, where $f \otimes f$ denotes the *tensor product* of two trade-off functions, $f = T(P, Q)$ and $g = T(P', Q')$, defined as

$$f \otimes g \coloneqq T(P \times P', Q \times Q').$$

The well-definition and the properties of the tensor product are proven in [38]. In our proofs, we will use that $\otimes$ is associative and commutative, and verifies $g \otimes f \geq g' \otimes f$ for all trade-off functions f and $g \geq g'$.

For instance, if a mechanism is μ-GDP, then its T-fold composition is $(\mu\sqrt{T})$-GDP [38]. We further analyze the composition behavior of GDP in Chapter 4.

Moreover, f-DP formulation facilitates the computation of quantities such as the total variation distance:

Definition 2.24. A mechanism $\mathcal{M}$ has total variation at most $\mathrm{TV}(\mathcal{M})$ if, for all neighboring datasets D_0, D_1,

$$\sup_{S \subseteq \Theta} |\Pr(\mathcal{M}(D_0) \in S) - \Pr(\mathcal{M}(D_1) \in S)| \leq \mathrm{TV}(\mathcal{M}).$$

For any $\mathcal{M}$ satisfying (ε, δ)-DP, its TV is bounded [66] as

$$\mathrm{TV}(\mathcal{M}) \leq \max_{\alpha \in [0,1]} (1 - f(\alpha) - \alpha) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}. \tag{2.5}$$

Moreover, if $\mathrm{TV}(\mathcal{M}_i) = \Delta$, then the T-adaptive sequential composition satisfies $\mathrm{TV}(\mathcal{M}) \leq 1 - (1-\Delta)^T$ [67]. When the mechanism characterization in terms of f-DP is known, this bound can be sharpened to $\max_\alpha(1 - f^{\otimes T}(\alpha) - \alpha)$ using Equation (2.5).

Hayes et al. [62] present the first ReRo bound for any f-DP mechanism:

$$\gamma \leq 1 - f(\kappa^{+}_{\pi,\ell}(\eta)). \tag{2.6}$$

which they showed empirically nearly tight for DP-SGD, the most known DP algorithm for private learning [49].

2.3.2. Correlation Assumption

As illustrated in previous section, most of the studies on attacks and risks on private data were developed under the assumption of statistical independence among data records. This is coherent with the initial assumptions and promises in original DP, which assumes that changing one person's record does not affect others and hence it suffices to consider the informed attacker and the database as an uniform random sample where each value is independent of the others. However, this assumption becomes quite unrealistic when we start to consider complex data structures where different types of dependencies are present, for instance in trajectory data as we further explain in Section 3.1, or social networks, where a change in one record can impact other records' information.

In fact, the limitations of DP in protecting datasets with dependencies among records have been highlighted in several formal works [29], [30], [37], [68]. More recently, these theoretical limitations were empirically confirmed by Humphries et al. [33]. In their study, they exploit correlations in the data to execute successful MIAs, achieving significantly higher empirical advantages than those guaranteed by DP for independent data (Eq. 2.2).

Intuitively, the DP neighboring-dataset definition fixes all records except one, denoted by D_{-}. This rigidity ignores the "propagation" of changes induced by correlations among records. When data are correlated, modifying a single individual can implicitly affect the distribution of the remaining records, so the impact of a change is no longer confined to one entry. For example, introducing an infected individual into a population does not only modify that person's record but also induces changes in others—previously healthy individuals may become infected due to the propagative nature of infectious diseases. To illustrate the effect of correlations more concretely, we consider the following toy example.

Example 2.25. Consider that Alice and Bob live in the same house. The sensitive secret is whether Alice is sick ($x_1 = 1$) or healthy ($x_1 = 0$).

The attacker observes a noisy count of sick family members, released via the Laplace mechanism for bounded DP:

$$\mathcal{M}(D) = \sum_{i \in \{1,2\}} x_i + Z, \text{ with } Z \sim \text{Lap}\left(\frac{1}{\varepsilon}\right).$$

Since changing one record value can affect the total sum a maximum of one, the sensitivity is $\Delta f = 1$. Hence, $\mathcal{M}$ is bounded ε-DP (or εd_H-private).

Moreover, denoting by θ the output of $\mathcal{M}$ and $s = x_1 + x_2$, we have:

$$\Pr_{\mathcal{M}}[\theta \geq t \mid s] = \begin{cases} 1 - \frac{1}{2}\mathrm{e}^{(t-s)\varepsilon} & \text{if } t \leq s, \\ \frac{1}{2}\mathrm{e}^{-(t-s)\varepsilon} & \text{if } t \geq s. \end{cases}$$

We denote by Π the dataset distribution and by π the marginal distribution of each records. We assume that a priori there is a uniform probability of being sick or healthy, i.e., $\pi[X_1 = 1] = \pi[X_1 = 0] = 0.5$. Since $\mathcal{M}$ is ε-DP, we can bound the normalized advantage of MIA, following [33]: under **independence assumption** (i.i.d data), the maximum advantage of an attacker that watches the output and tries to infer sick/healthy according to Equation (2.2) is:

$$\mathrm{Adv}^{Ind}_{MIA} = \mathrm{TPR} - \mathrm{FPR} \leq \frac{\mathrm{e}^{\varepsilon} - 1}{\mathrm{e}^{\varepsilon} + 1},$$

and this holds for *any* possible attacker (under independence).

However, the disease corresponds to a highly contagious virus. Therefore, living in the same house with an infected person increases the chances of infection to 0.9, i.e., $\Pr_{\Pi}[X_2 = 1 \mid X_1 = 1] = 0.9$ and symmetrically, $\Pr_{\Pi}[X_2 = 0 \mid X_1 = 0] = 0.9$. Now we consider the attacker that tries to infer Alice's heath status with the following strategy:

$$A(\theta, x_1) = \begin{cases} 1 & \text{if the observed output satisfies } \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

We compute the advantage for this attacker taking the correlation model into account:

$$\begin{aligned}
\mathrm{Adv}(A) =& \Pr_{\substack{\theta \sim \mathcal{M}(D) \\ D \sim \Pi}}[A(\theta) = 1 \mid X_1 = 1] - \Pr_{\substack{\theta \sim \mathcal{M}(D) \\ D \sim \Pi}}[A(\theta) = 1 \mid X_1 = 0] \\
=& 0.9 \Pr_{\theta \sim \mathcal{M}(1,1)}[A(\theta) = 1] + 0.1 \Pr_{\theta \sim \mathcal{M}(1,0)}[A(\theta) = 1] \\
& - 0.1 \Pr_{\theta \sim \mathcal{M}(0,1)}[A(\theta) = 1] - 0.9 \Pr_{\theta \sim \mathcal{M}(0,0)}[A(\theta) = 1] \\
=& 0.9\Big(\Pr_{\theta \sim \mathcal{M}(1,1)}[A(\theta) = 1] - \Pr_{\theta \sim \mathcal{M}(0,0)}[A(\theta) = 1]\Big),
\end{aligned}$$

where the last inequality holds since $\mathcal{M}(0,1)$ and $\mathcal{M}(1,0)$ are identically distributed since the output distribution only depends on the sum which in both cases is one. Following the attack definition (Eq. 2.7) we obtain:

$$\Pr[A(\theta) = 1 \mid x_1, x_2] = \Pr_{\mathrm{Lap}}[\theta \geq 1 \mid s] = \begin{cases} 1 - \frac{1}{2}\mathrm{e}^{(1-s)\varepsilon} & \text{if } 1 \leq s \\ \frac{1}{2}\mathrm{e}^{-(1-s)\varepsilon} & \text{if } 1 \geq s \end{cases}$$

Substituting in the advantage formula we have

$$\mathrm{Adv}(A) = 0.9\Big(1 - \frac{1}{2}\mathrm{e}^{(1-2)\varepsilon} - \frac{1}{2}\mathrm{e}^{-(1-0)\varepsilon}\Big) = 0.9(1 - \mathrm{e}^{-\varepsilon})$$

Now taking $\varepsilon = 1$,

$$\mathrm{Adv}^{Ind}_{MIA} \leq \frac{\mathrm{e}^{\varepsilon} - 1}{\mathrm{e}^{\varepsilon} + 1} \approx 0.46 < 0.57 \approx \mathrm{Adv}(A).$$

Therefore, in the correlated setting, the attacker exceeds the advantage bound established for the independent case. This demonstrates that correlations can strictly increase the

attacker's advantage, thereby violating existing DP protection bounds. Moreover, this violation is achieved by the so-called "weak" attacker [54], who, while targeting Alice's record, does not require access to the rest of the dataset and relies only on the overall data distribution. This result further highlights that the attack-resilience paradigm changes fundamentally once correlations are taken into account.

While there is ample evidence of the limitations of DP against correlation-based attacks, there is still a lack of precise interpretation or quantification of how correlations impact DP guarantees. Some recent extensions of DP aim to address this issue by explicitly accounting for correlations, such as *Bayesian DP* [37].

2.3.2.1. Bayesian Differential Privacy

Bayesian differential privacy (BDP) [37] extends the privacy guarantees of DP to settings with correlated data. Similar to DP, it assumes that the adversary is uncertain between two possible values of a target record, x_i and x_i'. However, instead of considering only an adversary who knows the entire remaining dataset, BDP accounts for all possible adversaries with arbitrary background knowledge.

To achieve this, BDP eliminates the traditional notion of neighboring databases. Formally, an adversary is denoted by (K, i), where the target is the record at position i and the adversary already knows the values of the subvector $\mathbf{x}_K$ in the database. For each such adversary, the *Bayesian leakage* is defined as follows:

Definition 2.26 (Adversary-specific BDPL [37])**.** Given $\mathcal{M} : \mathcal{X}^n \to \mathcal{D}(\Theta)$ a randomized mechanism, $\mathbf{X}$ the input random vector following the distribution Π, the targeted record index $i \in [n]$, and the known record indices $K \subseteq [n] \backslash \{i\}$, the *adversary-specific Bayesian differential privacy leakage* is[3]

$$\mathrm{BDPL}_{(K,i)} = \sup_{x_i, x_i', \mathbf{x}_K, S} \ln \frac{\Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i]}{\Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i']},$$

where the supremum is taken over all the possible target values $x_i, x_i' \in \mathcal{X}$, all the possible known vector values $\mathbf{x}_K \in \mathcal{X}^K$ and all the measurable sets $S \subseteq \Theta$.

When computing the adversary-specific BDPL, the correlation between the unknown and known records modifies the final leakage since given the unknown remaining indices U, we have

$$\Pr[Y \in S \mid \mathbf{x}_K, x_i] = \sum_{\mathbf{x}_U \in \mathcal{X}^u} \Pr[Y \in S \mid \mathbf{x}_K, x_i, \mathbf{x}_U] \Pr[\mathbf{x}_U \mid \mathbf{x}_K, x_i],$$

where $u = |U| = n - k - 1$. Note that the sum must be substituted by an integral in the continuous case.

While the adversary-specific BDPL only accounts for a particular case, we aim to protect against any possible adversary. Therefore, to compute the worst-case leakage we take the supremum:

[3] If both the numerator and denominator are zero, we conventionally set BDPL = 0.

Definition 2.27 (Bayesian DP [37])**.** A mechanism $\mathcal{M}$ satisfies ε-*Bayesian differentially private* if

$$\mathrm{BDPL}(\mathcal{M}) = \sup_{K,i} \mathrm{BDPL}_{(K,i)}(\mathcal{M}) \leq \varepsilon,$$

where the supremum is taken over all the possible set of indices $i \in [n]$ and $K \subseteq [n] \setminus \{i\}$. $\mathrm{BDPL}(\mathcal{M})$ is called *Bayesian differential privacy leakage*.

The BDPL has a similar role to the privacy leakage ε in DP: It measures the extent of a possible private information inference by comparing the difference in the output probabilities of mechanism $\mathcal{M}$. A lower BDPL corresponds to higher privacy because any adversary will be less likely to differentiate between any two target values $x_i, x_i' \in \mathcal{X}$. Particularly, ε-BDP implies ε-DP, hence it is stronger, and if X_i, X_j are mutually independent for all $i \neq j \in [n]$ then ε-DP and ε-BDP are equivalent [37].

While DP assumes the adversary knows all records except the target, BDP considers arbitrary priors, including those where unknown records are correlated. It ensures bounded changes in output distributions even when the target record is part of a correlated subset. When data is independent, BDP and DP coincide. Under correlation, however, BDP quantifies worst-case leakage by integrating the mechanism's output with the data distribution via Bayes' rule, capturing adversarial advantages that DP overlooks. Hence, BDP mitigates correlation-driven reconstruction attacks that breach DP's guarantees as empirically shown in [69].

Unfortunately, its practical applicability remains uncertain. The few mechanisms proposals are limited to specific correlation models or exhibit huge utility loss [37], [69]. Given the scarcity of mechanisms and their applicability restrictions, it remains unclear whether correlation-aware extensions of DP can serve as a usable privacy notion, motivating our further study in Chapter 6.

2.4. Measure Theory Results

In this section we present the disintegration theorem, a fundamental result in measure theory that plays a key role in the results established in this thesis.

In continuous probability spaces, events of the form $X = x$ have probability zero, so conditional probabilities defined via ratios are not well-defined. The disintegration theorem provides a rigorous substitute: any joint probability measure P on $\mathcal{X} \times \mathcal{Y}$ can be decomposed as

$$P(\mathrm{d}y\,\mathrm{d}x) = P_x(\mathrm{d}y)\, P_X(\mathrm{d}x),$$

where P_X is the marginal of X and P_x is a probability measure on $\mathcal{Y}$ representing the conditional law of Y given $X = x$. This decomposition allows conditional distributions to be defined pointwise (almost everywhere), despite conditioning on null events.

This intuition extends from Cartesian products to general measurable maps $a\colon \mathcal{X} \to \mathcal{Z}$, where disintegration allows one to define conditional measures P_z supported on the fibers $a^{-1}(z)$, providing a rigorous notion of conditioning on $a(x) = z$ even when $P(a^{-1}(z)) = 0$.

Theorem 2.28 (Disintegration Theorem)**.** *Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ be standard Borel spaces and let $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P)$ be a probability space. Let $a\colon \mathcal{X} \to \mathcal{Z}$ be measurable and denote $\nu = P \circ a^{-1}$ the pushforward of P through a. Then there exists a family of probability measures $\{P_z\}_{z\in\mathcal{Z}}$ on $\mathcal{X}$, uniquely determined for ν-almost every z, such that:*

1. *For ν-a.e. $z \in \mathcal{Z}$, P_z is supported on the fiber $a^{-1}(z)$, i.e.,*

$$P_z\left(\mathcal{X} \setminus a^{-1}(z)\right) = 0.$$

2. *For every measurable set $B \subseteq \mathcal{X}$,*

$$P(B) = \int_{\mathcal{Z}} P_z(B)\,\mathrm{d}\nu(z).$$

3. *For every integrable function $f \in L^1(\mathcal{X}, P)$,*

$$\int_{\mathcal{X}} f(x)\,\mathrm{d}P(x) = \int_{\mathcal{Z}} \left(\int_{\mathcal{X}} f(x)\,\mathrm{d}P_z(x) \right) \mathrm{d}\nu(z).$$

The map $z \mapsto P_z(B)$ is measurable for each measurable B, so $\{P_z\}$ is a regular conditional probability and is ν-a.e. unique.

Note that when $P = P_X \otimes P_Y$ is a product measure on $\mathcal{X} \times \mathcal{Y}$ and a is the projection onto $\mathcal{X}$, the conditional measures P_x may be taken equal to P_Y for P_X-almost every x, and the above reduces to the classical Fubini–Tonelli decomposition.

Remark 2.29. The disintegration theorem applies straightforwardly on discrete spaces. Let $\mathcal{X}$ and $\mathcal{Z}$ be finite (or countable) discrete sets and let $a\colon \mathcal{X} \to \mathcal{Z}$ be any measurable function. Let P be a probability measure on $\mathcal{X}$ with mass function π, so for any $B \subseteq \mathcal{X}$

$$P(B) = \sum_{x\in B} \pi(x).$$

Define the pushforward ν on $\mathcal{Z}$ by

$$\nu(z) = P(a^{-1}(z)) = \sum_{x:a(x)=z} \pi(x).$$

For z with $\nu(z) > 0$ define P_z on $\mathcal{X}$ by

$$P_z(B) = \frac{1}{\nu(z)} \sum_{x\in B} \mathbf{1}_{\{a(x)=z\}} \pi(x).$$

Then P_z is supported on the fiber $\mathcal{X}_z := a^{-1}(z)$ and, for any $B \subseteq \mathcal{X}$,

$$\int_{\mathcal{Z}} P_z(B)\,\mathrm{d}\nu(z) = \sum_{z\in\mathcal{Z}} P_z(B)\,\nu(z) = \sum_{z\in\mathcal{Z}} \sum_{x\in B} \mathbf{1}_{\{a(x)=z\}} \pi(x) = \sum_{x\in B} \pi(x) = P(B).$$

Moreover, for any nonnegative (or integrable) function $f\colon \mathcal{X} \to [0, \infty]$,

$$\begin{aligned}\sum_{x\in\mathcal{X}} f(x)\pi(x) &= \sum_{z\in\mathcal{Z}} \left(\int_{a^{-1}(z)} f(x)\ \mathrm{d}P_z(x) \right) \nu(z) \\ &= \sum_{z\in\mathcal{Z}} \left(\sum_{x\colon\, a(x)=z} f(x) \frac{\pi(x)}{\nu(z)} \nu(z) \right) \\ &= \sum_{z\in\mathcal{Z}} \sum_{x\colon\, a(x)=z} f(x)\pi(x)\end{aligned}$$

Theorem 2.28 is crucial to understand and prove novel properties on attack resilience of differential privacy that we present in Chapter 5.

3. Differential Privacy Challenges in Complex Data: A Trajectory Data Study

This chapter is based on the contributions:

- **Patricia Guerra-Balboa***, Àlex Miranda-Pascual*, Javier Parra-Arnau, Jordi Forne and Thorsten Strufe. Anonymizing trajectory data: limitations and opportunities. In: the AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI), 2023.
- Àlex Miranda-Pascual*, **Patricia Guerra-Balboa***, Javier Parra-Arnau, Jordi Forné, and Thorsten Strufe. "SoK: Differentially Private Publication of Trajectory Data". In: Proceedings on Privacy Enhancing Technologies (PoPETS), 2023, DOI: 10.56553/popets-2023-0065.
- Àlex Miranda-Pascual*, **Patricia Guerra-Balboa***, Javier Parra-Arnau, Jordi Forné, and Thorsten Strufe. "An overview of proposals towards the privacy-preserving publication of trajectory data". In: International Journal of Information Security (IJIS), 2024, DOI: 10.1007/s10207-024-00894-0.

The goal of this chapter is to survey and analyze the challenges and limitations that constrain the applicability of DP to complex data, understood as datasets with a large number of records, many variables, structures that are not readily representable in a standard tabular form, and intricate patterns and dependencies that require sophisticated models and methods of analysis [25]. Rather than attempting an exhaustive survey of DP mechanisms for all forms of complex data—which is inherently infeasible—we focus on trajectory data as a representative and challenging example. Trajectory analysis holds considerable promises, with applications ranging from improved traffic management and routing recommendations to infrastructure planning [2]. At the same time, learning users' paths is extremely privacy-invasive [70], [71], which creates a pressing need to protect trajectories in a way that preserves global properties that are useful for analysis while ensuring that specific, individual-level information remains inaccessible. Trajectories are particularly difficult to protect [72], as they are sequential, highly dimensional, strongly correlated, constrained by geophysical structure, and easily mapped to semantic points of interest, which exacerbates the risk of re-identification and sensitive inference [73], [74].

To this end, we survey existing attacks and threats against trajectory data, provide an in-depth analysis of DP adaptations and proposed granularity notions for trajectories, examine and elaborate on the state of the art in privacy-enhancing mechanisms and their

shortcomings, and expose the main limitations of current DP notions in the context of trajectory data. Particularly, we discuss and mathematically prove impossibility results that reveal which algorithms are not formally DP, thereby uncovering flaws in existing approaches. Summarizing, this chapter aims to answer the following research questions:

- How do known attacks manifest and impact real-world scenarios, and to what extent does the DP literature account for them, particularly in the presence of correlations?
- How has DP been adapted to trajectory data, and what are the implications of different granularity notions?
- What is the current landscape of DP masking mechanisms for trajectory data, and which are the actual formal guarantees that these approaches provide?
- What are the fundamental conceptual and technical limitations of DP adaptation in complex data structures such as trajectories?

We addressed each of these questions in Sections 3.2 to 3.5, respectively.

Related work. We succinctly describe the main differences between our work and prior surveys in the field. Primault et al. [75] provide a deep analysis of location-privacy protection mechanisms, including a division of the protection mechanisms into online and offline methods. However, the authors do not cover trajectory privacy extensively since their main focus is on the more general field of location privacy. Note that trajectory data is inherently more complex than simple location data: trajectories are not only comprised of visited locations but also include correlations and connections between them. In consequence, attacks, privacy-protection mechanisms, and limitations are notably different, even though these data types share a close relationship. Fiore et al. [76] offer a thorough overview and classification of attacks on trajectory databases, however they classify them from the syntactic point of view, focusing on threats that are more relevant for k-anonymity based notion than for DP, hence missing the coverage of data reconstruction or correlation-based attacks. Moreover, since they cover a broader spectrum of privacy notions their discussion of DP masking mechanisms is quite limited and in particular overlooks major limitation that we discuss in this work. Jin et al. [53] conduct a survey with an analysis and empirical evaluation of trajectory-privacy models to quantify their privacy and utility, but do not consider DP mechanisms in depth.

Our work entirely focuses on DP mechanisms for private database publication, which the aforementioned surveys do not fully explore. Other works focus on orthogonal topics, such as trajectory anonymization under syntactic notions [77] and location privacy (not comprising trajectories) [78].

Importantly, our survey of trajectory masking mechanisms corresponds to our published work [32], which was published in 2023 and therefore covers the literature up to that time. The impact of this work within the research community is evident in subsequent surveys that extend and build upon our results covering more recent literature, such as [79], [80]. In particular, Buchholz et al. [79] explicitly cites our work and leverages our

impossibility results to analyze newly proposed state-of-the-art solutions (after 2023), uncovering additional DP flaws.

3.1. Trajectories Data Structure

Trajectories correspond to a path or trace generated or drawn by a *moving object*, usually referred to as an *individual* or *user* (we will refer as such independently of what they are, e.g., a person walking, or a car carrying various people). Hence, they describe a particular case of *time series* [81].

Different types of trajectories exist. *Raw trajectories* consist of an ordered sequence of spatio-temporal points $T = \langle p_1, \ldots, p_m \rangle$, also written as $T = p_1 \to \cdots \to p_m$, where $|T| \coloneqq m$ denotes the *length* of T and $p_i = (x_i, y_i, t_i)$ corresponds to the location (x_i, y_i) at timestamp t_i. Trajectories respect the temporal order (i.e., t_{i+1} must happen strictly after t_i), which ensures there are no movements back in time, and no one is in two different locations at once. The term *subtrajectory* usually refers to a subset of a trajectory, including those formed by not necessarily consecutive locations, while *n-grams* (also called *subsequences*) are subtrajectories formed by n consecutive spatio-temporal points. The *prefixes* of a trajectory $T = \langle p_1, \ldots, p_m \rangle$ are the n-grams ($n \leq m$) starting at p_1, i.e., $\langle p_1, \ldots, p_n \rangle$.

Semantic trajectories are alternative representations where every spatio-temporal point contains additional *semantic meaning*, such as a name and description (e.g., "coffee shop" or "work"), possibly augmented with additional information such as the number of visitors or opening hours. In semantic trajectories, locations are called *point of interest* (POI). More complex trajectories, called *multiple aspect trajectories* [82], additionally consider any possible type of recordable information, like weather variations, transportation mode, or the current heart rate or emotions of individuals. Simplified trajectories have been suggested, such as $T = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$, where time is omitted and only the order of locations is retained [83], [84], [85], [86]. We will refer to the *spatial* and *temporal* aspects as *dimensions* of a trajectory, which are both commonly represented as numerical data. Semantic locations additionally have a *categorical dimension*.

Trajectory databases consist of one or multiple trajectories from individuals, usually over a shared region. We can represent them as collections of sequences, where each sequence contains the data of a single individual $D = \{T_i : \langle p_1^{(i)}, p_2^{(i)}, \cdots, p_{m_i}^{(i)} \rangle\}_{i \in [n]}$ where T_i denotes a trajectory belonging to user i. The length of each trajectory is denoted here by m_i and depends on each user. In some contexts, the same user can contribute multiple trajectories to the database. In this latter case, i is just a label of the trajectory and does not necessarily relate to a user.

Differences in structure between such databases exist. Some consist only of trajectories of equal length, and others assume that trajectories are *periodically recorded* (i.e., every trajectory has a spatio-temporal point for every time interval) [87], [88]. Further types include those with irregular recordings, with spatio-temporal points only included when the user is at a relevant location [89].

A particular scenario in trajectory publishing is the *data-stream scenario*, where a flow of information is received and published periodically. Therefore, a *streaming database* can be viewed as a sequence $D = \{S_1, \ldots, S_t, \ldots\}$, where each *update* S_i represents the information corresponding to time i:

$$D = \begin{cases} & S_1 & S_2 & \cdots \\ T_1: & p_1^{(1)} & p_2^{(1)} & \cdots \\ T_2: & p_1^{(2)} & p_2^{(2)} & \cdots \\ \vdots & \vdots & \vdots & \\ T_r: & p_1^{(r)} & p_2^{(r)} & \cdots \end{cases}.$$

The database at time t is denoted $D_t = \{S_1, \ldots, S_t\}$ and called a *stream prefix*. Note that since some databases consist of non-periodically recorded trajectories, "gaps" in this representation are possible, as shown in Figure 3.2. Hence, T_i may not have a location for time t, and remain empty in row i of S_t.

The structure of trajectory data and databases makes its protection exceptionally difficult. Long trajectories cause problems due to the *curse of dimensionality* [13], [90], and the sparseness and uniqueness of trajectories can aid in re-identification [72]. Another risk factor is the semantic meaning of points since this information can be enough to expose individuals.

A notorious statistical property of trajectory databases is the presence of correlation. Two conceptually different correlations are present in trajectory data:

Correlations between trajectories refers to the case when multiple users' records are correlated. In families' trajectories, for instance, we are bound to observe high correlations between their corresponding records as they engage in shared activities. Furthermore, an extreme case is regular repetitions of trajectories contributed by the same individual.

Correlations between attributes refers to the correlation in the data a single user contributes to the database. In the case of trajectories, it refers to the correlations within the spatio-temporal and semantic dimensions. A high-correlation level exists between close timestamps due to the laws of physics, route distribution, or social patterns. It is also termed *autocorrelation* for time series data.

We present in Section 2.3.2 the implications of correlation in privacy.

3.2. Attacks Against Trajectory Data

In this section, we review existing attacks on trajectory data that have been successfully implemented in practice. We highlight how specific properties of human mobility traces, especially their uniqueness, increase the risk of inference attacks.

In particular, we adopt the threat model described in Section 2.3. We assume an adversary who gains access to published trajectories, possesses potential auxiliary knowledge about users, and aims to extract sensitive information about individuals, such as reconstructing their paths or inferring indirect attributes, regardless of whether the users are explicitly identified.

Accordingly, we exclude other forms of inference, such as *sensitive location disclosure*. Although these attacks pose significant risks, they do not directly involve the leakage of user-specific private attributes, but rather the exposure of locations. Representative examples include the identification of previously undisclosed Israeli and U.S. military bases following the public release of soldiers' running trajectories through the Strava platform [91], [92].

Finally, even when DP mechanisms are correctly implemented, adversaries may exploit side channels—such as execution time or memory usage—to extract information about the underlying dataset [93]. While such attacks threaten the practical robustness of DP systems, they stem from implementation-dependent effects and hardware- or system-level leakage. Therefore, they fall outside the scope of this work.

We illustrate the tangible risks associated with a lack of privacy protection in trajectory data in the following examples. The New York City taxi dataset, which included around 173 million taxi trips and the corresponding tips [94], was published in 2013. Since then, plenty of attacks on this data, using *auxiliary knowledge* (see Section 2.3), quickly appeared: Tockar [94], [95] used paparazzi photos to link celebrities' identities to the corresponding trip in the data discovering where they went, which establishments they visited, and how much they tipped. Deneau [28] figured out that one could link stops with daily praying time to identify Muslim cab drivers. These examples are excellent representatives of important privacy risk associate with pseudo-anonymized traces (where only name and direct identifiers are removed).

To show the privacy risks in human traces, we expose the possible attacks and threats of the literature. We also provide examples, some of which have previously been extensively surveyed [53], [76].

Membership attacks aim to discover whether or not a specific individual is present in the database, regardless of whether their records can be directly identified (see Section 2.3). Learning merely the presence or absence of an individual in a trajectory database can be a direct privacy threat (e.g., consider a database of trajectories with traffic violations). Well-known examples include successful MIAs on trajectory data—such as those exploiting aggregate location statistics [96]—even when reducing the attacker's auxiliary knowledge to realistic assumptions, like zero prior traces [97].

In **attribute attacks** [52], an adversary learns additional information about the target without necessarily identifying their exact record in the database (see Section 2.3). In trajectory data, this includes the whereabouts and temporal information (e.g., when no one is at home). The disclosure of a user's *spatial* and *temporal information* [73] is inherently sensitive. Moreover, it can enable the indirect disclosure of additional attributes, as mobility patterns are often linked to semantic information and personal characteristics. For instance, presence at a hospital for extended amounts of time allows adversaries to infer a user's health status; while being at a place and time where a specific protest is happening may leak information about a user's political opinions. In the example of Muslim taxi drivers mentioned above, the attacker inferred an attribute: the victims' religion, indirectly from the stopping times.

Sui et al. [98] observe that 40% of records that cannot be immediately identified—and thus appear anonymous—directly disclose the shared attribute.

Users' most sensitive locations are another attribute that can be exposed, for example, point-clustering algorithms that can deterministically find them already exist [99]. Gambs et al. [100] demonstrate how this violates the privacy of sensitive attributes.

Group linkage attacks [53] discover connections between individuals. Relationships are particular attribute cases, and both social links and kinship can be inferred from correlated movement [71]. Their disclosure may entail different threats. Predisposition to hereditary diseases, communication between dissidents, homophily in friendships sharing religious and political views, or homosexual partnerships in certain jurisdictions are just a few prominent examples.

Due to the time-series structure of trajectories, predicting future attributes—such as a user's next locations (**prediction attacks**)–poses a significant threat. Attackers can thus discover destinations, potentially even before users arrive, or infer whether users are at home to plan attacks like robberies. For instance, Song et al. [70] demonstrate successful movement pattern predictions [100] with up to 93% accuracy in predicting mobility behavior.

In trajectory data, many **data reconstruction attacks** (see Section 2.3) were proposed trying to rebuilding trajectories in the database. For example, Buchholz et al. [74] introduce a reconstruction algorithm that can construct trajectories closer to the original data than the perturbed one. Similarly, *filtering attacks* [31] also aim at reducing noise added. On the other hand, Xu et al. [101] develop an iterative attack that can exploit the uniqueness and regularity of human mobility to step-by-step recover individual's trajectories from mobility data without using any auxiliary knowledge.

Finally we find **re-identification attacks** that attempt to infer the record index of a target in a public data set. In the case of trajectory data this translates to infer the exact individual trajectory in the dataset. Re-identification is often not considered a threat itself but an intermediate strategy to perform DRAs and AIAs: If you link a user to a trajectory, in particular you exactly reconstruct the trace and attributes of it, such as most visited locations or speed. These attacks usually utilize auxiliary information, i.e., information exposed through other means and thus available to the adversary. In particular, *personal context linking attacks* [73], [102] use known information about a victim (e.g., they have been to a coffee shop) to discover their trajectory in the database. Some record linkage attacks aim to discover uniquely identifiable traits to reconstruct the victim's path. In the case of trajectories, little information suffices to do so. For example, in the empirical study by de Montjoye et al. [11], the authors demonstrate that knowledge of only four spatio-temporal points is sufficient to uniquely identify 95% of individual trajectories within a dataset of 1.1 million users. In other words, with minimal auxiliary information, an adversary can narrow the set of candidate trajectories corresponding to a target individual down to a single record. Furthermore, if using highly accurate GPS data, two points are empirically sufficient to uniquely identify all individuals in the database [103], hence two points of auxiliary knowledge enable the recovery of their whole path.

Attack models can be designed to use location probability distributions, mobility preferences and patterns, exposed locations, and physical encounters in order to detect the unique traits more successfully [76]. Along this line, De Mulder et al. [12] show that human movement is characterized by strong regularities and can link 80% of users in real databases. Freudiger et al. [104] exploit the uniqueness of home and work locations to design an attack model that identifies trajectories of real databases. Rossi et al. [103] show that one can uniquely identify up to 95% of users when using movement data such as traveled distance, speed, and direction. In addition, location traces can reveal speed and acceleration patterns that can identify the type of vehicle that generated the trajectory (car, truck, or motorcycle) [105]; and knowing the physical dimension of the vehicle can also help identify trajectories in vehicular data [106]. A comprehensive list of similar attacks is presented in Fiore et al.'s survey [76].

Correlation impact in attacks against trajectory data. As previously explained in Section 2.3, DP faces strong limitations when protecting against correlation-based attacks that exploit the inherent dependencies in the data. In trajectory data this is of major relevance since, as we saw in Section 3.1, not only dependencies among different trajectories exists but also inside of the same trajectory we find spatio-temporal correlations that determine the next movements.

Most of previous attacks would not be possible without exploiting such correlations, such as the prediction attacks [70] or reconstruction attacks [36], [74], [101]. Particularly, in [74] they show a successful reconstruction of DP anonymized trajectories across different methods hence breaking the post-processing property of DP due to the dependencies in the data. Moreover, DP mechanisms typically rely on adding noise (see Section 2.1), which increases the uncertainty about the underlying record when observing the noisy output. However, when this noise is added independently at each location, without considering the autocorrelation, applying time-series filters, such as the Kalman or Wiener filters, effectively removes the noise added by sanitation mechanisms, as shown by Wang et al. [31].

From a formal point of view, Cao et al. [36] demonstrate how this problem specially affects protection under event-level privacy—which aims to protect the existence of each spatio-temporal point in the database (see Section 3.3 for technical details). Particularly, they show the limitation of such approaches due to the spatio-temporal autocorrelation between nearby spatial points. As we see in Figure 3.1, each spatio-temporal point affects other nearby points, simply due to the laws of physics and external limitations, such as road networks. However, if the attacker uses autocorrelation knowledge, then the difference between the output distributions of Definition 2.1, conditioned to whether the target spatio-temporal point is in the database or not, will not be bounded by ε anymore. Intuitively, the change to a single location is bounded by ε. However, in Figure 3.1, modifying only the green location is impossible without altering the entire green path (which contains m locations) to maintain a feasible trajectory. By applying Proposition 2.3, this change is therefore no longer bounded by ε, but by $m\varepsilon$, leading to increased privacy leakage. Consequently, an attacker may be able to infer whether the point was originally in the database simply by examining the output.

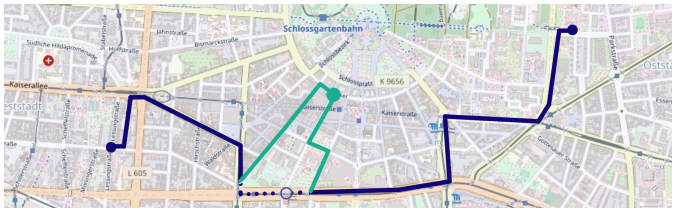


Figure 3.1.: The green location is naturally no sensible alternative for the original blue point. Jumping from one location to another far away in seconds is not possible in real life, which is easily modeled with correlations. Changing that location also would imply changing the nearby points. Map screenshot from © OpenStreetMap contributors [107].

While correlations possess a threat to general DP systems, we see a huge impact, both theoretically and practically, when applied to human traces.

Conclusions. DP has been primarily analyzed in the context of MIAs (see Section 2.3). However, our survey of attacks on trajectory data reveals a much broader range of threats that are both more relevant in practice and potentially more damaging.

Furthermore, many attacks against trajectory data successfully undermine DP guarantees due to correlations in the data. Because most DP analyses assume independence, there is a substantial gap in the literature regarding the impact of data correlations and how to select privacy parameters to mitigate these effects. No general solution currently exists to address this challenge.

3.3. Granularity Notions in Trajectory Data

The original DP notion aims to protect the entire existence of an individual's records, thus assuming a one-to-one correspondence between record and individual, consistently with tabular datasets (original domain of DP definition). As discussed in Section 2.1, DP guarantees that, just observing the output, an adversary cannot reliably distinguish between two neighboring datasets: one in which the target record is included and one in which it is not—either because it has been removed (unbounded DP) or replaced with another record (bounded DP).

However, with the extension of DP to complex data structures such as graphs, time series, or images, the one-to-one correspondence between record and individual is not that clear anymore. Hence, various adaptations of the concept of *neighborhood* (i.e., what is considered a single entry in the database) have been suggested in the literature. We refer to the neighborhood definition as the *level of granularity* [16] of a DP notion.

In trajectory data, where each individual contributes multiple spatio-temporal points—potentially spanning different dimensions and carrying distinct semantics—the notion of granularity becomes particularly relevant. The neighborhood definition directly impacts the privacy guarantee offered. Therefore, we explore the most common granularity notions in the following paragraphs and provide a quick summary of these in Table 3.1.

Granularity	Difference between neighboring databases
User-level	A user's whole trajectories
Event-level	A spatio-temporal point visited by a user (an event)
w-event	A window of events over w consecutive timestamps
ℓ-trajectory	A sequence of ℓ consecutive events from a single user
Element-level	A user's set of points belonging to the same cluster

Table 3.1.: Granularity notions and their concept of neighborhood.

User-level privacy corresponds to the original idea of DP. We consider two databases D and D' to be *user-level neighboring* if they only differ in the information attributed to a single user. For instance, if each user contributes a single trajectory, then two databases D and D' are considered user-level neighbors if they differ in one user's entire trajectory, either by removing/adding their trajectory (*unbounded* DP) or by exchanging it with another user's trajectory (*bounded* DP). When a user contributes multiple trajectories to the database, the definition of neighboring datasets naturally extends to include all trajectories belonging to that user. Under this distinction, **trajectory-level** privacy protects an individual trajectory, while user-level privacy provides protection for the complete collection of trajectories associated with a given user.

Event-level privacy adapts DP to streaming settings, where data are continuously generated and processed [16]. In trajectory data, where individuals produce sequences of spatio-temporal points, applications like real-time traffic monitoring or traffic jam prediction show how such data naturally arrive as streams, making streaming-based privacy solutions essential.

When applied to *data streams*, DP rise to the *continual observation* setting [16]. In this setting, user data arrives continuously, and systems are required to provide low-latency responses. Consequently, each event must be protected and transmitted as it occurs, with no opportunity to consider future events nor to change previously sent outputs—which makes the design of privacy-preserving protocols particularly challenging.

Event-level granularity was presented as a solution in streaming applications. This notion aims to protect a spatio-temporal point visited by a user (an event), i.e., to hide the presence or absence of a single event from a sequence of observations contributed by an individual. Formally,

Definition 3.1 (Event-neighborhood)**.** For trajectory data, two streaming databases D and D' are *event-neighboring* if we obtain one from the other by changing a single spatio-temporal point.

We illustrate event-neighboring time series with the following example:

$$D = \begin{cases} & S_1 & S_2 & \cdots & S_m \\ T_1: & p_1^{(1)} & p_2^{(1)} & \cdots & p_m^{(1)} \\ T_2: & p_1^{(2)} & p_2^{(2)} & \cdots & p_m^{(2)}, \\ & \vdots & \vdots & & \vdots \\ T_r: & p_1^{(r)} & p_2^{(r)} & \cdots & p_m^{(r)} \end{cases} \quad D' = \begin{cases} & S_1 & S_2 & \cdots & S_m \\ T_1: & p_1^{(1)} & p_2^{(1)} & \cdots & p_m^{(1)} \\ T_2: & p_1^{(2)} & \hat{p}_2^{(2)} & \cdots & p_m^{(2)}. \\ & \vdots & \vdots & & \vdots \\ T_r: & p_1^{(r)} & p_2^{(r)} & \cdots & p_m^{(r)} \end{cases}$$

In this context, event-level privacy aims to make it more difficult to determine whether a particular spatio-temporal point has been visited by a given user.

Under the assumption of independence between records (see Section 2.3.2), event-level DP ensures that each point in the database remains indistinguishable (up to ε) to an attacker. Because the neighborhood definition is restricted, the sensitivity of a query is never larger—and often smaller—than the user-level sensitivity. As a result, for the same ε, mechanisms such as the Laplace mechanism require less noise, yielding higher utility. Additionally, event-level DP naturally extends to infinite streaming scenarios and addresses the need to protect each event individually in continual observation settings.

This notion, however, comes with its own drawbacks. In particular, it introduces a risk of membership disclosure. If an attacker knows $r > 1$ points from a target's trajectory, the change of these r points is no longer bounded by ε, but by $r\varepsilon$ [16]. As a result, the attacker's ability to infer membership can increase significantly, leaving the individual vulnerable to MIAs. Since real-world trajectories often contain hundreds of spatio-temporal points per person, the risk can be substantial.

Additionally, event-level DP does not fully protect against attribute disclosure. If a location is visited multiple times by the same user, the attribute "the target visited this location" is less protected, because it is represented across several events rather than a single one. For example, if a user visits a hospital multiple times, the attribute "has been to the hospital" may still be inferred, despite event-level guarantees.

Finally, an inherent limitation of event-level DP is its vulnerability to correlation attacks (see Section 2.3.2). Event-level guarantees can improve utility in scenarios where only the exact data point is sensitive—for example, the combination of time and place ("the user is at home right now") rather than a coarser fact ("the user has been at home at some point during the day"). However, its guarantees can be undermined by correlations among data points. This is a significant issue for trajectory data, which exhibit strong autocorrelation (see Section 3.1) and for which the limitations of event-level DP have been highlighted in prior work [31], [36]. Intuitively, because altering a single event requires modifying a span of m events (see Figure 3.1) to maintain a feasible trajectory, the effective indistinguishability bound for one event becomes $m\varepsilon$, increasing privacy leakage and potentially allowing an attacker to infer a specific spatio-temporal point.

w-event privacy builds on top of event-level to address attribute disclosure that arises when a sensitive attribute is distributed across multiple events [108]. Particularly, Kellaris et al. argue that, in many practical scenarios, sensitive information is not disclosed by a single event in isolation, but rather by a sequence of events occurring over *consecutive*

time steps. To address this issue, they define w-event privacy, which guarantees indistinguishability (up to ε) for any sequence of events within a sliding window of w successive time instants. Its definition of neighboring databases is the following:

Definition 3.2 (w-neighborhood)**.** Let w be a positive integer. Two stream prefixes $D_t = \{S_1, \ldots, S_t\}$ and $D'_t = \{S'_1, \ldots, S'_t\}$ are *w-neighboring*, if, for all $i \leq t$, S_i and S'_i are either equal or we can obtain one from the other by changing an entry of S_i, and all pair of indexes i, j corresponding to the latter case verify that $|i - j| < w$.

Intuitively, this definition imposes that all the differing S_i and S'_i of the stream must fit in a *w-window* (see Figure 3.2).

This definition captures settings where sensitive information is disclosed from a sequence of events of length w. It does not only protect the locations visited by a single user over w consecutive timestamps but also can protect those of different users. In terms of privacy, for values of w close to 1, w-event privacy approximates to event-level privacy, and for large values, it converges to user-level privacy. In terms of sensitivity, its lower bound is the event-level sensitivity, and its upper bound is the user-level one. Therefore, this notion protects more information than event-level privacy while allowing less noise addition than user-level, even though some of its deficiencies remain present.

The notion still leaks attributes (e.g., "Being at the hospital"), when these cannot be protected by the same w-window. For example, assume user u_1 in Figure 3.2 (where $w = 3$) is a compulsive gambler who visits the casino (red dot) multiple *nonconsecutive* times a day. The sensitive information that u_1 has been at the casino is not protected as the red dots cannot fit into a unique w-window. Also, the user's participation is still unprotected if the attacker's knowledge exceeds the window.

Given that consecutive spatial points are usually more correlated, this new notion is also superior to event-level privacy against correlation attacks. However, we highlight the importance of a correct selection of the w value, that must be larger than the mixing time of the considered time series in order to ensure that samples in different blocks are almost uncorrelated and thus to prevent correlation-based attacks on successive releases [36]. Here, the mixing time refers to the number of time steps required for the stochastic process to become close (e.g., in total variation distance) to its stationary distribution, so that temporal dependencies between sufficiently distant observations become negligible [81].

Importantly, the assumption of w-event privacy that trajectories are periodically recorded, may overestimate the number of consecutive protected locations. For instance, in Figure 3.2, where we have non-periodically recorded trajectories, the 3-window 5–7 cannot protect more than two locations of a single user.

***ℓ*-trajectory privacy.** Cao and Yoshikawa [89] aim to overcome this last deficiency of w-event privacy, especially when users' trajectories are not periodically recorded. To tackle this issue, they extend the previous model to introduce ℓ-trajectory privacy.

Definition 3.3 (ℓ-trajectory neighborhood)**.** We say two databases are *ℓ-trajectory neighboring* if one is obtained from the other by only modifying all locations in a single

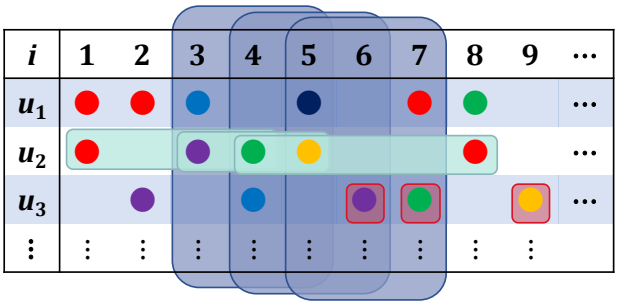


Figure 3.2.: An example of a non-periodically recorded streaming database, colored dots represent different locations. The rounded boxes represent protection scopes of event-level (red), w-event (blue), and ℓ-trajectory privacy (green), for $w = \ell = 3$. Observe that the blue box (w-window) always spans w timestamps independent of how many points they include, and that the green box (ℓ-trajectory) always includes ℓ points independently of the number of timestamps it spans.

ℓ-trajectory. Here, an *ℓ-trajectory* is defined as a sequence of ℓ successive spatio-temporal data points produced by the same user (see Figure 3.2).

The goal of the ℓ-trajectory privacy notion is to protect each sequence of ℓ points from the same user independently of the number of timestamps they span. Clearly, varying ℓ allows us to move closer to event-level ($\ell = 1$) or to user-level privacy ($\ell \to \infty$).

Although this notion overcomes the problem of w-event privacy of assuming periodically recorded trajectories, it does not address its other deficiencies.

Element-level privacy. The high privacy cost and resulting utility loss of user-level privacy motivated the introduction of **element-level** privacy [109], originally proposed in the context of text privacy. The authors argue that producing text is not inherently sensitive; rather, sensitive attributes may arise from specific elements within the text, such as the use of swear words.

This philosophy naturally extends to the trajectory domain: participation in a database (user-level privacy) may not be sensitive. For instance, in a traffic study, participating in the database only discloses information such as "Having a car" or "Living in the area", which users can regard as insensitive. However, one may wish to prevent attribute disclosure. Continuing the previous example, consider a person who visits a hospital multiple times throughout the day and may want to keep the mere fact of visiting the hospital private, as it directly reveals sensitive information about their health status. In this case, event-level privacy is not sufficient for the previously mentioned reasons, and neither w-event nor ℓ-trajectory privacy if the visits do not fit into the w-window or ℓ-trajectory. To address this situation, the authors propose element-level privacy.

The original proposal [109] models the data of a user u as a multiset of values $x^{(u)} = \{x_1^{(u)}, x_2^{(u)}, \ldots, x_{m_u}^{(u)}\}$, where each $x_i^{(u)}$ belongs to the universe of possible values $\mathcal{X}$. Then it considers a K-partition of the universe $\mathcal{X}$ into the clusters $c_1, \ldots, c_K$. These

clusters are viewed as the *elements* to be protected. By definition, each $x_i^{(u)}$ belongs to one cluster c_j.

Definition 3.4 (Element-neighborhood)**.** Two databases D, D' are *element-neighboring* if they are equal except for a pair of users' data $x^{(u)} = \{x_1, \ldots, x_{m_u}\} \in D$, $x^{(u')} = \{x'_1, \ldots, x'_{m_{u'}}\} \in D'$ such that

$$d_{\text{user}}(x^{(u)}, x^{(u')}) \coloneqq \sum_{k=1}^{K} \mathbf{1}_{\{\{x_i | x_i \in c_k\} \neq \{x'_i | x'_i \in c_k\}\}} \leq 1,$$

where $\mathbf{1}_{\{\{x_i | x_i \in c_k\} \neq \{x'_i | x'_i \in c_k\}\}}$ denotes the indicator function that outputs 1 when the inequality holds and 0 in other case, with $\{x_i \mid x_i \in c_k\}$ being implicitly multisets.

Observe that by modifying the cluster selection, we can achieve user-level granularity by taking only one cluster, $c_1 = \mathcal{X}$.

The interpretation of ensuring element-level privacy is that we are hiding that each user has elements belonging to the cluster, independently of how many elements it includes.

We believe that this notion can be adapted to trajectory data. In the case of raw trajectories, we can cluster data points according to geographical zones and times. And in the case of semantic trajectories, we can choose the clusters according to semantic values, e.g., having a cluster for all health-related locations.

A challenge here is how to establish the clusters to provide real protection that covers all possible scenarios regarding the user's privacy desire. For instance, if we choose spatial areas as clusters, a question arises about the size we should take and which privacy guarantees we would have according to our selection. This extends when considering semantic trajectories: If we reduce a cluster to a specific hospital, we will protect the visits to this hospital, but if we instead include all hospitals in the cluster, we will then be protecting any visit to any hospital.

This notion is relatively recent. Hence, there have not been many mechanisms achieving it and no adaptations specifically to trajectory data. Moreover, no comparison has been conducted against other granularity notions regarding utility and privacy.

Conclusions on granularity notions. In terms of privacy protection, user-level privacy is the strongest, followed by element-level, ℓ-trajectory, and w-event privacy. However, user-level privacy may result in excessive loss of utility in the complex field of trajectory publication, ending in unbounded sensitivities, huge privacy degradations due to sequential composition and infeasibility in infinite streaming context or heterogeneous datasets with different streaming lengths.

With the exception of user-level privacy, all proposed granularities permit membership disclosure. Importantly, we consider event-level privacy to be insufficient for trajectory data, since—even with respect to its original goal of preventing exact spatiotemporal point reconstruction—it remains vulnerable to correlation-based attacks.

Both w-event privacy and ℓ-diversity offer improvements in this regard, but only when their parameters are carefully selected according to the underlying correlation model. Moreover, neither provides effective protection against arbitrary attribute disclosure,

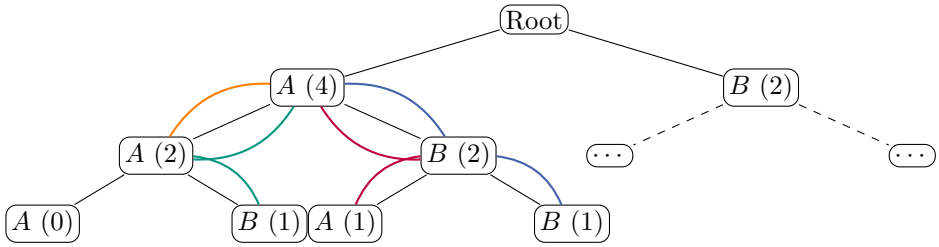


Figure 3.3.: Example of an exploration prefix tree encoding a trajectory dataset [84] for a universe with two possible locations, $\mathcal{X} = \{A, B\}$. Each colored line represents a trajectory, while dashed lines indicate omitted parts of the tree. The counts show how many trajectories share a given prefix.

such as revealing the length of a trajectory. Even if participation in the database is not itself sensitive, the leakage of user attributes constitutes a significant privacy risk.

Element-level privacy may represent a promising approach to mitigating attribute disclosure. However, it has not yet been adapted to trajectory data, making it difficult to evaluate its impact on the utility of anonymized trajectories.

3.4. DP Masking Mechanisms

Next, we examine masking algorithms that adapt trajectory databases for publication with DP guarantees. The corresponding state of the art we review in this work covers the static-context publication in which the sanitized database is released just once in its entirety, without subsequent updates[1]. We classify them according to their fundamental concept. We provide an overview of our classification in Table 3.2, including information on the privacy notion they satisfy and their properties. Observe that most of the reviewed proposals aim to achieve user-level DP.

DP algorithms require a randomized approach since deterministic algorithms cannot achieve DP guarantees [16]. The two classical mechanisms to provide DP are the *Laplace* and *exponential mechanisms* [16]. Nearly all the algorithms presented in this section leverage these mechanisms in some way.

3.4.1. Noisy Counts

We categorize here the approaches that encode the trajectory dataset in a structured format—often a tree—apply Laplace noise to the counts associated with each element—often a node of the three—within the transformed representation, and then invert the transformation to reconstruct a sanitized dataset.

[1]Note that this survey corresponds to our previously published work [32] and therefore does not include proposals introduced after 2023.

Privacy notion		Classification		Ref.	Correct DP notion	Mech. (Δ)		Properties		
						Laplace	Exponential	Considers time	Unb. loc. univ.	Realism
User-level	ε-DP*	Noisy counts	Exploration tree	[110]		l_{max}	○		✓	
	ε-DP	Noisy counts	Exploration tree	[84]		1	○			
		Noisy counts	Exploration tree	[83]		l_{max}	○			
		Noisy counts	Exploration tree	[111]		l_{max}	○			
		Noisy counts	Sequence tree	[112]	✗	•	○			
		Noisy counts	Sequence tree	[113]	✗	•	○			
		Noisy counts	Sequence tree	[114]	✗	•	○			
		Noisy counts	Trajectory count	[115]	✗	•	○		✓	✓
		Noisy counts	Tree + Markov / Random centroid	[116]	✗	•	○		✓	
		Clustering	Exp.: k-means	[117], [118]	✗	•	•		✓	
		Clustering	Exp.: k-means	[119]	✗	•	•		✓	
		Clustering	Exp.: Hilbert curves	[120]	✗	•	•		✓	
		Clustering	Universal clustering	[121]	✗	•	○		✓	
Event-level	$(0, \delta)$-DP	Sampling + interp.		[122]		○	○		✓	
User-level	(ε, δ)-DP	Sampling + interp.		[123]		ΔX	○	✓	✓	
	ε-LDP	Perturbation		[124]		○	Δd_w	✓	✓	✓

Table 3.2.: Summary of explored DP-based mechanisms according to our classification and exact privacy notion they satisfy. "Correct DP notion" labels mechanisms that incorrectly claim DP. *It provides ε-DP only when restricted to certain spatial areas.

3.4.1.1. Exploration Tree

Chen et al. [84] encode the trajectory dataset as weighted counts on a prefix tree over the location domain. Since there is a one-to-one correspondence between the tree and the original dataset, adding noise to the counts at each node enables the reconstruction of a sanitized version of the dataset.

Formally, in the exploration prefix tree each node is labeled with a possible location, which can only be an element of a predefined finite set of locations (the *universe of locations*). Every possible prefix trajectory is represented uniquely as a walk from the root node to another (i.e., we represent prefix $\langle p_1, \ldots, p_m \rangle$ by the node obtained after walking through the tree following the labels: $root \to p_1 \to \cdots \to p_m$). This node stores

the number of times (i.e., the *counts*) the prefix appeared in the database. The tree includes all the possible trajectories drawn from the universe of locations, including those not present in the database (i.e., those that are zero). By definition of a prefix, the count of any node must be less than or equal to the count of its parent. This way the count of each prefix of length n is stored at the nth level of the tree. We show an example of this model in Figure 3.3 covering a universe of two locations, $\mathcal{X} = \{A, B\}$, and a database with six trajectories.

To enforce DP, Laplace noise is added to the count of each node (including the 0 ones, potentially creating sequences not contained in the original data). Since each trajectory has only one prefix of length n, the sensitivity of the counting is 1. For consistency, any node with a noisy count of 0 is considered a leaf, which is equivalent to assigning a count of 0 to all of its children.

Then, to release the trajectory database, we only need to explore the resulting tree. Based on the noisy prefix tree, we can draw the sanitized database by traversing it once, calculating the number r of trajectories terminated at each node, and appending r copies of the prefix saved in that node to the output. Since creating and exploring the tree are inverse operations, there is a one-to-one correspondence between the database and the prefix tree. Note we need a post-processing module to maintain the tree consistency (i.e., the sum of counts of descendant nodes cannot be higher than that of their ancestors).

One important limitation is the preservation of common transitions or sequences, especially when they occur in different long trajectories. Because each prefix receives independent noise, the total noise for a frequent pattern accumulates proportionally to the number of longer sequences that include that pattern.

To address this limitation, in subsequent work, the same authors improved this approach by introducing an n-gram exploration tree [83] that looks at n-grams instead of prefixes, and they empirically show improved utility. Note that, the change from prefixes to n-grams leads to higher counts in each node and higher sensitivity. In this case, each trajectory could add its total length to a node count. Therefore, the sensitivity is the maximum trajectory length, l_{max}, allowed in the database (any trajectory longer than l_{max} is cut before introducing it in the data). The authors also add Laplace noise on the n-grams counts. This way, a frequent n-gram is directly perturbed with the same amount of noise, independently of the length of the trajectory it belongs to, effectively reducing the error. Once again, by exploring the tree, we recover the perturbed version of the original trajectories, obtaining a sanitized trajectory database.

Other proposals modify these algorithms in various ways. Firstly, Wang and Kankanhalli [110] define sensitive zones and apply Chen et al.'s method [83] only to these zones, which provides better utility. However, their privacy notion is weaker since they do not provide DP for the whole database but only for sensitive zones.

DPLG [111] constructs the same noisy n-gram tree (therefore, the sensitivity of each node count is l_{max}) but provides a non-uniformly distributed privacy level by regulating the amount of noise added, so the location will be more or less protected depending on the area of the map it is.

All exploration-tree–based methods suffer from several common limitations. First, they require a fixed and discrete universe of possible locations and setting a maximum trajectory length. These strong assumptions are necessary to bound the sensitivity of the mechanisms. Second, the size of the trees grows exponentially with the number of locations and the allowed trajectory length, i.e., $\mathcal{O}(|\mathcal{X}|^{\ell_{\max}})$, which leads to both computational and storage challenges. Limiting the trajectory length, however, can significantly reduce utility. As a result, these methods can only be applied to a small universe of locations—for example, the experimental setting in [83] considers at most 342 locations. In contrast, real-world applications for traffic analysis often involve GPS data (i.e., continuous data domains with $|\mathcal{X}| = \infty$) or city road networks with thousands of streets (see Section 5.5 and [88], [125]).

Moreover, these mechanisms retain only spatial information and counts, discarding temporal information. This further limits utility and renders them unsuitable for tasks such as traffic monitoring, scheduling and planning, or traffic prediction—any application that requires knowledge of when particular locations are most visited.

Finally, it is important to note that none of the surveyed approaches take into account the correlations between spatio-temporal points, nor the regularity and self-similarity of trajectories belonging to the same or related users (see Sections 2.3 and 3.1). When applying the Laplace mechanism, noise added to originally zero counts may generate impossible or highly unlikely sequences. By incorporating road-map constraints and physical movement laws, a simple stochastic model can detect unrealistic trajectories and leverage correlations to compromise the claimed privacy guarantees [74].

3.4.1.2. Sequence Tree

More recent approaches try to build trees storing the counts of subsequences in each node instead of only one location (i.e., *sequence trees*). This is the case of NTPT [112]. This mechanism first tries to overcome data sparseness by simplifying the trajectories. By performing an optimal segmentation process, the trajectories are divided into sequences, and then, it constructs a prefix tree where each node stores a sequence. Afterward, it adds Laplace noise to the counts of each node.

Related approaches are presented in [113], [114], with the difference that they rely on a similarity factor. More specifically, they save sequences of spatio-temporal points in a tree structure according to the number of location points they have in common. Analogously to previous approaches, they add Laplace noise to the count of each sequence node and reconstruct the dataset from the noisy tree.

3.4.1.3. Trajectory Count

Finally, one work considers the correlation between individuals in the database [115]. Here, the authors measure the correlation coefficients between the different trajectories in the database, which translate into privacy risk: the more correlated trajectories are, the more risk they pose. Therefore, they allocate different privacy budgets adding more

Laplace noise to the counts of the risky ones. Note that this approach adds noise directly to the count of the entire trajectory, without encoding it as a tree.

3.4.1.4. Correctness of DP in Noisy-Counts Mechanisms

We would like to note that *all* the sequence tree and trajectory count suggestions (namely, [112], [113], [114], [115]) suffer from a common formal mistake and do not provide DP. They output perturbed counts of only those segments, subsequences, or trajectories present in the original database, but do not change the output of hypothetical sequences with zero counts, as in the exploration-tree–based methods we discussed. These design choices contradict the definition of DP, and thus the final protocol cannot provide DP. We show in the following proof that a meaningful DP mechanism cannot simply change the counts of the elements in the database:

Proposition 3.5. *Let $\mathcal{M}$ be a randomized algorithm with domain $\mathbb{D}$[2]. Suppose $\mathcal{M}$ changes the counts of the rows of $D \in \mathbb{D}$ (where it is possible to change a positive count into 0, but not the other way around). If $\mathcal{M}$ is ε-DP, then $\mathcal{M}$ is the void algorithm (i.e., it outputs the empty set independently of the input).*

Proof. Let $\mathcal{M}$ be an ε-DP algorithm, as described in the statement. By definition, the output domain of $\mathcal{M}$ is a subset $\Theta \subseteq \mathbb{D}$.

Fix $D \in \mathbb{D}$. For every $x \in D$, denote $k_x < \infty$ as the number of times x appears in D and D_x as the database obtained after removing all elements x from D. For every $x \in D$, there exists a sequence of neighboring databases of $\mathbb{D}$:

$$D = D_0 \sim D_1 \sim \cdots \sim D_{k_x - 1} \to D_{k_x} = D_x,$$

i.e., D_{i-1} and D_i are neighboring for all $i \in \{1, \ldots, k_x\}$. Then, since $\mathcal{M}$ is ε-DP, we obtain for all measurable $S \subseteq \Theta$ and $x \in D$ that

$$\begin{aligned}
\Pr[\mathcal{M}(D) \in S] &\leq \mathrm{e}^{\varepsilon} \Pr[\mathcal{M}(D_1) \in S] \\
&\leq \mathrm{e}^{2\varepsilon} \Pr[\mathcal{M}(D_2) \in S] \leq \cdots \\
&\leq \mathrm{e}^{(k_x - 1)\varepsilon} \Pr[\mathcal{M}(D_{k_x - 1}) \in S] \\
&\leq \mathrm{e}^{k_x \varepsilon} \Pr[\mathcal{M}(D_x) \in S] = 0.
\end{aligned}$$

Let $\Theta_D \subseteq \Theta$ be the set of all possible outputs of $\mathcal{M}(D)$, by definition $\Pr[\mathcal{M}(D) \in \Theta_D] = 1$. Furthermore, Θ_D is contained in the discrete set $\{S \text{ multiset} \mid \text{for all } x \in S,\ x \in D\}$, and therefore Θ_D is discrete, and

$$\Pr[\mathcal{M}(D) \in \Theta_D] = \sum_{\theta \in \Theta_D} \Pr[\mathcal{M}(D) = \theta].$$

[2]We will use Dwork and Roth's definition of database [16], defined as a multiset drawn from $\mathcal{X}$, the universe of database rows (represented too by their histograms from $\mathbb{N}^{|\mathcal{X}|}$). To simplify notation, we use $\mathbb{D}$ to denote a set of finite databases.

For every non-empty output dataset $\theta \in \Theta_D$, we select an element $x \in \theta$. By the previous inequalities, we obtain that

$$\Pr[\mathcal{M}(D) = \theta] \leq \mathrm{e}^{k_x \varepsilon} \Pr[\mathcal{M}(D_x) = \theta] = 0,$$

since $x \notin D_x$ and $x \in \theta$. Therefore,

$$1 = \Pr[\mathcal{M}(D) \in \Theta_D] = \sum_{\theta \in \Theta_D} \Pr[\mathcal{M}(D) = \theta] = \Pr[\mathcal{M}(D) = \varnothing].$$

Since $\mathcal{M}(D)$ is a discrete random variable, it proves that it can only output the empty set. Then, we repeat the proof for every possible database $D \in \mathbb{D}$, proving that $\mathcal{M}$ is the void algorithm. □

This issue is not reflected in the privacy analyses of sequence tree and trajectory count methods [112], [113], [114], [115]. While the authors provide proofs of DP for the individual mechanisms they employ (e.g., the Laplace mechanism), they overlook the fact that embedding a DP mechanism within a more complex system can compromise the overall privacy guarantees. Particularly, if the count of the victim's trajectory is positive after perturbation, and this trajectory contains a quasi-identifier known by the attacker, such as their home or work, the victim and the rest of its path can still be identified.

This can be generalized to the requirement that any DP mechanism must be able to output any possible value of the range independently of the database. We formalize this statement with the precise hypotheses in Proposition 3.6.

Proposition 3.6. *Let $\mathcal{M}$ be a randomized algorithm that satisfies bounded ε-DP, $\mathcal{X}^n$ its domain, and $\Theta = \mathrm{Range}(\mathcal{M})$ the set of all possible outputs of $\mathcal{M}$. Then, given any measurable $S \subseteq \Theta$, if there exist $D \in \mathcal{X}^n$ such that $\Pr[\mathcal{M}(D) \in S] > 0$, it is also true for all other $D' \in \mathbb{D}$.*

Proof. Consider a measurable $S \subseteq \Theta$ such that there exist $D \in \mathbb{D}$ in a way that $\Pr[\mathcal{M}(D) \in S] > 0$. We then proceed by *reductio ad absurdum*: that is, we assume that there exists $D' \in \mathbb{D}$ such that $\Pr[\mathcal{M}(D') \in S] = 0$ and we will end in a contradiction.

Since we assume all databases are finite, there exists a finite sequence of neighboring databases from D to D' of length k. As in the proof of Proposition 3.5, we obtain

$$\Pr[\mathcal{M}(D) \in S] \leq \mathrm{e}^{k\varepsilon} \Pr[M(D') \in S] = 0.$$

This contradicts that $\Pr[\mathcal{M}(D) \in S] > 0$. □

Note that this proof can be extended to any connected privacy space, i.e., $(\mathbb{D}, d_{\mathbb{D}})$ such that $d(D, D') < \infty$ for all $D, D' \in \mathbb{D}$ (see Section 2.1).

Conclusions on noisy counts. We conclude that the only noisy-count mechanisms that achieve acceptable privacy guarantees are the original exploration-tree approaches [83], [84], [110], [111]. However, due to their high computational and storage cost for large databases, we only see these methods used for cases with reduced universes, such as the analysis of public-transport lines of a city. Moreover, not any DP proposals offers protection against correlation-based attacks.

3.4.2. Clustering

In this category, we consider all approaches that cluster locations and subsequently release trajectories through these clusters with some perturbation to guarantee privacy.

They follow a common structure that consists of two privacy mechanisms: A generalization mechanism $\mathcal{M}_1$, which generalizes the set of locations by grouping them into clusters, and a releasing mechanism $\mathcal{M}_2$, which outputs resulting trajectories drawn from the generalized sets. To achieve DP publication, both $\mathcal{M}_1$ and $\mathcal{M}_2$ must ensure DP (as we prove in Section 4.5).

3.4.2.1. Exponential Clustering

Hua et al. [117] is the first proposal using clustering. Their idea for $\mathcal{M}_1$ is to cluster and merge concurrent locations from different trajectories, following a probabilistic partitioning based on the exponential mechanism. Then, using the Laplace mechanism, $\mathcal{M}_2$ connects the merged locations and forms the final generalized trajectories.

Specifically, the authors propose a noisy clustering approach to partition the set of locations at each time step into m groups. Given the set of *reported* locations at time i, denoted by Γ_i, they assign a score to each possible partition of Γ_i, $p_i \in \mathcal{P}(\Gamma_i)$, based on the average distance between the trajectories whose locations are grouped together. The final partition is then selected according to this score using a exponential mechanism procedure, i.e., they choose $p_i \in \mathcal{P}(\Gamma_i)$ with probability proportional to $\exp\left(\varepsilon \frac{u(D,p_i)}{2\Delta u}\right)$ where u is the desired scored function.

Finally, the locations in each subset are clustered and replaced by their corresponding centroid. Hence, from a initial set of reported locations Γ_i they obtain a smaller set (after clustering) $\tilde{\Gamma}_i$.

Finally, they build the new trajectories from this reduced set $\tilde{\Gamma}_i$ using the mechanism $\mathcal{M}_2$, which draws sequences from $\tilde{\Gamma}_i$ at random. The counts are attributed following the Laplace mechanism until obtaining a sanitized database of the same size as the original.

Subsequent approaches build upon this technique. For instance, Chen et al. [118] incorporate Hua et al.'s method [117] as the final step of their protocol. Later, Li et al. [119] enhance the utility of the $\mathcal{M}_2$ algorithm by using bounded Laplace noise. Additionally, Han et al. [120] refine the private clustering strategy in $\mathcal{M}_1$ with Hilbert curves, eliminating the need to predefine the number of clusters.

Correctness of DP in exponential-clustering mechanisms. After studying these approaches [117], [118], [119], [120], we observe an issue when applying the exponential mechanism. The exponential mechanism [16] selects the best element of a certain given set $\mathcal{R}$, the range of this mechanism. The best assignments for each database are chosen using a *score function* u, which associates scores to each element in the database: the higher the score, the higher its chances to be chosen. More formally, given a database $D \in \mathbb{D}$, the exponential mechanism outputs $r \in \mathcal{R}$ with probability proportional to $\exp\left(\varepsilon \frac{u(D,r)}{2\Delta u}\right)$, where $u\colon \mathbb{D} \times \mathcal{R} \longrightarrow \mathbb{R}$ is the score function and

$$\Delta u := \max_{r \in \mathcal{R}} \max_{D \sim D'} |u(D,r) - u(D',r)|,$$

is its sensitivity. Note, in particular, that $\mathcal{R}$ needs to be data independent.

In the original framework [117], that all the others build on, the exponential mechanism is used to output the centroids of the partitions of the location set at every timestamp i. In this work, the score function is defined as $u\colon \mathbb{D} \times \tau \longrightarrow \mathbb{R}$, with τ being the set of partitions of the locations set at time i of a specific database D. The previous expression is not well-defined, since τ depends on the chosen element $D \in \mathbb{D}$, and varies when changing to another D, as mentioned in the paper. As a direct consequence, Δu is not theoretically computable (even if fixing D, since the definition compares two different databases), and an exponential mechanism cannot be defined. Hence, one cannot claim the algorithm ensures DP via the exponential mechanism.

This error leads to formal mistakes in the suggested proposal. First, the cluster size does not affect the privacy guaranteed: i.e., we can choose to partition into sets of size 1, which would simply be a mechanism outputting the original unmodified database, providing no privacy. Secondly, $u(D, r) \leq 1$ for all possible combinations, would imply that the absolute difference between any possible score function is at most 1. If the exponential mechanism were correctly applied, it would mean that changing the whole database has the same effect as changing one record, which is highly improbable.

Having explained why the mechanism is not the exponential mechanism, we discuss why it is not DP. We know that given two different sets, S and S', their sets of partitions into m groups, $\mathcal{P}_S^m$ and $\mathcal{P}_{S'}^m$, then $\mathcal{P}_S^m \cap \mathcal{P}_{S'}^m$ are disjoint. Since a partition covers exactly the elements of its underlying set, two distinct sets cannot share a common partition.

For example, consider $S = \{1, 2, 3\}$ and $S' = \{1, 2\}$. The only partition of S' into two clusters is $P_{S'} = \{\{1\}, \{2\}\}$, while for S we have $P_S^{(1)} = \{\{1, 2\}, \{3\}\}$, $P_S^{(2)} = \{\{1, 3\}, \{2\}\}$ or $P_S^{(3)} = \{\{2, 3\}, \{1\}\}$. It is then easy to see that $\mathcal{P}_S^2 \cap \mathcal{P}_{S'}^2 = \varnothing$.

More formally, consider two neighboring databases $D, D' \in \mathbb{D}$ and their respective location set at time i, Γ_i and Γ_i'. Let $\mathcal{P}, \mathcal{P}' \subseteq \text{Range}(\mathcal{M})$ be the set of all possible partitions of Γ_i and Γ_i', respectively, into m groups. As mentioned, $\mathcal{P} \cap \mathcal{P}' = \varnothing$, hence

$$1 = \Pr[\mathcal{M}(D) \in \mathcal{P}] \leq \mathrm{e}^{\varepsilon} \Pr[\mathcal{M}(D') \in \mathcal{P}] = 0,$$

resulting in a contradiction with the definition of DP. As we already proved in Proposition 3.6, if an output is possible for a database, it needs to be possible for all the remaining ones, which simply cannot happen if the range of outputs is data dependent.

In summary, the protection mechanisms ($\mathcal{M}_1$, in particular) of the exponential clustering approaches [117], [118], [119], [120] do not provide DP because they do not use correct exponential mechanisms, since their abstract range of outputs is completely dependent on the input database. Furthermore, we can construct an example attack that shows how DP breaks in this case:

Example 3.7. We propose a simple trajectory database consisting of three users, and we focus on any specific timestamp. We are working with a set of three locations $\Gamma = \{l_1 = (2, 2), l_2 = (5, 2), l_3 = (5, 5)\}$. The mechanism $\mathcal{M}$ clusters databases into two and outputs its centroid according to the Euclidean distance (essentially, Hua et al.'s proposal [117] with $m = 2$). Assume a strong attacker that knows all the values

except their target l_1 (consistent with DP definition). We show that, with the released information, the attacker obtains the target data with total accuracy.

The attacker knows l_2 and l_3, and that the possible clusters that $\mathcal{M}$ can compute are $\mathcal{P} = \{P_1, P_2, P_3\}$ with $P_1 = \{\{l_1\}, \{l_2, l_3\}\}$, $P_2 = \{\{l_1, l_2\}, \{l_3\}\}$ and $P_3 = \{\{l_2\}, \{l_1, l_3\}\}$. So, the attacker knows also a priori that the mechanism will output one of the following centroid sets: $C(P_1) = \{?, (5, 3.5)\}$, $C(P_2) = \{?, (5, 3)\}$, or $C(P_3) = \{(5, 2), ?\}$, with ? denoting the coordinates they cannot predict. Suppose the mechanism computes $\{\{l_2\}, \{l_1, l_3\}\}$ (unknown to the attacker) and their centroids, and then releases $C = \{(5, 2), (3.5, 3.5)\}$. Now, the attacker can compare their computation with the released centroid set, and conclude that the only possible partition is P_3. Additionally, since $(3.5, 3.5)$ is released and corresponds to the middle point between $l_1 = (x_1, y_1)$ and $l_3 = (x_3, y_3)$, the attacker can easily recover l_1:

$$(3.5, 3.5) = \left(\frac{x_1 + x_3}{2}, \frac{y_1 + y_3}{2}\right) = \left(\frac{x_1 + 5}{2}, \frac{y_1 + 5}{2}\right).$$

Therefore,

$$l_1 = (x_1, y_1) = (2 \cdot 3.5 - 5, 2 \cdot 3.5 - 5) = (2, 2),$$

which allows the attacker to reconstruct the original database in its entirety.

The only way to avoid this issue would be to define a data-independent universe of locations, for instance, based on the city map, and output a partition of this universe. This way, the mechanism could achieve DP. Being independent of the actual patterns in the data could incur a significant utility loss in some scenarios.

Universal clustering. Recently, Zhao et al. [121] introduce a protection proposal independent of the specific clustering algorithm. It allows one to choose any preferred clustering and run it on the database without modification. They add Laplace noise to location coordinates (using the polar form) and to the counts of these data in the cluster. Finally, the authors calculate the noisy centroid according to the noisy counts and locations and release these centroids. The noisy count algorithm they use is the same as in the works [112], [113], [114] that we have shown to lack DP guarantees in Section 3.4.1. Furthermore, following this scheme, we cannot release more than the corresponding centroids since there is no private way of establishing connections between centroids and thus forming trajectories without using the original data. The authors do not propose any mechanism for trajectory release ($\mathcal{M}_2$).

3.4.2.2. Random Centroid

Finally, we find DPTD [116], which introduces a generalization module that clusters the locations without consuming privacy budget (the proposed solution chooses a random location instead of the centroid). For the release method M_2, the authors adapt the noisy prefix tree structure by Chen et al. [83] to reduce the consumption of the privacy budget and provide higher utility. Instead of adding Laplace noise to the odd layers of the tree, they predict the new count with a Markov process. This Markov process

uses the frequencies of the original database, apparently without protection (i.e., no noise or perturbation added to the frequencies). Although the authors attempt to reduce the privacy budget consumed, the generalization step indirectly uses the database in its election of the centroid, thus breaking DP. The publications also contain neither analyses nor proofs of privacy, so the actual protection achieved remains unclear.

3.4.2.3. General Problems

Apart from the privacy issues we have explained in each proposal, we find general problems. First, the generation of impossible trajectories challenges the utility of the resulting output. Specifically, the presented methods can create trajectories in which two consecutive locations are unreachable in the given time and unrealistic centroids placed at impossible locations, such as in the middle of a river or on top of a building.

Another limitation is that the used score function of the exponential mechanism only depends on physical distance and therefore does not consider time. These proposals are thus inapplicable for non-periodically recorded and variable-length trajectories, which represent a majority of real-world databases.

Conclusions on clustering. This category of approaches overcomes the applicability problem of those using trees (see Section 3.4.1), as they do not need to assume a small universe of locations. However, we can still identify several deficiencies: merging without considering time can yield to unrealistic patterns and facilitate correlation-based attacks. Also, as mentioned above, all of these proposals contain erroneous DP analyses or proofs. It hence remains unclear which protection they provide.

3.4.3. Sampling and Interpolation

Another group of mechanisms is based on point sampling and interpolation [122], [123]. The sampling technique consists of selecting a subset of the database (in this case, trajectory points), while interpolation is used to counteract the size reduction due to sampling by reconstructing intermediate points of the trajectories. The sampling techniques used provide (ε, δ)-DP, and interpolation is conducted as post-processing, without affecting the privacy guarantees.

Shao et al. [122] present two mechanisms, SFI and IFS, for ship-trajectory privacy based on these techniques. SFI first randomly samples points over each trajectory and then redraws trajectories using a cubic Bézier interpolation (the "a priori" mechanism). IFS first interpolates and then samples (the "a posteriori" mechanism). The mechanisms are proven to achieve event-level $(0, \delta)$-DP. In their experimentation, the authors conclude that SFI works better than IFS for small values of δ and not-so-smooth trajectories. Note that the interpolation technique is specifically designed for smooth trajectories without geo-spatial constraints, such as those of ships at sea. Adapting this approach to other trajectory domains, such as road traffic analysis, would require modifications to the interpolation strategy, which have not yet been explored.

Similar to the mechanisms discussed in the previous sections, this algorithm ignores the temporal dimension, which can result in infeasible trajectories where two consecutive

points are not physically reachable within the given time. Furthermore, both SFI and IFS are designed for a very specific domain (ships at sea) and do not trivially generalize to other types of trajectories, such as those of pedestrians or road vehicles, which involve sharper turns and must conform to a road network.

Another proposal is VTDP [123], which consists of a three-phased sampling with a final interpolation step and satisfies (ε, δ)-DP.

The VTDP protocol performs trajectory sanitization through three sequential conditional sampling phases followed by a final interpolation step. In Phase I, discretized vehicle positions are perturbed using noisy counts calibrated with the Laplace mechanism. Phase II samples speeds and accelerations conditioned on the sanitized positions via a Dirichlet-Multinomial model, and Phase III incorporates timestamps dependent on the outputs of the previous phases. VTDP provides approximate DP by independently calibrating noise at each phase and bounding the overall privacy loss through sequential composition, i.e., $\varepsilon = \sum_i \varepsilon_i$. Prior preprocessing steps, including geo-temporal discretization and outlier removal, reduce sensitivity, while a client-side physics-based interpolation enhances trajectory realism and density without incurring additional privacy cost due to the post-processing property of DP.

Unlike prior approaches, VTDP explicitly incorporates the temporal dimension and enforces geo-physically consistent trajectories. However, its utility evaluation is restricted to a very limited setting, namely a small section of an arterial road. A more extensive experimental assessment would therefore be necessary to determine its applicability to real-world traffic monitoring scenarios.

3.4.4. Local Perturbation

While LDP proposals for location privacy have been explored [126], we only find one protection mechanism [124] that perturbs trajectories to satisfy ε-LDP. Recall that these trajectories are a time-ordered sequence of POIs visited by a user. The authors integrate public knowledge to improve the utility without affecting the privacy budget ε. The proposed mechanism utilizes this public knowledge to partition the set of all POIs into spatio-tempo-categorical regions, such that each contains some number of POIs.

The mechanism is divided into four parts: first, it generalizes every POI into the corresponding region; it partitions these new trajectories into n-grams, which are then individually perturbed following the exponential mechanism to ensure ε-LDP, where the score function is a distance function d_w defined over the spatial, temporal and categorical dimensions; then trajectories are reconstructed by minimizing the distance function; and finally, the mechanism returns to the initial domain by randomly picking a POI for each section, making sure that consecutive locations in a trajectory are reachable in the corresponding time.

This mechanism demonstrates several advantages over those described above. First of all, ε-LDP is stricter trust model than ε-DP since there is no need for a trusted curator. Furthermore, it does consider the temporal dimension (and the categorical dimension of the trajectories). It also takes into consideration publicly available information to

improve the overall utility of the mechanism, without any effects on the privacy budget, and ensures that the published data is realistic.

However, it also faces some challenges: First, to adapt the mechanism to a multiple-release setting (i.e., the same user contributing more than one trajectory), the user needs to know in advance how many trajectories they want to share, to divide the overall privacy budget by this number [124].

Second, the sensitivity of the exponential mechanism, Δd_w, depends on the fixed data universe. As the number of possible locations and time intervals increases, the amount of injected noise grows sharply. While the mechanism may perform reasonably in highly constrained settings (e.g., the largest domain considered by the authors was a 4×4 grid of possible locations), its utility deteriorates rapidly as the spatial and temporal domain expands—for instance, in city-level analyses with richer spatio-temporal resolution.

Moreover, due to sequential composition, the privacy loss accumulates with trajectory length. The authors' utility analysis confirms that the error increases with longer trajectories, with the maximum evaluated length being $\ell_{\max} = 8$ points.

Overall, this approach may be suitable for societal contact-tracing applications characterized by a limited number of POIs and relatively short trajectories—such as those derived from credit card transactions, where location information is only recorded when a payment occurs and thus may be generated only a few times per day. However, in other use cases—such as traffic management, driving behavior analysis, or large-scale traffic flow monitoring—each user may generate thousands of spatio-temporal data points across extensive geographic areas. In such settings, the noise required to preserve privacy would lead to a substantial degradation in utility, rendering the approach impractical.

3.4.5. A Note on Synthetic Trajectory Generation

A common approach for privacy-preserving data analysis, which lies outside the scope of this chapter, is synthetic database generation. Rather than applying masking techniques—i.e., releasing a modified version of the original dataset [17]—synthetic data methods generate entirely new (artificial) datasets designed to preserve the statistical properties of the original data.

Although synthetic data does not establish a one-to-one correspondence with individual users, it can still compromise the privacy of those whose records were used for training. The generator learns from—and attempts to replicate—the underlying structure of the training dataset. As a result, adversaries can exploit machine learning phenomena such as *overfitting* [54] and conduct MIAs or DRAs [74], [127] to successfully recover sensitive information, thereby threatening the privacy of users whose records were included in the training dataset. To mitigate these risks, several works propose enforcing DP during the training phase of the generative model.

Building up on our work, Buchholz et al. [79] subsequent research has further examined DP synthetic trajectory generation, including approaches proposed after 2023. Using our impossibility results, they show that many of these proposals either do not claim any formal guarantee or fail in the DP proof.

3.5. Conclusions and Problem Statement

Our global conclusion is that DP fails as practical and applicable tool in trajectory data, with most of proposals being factually incorrect and the foundations of DP not trivially generalizable to more complex data structures. Some limitations of DP were already present in simple data structures such as the lack of interpretation but they considerably intensify when adding new granularities. Others appear as a consequence of the extension of DP to complex data, namely the new granularities and the correlation threat.

3.5.1. New Granularities, New Challenges

The emergence of new granularity notions that adapt original DP definition to complex data domains is a necessary step. To adapt DP to time-series, such as trajectories, many new granularities have been defined as surveyed in Section 3.3. This paradigm extends to general complex data structures with the increasing number of proposed DP granularities [18]. However, they come at the cost of new challenges in interpretability and extension of the known DP properties such as composability.

Understanding the actual privacy protection offered by a given choice—and whether it is appropriate for a specific application—has become increasingly challenging. While finer granularities are often necessary to capture the structure of complex data, they exacerbate an interpretability issue that already exists in DP. Moreover, recent efforts to improve the interpretability of DP, such as Gaussian differential privacy, which allows privacy parameters to be directly interpreted in terms of attack mitigation, have so far been confined to the original bounded DP setting. These advances have not yet been extended to alternative granularities and therefore remain inapplicable to structured domains such as graph or trajectory data.

Furthermore, important properties of DP such as composition do not trivially extend to new granularity definitions (cf. Section 2.2 and Example 2.16). Composability enables the combination of multiple DP mechanisms while tracking cumulative privacy loss. Complex data publishing in a DP-compliant manner requires multiple mechanisms, making composability essential. However, each release inevitably leaks some information about individuals in the database, and accurately quantifying cumulative privacy risk remains challenging. However, existing composition theorems (e.g., parallel composition) do not generalize to all composition protocols, data domains, or granularities.

Since both new granularities and composition are central to extending DP to complex data structures, a deeper understanding of composition in these settings is urgently needed. In particular, new theoretical results are required to establish composition guarantees that hold across richer data domains and privacy definitions. We focus on addressing this important problem in Chapter 4.

3.5.2. Flaws in DP Formalization and Implementation

While failures in formal DP proofs are not a limitation of DP itself, but rather of its human application, it would be unrealistic to ignore human error as a practical concern.

This is especially true given that we have repeatedly identified such mistakes in peer-reviewed work (see Section 3.4). As a consequence, mechanisms and analyses that rely exclusively on lengthy or intricate proofs are inherently susceptible to undetected errors.

In this context, the impossibility results presented in Section 3.4 play a crucial role. Beyond their theoretical significance, they offer simple and robust criteria for double-checking correctness and enabling early detection of inconsistencies—without requiring a full inspection of the underlying proofs. While these results provide a practical tool, they do not cover every potential failure.

This underscores the growing need for systematic DP auditing. Rather than assuming correctness by construction, auditing frameworks seek to verify whether claimed privacy guarantees hold in practice. Impossibility results and audit-oriented tools provide complementary approaches to bolster confidence in DP deployments, particularly in complex or novel settings.

We address this point by developing a novel auditing framework in Chapter 5.

3.5.3. Lack of Privacy Parameters Interpretation

In Section 3.2, we surveyed a wide range of threats and attack models discussed in the literature, highlighting numerous cases where public trajectory data releases led to successful privacy breaches. These incidents show that sensitive information—ranging from home addresses to religious affiliations—can be inferred in practice.

DP was introduced as a principled response to such threats, improving on earlier syntactic privacy notions. However, its effectiveness in practice largely hinges on a single abstract parameter: the privacy budget ε. While critical, ε is difficult to interpret. A large ε may allow substantial information leakage, whereas a very small value can render the data practically useless for analysis [56]. In other words, the formal guarantee that DP limits the influence of any single individual on a mechanism's output does not straightforwardly translate into protection against specific real-world attacks.

Efforts to interpret DP, particularly against membership inference attacks (MIAs), have provided some clarity [128], but MIAs represent only one type of threat. In trajectory data, knowing whether someone is in the dataset is often less critical than inferring sensitive attributes. Two mechanisms might perform similarly against MIAs yet differ substantially in protecting attribute privacy. Focusing solely on MIAs can therefore obscure meaningful differences between privacy mechanisms.

We address these open questions about DP interpretation for more general threats, such as AIAs and DRAs, in Chapter 5.

3.5.4. Correlation Threat

A well-known statistical characteristic of trajectory databases is the presence of correlations and dependencies (see Section 3.1). This is not an isolated property from human traces but rather a common factor in all complex data structures: correlations among data records are common in real-world databases, such as those induced by friendships

in social networks [34], genetic similarities among family members [35] or pixels in an image [14].

DP has been design and analyzed for i.i.d. data, hence, the actual privacy implications and adversarial bounds of DP when applied to correlated data remain an open question in the literature. Recent theoretical studies suggest that DP does not provide strong guarantees in such cases [29], [36], [37], [68]. Furthermore, empirical attacks confirm this hypothesis in different applications, breaking DP expected guarantees due to the existence of correlations [14], [33] To address this limitation, Bayesian DP has emerged as a novel approach that explicitly accounts for correlation. However, given that it is already challenging to achieve good utility under standard DP, it remains unclear whether Bayesian DP can be practically viable.

Motivated by the prevalence of correlations in complex data, we explore this question, in Chapter 6.

4. Unlocking the Potential of Composition for Metric Privacy

This chapter is based on the contributions:

- **Patricia Guerra-Balboa**⋆, Àlex Miranda-Pascual⋆, Javier Parra-Arnau, and Thorsten Strufe. "Composition in Differential Privacy for General Granularity Notions". In: IEEE Computer Security Foundations Symposium (CSF), 2024, DOI: 10.1109/CSF61375.2024.00004

The composition properties of DP are essentially governed by a binary notion of domain interaction: mechanisms compose sequentially when they act on overlapping data, and in parallel when they act on disjoint subsets, leading to additive (for sequential) or max-type (for parallel) bounds on the privacy loss (see Section 2.2). Both theorems were originally stated for tabular databases in the *unbounded* [29] scenario. Nowadays, the literature works with different *database domains*, such as graphs, images, text, or human traces [18], and with different *neighborhood definitions* [16], such as *bounded DP*, *edge-DP*, and *w-event DP* [18], or, more generally, metric privacy [47]. Even within a single data domain, there is often a variety of granularities, as we analyze in Section 3.3 for the case of trajectory data.

Nevertheless, existing composition theorems do not necessarily extend to these settings. For instance, Li et al. [27] show that the classical parallel composition theorem [19] fails when unbounded DP is replaced by bounded DP. Intuitively, partitioning the data into disjoint subsets does not preserve the bounded neighborhood relation, hence, the mechanisms do not protect them under bounded DP. Consequently, parallel composition does not hold for general granularity notions without additional assumptions.

This naturally raises several open questions. For instance, can the parallel composition bound be obtained for granularities beyond the unbounded case, thereby enabling improved utility in complex data domains with heterogeneous granularities? If so, under what structural conditions does such a bound hold?

Motivated by this gap, the goal of this chapter is to develop a unified setting in which arbitrary granularities can be composed, and where the resulting privacy loss can be systematically characterized and directly compared to the original guarantee.

To address this challenge, rather than analyzing composition separately for each specific granularity notion, we study composition within the framework of metric privacy [47]—a general mathematical formalism that captures arbitrary notions of granularity (see Sec-

tion 2.1). Working in the metric privacy setting provides a unifying perspective, while also simplifying notation and proofs. Moreover, leveraging the correspondence between metrics and granularities established in Section 2.1, any composition bound for metric privacy immediately yields composition bounds for any domain and granularity notion, whether existing or yet to be defined. The framework even enables the composition of mechanisms defined over different domains and under different granularity notions. In this way, we advance the understanding of how granularity choices influence composition in DP, viewed through the lens of metric privacy.

We explore this framework to investigate whether composition can be characterized directly in terms of the underlying metric and its interaction with intermediate preprocessing functions. Rather than asking whether neighboring datasets intersect—as in original DP composition, we try to capture how preprocessing functions distort distances in the metric space and how these distortions propagate through successive mechanisms.

Ideally a composition principle should reflect the full structural richness of metric privacy, allowing amplification or attenuation effects induced by preprocessing to be explicitly accounted for. Therefore, our goal is to provide general composition results where privacy degradation is no longer determined solely by the combinatorial structure of data access (sequential vs. parallel), but by the way in which the metric is transformed along the computation pipeline. Hence, facilitating a more accurate calculation of the privacy loss upon any possible composition of metric private mechanisms and showcase the effect that preprocessing has on the computation.

Moreover, we revisit the motivating question of this work: can parallel composition be extended to arbitrary metrics, and thus to arbitrary granularities? In particular, we aim to identify sufficient conditions on both the underlying metric and the partitioning (i.e., preprocessing) function under which the parallel composition bound can be achieved. Previously, beyond the unbounded setting, one typically had to rely on sequential composition, often incurring substantial utility loss due to increased noise. Therefore, it is crucial to determine whether, even for more general metrics and granularities, the conditions for parallel composition can be verified, thereby significantly reducing the required noise injection.

Finally, we aim to enable the composition of metric-private mechanisms to be interpreted directly in terms of attacks. To this end, we analyze how to define a metric generalization of Gaussian DP (GDP) that extends the benefits of GDP composition and its interpretability to arbitrary domains (cf. Section 2.3.1.1). We then extend our analysis of composability to this new notion, investigating whether tighter bounds can be established for a metric version of GDP.

More specifically, this chapter aims to answer the following research questions:

- Can sequential and parallel composition be unified within a general metric privacy framework that extends to arbitrary metrics and allows computation of tighter privacy loss for any existing granularity?

- Is it possible to extend the benefits of parallel composition to metric privacy and general granularities?

- How can we define and analyze a metric generalization of μ-GDP, that extends the benefits of GDP composition and interpretability to every domain?

Related work. Li et al. [27] analyze the composition theorems in unbounded and bounded DP, and find out that the parallel composition theorem does not necessarily hold for bounded DP mechanisms. However, they do not explore other granularities of the state of the art or attempt to provide a solution for the bounded problem. McSherry [19] gives the first distance-based formulation of DP, later generalized by Chatzikokolakis et al. [47] with the definition of $d_{\mathbb{D}}$-privacy, which we use to set the general framework for composition. However, only sequential composition has been explored for $d_{\mathbb{D}}$-privacy [129]. Therefore, the generalization of other composition settings, such as parallel, to other granularities (metrics) is still an open question, and to the best of our knowledge, there is no work in the literature, either for DP or for d-privacy, that, in a general manner, computes an accurate privacy loss bound when we have other metrics, domains and composition rules.

4.1. Composition in Metric Privacy

In this section, we present our general composition results for metric privacy. We adopt the adaptive composition framework introduced in Section 2.2. Note that adaptive-composed mechanisms are more general than independent-composed mechanisms, corresponding to the case where $\mathcal{M}_i$ are mutually independent and, in particular, constant over previous outputs. Consequently, all the results presented in this chapter include the independent composition as a particular case.

To capture both sequential and parallel composition within a unified setting, we move beyond the standard distinction based on whether mechanisms access the entire dataset or disjoint subsets. Instead, we introduce a more general model based on preprocessing functions, which abstract how data are partitioned or transformed prior to the application of each mechanism. This perspective allows us to treat sequential and parallel composition as special cases of a single, coherent framework.

Formally, we consider $s_i = f_i(D)$ representing the output of a query f_i on database D (with f_i possibly being the identity) and we compose k d_i-private mechanisms $\mathcal{M}_i^* \colon \overline{\Theta}_i \times \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ such that $\theta_i \sim \mathcal{M}_i^*(f_i(D), \theta_1, \ldots, \theta_{i-1})$. Note that $\mathcal{M}_i \coloneqq \mathcal{M}_i^* \circ f_i$ defines a mechanism over $\mathbb{D}$ for all $i \in [k]$. We provide a visualization of this framework in Figure 4.1. Therefore, the question arises whether the composition of the mechanisms $\mathcal{M}$ such that

$$\mathcal{M}(D) = (\mathcal{M}_1^*(f_1(D)), \ldots, \mathcal{M}_k^*(f_k(D), \theta_1, \ldots, \theta_{k-1}))$$

for all $D \in \mathbb{D}$ is $d_{\mathbb{D}}$-private, and what privacy $d_{\mathbb{D}}$ implies.

To the best of our knowledge, there is no general result stating the bound of such general composition for any metric or granularity that takes the pre-processing functions into account. Note that, generalization to metric privacy or simply to other neighboring definitions may seem a natural step but is not trivial at all. As we saw in Example 2.16, when f defines a partition, parallel composition result (only proved for unbounded DP)

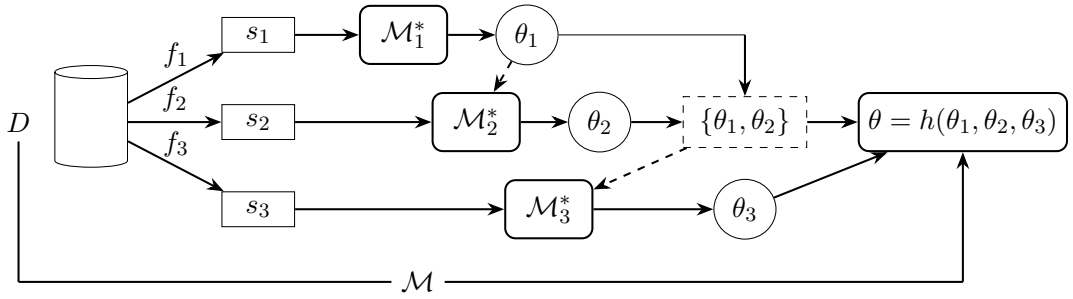


Figure 4.1.: Generalized Adaptive Composition Scheme. The independent counterpart becomes the particular case without dashed lines.

completely breaks for other granularities. To answer this question, we state and prove the general composition Theorem 4.1.

Theorem 4.1 (AC theorem)**.** *Let $\mathbb{D}$ be a database class, and, for all $i \in [k]$, let $(\mathbb{D}_i, d_i)$ be a privacy space, $f_i \colon \mathbb{D} \to \mathbb{D}_i$ a deterministic map and $f_i^* = \mathrm{id}_{\overline{\Theta}_i} \times f_i$ (with $f_1^* = f_1$).*

For $i \in [k]$, let $\mathcal{M}_i^* \colon \overline{\Theta}_i \times \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_i^*(\overline{\theta}_i, \cdot) \colon \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ satisfies d_i-privacy for any $\overline{\theta}_i \in \overline{\Theta}_i$.

Then mechanism $\mathcal{M} = (\mathcal{M}_1^* \circ f_1^*, \dots, \mathcal{M}_k^* \circ f_k^*)$ is $d_{\mathbb{D}}$-private with

$$d_{\mathbb{D}}(D, D') \coloneqq \sum_{i=1}^{k} d_i(f_i(D), f_i(D')) \quad \textit{for all } D, D' \in \mathbb{D}.$$

Proof. Note that $d_{\mathbb{D}}$ is a well-defined metric since it is the sum of metrics.

We prove the statement by induction over k. The result is trivial for $k = 1$, and we consider the case $k = 2$.

Denote $\Theta \coloneqq \Theta_1 \times \Theta_2$ and fix $D, D' \in \mathbb{D}$. For $i \in [2]$, denote $D_i = f_i(D)$, $D_i' = f_i(D')$, and $d_i \coloneqq d_i(f_i(D), f_i(D'))$ to simplify the notation.

Note that for every $\mathcal{M}(D)$ and measurable set $S \subseteq \Theta$, $P_{\mathcal{M}(D)}(S) = \Pr[\mathcal{M}(D) \in S]$ defines a measure. This can also be defined with an integral, i.e.,

$$\Pr[\mathcal{M}(D) \in S] = \int_S \mathrm{d}P_{\mathcal{M}(D)},$$

known as the *Lebesgue–Stieltjes integral*. We will use the Lebesgue–Stieltjes integral because it allows us to generalize our results to any random element, such as discrete, continuous, and mixed random variables or random vectors.

By the law of total probability (see Remark A.9), for any measurable $S \subseteq \Theta$, we have

$$\begin{aligned}
\Pr[\mathcal{M}(D) \in S] &= \Pr[(\mathcal{M}_1^*(D_1), \mathcal{M}_2^*(\mathcal{M}_1^*(D_1), D_2)) \in S] \\
&= \int_{\Theta_1} \Pr[(\mathcal{M}_1^*(D_1), \mathcal{M}_2^*(\mathcal{M}_1^*(D_1), D_2)) \in S \mid \mathcal{M}_1^*(D_1) = s_1] \, \mathrm{d}P_{\mathcal{M}_1^*(D_1)}(s_1) \\
&= \int_{\Theta_1} \Pr[(s_1, \mathcal{M}_2^*(s_1, D_2)) \in S] \, \mathrm{d}P_{\mathcal{M}_1^*(D_1)}(s_1)
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{(i)}}{\leq} \mathrm{e}^{d_1} \int_{\Theta_1} \Pr[(s_1, \mathcal{M}_2^*(s_1, D_2)) \in S] \, \mathrm{d}P_{\mathcal{M}_1^*(D_1')}(s_1) \\
&\overset{\text{(ii)}}{\leq} \mathrm{e}^{d_1} \int_{\Theta_1} \mathrm{e}^{d_2} \Pr[(s_1, \mathcal{M}_2^*(s_1, D_2')) \in S] \, \mathrm{d}P_{\mathcal{M}_1^*(D_1')}(s_1) \\
&= \mathrm{e}^{d_1+d_2} \Pr[\mathcal{M}(D') \in S],
\end{aligned}$$

where

(i) uses Lemma A.10,

(ii) uses the fact that $\mathcal{M}_2^*$ is d_2-private.

Taking $d_{\mathbb{D}} = d_1 + d_2$ proves the case $k = 2$. Now suppose the statement is true for $k-1$ fixed and we prove it for k. Consider the mechanism $\mathcal{M}' = (\mathcal{M}_1^*, \ldots, \mathcal{M}_{k-1}^*)$ with domain $\mathbb{D}$. By the induction hypothesis, $\mathcal{M}'$ is $d'_{\mathbb{D}}$-private with

$$d'_{\mathbb{D}}(D, D') = \sum_{i=1}^{k-1} d_i(f_i(D), f_i(D'))$$

for all $D, D' \in \mathbb{D}$. Then, we have that

$$\mathcal{M}(D) = (\mathcal{M}'(D), \mathcal{M}_k^*(\mathcal{M}'(D), f(D)_k))$$

for all $D \in \mathbb{D}$. We can easily check that we are in the conditions of the case $k = 2$ by taking $\mathcal{M}'$ as $\mathcal{M}_1^*$ and $\mathcal{M}_k$ as $\mathcal{M}_2^*$. Therefore, we obtain that $\mathcal{M}$ is $d_{\mathbb{D}}$-private with

$$d_{\mathbb{D}}(D, D') = \sum_{i=1}^{k} d_i(f_i(D), f_i(D'))$$

for all $D, D' \in \mathbb{D}$. □

Since this theorem does not impose any condition on the privacy metric of the initial $\mathcal{M}_i$, our result can be used for any privacy space and any possible composition strategy.

Importantly, although the theorem is stated in the language of metric privacy, it applies directly to granularity-based notions. Indeed, by Proposition 2.6, every granularity notion $\mathcal{G}$ induces a canonical metric $d^{\mathcal{G}}$, defined as the shortest-path metric over neighboring changes. Furthermore, Proposition A.7 shows that any two granularities $\mathcal{G}_1, \mathcal{G}_2$ can be related by bounding

$$\max_{D \sim_{\mathcal{G}_2} D'} d^{\mathcal{G}_1}(D, D').$$

Together, these results allow our metric-based composition theorem to be instantiated for arbitrary granularity notions and to systematically compare their corresponding privacy guarantees. Formally, when we apply this theorem to $\mathcal{M}_i$ that satisfy $\mathcal{G}_i$ ε_i-DP mechanism we obtain that the composed mechanism $\mathcal{M}$ is $\mathcal{G}$ ε-DP with

$$\varepsilon = \max_{D \sim_{\mathcal{G}} D'} \sum_{i=1}^{k} \varepsilon_i d_{\mathbb{D}_i}^{\mathcal{G}_i}(f_i(D), f_i(D')) \leq \max_{D \sim_{\mathcal{G}} D'} \sum_{i \,:\, f_i(D) \neq f_i(D')} r_i \varepsilon_i,$$

where $r_i \coloneqq \max_{D \sim_{\mathcal{G}} D'} d_{\mathbb{D}}^{\mathcal{G}_i}(f_i(D), f_i(D'))$ for any well-defined granularity $\mathcal{G}$ in the domain $\mathbb{D}$. Hence, obtaining a tighter privacy loss estimate than Theorem 2.14.

The significance of the AC Theorem 4.1 lies in its ability to subsume all possible composition mechanisms within a single, unified framework. Without assuming any particular metric or pre-processing strategy, the AC Theorem determines the privacy level of the resulting mechanism by construction. In cases where valid composition cannot be achieved (as in Example 2.16), the framework correctly yields degenerate values ($d_{\mathbb{D}}(D, D') = \infty$) detecting the absence of privacy. Nevertheless, in the general case—when the combination of pre-processing with d satisfies natural privacy properties—the framework provides meaningful and interpretable results. Consequently, this theorem serves as a single, unifying result that enables systematic reasoning about both standard and non-standard composition scenarios. This is precisely the purpose of the present section: to use the AC theorem as a foundation to derive insightful results about composition, moving from the most straightforward case of sequential composition to the more challenging extension of parallel composition.

Particularly, if we impose $f_i = \mathrm{id}$ and $\mathbb{D} = \mathbb{D}_i$ for all $i \in [k]$, we obtain a generalization of the sequential setting:

Corollary 4.2 (Generalized Sequential Compostion)**.** *Let $\{(\mathbb{D}, d_i)\}_{i\in[k]}$ be a set of privacy spaces. For $i \in [k]$, let $\mathcal{M}_i \colon \overline{\Theta}_i \times \mathbb{D} \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_k(\overline{\theta}_i, \cdot) \colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ is d_i-private for all $\overline{\theta}_i \in \overline{\Theta}_i$. Then $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ is $(\sum_{i=1}^k d_i)$-private.*

Proof. Direct from Theorem 4.1 by taking $\mathbb{D}_i = \mathbb{D}$ and $f_i = \mathrm{id}$. □

Note that by choosing $d_i = \varepsilon_i d$, we obtain that $\mathcal{M}$ is εd-private with $\varepsilon = \sum_{i=1}^k \varepsilon_i$ (first proven in [129]). Furthermore, by selecting d as $d_{\mathbb{D}}^{\mathcal{G}}$, we obtain the sequential composition theorem for every granularity: If $\mathcal{M}_i \colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ are mutually independent $\mathcal{G}$ ε_i-DP mechanisms, then $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ is $\mathcal{G}$ $(\sum_{i=1}^k \varepsilon_i)$-DP. This shows that sequential composition works as expected for every granularity.

Beyond recovering standard sequential composition, our Theorem 4.1 enables the analysis of substantially richer scenarios, including heterogeneous privacy requirements and combinations of queries that yield strictly tighter guarantees than those provided by classical sequential composition. We illustrate these scenarios through a series of examples and conclude by revisiting the parallel composition case.

The following example illustrates a setting in which incorporating the preprocessing function yields a strictly tighter privacy bound than that obtained via sequential composition alone.

Example 4.3 (Distributed Network Analysis)**.** We consider a social network $G = (V, E)$ in which each vertex x_i represents an individual and edges, $e = (x_i, x_j)$, represent connections between individuals. Since these connections are considered private, we aim to protect them.

Formally, let $\mathbb{D}$ denote the set of simple undirected graphs on a fixed vertex set $V = \{x_1, \ldots, x_n\}$, endowed with the edge-level DP, $\sim_{\mathcal{E}}$, where two graphs are neighboring

if they differ by the addition or removal of a single edge. Let $d_{\mathbb{D}}^{\mathcal{E}}$ denote the corresponding canonical metric, defined as the minimum number of edge additions or removals required to transform one graph into another (see Section 2.1).

Our network analysis task consists of computing the maximum degree of the graph. This is a classical graph statistic [130], as it captures the extent of the highest connectivity in the network and the heterogeneity of the degree distribution.

Since there is not any trusted entity, the computation of the maximum degree is a distributed task, in which each node has access to their connections, and therefore their degree, but not to the others information. For each $i \in [n]$, define the function

$$f_i \colon \mathbb{D} \to \mathbb{N}, \qquad f_i(G) \coloneqq \deg_G(x_i).$$

The maximum degree can be computed as $h(f_1(G), \dots, f_n(G)) = \max_{i\in[n]} f_i(G)$. To provide privacy we consider, $\mathcal{M}^* \colon \mathbb{N} \to \mathcal{D}(\mathbb{R})$, the Laplace mechanism

$$\mathcal{M}^*(x) = x + Z, \qquad Z_i \sim \mathrm{Lap}\left(\tfrac{1}{\varepsilon}\right),$$

so that $\mathcal{M}^*$ is $(\varepsilon\ell_1)$-private over $\mathbb{N}$ (see Definition 2.9).

Note that since removing/adding one node affects the degree at most by one unit we have that

$$|f_i(G) - f_i(G')| \leq d_{\mathcal{E}}(G, G').$$

Consider the composed mechanism $\mathcal{M} = (\mathcal{M}_i)_{i\in[n]}$, together with the post-processing that outputs the maximum of the reported values. Since $\Delta f_i = 1$ for all $i \in [n]$, according to Proposition A.3, $\mathcal{M}_i$ are event-level ε-DP and applying sequential composition, one obtains that $\mathcal{M}$ is $(n\varepsilon)$-edge DP, or analogously using Corollary 4.2, $\mathcal{M}$ is $nd_{\mathcal{E}}$-private. However, Theorem 4.1 yields a tighter bound:

The modification of a single edge affects the degree of exactly its two incident vertices. Therefore, $\mathcal{M}$ is d-private in $\mathbb{D}$ with

$$d(G, G') = \sum_{i=1}^{n} |f_i(G) - f_i(G')| = 2d_{\mathcal{E}}(G, G') << nd_{\mathcal{E}}(G, G').$$

In particular, using Proposition 2.6 the composed mechanism $\mathcal{M}$ is 2ε-edge DP. Hence, we obtain a substantially tighter bound compared to the sequential composition.

More generally, rather than bounding privacy loss via a simplistic linear decay, the AC theorem (Th. 4.1) provides a composition principle that captures the full structural richness of metric privacy, explicitly accounting for amplification or attenuation effects induced by preprocessing. We illustrate this behavior in the following example:

Example 4.4 (k-Clique Counting)**.** Consider an undirected simple social network $G = (V, E)$ with $|V| = n$ nodes and the edges, $e \in E$, represent social relationships. We are interested in releasing statistics of k-cliques in the network, i.e., complete subgraphs with k nodes. To this end, one may apply the composition protocol:

$$\mathcal{M}(G) = (\mathcal{M}^*(f_3(G)), \dots, \mathcal{M}^*(f_k(G))),$$

where each $\mathcal{M}^*$ adds Laplace noise with scale $b = \frac{1}{\varepsilon}$ to the count of i-cliques, $f_i(G)$.

Under the additional assumption that the maximum clique containing each edge is unique, using Theorem 4.1, we obtain that $\mathcal{M}$ is a d-private protocol with

$$d(G, G') = \sum_{i=3}^{k} \varepsilon |f_i(G) - f_i(G')| \leq \varepsilon \sum_{e \in E \triangle E'} \sum_{i=3}^{k} \binom{r(e) - 2}{i - 2} = \varepsilon \sum_{e \in E \triangle E'} (2^{r(e)-2} - 1),$$

where $r(e)$ denotes the size of the largest clique containing edge e. Hence, there is a clear propagation of the loss through the composition. For instance, removing $r(e) = 3$ composes in parallel (i.e., the privacy loss does not increase with respect to the initial ε), while higher r values propagate the loss through the whole composition process. Hence, this protocol protects connections according to the group size to which they belong. For example, if G and G' differ by a single edge that belongs to a clique of size $r(e) = 2$ (i.e., an isolated edge), then $d(G, G') = 0$, and the relationship between the nodes is completely indistinguishable. In contrast, if the edge belongs to a clique of size $r(e) = 5$, then $d(G, G') = \varepsilon\, 7$.

This metric formalizes the intuition that edges in small groups are far more sensitive than edges in large cliques which is supported by empirical studies on social network privacy: In large cliques, the presence of an edge between two individuals is often predictable from shared social roles, such as workplace, class, or community membership. Conversely, in very small cliques, the edge itself largely defines the context, rendering its disclosure highly informative [131], [132].

Finally, the setting in which the mechanisms take as input disjoint subsets of the initial database (parallel composition) does not generally yield analogous results to Theorem 2.15 as we show in Example 2.16. That means, unlike sequential, parallel composition does not trivially generalize to arbitrary metric spaces.

To address this issue, we model the parallel composition setting as a particular case of Theorem 4.1, in which the pre-processing functions induce a partition of the data domain. By applying Theorem 4.1, we derive the necessary conditions relating the metric and the partitioning function that ensure the preservation of parallel composition bounds, thereby addressing this gap in the literature.

4.2. Parallel Composition in Metric Privacy

In this section, we analyze the conditions under which composition results similar to the classical parallel theorem can be obtained in metric spaces. To analyze parallel composition as a special case within our general composition framework, we simply represent each mechanism as operating on a disjoint subdataset, with the preprocessing functions encoding the underlying partition.

Formally, we define a *k-partitioning function* $p = \{p_1, \ldots, p_k\}$ as a function where $p_i \colon \mathbb{D} \to p_i(\mathbb{D}) =: \mathbb{D}_i$ such that $p_i(D) \subseteq D$ with $p_i(D) \cap p_j(D) \neq \varnothing$ for $i \neq j$[1]. Note that

[1]We do not require that $D = \bigcup_{i=1}^{k} p_i(D)$, i.e., our partition can be *non-exhaustive*.

the domains $\mathbb{D}_i$ of $\mathcal{M}_i$ might be different in this setting by construction. Let us see an example of a partitioning function, based on that of [27].

Example 4.5 (Partitioning function for $\mathbb{D} \subseteq \mathbb{D}_{\mathcal{X}}$)**.** Let $\mathbb{D} \subseteq \mathbb{D}_{\mathcal{X}}$. A partition $\{\mathcal{X}_i\}_{i\in[k]}$ of $\mathcal{X}$, extends naturally as a partition of the elements $D \in \mathbb{D}$, i.e., $p_i(D) \subseteq D$ is the multiset such that element $x \in D$ has multiplicity $m_{p_i(D)}(x) = \mathbf{1}_{\mathcal{X}_i}(x)\, m_D(x)$. For instance, given a geographical area and a grid that divides it into k regions, a database of locations can be automatically partitioned into subdatasets corresponding to each region. In this case, the partitioning function p uses only x to compute the value of $p(x)$, and therefore the result is independent of the other records.

Partitioning preprocessing functions model scenarios where analyses require computing aggregates over different subpopulations [19]. For example, in a medical study, one might be interested in the average blood pressure across different age groups, such as younger and older participants. Since each individual belongs to exactly one age group, this naturally induces a partition of the population. Although such operations can be analyzed using sequential composition (Corollary 4.2), in which the privacy guarantee scales with the number of subpopulations considered, our general theorem enables more detailed analyses while incurring only a modest privacy cost. Particularly, in this setting, Theorem 4.1 yields that $\mathcal{M}$ is $d_{\mathbb{D}}$-private with

$$d_{\mathbb{D}}(D, D') = \sum_{i=1}^{k} d_i(p_i(D), p_i(D')) \leq I_p(D, D')(\max_{i\in[k]} \Delta p_i\, d_i(D, D')), \tag{4.1}$$

for all $D, D' \in \mathbb{D}$, where $I_p(D, D') := \#\{i \mid p_i(D) \neq p_i(D')\}$.

This fact is coherent with what we know: Assuming a partitioning function of Example 4.5, if we select unbounded DP mechanisms , i.e., $\varepsilon_i d^{\mathcal{U}}_{\mathbb{D}_i}$, then $d_{\mathbb{D}} \leq (\max_{i\in[k]} \varepsilon_i) d^{\mathcal{U}}_{\mathbb{D}}$, since $\Delta p_i = \varepsilon_i$ and $I_p(D, D') = 1$ for all $D \sim_{\mathcal{U}} D'$.

If we select $d_i = \varepsilon_i d^{\mathcal{B}}_{\mathbb{D}_i}$, there may exist $D, D' \in \mathbb{D}$, as we saw in Example 2.16, such that $d_i(D, D') = d^{\mathcal{B}}_{\mathbb{D}_i}(p_i(D), p_i(D')) = \infty$ for some i and therefore $d_{\mathbb{D}}(D, D') = \infty$. In general, we have no better expression for $d_{\mathbb{D}}$ unless we add extra conditions. To address this point, we will explore conditions to achieve the best bound in the following.

The first case we consider is a metric-type d^* that is well-defined over $\mathbb{D}$ and $\mathbb{D}_i$ for all $i \in [k]$[2]. We can give a sufficient condition for obtaining the best bound: We say that metric d^* *commutes with the partition given by* p if, for all $D, D' \in \mathbb{D}$,

$$\sum_{i=1}^{k} d^*_{\mathbb{D}_i}(p_i(D), p_i(D')) = d^*_{\mathbb{D}}\Big(\bigcup_{i=1}^{k} p_i(D), \bigcup_{i=1}^{k} p_i(D')\Big) \leq d^*_{\mathbb{D}}(D, D'). \tag{4.2}$$

By Theorem 4.1, if d^* commutes with p and $\mathcal{M}_i$ are $\varepsilon_i d^*_{\mathbb{D}_i}$-private, then $\mathcal{M}$ is $(\max_{i\in[k]} \varepsilon_i) d^*_{\mathbb{D}}$-private. For example, we prove in Proposition A.11 that $d^{\triangle}$ commutes with all partitions p of Example 4.5, which relates to the original result of McSherry [19].

[2]This means that metrics $d^*_{\mathbb{D}}$ and $d^*_{\mathbb{D}_i}$ are well-defined metrics and that $d^*(D, D')$ is constant for all domains containing $D, D' \in \mathbb{D}$. Examples include $d^{\triangle}$, which is well-defined for all $\mathbb{D} \subseteq \mathbb{D}_{\mathcal{X}}$.

However, verifying commutativity is not always feasible. In particular, for a fixed granularity notion $\mathcal{G}$, different domains $\mathbb{D}_i$ induce different canonical metrics associated with $\mathcal{G}$ (see Section A.2). As a result, there may be no single metric d^* well defined over all $\mathbb{D}_i$, and checking commutativity of a canonical metric is therefore not possible. In this setting, the corresponding condition translates into

$$\sum_{i=1}^{k} d^{\mathcal{G}}_{\mathbb{D}_i}(p_i(D), p_i(D')) \;=\; d^{\mathcal{G}}_{\mathbb{D}}(D, D').$$

Although this equality can be difficult to verify in general, it holds under suitable structural conditions on the partition. In particular, it is satisfied if the partition verifies:

- $d^{\mathcal{G}}_{\mathbb{D}}$*-compatibility*: For all $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$, there exists at most one $j \in [k]$ such that $p_i(D) = p_i(D')$ for all $i \neq j$, i.e., $I_p(D, D') = 1$ for all $D \sim_{\mathcal{G}} D'$; and
- $\mathcal{G}$ is also well-defined over $\mathbb{D}_i$ and the sensitivity of p_i with respect to $d^{\mathcal{G}}_{\mathbb{D}}$ and $d^{\mathcal{G}}_{\mathbb{D}_i}$ is $\Delta p_i \leq 1$ (i.e., $d^{\mathcal{G}}_{\mathbb{D}_i}(p_i(D), p_i(D')) \leq 1$ if $d^{\mathcal{G}}_{\mathbb{D}}(D, D') = 1$).

Under these conditions, we obtain the desired result (where $\mathcal{M}^*_i$ can have different domains) as we prove in the following result:

Theorem 4.6 (Best bound for disjoint inputs)**.** *Let $\mathbb{D}$ be a database class and $\mathcal{G}$ a granularity over $\mathbb{D}$. Let p be a $d^{\mathcal{G}}_{\mathbb{D}}$-compatible k-partitioning function such that $\Delta p_i \leq 1$, and $p^*_i = \mathrm{id}_{\overline{\Theta}_i} \times p_i$ (with $p^*_1 = p_1$). For $i \in [k]$, let $\mathcal{M}^*_i \colon \overline{\Theta}_i \times \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}^*_i(\overline{\theta}_i, \cdot) \colon \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ satisfies $\varepsilon_i d^{\mathcal{G}}_{\mathbb{D}_i}$-privacy for any $\overline{\theta}_i \in \overline{\Theta}_i$. Then mechanism $\mathcal{M} = (\mathcal{M}^*_1 \circ p^*_1, \ldots, \mathcal{M}^*_k \circ p^*_k)$ is $\varepsilon d^{\mathcal{G}}_{\mathbb{D}}$-private with $\varepsilon = \max_{i \in [k]} \varepsilon_i$.*

Proof. From Proposition 2.6, it is equivalent to see that $\mathcal{M}$ is $\mathcal{G}$ ε-DP with $\varepsilon = \max_{i \in [k]} \varepsilon_i$, i.e., that for all $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$ and measurable $S \subseteq \Theta$,

$$\Pr[\mathcal{M}(D) \in S] \leq \mathrm{e}^{\varepsilon} \Pr[\mathcal{M}(D') \in S].$$

Applying the AC theorem (4.1), we obtain that $\mathcal{M}$ is d-private with

$$d(D, D') = \sum_{i=1}^{k} \varepsilon_i d^{\mathcal{G}}_{\mathbb{D}_i}(p_i(D), p_i(D')).$$

Now suppose that $D, D' \in \mathbb{D}$ are $\mathcal{G}$-neighboring. By definition of $d^{\mathcal{G}}_{\mathbb{D}}$-compatibility, there exist $j \in [k]$ such that $p_i(D) = p_i(D')$ for all $i \neq j$. Consequently, for all $i \neq j$, $d^{\mathcal{G}}_{\mathbb{D}_i}(p_i(D), p_i(D')) = 0$. Moreover, by preprocessing (Proposition A.3), we have that $d^{\mathcal{G}}_{\mathbb{D}_j}(p_j(D), p_j(D')) \leq \Delta p_j d^{\mathcal{G}}_{\mathbb{D}}(D, D') \leq 1$ since $D \sim_{\mathcal{G}} D'$ and $\Delta p_j \leq 1$. Therefore,

$$d(D, D') = \sum_{i=1}^{k} \varepsilon_i d^{\mathcal{G}}_{\mathbb{D}_i}(p_i(D), p_i(D')) = \varepsilon_j d^{\mathcal{G}}_{\mathbb{D}_j}(p_j(D), p_j(D')) \leq \varepsilon_j.$$

Consequently, since $\mathcal{M}$ is d-private, for all measurable $S \subseteq \Theta$,

$$\Pr[\mathcal{M}(D) \in S] \leq \mathrm{e}^{\varepsilon_j} \Pr[\mathcal{M}(D') \in S].$$

Since $j \in [k]$ depends on the choice of the $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$, it is sufficient to choose $\varepsilon = \max_{i \in [k]} \varepsilon_i$ to cover all cases. In conclusion, $\mathcal{M}$ is $\mathcal{G}$ ε-DP. □

Even though Theorem 4.6 is stated for any granularity, $d_{\mathbb{D}}^{\mathcal{G}}$-compatibility is a strict condition. For example, no partitioning function of Example 4.5 (with $k > 1$) is $d_{\mathbb{D}_{\mathcal{X}}}^{\mathcal{B}}$-compatible (see Proposition A.12).

Nevertheless, we can construct compatible partitioning functions to certain bounded metrics $d_{\mathbb{D}}^{\mathcal{B}}$, as shown in the following result:

Proposition 4.7 (Compatible order-based partitions for bounded DP)**.** *Consider a database D with ordered elements, i.e., every element $(n, x) \in D$ consists of a record value $x \in \mathcal{X}$ and an unique identifier $n \in [|D|]$. Let $\mathbb{D}_{\mathcal{X}}^{\mathrm{ord}}$ denote class of all such databases.*

Let p be a k-partitioning function of $\mathbb{N}$, which induces a partition of the elements of $\mathbb{D} \subseteq \mathbb{D}_{\mathcal{X}}^{\mathrm{ord}}$ that divides the databases only taking the order into account, i.e., such that $p(n, x) = p(n, y)$ for all $x, y \in \mathcal{X}$. Then p is $d_{\mathbb{D}}^{\mathcal{B}}$-compatible and $\Delta p_i \leq 1$ for all $i \in [k]$.

Proof. Due to the databases being ordered, two databases $D, D' \in \mathbb{D}$ are bounded neighboring if and only if we obtain one from the other by changing the record with identifier $n \in [|D|] = [|D'|]$.

Let $D, D' \in \mathbb{D}$ be bounded neighboring databases. Since $p(n, x) = p(n, y)$ for all $n \in [k]$, there exists $j \in [k]$ such that $p_i(D) = p_i(D')$ for all $i \neq j$. In conclusion, p is $d_{\mathbb{D}}^{\mathcal{B}}$-compatible. Moreover, $p_j(D) \triangle p_j(D') = \{(j, x), (j, y)\}$, so in particular $p_j(D)$ and $p_j(D')$ are also bounded neighboring. Therefore,

$$\Delta p_i \coloneqq \max_{D \sim_{\mathcal{B}} D'} d^{\mathcal{B}}(p_i(D), p_i(D')) \leq 1$$

for all $i \in [k]$, since it holds independently of the choice of D and D'. □

Previous results allow parallel composition to be applied for bounded DP/Hamming distance privacy in distributed systems where the partitioning depends on the user (i.e., the data index) rather than on the data values themselves. However, this setting is rather restrictive and does not fully address the broader applicability of parallel composition. In many practical scenarios, we wish to enforce parallel composition across different queries. In this scenario, the partition naturally depends on the query outcome and not solely on the database index. This limitation motivates our search for a more applicable solution, which we develop in the following section.

Summarizing, we have shown that composition can be treated as a single, unified operation, which enables reasoning about the final privacy guarantees within a common and tighter theorem that eliminates the traditional separation between parallel and sequential composition. Both sequential and parallel composition emerge as special cases of our general result.

In particular, we characterize how different metrics behave with respect to parallel composition and show that, even for a fixed metric, different partitioning strategies may lead to different privacy bounds. Moreover, certain metrics—such as the symmetric difference metric—are more robust under general partitioning schemes, whereas others—such as the Hamming distance—are more sensitive to the chosen partition. This highlights the fundamental role of the metric in determining the effectiveness of parallel composition.

Up to this point, we have assumed that each mechanism $\mathcal{M}_i^*$ is d_i-private and analyzed how the preprocessing function f_i affects their composition. However, even in the standard sequential composition setting—where we simply consider mechanisms $\mathcal{M}_i(D)$ that are d_i-private—it is important to recognize that such mechanisms typically evaluate internal queries whose sensitivity directly influences the overall privacy loss. These internal computations are often abstracted away in composition analyses and therefore not explicitly accounted for.

In the next section, we address this important aspect by incorporating the role of the underlying query into the composition framework. In particular, we study how the sensitivity of the query computed by the privacy mechanism contributes to the composition privacy guarantees.

4.3. Common-Domain Setting

In this section, we study the composition of mechanisms $\mathcal{M}_i$ that access the same dataset, i.e., share the same input domain $\mathbb{D}$. From the previous section—most notably Corollary 4.2—we know that their adaptive composition satisfies d-privacy with $d = \sum_{i\in[k]} d_i$. However, in this section, we prove that this bound can be further improved.

Our key observation is the following: A private mechanism $\mathcal{M}_i \colon \mathbb{D} \to \mathcal{D}(\Theta)$, unless it is intended to release the entire dataset, typically operates by first computing a query that extracts or aggregates the relevant information from D, and then randomizing the result to ensure privacy. Standard composition theorems abstract away this internal structure and treat each mechanism as a black box. We show, however, that explicitly incorporating the structure of the underlying query into the analysis can lead to strictly tighter privacy bounds. We illustrate this idea with the following example.

Example 4.8. Consider the Laplace mechanism for a counting query f, for instance, f_i number of inhabitants of a certain neighborhood i. Formally, $\mathcal{M} \colon \mathcal{X} \to \mathcal{D}(\mathbb{R})$ such that $\mathcal{M}_i(D) = f_i(D) + Z$ where $Z \sim \mathrm{Lap}(\frac{1}{\varepsilon})$ [16]. These mechanism satisfy unbounded ε-DP, because $\Delta f_i = 1$, and their privacy estimation is tight. When we combine both mechanism with sequential composition we obtain 2ε-DP.

Note that we can decompose $\mathcal{M}$ into an internal mechanism $\mathcal{M}^*$ and a deterministic function f. Precisely, the mechanism $\mathcal{M}^* \colon \mathbb{N} \to \mathcal{D}(\mathbb{R})$ that adds Laplace noise to any given number in the domain, i.e., $\mathcal{M}^*(y) = y + Z$ with $Z \sim \mathrm{Lap}(\frac{1}{\varepsilon})$ and f the initial numerical function, verify that $\mathcal{M} = \mathcal{M}^* \circ f$. But in this case, $\mathcal{M}^*$ is not DP in $\mathbb{N}$.

The purpose of Example 4.8 is to illustrate that, while pre-processing functions can be attached to DP mechanisms within a composition process to reduce privacy loss—as

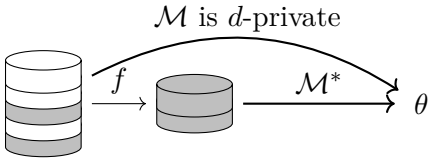


Figure 4.2.: Common Domain Setting

discussed in Theorem 4.1—it is often the case that the DP mechanisms under consideration already admit a natural decomposition into a pre-processing step followed by a subsequent randomized mechanism.

In this scenario, while $\mathcal{M}_i$ provides privacy guarantees for every database $D \in \mathcal{D}$, its computation depends exclusively on the information contained in $f_i(D)$, rather than on the full content of D. This exclusive dependence of $\mathcal{M}_i$ on specific information leads to improved privacy guarantees and tighter privacy-loss bounds under composition. To analyze composition in this setting, we introduce a coherent formalization of what it means for a mechanism to "depend exclusively on $f_i(D)$" under the notion of *dependency*:

Definition 4.9 (Dependency)**.** Let $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$ be a randomized mechanism, and let f be a deterministic map with domain $\mathbb{D}$. We say that $\mathcal{M}$ is *f-dependent* if there exists $\mathcal{M}^*\colon f(\mathbb{D}) \to \mathcal{D}(\Theta)$ such that $\mathcal{M} = \mathcal{M}^* \circ f$.

This definition implies that $\Pr[\mathcal{M}^*(f(D)) \in S] = \Pr[\mathcal{M}(D) \in S]$ for all measurable $S \subseteq \Theta$. Since $\mathcal{M}^*(f(D))$ depends exclusively on $f(D)$, consequently $\mathcal{M}(D)$ depends exclusively on the information in $f(D)$ for all $D \in \mathbb{D}$ (i.e., only data in $f(D)$ affects the output of $\mathcal{M}(D)$). Besides, this concept is well-defined as we prove in Remark A.15

One may ask why it is important to consider the internal structure of each mechanism $\mathcal{M}_i$. At first glance, the dependence of $\mathcal{M}_i$ on its internal query f_i might seem irrelevant, since the mechanism's privacy guarantees are already calibrated to account for this dependence—e.g., via the sensitivity of f_i when using Laplace or Gaussian noise.

However, these internal queries can have a significant impact on the overall composition. Intuitively, as illustrated in Example 4.8, changing a single individual in the dataset may affect only one query (e.g., incrementing the count of inhabitants in a neighborhood) while leaving other queries unaffected. In this case, the sensitivity of each query is tight and cannot be improved. Yet, similar to the effect of preprocessing functions, a single change in the dataset may influence some queries but not others simultaneously. This observation implies that the standard sequential composition bound may be overly pessimistic, and tighter privacy estimates are possible by accounting for the query structure. We formalize this phenomenon in the following theorem.

Theorem 4.10 (AC theorem for common domain)**.** *For $i \in [k]$, let $(\mathbb{D}, d_i)$ be a privacy space, and let f_i be a deterministic map over $\mathbb{D}$. For $i \in [k]$, let $\mathcal{M}_i\colon \overline{\Theta}_i \times \mathbb{D} \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_k(\overline{\theta}_i, \cdot)\colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ satisfies d_i-privacy and f_i-dependency for*

any $\overline{\theta}_i \in \overline{\Theta}_i$*. Then mechanism* $\mathcal{M} = (\mathcal{M}_1, \dots, \mathcal{M}_k)$ *is* $d_{\mathbb{D}}$*-private* with*

$$d_{\mathbb{D}}(D, D') \coloneqq \sum_{i=1}^{k} \min_{\substack{\tilde{D}, \tilde{D}' \in \mathbb{D} \\ f(\tilde{D}) = f(D) \\ f(\tilde{D}') = f(D')}} d_{\mathbb{D}}(\tilde{D}, \tilde{D}') \equiv \sum_{i=1}^{k} d_i^{f_i}(D, D').$$

Proof. Applying Proposition A.16, we obtain that $\mathcal{M}_i$ are $d_i^{f_i}$-private*. Then the result follows from an analogous proof of Theorem 4.1. □

Note that $d_{\mathbb{D}}^{f}$ is not necessarily a metric[3] (thus we call it *d-privacy**). However, it gives an accurate value for the distance between the probability distributions of the output given two input databases. Moreover, this results leads to the composition upper bound

$$d_{\mathbb{D}}(D, D') = \sum_{i=1}^{k} d_i^{f_i}(D, D') \leq \sum_{i \,:\, f_i(D) \neq f_i(D')} d_i(D, D'),$$

which provide better bounds than $\sum_{i=1}^{k} d_i$ given by the AC theorem (4.1). Translating this result to the case of granularities, if we take $\mathcal{M}_i$ to be $\mathcal{G}$ ε_i-DP (i.e., $\varepsilon_i d_{\mathbb{D}}^{\mathcal{G}}$-private), we obtain that $\mathcal{M}$ is $\mathcal{G}$ ε-DP (i.e., $\varepsilon d_{\mathbb{D}}^{\mathcal{G}}$-private) with

$$\varepsilon = \max_{D \sim_{\mathcal{G}} D'} \sum_{i \,:\, f_i(D) \neq f_i(D')} \varepsilon_i.$$

Theorem 4.10 allows us to obtain the corresponding cases, corollaries, and examples to those we obtained from the AC theorem (4.1) for this new setting. In some cases, such as taking $f_i = \mathrm{id}$ for all $i \in [k]$, correspond to the same result (Corollary 4.2), since $d_{\mathbb{D}}^{\mathrm{id}} = d_{\mathbb{D}}$. In others, however, the change of setting leads to a different scenario and results, such as when trying to find the best bound for disjoint inputs (i.e., the counterpart of Section 4.2).

In the common domain, the parallel composition problem (corresponding to Section 4.2) can be stated as follows: Given k mechanisms $\mathcal{M}_i \colon \mathbb{D} \to \mathcal{D}(\Theta)$ that are d_i-private with $d_i = \varepsilon_i d$ for a metric d over $\mathbb{D}$ and p_i-dependent with p an arbitrary partitioning function, we are interested in studying the conditions such that $\mathcal{M} = (\mathcal{M}_1, \dots, \mathcal{M}_k)$ is $d_{\mathbb{D}}$-private with $d_{\mathbb{D}} = (\max_{i \in [k]} \varepsilon_i) d$.

The natural approach is to check when metric d verifies

$$\sum_{i=1}^{k} d^{p_i}(D, D') = d(D, D') \tag{4.3}$$

for all $D, D' \in \mathbb{D}$, since then $d_{\mathbb{D}} = \max_{i \in [k]} \varepsilon_i d$ follows from Theorem 4.10.

Equation (4.3) can be hard to check directly, but we can give sufficient conditions for it when $d = d_{\mathbb{D}}^{\mathcal{G}}$, the canonical distance of a granularity notion. Here, it is sufficient to ask that the partition is $d_{\mathbb{D}}^{\mathcal{G}}$-compatible.

[3] It does not generally fulfill the triangle inequality.

Theorem 4.11 (AC best bound for disjoint inputs (common domain))**.** *Let $\mathbb{D}$ be a database class and $\mathcal{G}$ a granularity over $\mathbb{D}$. Let p be a $d_{\mathbb{D}}^{\mathcal{G}}$-compatible k-partitioning function. For $i \in [k]$, let $\mathcal{M}_i \colon \overline{\Theta}_i \times \mathbb{D} \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_k(\overline{\theta}_i, \cdot) \colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ satisfies $\varepsilon_i d_{\mathbb{D}}^{\mathcal{G}}$-privacy and p_i-dependency for any $\overline{\theta}_i \in \overline{\Theta}_i$. Then mechanism $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ is $\varepsilon d_{\mathbb{D}}^{\mathcal{G}}$-private with $\varepsilon = \max_{i \in [k]} \varepsilon_i$.*

Proof. From Proposition 2.6, it is equivalent to see that $\mathcal{M}$ is $\mathcal{G}$ ε-DP with $\varepsilon = \max_{i \in [k]} \varepsilon_i$, i.e., that for all $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$ and measurable $S \subseteq \Theta$,

$$\Pr[\mathcal{M}(D) \in S] \leq \mathrm{e}^{\varepsilon} \Pr[\mathcal{M}(D') \in S].$$

Applying Theorem 4.10, we obtain that $\mathcal{M}$ is d-private* with

$$d(D, D') = \sum_{i=1}^{k} (\varepsilon_i d_{\mathbb{D}}^{\mathcal{G}})^{p_i}(D, D') = \sum_{i=1}^{k} \varepsilon_i d_{\mathbb{D}}^{\mathcal{G}, p_i}(D, D').$$

Now suppose $D, D' \in \mathbb{D}$ are $\mathcal{G}$-neighboring. By definition of $d_{\mathbb{D}}^{\mathcal{G}}$-compatibility, there exist $j \in [k]$ such that $p_i(D) = p_i(D')$ for all $i \neq j$. Consequently, for all $i \neq j$, $d_{\mathbb{D}}^{\mathcal{G}, p_i}(D, D') \leq d_{\mathbb{D}}^{\mathcal{G}}(D, D) = 0$, since we can select D as both $\tilde{D}$ and $\tilde{D}'$ in the definition (see Proposition A.16). Therefore,

$$d(D, D') = \sum_{i=1}^{k} \varepsilon_i d_{\mathbb{D}}^{\mathcal{G}, p_i}(D, D') = \varepsilon_j d_{\mathbb{D}}^{\mathcal{G}, p_j}(D, D') \leq \varepsilon_j d_{\mathbb{D}}^{\mathcal{G}}(D, D') \leq \varepsilon_j,$$

where the last inequality comes from the fact that D and D' are $\mathcal{G}$-neighboring. Consequently, since $\mathcal{M}$ is d-private*, for all measurable $S \subseteq \Theta$,

$$\Pr[\mathcal{M}(D) \in S] \leq \mathrm{e}^{\varepsilon_j} \Pr[\mathcal{M}(D') \in S].$$

Since $j \in [k]$ depends on the choice of the $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$, it is sufficient to choose $\varepsilon = \max_{i \in [k]} \varepsilon_i$ to cover all cases. In conclusion, $\mathcal{M}$ is $\mathcal{G}$ ε-DP. □

Observe that, in this setting, it is no longer necessary to impose the condition "$\Delta p_i \leq 1$", which was required in our previous theorem (4.6).

Therefore, our analysis shows not only that we can improve upon standard sequential composition—recovering bounds analogous to parallel composition even in the absence of explicit preprocessing—but also that the presence of an embedded query computation can further tighten the privacy guarantees. In particular, the resulting privacy behavior can be strictly better than that predicted by classical composition, and the conditions required to obtain these improved bounds are less restrictive than those previously assumed.

In particular, while arbitrary partitioning functions can lead to severe composition degradation—potentially breaking all guarantees under bounded DP—we show that the situation improves significantly in the common-domain setting, where all mechanisms operate on the same dataset.

We thus provide a solution to the problem posed by Li et al. [27], obtaining a tight bound for composition over disjoint databases in bounded DP (when taking a partition of Example 4.5), which was previously missing.

Corollary 4.12. *Let p be a k-partitioning function of Example 4.5. For all $i \in [k]$, let $\mathcal{M}_i \colon \overline{\Theta}_i \times \mathbb{D} \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_k(\overline{\theta}_i, \cdot) \colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ satisfies bounded ε_i-DP for all $\overline{\theta}_i \in \overline{\Theta}_i$ and p_i-dependent. Then mechanism $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ with domain $\mathbb{D}$ is bounded ε-DP with $\varepsilon = \max_{i,j \in [k];\, i \neq j}(\varepsilon_i + \varepsilon_j)$.*

Proof. From Proposition 2.6, it is equivalent to see that $\mathcal{M}$ is bounded ε-DP with $\varepsilon = \max_{i,j \in [k];\, i \neq j}(\varepsilon_i + \varepsilon_j)$, i.e., that for all bounded-neighboring $D, D' \in \mathbb{D}$ and measurable $S \subseteq \Theta$,

$$\Pr[\mathcal{M}(D) \in S] \leq \mathrm{e}^{\varepsilon} \Pr[\mathcal{M}(D') \in S].$$

Applying Theorem 4.10, we obtain that $\mathcal{M}$ is d-private* with

$$d(D, D') = \sum_{i=1}^{k} (\varepsilon_i d_{\mathbb{D}}^{\mathcal{B}})^{p_i}(D, D') = \sum_{i=1}^{k} \varepsilon_i d_{\mathbb{D}}^{\mathcal{B}, p_i}(D, D').$$

Now suppose $D, D' \in \mathbb{D}$ are bounded-neighboring. We know there exists $x \in D$ and $x' \in D'$ such that $D \triangle D' = \{x, x'\}$. Then, we have the following possibilities:

(a) $x, x' \in \mathcal{X}_j$ for a $j \in [k]$. This implies that $p_i(D) = p_i(D')$ for all $i \neq j$.

(b) $x \in \mathcal{X}_j$ and $x' \in \mathcal{X}_l$ for different $j, l \in [k]$. This implies that $p_i(D) = p_i(D')$ for all $i \neq j, l$.

(c) $x \in \mathcal{X}_j$ for $j \in [k]$ and $x' \notin \mathcal{X}_l$ for any $l \in [k]$ (or vice-versa). This implies that $p_i(D) = p_i(D')$ for all $i \neq j$.

(d) $x, x' \notin \mathcal{X}_l$ for any $l \in [k]$. Then $p_i(D) = p_i(D')$ for all $i \in [k]$.

In the worst case scenario, there are at most two subindices $j, l \in [k]$ such that $p_i(D) = p_i(D')$ for all $i \neq j, l$. For these subindices, $d_{\mathbb{D}}^{\mathcal{B}, p_j}(D, D'), d_{\mathbb{D}}^{\mathcal{B}, p_l}(D, D') \leq d_{\mathbb{D}}^{\mathcal{B}}(D, D') \leq 1$, since D and D' are bounded-neighboring. Therefore,

$$d(D, D') = \sum_{i=1}^{k} \varepsilon_i d_{\mathbb{D}}^{\mathcal{B}, p_i}(D, D') \leq \max_{j,l \in [k];\, j \neq l} (\varepsilon_j + \varepsilon_l) = \varepsilon$$

for all bounded-neighboring $D, D' \in \mathbb{D}$. □

Note that this result is stated for common domain, and that the non–common-domain counterpart cannot be defined as we prove in Example 2.16.

In summary, Theorems 4.1 and 4.10 provide a characterization of the privacy loss under composition for metric privacy while explicitly accounting for pre-processing functions. This characterization applies both when the pre-processing is applied prior to the mechanism (as in Theorem 4.1) and when it is embedded as part of the mechanism itself (as in Theorem 4.10).

Beyond offering a unifying framework for composition in metric privacy, these theorems also identify specific scenarios in which the privacy loss under composition can be improved relative to standard sequential composition. Such compatibility properties between pre-processing functions and metrics can be directly leveraged by practitioners to optimize their query strategies, enabling them to extract as much information as possible while incurring the minimal privacy cost.

4.4. Composition for Metric Gaussian DP

In the previous section, we improved composition bounds for metric privacy by explicitly accounting for pre-processing steps. Metric privacy generalizes DP to arbitrary metric spaces and granularities, hence represents a step forward on improving the interpretability and applicability of the DP framework controlling the exact information that must be indistinguishable. Nevertheless, metric privacy inherits from DP a limited interpretability of its parameters. In particular, there is no clear consensus on how to select the privacy parameter ε in DP. Since the metric d generalizes the role of ε, there is likewise no consensus on which values of $d(D, D')$ are acceptable, and it is often difficult to relate these choices to concrete adversarial capabilities [133].

As a response to these limitations, GDP provides a significantly more interpretable framework. By adopting a hypothesis-testing perspective, GDP characterizes privacy in terms of the optimal power of an adversary's test attempting to distinguish a target's participation—precisely capturing the classical notion of MIAs. We leverage this interpretation to derive novel bounds on attack mitigation for AIAs and DRAs in Chapter 5, using the general f-DP framework and, in particular, GDP. This strongly motivates a careful study of GDP composition properties here, as they are essential for the applicability of our later results (see Example 5.17).

Beyond interpretability, GDP also enjoys substantially tighter sequential composition bounds than the original DP framework [38]. This naturally suggests that analogous improvements should be achievable in the metric GDP setting, providing a clear motivation for extending our composition results to metric GDP.

4.4.1. Metric Gaussian Privacy

In this section we extend metric privacy to Gaussian DP (GDP) [38]. GDP uses the hypothesis testing interpretation of DP to bound the privacy loss (cf. Chapter 2). This way, it provides a more interpretable intuition of DP guarantees with respect to a real attacker, directly establishing a relation between the privacy parameters (noise injection) and the maximum power an attacker can achieve when trying to distinguish between two databases differing in one record. This intuition directly extends to any granularity:

Definition 4.13 ($\mathcal{G}$ Gaussian DP)**.** Let $\mu \geq 0$. A mechanism $\mathcal{M}$ with domain $\mathbb{D}$ is said to be $\mathcal{G}$ *μ-GDP* if, for all $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$,

$$T(\mathcal{M}(D), \mathcal{M}(D'))(\alpha) \geq T(\mathcal{N}(0,1), \mathcal{N}(\mu,1))(\alpha)$$

for all $\alpha \in [0,1]$. We denote $G_\mu \coloneqq T(\mathcal{N}(0,1), \mathcal{N}(\mu,1))$.

GDP satisfies a group privacy property that establishes that privacy degrades linearly with respect to the number of changes between the two databases [38]. Consequently, we use this property to define the $d_{\mathbb{D}}$-privacy adaptation of GDP:

Definition 4.14 ($d_{\mathbb{D}}$-Gaussian privacy)**.** Let $d_{\mathbb{D}} \colon \mathbb{D}^2 \to [0, \infty]$ be a metric. A mechanism $\mathcal{M}$ with domain $\mathbb{D}$ is said to be *$d_{\mathbb{D}}$-Gprivate* if, for all $D, D' \in \mathbb{D}$,

$$T(M(D), M(D')) \geq G_{d_{\mathbb{D}}(D,D')},$$

where $G_\infty(\alpha) \coloneqq \lim_{\mu\to\infty} G_\mu(\alpha) = 0$.

Our definition of $d_{\mathbb{D}}$-Gprivacy generalizes the original notion of GDP:

Theorem 4.15. *Let $\mathcal{G}$ be a granularity notion over the database class $\mathbb{D}$. Then, a mechanism $\mathcal{M}$ with domain $\mathbb{D}$ is $\mu d_{\mathbb{D}}^{\mathcal{G}}$-Gprivate if and only if it is $\mathcal{G}$ μ-GDP.*

Proof. First, we see that $\mu d_{\mathbb{D}}^{\mathcal{G}}$-Gprivacy implies $\mathcal{G}$ μ-GDP. Suppose that $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$ is $\mu d_{\mathbb{D}}^{\mathcal{G}}$-Gprivacy. Then, for any $\mathcal{G}$-neighboring databases $D, D' \in \mathbb{D}$, we have that

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_{\mu d_{\mathbb{D}}^{\mathcal{G}}(D,D')}.$$

By construction of the canonical metric, $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = 1$ since D and D' are $\mathcal{G}$-neighboring, and therefore $\mathcal{M}$ is $\mathcal{G}$ μ-GDP.

Now we prove the other implication. Suppose $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$ is $\mathcal{G}$ μ-GDP. We want to see that for all $D, D' \in \mathbb{D}$

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_{\mu d_{\mathbb{D}}^{\mathcal{G}}(D,D')}.$$

We now prove this by induction over $d_{\mathbb{D}}^{\mathcal{G}}(D, D')$. For $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = 1$ we have $D \sim_{\mathcal{G}} D'$ and thus

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_\mu.$$

We now assume that the statement holds for $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = k - 1$ and we prove for $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = k$. Since $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = k$, there exists $D_1, \ldots, D_{k-1} \in \mathbb{D}$ such that

$$D \sim_{\mathcal{G}} D_1 \sim_{\mathcal{G}} \cdots \sim_{\mathcal{G}} D_{k-1} \sim_{\mathcal{G}} D'.$$

By the induction hypothesis, we have that

$$T(\mathcal{M}(D_{k-1}), \mathcal{M}(D')) \geq G_\mu$$

and

$$T(\mathcal{M}(D), \mathcal{M}(D_{k-1})) \geq G_{\mu(k-1)}.$$

Then, by Lemma A.5 in [38], we have that

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_\mu(1 - G_{\mu(k-1)}(\alpha)).$$

Therefore, in conclusion,

$$\begin{aligned} G_\mu(1 - G_{\mu(k-1)}(\alpha)) &= \Phi(\Phi^{-1}(G_{\mu(k-1)}(\alpha)) - \mu) \\ &= \Phi(\Phi^{-1}(1-\alpha) - \mu - (1-k)\mu) \\ &= G_{\mu+\mu(k-1)}(\alpha) \\ &= G_{\mu k}(\alpha). \end{aligned}$$

This proves the result for all $d_{\mathbb{D}}^{\mathcal{G}}(D, D') \in \mathbb{N}$. Note that the case $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = \infty$ holds trivially since $G_\infty \equiv 0$. □

4.4.2. General Composition for Metric Gaussian Privacy

In this section, we extend our unified composition framework to metric Gprivacy. We consider the same setting as in the metric privacy case: We begin by proving a general theorem from which all subsequent results follow, including the improvements obtained when accounting for embedded query functions in the common-domain setting. Finally, we show how parallel composition translates into the metric $\mathcal{G}$-privacy framework.

Theorem 4.16 (Gaussian AC theorem)**.** *Let $\mathbb{D}$ be a database class and, for all $i \in [k]$, let $(\mathbb{D}_i, d_i)$ be a privacy space, and $f_i \colon \mathbb{D} \to \mathbb{D}_i$ a deterministic map and $f_i^* = \mathrm{id}_{\overline{\Theta}_i} \times f_i$ (with $f_1^* = f_1$).*

For $i \in [k]$, let $\mathcal{M}_i^* \colon \overline{\Theta}_i \times \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_i^*(\bar{\theta}_i, \cdot) \colon \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ satisfies d_i-Gprivacy for any $\bar{\theta}_i \in \overline{\Theta}_i$. Then mechanism $\mathcal{M} = (\mathcal{M}_1^* \circ f_1^*, \ldots, \mathcal{M}_k^* \circ f_k^*)$ is $d_{\mathbb{D}}$-Gprivate with

$$d_{\mathbb{D}}(D, D') \coloneqq \sqrt{\sum_{i=1}^{k} d_i(f_i(D), f_i(D'))^2} \quad \text{for all } D, D' \in \mathbb{D}.$$

Proof. Note that $d_{\mathbb{D}}$ is a well-defined metric since the square root of the sum of squared distances is still a distance (i.e., the ℓ_2-norm).

Now we need to prove for all $D, D' \in \mathbb{D}$ that

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_{d_{\mathbb{D}}(D,D')}.$$

We prove the result by induction over k. For $k = 1$, the result is trivial. Therefore, fixing k, we suppose it is true for $k-1$ and we prove for k.

Let $\overline{\mathcal{M}} = (\mathcal{M}_1^* \circ f_1^*, \ldots, \mathcal{M}_{k-1}^* \circ f_{k-1}^*)_{\mathrm{adapt}}$. By the induction hypothesis, for all $D, D' \in \mathbb{D}$,

$$T(\overline{\mathcal{M}}(D), \overline{\mathcal{M}}(D')) \geq G_{\bar{d}(D,D')}$$

with $\bar{d} = \sqrt{d_1^2 + \cdots + d_{k-1}^2}$. We can also rewrite $\mathcal{M}$ as a function of $\overline{\mathcal{M}}$ and $\mathcal{M}_k$ as

$$\mathcal{M}(D) = (\overline{\mathcal{M}}(D), \mathcal{M}_k(\overline{\mathcal{M}}(D), f_k(D)))$$

for all $D \in \mathbb{D}$.

We fix $D, D' \in \mathbb{D}$. Since $\mathcal{M}_k(\bar{s}_k, \cdot)$ is d_k-Gprivate for all $\bar{s}_k \in \overline{\Theta}_k$, we have that

$$T(\mathcal{M}(D), \mathcal{M}(D')) = T(\overline{\mathcal{M}}(D), \overline{\mathcal{M}}(D')) \otimes G_{d_k(D,D')}.$$

This fact follows from Lemma C.1 in [38] as explained in their proof of Lemma C.3. Since $\overline{\mathcal{M}}$ is $\bar{d}$-private, we obtain

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_{\bar{d}(D,D')} \otimes G_{d_k(D,D')}.$$

by the properties of $\otimes$ (see Remark 2.23). Finally, by Proposition D.1 in [38], we obtain

$$G_{\bar{d}(D,D')} \otimes G_{d_k(D,D')} = G_{\sqrt{\bar{d}(D,D')^2 + d_k(D,D')^2}}$$

where

$$\sqrt{\bar{d}(D,D')^2 + d_k(D,D')^2} = \sqrt{\sum_{i=1}^{k} d_i(D,D')^2}. \qquad \square$$

Note that unlike the AC theorem (4.1), $d_{\mathbb{D}}$ is not the sum of the distances (i.e., the ℓ_1-norm), but actually the sum of the squares of the distances (i.e., the ℓ_2-norm). Recall that $\|(d_1,\dots,d_k)\|_2 \leq \|(d_1,\dots,d_k)\|_1$. In this case, we can notice improvements in GDP to the composition results.

As in the previous sections, we recover the generalized ASC results when $f_i = \mathrm{id}$:

Corollary 4.17 (Generalized Gaussian ASC)**.** *Let $\mathbb{D}$ be a database class, and d a metric defined in $\mathbb{D}$. For $i \in [k]$, let $\mathcal{M}_i^*\colon \overline{\Theta}_i \times \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_i^*(\overline{\theta}_i,\cdot)\colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ satisfies d_i-Gprivacy for any $\overline{\theta}_i \in \overline{\Theta}_i$. Then mechanism $\mathcal{M} = (\mathcal{M}_1^*,\dots,\mathcal{M}_k^*)$ is $d_{\mathbb{D}}$-Gprivate with with $d_{\mathbb{D}} = \sqrt{d_1^2 + \cdots + d_k^2}$.*

Proof. Direct from Theorem 4.16 by taking $\mathbb{D}_i = \mathbb{D}$ and $f_i = \mathrm{id}$. $\square$

Choosing $d_i = \mu_i d_{\mathbb{D}}^{\mathcal{G}}$, we obtain from this theorem the already-known [65] sequential bound $\|(\mu_1,\dots,\mu_k)\|_2$.

If, instead of considering preprocessing functions, we account for the embedded queries of the composed mechanisms—i.e., we work in the common-domain setting—we obtain a result analogous to the metric privacy case. In particular, the overall privacy loss can again be reduced. However, in this setting the bound explicitly reflects the specific features of metric $\mathcal{G}$-privacy composition.

Theorem 4.18 (Gaussian AC theorem for common domain)**.** *For $i \in [k]$, let $(\mathbb{D}_i, d_i)$ be a privacy space, and let f_i be a deterministic map over $\mathbb{D}$. For $i \in [k]$, let $\mathcal{M}_i\colon \overline{\Theta}_i \times \mathbb{D} \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_k(\overline{\theta}_i,\cdot)\colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ satisfies d_i-Gprivacy for all $\overline{\theta}_i \in \overline{\Theta}_i$ and f_i-dependency for any $\overline{\theta}_i \in \overline{\Theta}_i$. Then mechanism $\mathcal{M} = (\mathcal{M}_1,\dots,\mathcal{M}_k)$ is $d_{\mathbb{D}}$-Gprivate* with $d_{\mathbb{D}} \coloneqq \sqrt{\sum_{i=1}^{k}(d_i^{f_i})^2}$.*

Proof. We just need to prove that $\mathcal{M}_i$ are $d_i^{f_i}$-private* and it follows from an analogous proof of Theorem 4.16.

By hypothesis, $\mathcal{M}_i$ are d_i-Gprivate, which means that for all $D, D' \in \mathbb{D}$,

$$T(\mathcal{M}_i(D), \mathcal{M}_i(D')) \geq G_{d_i(D,D')}.$$

Using Remark A.15, we obtain that

$$T(\mathcal{M}_i(D), \mathcal{M}_i(D')) = T(\mathcal{M}_i(\tilde{D}), \mathcal{M}_i(\tilde{D}'))$$

for all $\tilde{D}, \tilde{D}' \in \mathbb{D}$ such that $f(D) = f(\tilde{D})$ and $f(D') = f(\tilde{D}')$. In conclusion,

$$T(\mathcal{M}_i(D), \mathcal{M}_i(D')) \geq G_{d_i^f(D,D')}$$

with

$$d_i^f(D,D') = \min_{\substack{\tilde{D},\tilde{D}' \in \mathbb{D} \\ f(\tilde{D})=f(D) \\ f(\tilde{D}')=f(D')}} d_i(\tilde{D},\tilde{D}'). \qquad \square$$

These theorems enable a detailed characterization of privacy-loss propagation in metric privacy, while incorporating the improved composition properties of GDP. These improvements are particularly pronounced in the parallel composition setting. In contrast to standard metric privacy—where the strongest parallel composition bound is given by $\max_i d_i$—we obtain novel strictly tighter privacy-loss estimates in the metric Gaussian case, as we demonstrate in the following section.

4.4.3. Parallel Composition in Metric Gaussian Privacy

For d-Gprivacy, as for metric privacy, it is interesting to find cases where we can obtain better bounds than the sequential one using our result. We explore these cases in the following corollaries. For example, we can also obtain the best bound for when f defines a partitioning function:

Theorem 4.19 (Parallel for Metric Gprivacy)**.** *Let $\mathbb{D}$ be a database class, and let p be k-partitioning function of $\mathbb{D}$ in $\mathbb{D}_i$ and $p_i^* = \mathrm{id}_{\overline{\Theta}_i} \times p_i$ (with $p_1^* = p_1$). Let d^* be well-defined over $\mathbb{D}$ and $\mathbb{D}_i$. For $i \in [k]$, let $\mathcal{M}_i^* \colon \overline{\Theta}_i \times \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_i^*(\overline{\theta}_i, \cdot) \colon \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ satisfies $\mu_i d^*_{\mathbb{D}_i}$-Gprivacy for all $\overline{\theta}_i \in \overline{\Theta}_i$. If d^* commutes with p then mechanism $\mathcal{M} = (\mathcal{M}_1^* \circ p_1^*, \ldots, \mathcal{M}_k^* \circ p_k^*)$ is $\tilde{d}_{\mathbb{D}}$-Gprivate with*

$$\tilde{d}_{\mathbb{D}}(D, D') \coloneqq \sqrt{\sum_{i=1}^{k} (\mu_i d^*_{\mathbb{D}_i}(p_i(D), p_i(D')))^2} \leq \max_{i \in [k]} \mu_i d^*_{\mathbb{D}}(D, D'). \tag{4.4}$$

Proof. By Theorem 4.16, we have that $\mathcal{M}$ is $d_{\mathbb{D}}$-Gprivate for

$$d_{\mathbb{D}}(D, D') = \sqrt{\sum_{i=1}^{k} \mu_i^2 d^*_{\mathbb{D}_i}(p_i(D), p_i(D'))^2} \leq \max_{i \in [k]} \mu_i \sqrt{\sum_{i=1}^{k} d^*_{\mathbb{D}_i}(p_i(D), p_i(D'))^2}.$$

Then, we have that

$$\begin{aligned}
\sum_{i=1}^{k} d^*_{\mathbb{D}_i}(p_i(D), p_i(D'))^2 &\overset{\text{(i)}}{\leq} \Big(\sum_{i=1}^{k} d^*_{\mathbb{D}_i}(p_i(D), p_i(D')) \Big)^2 \\
&\overset{\text{(ii)}}{=} d^*_{\mathbb{D}} \Big(\bigcup_{i=1}^{k} p_i(D), \bigcup_{i=1}^{k} p_i(D') \Big)^2 \\
&\overset{\text{(ii)}}{\leq} d^*_{\mathbb{D}}(D, D')^2,
\end{aligned}$$

where

(i) comes from the fact that $\sum_{i=1}^{k} a_i^2 \leq (\sum_{i=1}^{k} a_i)^2$ for all $a_i \geq 0$,

(ii) is due to the commutativity of d^* with respect to p.

Since the square root is a monotonically increasing function, we have the result. □

Note that the inequality is in fact an equality only when $d^*_{\mathbb{D}_i}(p_i(D), p_i(D')) = 0$ for all but one $i \in [k]$. Therefore, in some cases, the Gaussian AC theorem (4.16) can give us a tighter bound than $\max_{i\in[k]} \mu_i d^*_{\mathbb{D}}$—which is not possible for metric privacy. We see this in the following example:

Example 4.20. Let $\mathbb{D} \subseteq \mathbb{D}_{\mathcal{X}}$, let $\mathbb{D}_i = \mathbb{D}_{\mathcal{X}_i}$ where $\{\mathcal{X}_i\}_{i\in[k]}$ defines a partition, and consider $d^{\triangle}$, which commutes with the previous partition (see Proposition A.11). If $\mathcal{M}_i \colon \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ are $d^{\triangle}_{\mathbb{D}_i}$-Gprivate, then mechanism $\mathcal{M} = (\mathcal{M}_1, \dots, \mathcal{M}_k)$ is $\tilde{d}_{\mathbb{D}}$-Gprivate with $\tilde{d}_{\mathbb{D}} \leq d^{\triangle}_{\mathbb{D}}$. For instance, if $D = D' \backslash \{x_i, x_j\}$ with $x_i \in \mathcal{X}_i$ and $x_j \in \mathcal{X}_j$ $(i \neq j)$, we have that $d^{\triangle}_{\mathbb{D}}(D, D') = 2$, while

$$\begin{aligned}\tilde{d}_{\mathbb{D}}(D, D') &= \sqrt{d^{\triangle}_{\mathbb{D}_i}(p_i(D), p_i(D'))^2 + d^{\triangle}_{\mathbb{D}_j}(p_j(D), p_j(D'))^2} \\ &= \sqrt{|\{x_i\}|^2 + |\{x_j\}|^2} = \sqrt{1+1} = \sqrt{2} < 2 = d^{\triangle}_{\mathbb{D}}(D, D').\end{aligned}$$

The previous example highlights the advantages of the metric privacy composition relative to the original DP setting. While in bounded ε-GDP the best bound after composition is $\max_i \varepsilon_i$, for metric Gprivacy we obtain a better bound than $\max_i d_i(D, D')$.

The Gaussian version of Theorem 4.6 also holds. However, in this case, a compatible partition implies $d^{\mathcal{G}}_{\mathbb{D}_i}(p_i(D), p_i(D')) = 0$ for all but one $i \in [k]$, so the inequality in Equation (4.4) becomes an equality.

Theorem 4.21 (Gaussian best bound for disjoint inputs)**.** *Let $\mathbb{D}$ be a database class and $\mathcal{G}$ a granularity over $\mathbb{D}$. Let p be a $d^{\mathcal{G}}_{\mathbb{D}}$-compatible k-partitioning function such that $\Delta p_i \leq 1$, and $p^*_i = \mathrm{id}_{\overline{\Theta}_i} \times p_i$ (with $p^*_1 = p_1$). For $i \in [k]$, let $\mathcal{M}^*_i \colon \overline{\Theta}_i \times \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}^*_i(\overline{\theta}_i, \cdot) \colon \mathbb{D}_i \to \mathcal{D}(\Theta_i)$ satisfies $\mu_i d^{\mathcal{G}}_{\mathbb{D}_i}$-Gprivacy for any $\overline{\theta}_i \in \overline{\Theta}_i$. Then mechanism $\mathcal{M} = (\mathcal{M}^*_1 \circ p^*_1, \dots, \mathcal{M}^*_k \circ p^*_k)$ is $\mu d^{\mathcal{G}}_{\mathbb{D}}$-Gprivate with $\mu = \max_{i\in[k]} \mu_i$.*

Proof. From Theorem 4.15, it is equivalent to see that $\mathcal{M}$ is $\mathcal{G}$ μ-GDP with $\mu = \max_{i\in[k]} \mu_i$, i.e., that for all $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$,

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_\mu$$

Applying Theorem 4.16, we obtain that $\mathcal{M}$ is d-Gprivate with

$$d_{\mathbb{D}}(D, D') = \sqrt{\sum_{i=1}^{k} \mu_i^2 d^{\mathcal{G}}_{\mathbb{D}_i}(p_i(D), p_i(D'))^2}.$$

Now suppose $D, D' \in \mathbb{D}$ are $\mathcal{G}$-neighboring. By definition of $d^{\mathcal{G}}_{\mathbb{D}}$-compatibility, there exist $j \in [k]$ such that $p_i(D) = p_i(D')$ for all $i \neq j$. Consequently, for all $i \neq j$, $d^{\mathcal{G}}_{\mathbb{D}_i}(p_i(D), p_i(D')) = 0$. Moreover, by preprocessing (Proposition A.3), we have that $d^{\mathcal{G}}_{\mathcal{D}_j}(p_j(D), p_j(D')) \leq \Delta p_j d^{\mathcal{G}}_{\mathbb{D}}(D, D') \leq 1$ since $D \sim_{\mathcal{G}} D'$ and $\Delta p_j \leq 1$. Therefore,

$$d_{\mathbb{D}}(D, D') = \sqrt{\sum_{i=1}^{k} \mu_j^2 d^{\mathcal{G}}_{\mathcal{D}_j}(p_j(D), p_j(D'))^2} = \mu_j d^{\mathcal{G}}_{\mathcal{D}_j}(p_j(D), p_j(D')) \leq \mu_j.$$

Consequently, since $\mathcal{M}$ is d-Gprivate,

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_{\mu_j}.$$

Since $j \in [k]$ depends on the choice of the $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$, it is sufficient to choose $\mu = \max_{i\in[k]} \mu_i$ to cover all cases. In conclusion, $\mathcal{M}$ is $\mathcal{G}$ μ-GDP. □

Similarly to the pure metric case, we obtain an improved result for common domain.

Theorem 4.22 (Gaussian best bound for disjoint inputs (common domain))**.** *Let $\mathbb{D}$ be a database class and $\mathcal{G}$ a granularity over $\mathbb{D}$. Let p be a $d_{\mathbb{D}}^{\mathcal{G}}$-compatible k-partitioning function. For $i \in [k]$, let $\mathcal{M}_i \colon \overline{\Theta}_i \times \mathbb{D} \to \mathcal{D}(\Theta_i)$ be a mechanism such that $\mathcal{M}_k(\overline{\theta}_i, \cdot) \colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ satisfies $\mu_i d_{\mathbb{D}}^{\mathcal{G}}$-Gprivacy and p_i-dependency for any $\overline{\theta}_i \in \overline{\Theta}_i$. Then mechanism $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ is $\mu d_{\mathbb{D}}^{\mathcal{G}}$-Gprivate with $\mu = \max_{i\in[k]} \mu_i$.*

Proof. From Theorem 4.15, it is equivalent to see that $\mathcal{M}$ is $\mathcal{G}$ μ-GDP with $\mu = \max_{i\in[k]} \mu_i$, i.e., that for all $\mathcal{G}$-neighboring $D, D' \in \mathbb{D}$,

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_{\mu}.$$

From Theorem 4.18, we have that $\mathcal{M}$ is $d_{\mathbb{D}}$-Gprivate* with

$$d_{\mathbb{D}}(D, D') = \sqrt{\sum_{i=1}^{k} \mu_i^2 d_{\mathbb{D}}^{\mathcal{G},p_i}(D, D')^2}.$$

Since $d_{\mathbb{D}}^{\mathcal{G}}$-compatible, there exist only one $j \in [k]$ such that $p_j(D) \neq p_j(D')$. Consequently, for all $i \neq j$, $d_{\mathbb{D}}^{\mathcal{G},p_i}(D, D') \leq d_{\mathbb{D}}^{\mathcal{G}}(D, D) = 0$, since we can select D as both $\tilde{D}$ and $\tilde{D}'$ in the definition (see Proposition A.16). Therefore,

$$d_{\mathbb{D}}(D, D') = \sqrt{\sum_{i=1}^{k} \mu_i^2 d_{\mathbb{D}}^{\mathcal{G},p_i}(D, D')^2}$$

$$= \sqrt{\mu_j^2 d_{\mathbb{D}}^{\mathcal{G},p_j}(D, D')^2 + 0} = \mu_j d_{\mathbb{D}}^{\mathcal{G},p_j}(D, D') \leq \mu_j d_{\mathbb{D}}^{\mathcal{G}}(D, D') \leq \mu_j,$$

where the last inequality comes from the fact that $D \sim_{\mathcal{G}} D'$. Since j depends on the choice of D and D', it is sufficient to take $\mu = \max_{i\in[k]} \mu_i$ to cover all possible cases. □

Moreover, this theorem leads to solving parallel for bounded GDP (which is actually the only granularity for which GDP was defined until now):

Corollary 4.23. *Let p be a k-partitioning function of Example 4.5. For all $i \in [k]$, let $\mathcal{M}_i \colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ be mutually independent bounded μ_i-Gprivacy mechanisms that are p_i-dependent. Then mechanism $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)_{\text{ind}}$ with domain $\mathbb{D}$ is bounded μ-GDP with $\mu = \max_{i,j\in[k];\, i\neq j} \sqrt{\mu_i^2 + \mu_j^2}$.*

Proof. From Theorem 4.15, it is equivalent to see that $\mathcal{M}$ is bounded μ-GDP ($\mu = \max_{i,j\in[k];\, i\neq j} \sqrt{\mu_i^2 + \mu_j^2}$), i.e., that for all bounded-neighboring $D, D' \in \mathbb{D}$,

$$T(\mathcal{M}(D), \mathcal{M}(D')) \geq G_\mu.$$

Applying Theorem 4.18, we obtain that $\mathcal{M}$ is $d_{\mathbb{D}}$-Gprivate* with

$$d_{\mathbb{D}}(D, D') = \sqrt{\sum_{i=1}^{k} \mu_i^2 d_{\mathbb{D}}^{\mathcal{B},p_i}(D, D')^2}.$$

Now suppose $D, D' \in \mathbb{D}$ are bounded-neighboring. We know there exists $x \in D$ and $x' \in D'$ such that $D \triangle D' = \{x, x'\}$. Then, we have the following possibilities:

(a) $x, x' \in \mathcal{X}_j$ for a $j \in [k]$. This implies that $p_i(D) = p_i(D')$ for all $i \neq j$.

(b) $x \in \mathcal{X}_j$ and $x' \in \mathcal{X}_l$ for different $j, l \in [k]$. This implies that $p_i(D) = p_i(D')$ for all $i \neq j, l$.

(c) $x \in \mathcal{X}_j$ for $j \in [k]$ and $x' \notin \mathcal{X}_l$ for any $l \in [k]$ (or vice-versa). This implies that $p_i(D) = p_i(D')$ for all $i \neq j$.

(d) $x, x' \notin \mathcal{X}_l$ for any $l \in [k]$. Then $p_i(D) = p_i(D')$ for all $i \in [k]$.

In the worst case scenario, there are at most two subindices $j, l \in [k]$ such that $p_i(D) = p_i(D')$ for all $i \neq j, l$. For these subindices, $d_{\mathbb{D}}^{\mathcal{B},p_j}(D, D'), d_{\mathbb{D}}^{\mathcal{B},p_l}(D, D') \leq d_{\mathbb{D}}^{\mathcal{B}}(D, D') \leq 1$, since D and D' are bounded-neighboring. Therefore,

$$d_{\mathbb{D}}(D, D') = \sqrt{\sum_{i=1}^{k} \mu_i^2 d_{\mathbb{D}}^{\mathcal{B},p_i}(D, D')^2} \leq \max_{j,l\in[k];\, j\neq l} \sqrt{\mu_j^2 + \mu_l^2} = \mu$$

for all bounded-neighboring $D, D' \in \mathbb{D}$. In conclusion, $\mathcal{M}$ is bounded μ-GDP since it is $d_{\mathbb{D}}$-Gprivate*. □

Overall, this section shows that metric Gprivacy exhibits particularly favorable composition properties and yields novel insights in the parallel composition setting that had not been previously identified.

4.5. Reciprocal Results

In this section, we elaborate on the reciprocal results of our theorems. Reciprocal results do not help to design new composition strategies, however, similarly to our impossibility results presented in Chapter 3, they allow the early detection of formal flaws and failures when designing complex DP protocols.

To prove reciprocals, we first need to understand the post-processing properties of metric privacy and metric Gprivacy, as they are the key tool in the following proofs. Particularly, both $d_{\mathbb{D}}$-privacy and $d_{\mathbb{D}}$-Gprivacy, are robust to post-processing:

Proposition 4.24 (Post-processing)**.** *The privacy notions of $d_{\mathbb{D}}$-privacy and $d_{\mathbb{D}}$-Gprivacy are robust to post-processing.*

Proof. We need to prove that if $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$ is $d_{\mathbb{D}}$-private, then $g \circ \mathcal{M}\colon \mathbb{D} \to g(\Theta)$ is $d_{\mathbb{D}}$-private for all deterministic functions $g\colon \Theta \to g(\Theta)$; and analogously for $d_{\mathbb{D}}$-Gprivacy. By construction do note that $\text{Range}(g \circ \mathcal{M}) = g(\text{Range}(\mathcal{M})) =: g(\Theta)$.

For $d_{\mathbb{D}}$-privacy the proof follows directly from the fact that $\Pr[\mathcal{M}(D) \in S] = \Pr[g(\mathcal{M}(D)) \in g(S)]$ for all measurable $S \subseteq \Theta$ and $D \in \mathbb{D}$.

Finally, from Lemma 2.9 in [38], we obtain the following inequality:

$$T(g(\mathcal{M}(D)), g(\mathcal{M}(D'))) \geq T(\mathcal{M}(D), \mathcal{M}(D')) \geq d(D, D').$$

This proves the result for $d_{\mathbb{D}}$-Gprivacy. □

Moreover, we obtain reciprocal results for the composition theorems for common domain for any privacy notion $\mathfrak{P}$ that is robust to post-processing. More precisely, Theorem 4.10 has a reciprocal result.

Theorem 4.25 (Reciprocal to the IC theorem (common domain))**.** *Let $\mathfrak{P}$ be a privacy notion that is robust to post-processing. For all $i \in [k]$, let $\mathcal{M}_i\colon \mathbb{D} \to \mathcal{D}(\Theta_i)$ be mutually independent randomized mechanisms. Let $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)_{\text{ind}}$ be a mechanism that satisfies $\mathfrak{P}$. Then $\mathcal{M}_i$ must satisfy $\mathfrak{P}$ for all $i \in [k]$.*

Proof. Fix $i \in [k]$. Consider the deterministic projection to the ith coordinate π_i. In this case, $\mathcal{M}_i = \pi_i \circ \mathcal{M}$. Since $\mathcal{M}$ satisfies $\mathfrak{P}$ and $\mathfrak{P}$ is robust to post-processing, $\mathcal{M}_i$ satisfies $\mathfrak{P}$ too. □

Even though it is not useful in constructing new mechanisms, this result makes it clear that we cannot obtain a $\mathfrak{P}$ mechanism by independently composing mechanisms that do not satisfy $\mathfrak{P}$, and can serve as a first check to ensure whether a mechanism satisfies $\mathfrak{P}$ or not. For instance, Example 2.16 fails because $\mathcal{M}_i = \mathcal{M}_i^* \circ f_i$ do not satisfy $\mathfrak{P}$. Also, for the adaptive case, we have the following result:

Theorem 4.26 ("Reciprocal" to the AC theorem (common domain))**.** *Let $\mathfrak{P}$ be a privacy notion that is robust to post-processing. Let $\mathcal{M}_i\colon \overline{\Theta}_i \times \mathbb{D} \to \mathcal{D}(\Theta_i)$ for $i \in [k]$ be randomized mechanisms. Let $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$ be a mechanism satisfying $\mathfrak{P}$. Recall that by definition $\mathcal{M}(D) = (\mathcal{N}_1(D), \ldots, \mathcal{N}_k(D))$ for all $D \in \mathbb{D}$, where $\mathcal{N}_i(D)$ are defined recursively as $\mathcal{N}_i(D) = \mathcal{M}_i(\mathcal{N}_{i-1}(D), \ldots, \mathcal{N}_1(D), D)$ for $i \in [k]$. Then $\mathcal{N}_i$ must satisfy $\mathfrak{P}$ for all $i \in [k]$.*

Proof. Fix $i \in [k]$. Consider the deterministic projection to the ith coordinate π_i. In this case, $\mathcal{N}_i = \pi_i \circ \mathcal{M}$. Since $\mathcal{M}$ satisfies $\mathfrak{P}$ and $\mathfrak{P}$ is robust to post-processing, $\mathcal{N}_i$ satisfies $\mathfrak{P}$ too. □

Note that this result tells us that all $\mathcal{N}_i$ satisfy $\mathfrak{P}$, but this is not the exact reciprocal of Theorem 4.10. Given the same hypotheses, it is not necessarily true that $\mathcal{M}_i(\overline{\Theta}_i, \cdot)$ satisfy $\mathfrak{P}$ for all $\overline{\theta}_i \in \overline{\Theta}_i$.

Furthermore, no result for $\mathcal{M}_i^*$ can be generally stated, since the reciprocal of the pre-processing property is not generally true. For example, in Remark A.4, we provide a case where $\mathcal{M}_i^* \circ f_i$ is LDP while $\mathcal{M}_i^*$ is not.

4.6. Conclusions

In this chapter, we study the composability properties of DP under more interpretable extensions, namely metric privacy and GDP. Metric privacy generalizes DP by allowing arbitrary granularities and data domains, enabling a finer understanding of which information is protected when adapting DP to complex data types such as graphs. We further focus on Gaussian DP—and our novel extension, metric Gaussian privacy—which provides a more interpretable framework for parameter selection once the underlying metric or granularity is specified.

We show that composability can be defined independently of the neighborhood definition. Our results can be used to directly obtain specific composition rules when new granularity notions (or metrics) are proposed over any desired data domain and even under mixed privacy requirements, which was not previously defined.

Our main contribution is a unified composition framework, together with the corresponding results Theorems 4.1 and 4.16, which reduces privacy loss analysis to the interaction between metrics and preprocessing functions (i.e., their sensitivities). This perspective removes the need to distinguish between sequential and parallel composition: composition becomes a modular operation determined solely by the metric and the preprocessing, and applies uniformly across domains and levels of granularity. Under the lens of metric privacy, both sequential and parallel composition reduce to summing the metrics induced by the preprocessing, yielding tight privacy loss estimates without additional case distinctions.

Beyond providing a unifying framework that facilitates the computation of the final privacy guarantee, our theorems yield tighter bounds than those previously reported in the literature when the effect of preprocessing is taken into account, including intermediate settings between sequential and parallel composition. Particularly, we derive necessary conditions to obtain parallel composition bounds in arbitrary data domains and granularities.

Furthermore, we extend our results to GDP, extending the interpretability of GDP to composition settings while offering tighter privacy estimates. Particularly, we show that the parallel composition metric bound can be improved in $d_{\mathbb{D}}$-Gprivacy, providing a better result than the maximum privacy loss as expected from the DP original setting.

Finally, we discuss reciprocal versions of the composition, which can be used to check when a mechanism fails to guarantee DP, hence avoid common pitfalls in the literature as previously exposed in Chapter 3.

5. Understanding Disclosure Risk in Differential Privacy

This chapter is based on the contributions:

- **Patricia Guerra-Balboa**, Annika Sauer, Héber Arcolezi, and Thorsten Strufe. "Understanding Disclosure Risk in Differential Privacy with Applications to Noise Calibration and Auditing". In: Proceedings of the VLDB, 2026. DOI: 10.14778/3801059.3801069
- **Patricia Guerra-Balboa**, Annika Sauer, and Thorsten Strufe. "Analysis and Measurement of Attack Resilience of Differential Privacy". In: ACM Workshop on Privacy in the Electronic Society (WPES), 2024, DOI: 10.1145/3689943.3695046.

As we discussed in Sections 2.3 and 3.5, despite the solid theoretical foundation of DP, a central practical question remains: How do these formal parameters, especially ε, translate into concrete protection against real-world attacks? [128] This question is critical for calibrating ε: if set too high, sensitive information may be exposed; if too low, utility is unnecessarily compromised. Furthermore, understanding this relationship is essential for DP auditing, which seeks to empirically estimate privacy [134], evaluate the tightness of DP mechanisms [60], and identify implementation errors [135]. The significance of DP auditing extends beyond identifying deliberate misbehavior—for instance, organizations that falsely claim to protect user data with DP—by also revealing unintended errors and design oversights, which are prevalent in practice, as evidenced by our trajectory privacy survey presented in Chapter 3.

Motivated by its applications in noise calibration and auditing, there is growing interest in the data management community in risk assessment for DP mechanisms [56], [136], [137], [138]. As discussed in Section 2.1, the definition of DP aims to make the scenario in which a particular target record is included in a dataset indistinguishable from the scenario in which it is replaced by another record. This establishes a direct connection between DP parameters and *membership inference attacks* (MIAs), whose objective is precisely to infer whether a target record participated in the dataset. Consequently, significant progress has been made in linking DP guarantees to the risk of MIAs. [33], [54], [56], [59], even enabling direct noise calibration for desired MIA risk levels [139] without explicitly choosing ε. However, MIAs capture only one aspect of privacy risk and may be less relevant in real use-cases as discussed in Section 3.2. In particular, *attribute inference attacks* (AIAs) [54], which can expose sensitive information even when membership is public [39], remain less understood. Recently, *data reconstruction attacks*

(DRAs) [39] were proposed as a unifying framework subsuming both MIAs and AIAs, while also accounting for partial or imperfect reconstruction, e.g., revealing a car's license plate may suffice to compromise privacy even if the background image is inaccurate.

Balle et al. [39] introduced the first metric for DRAs, *reconstruction robustness* (ReRo), providing a pioneering unified view of DP attack resilience. ReRo was foundational, but has limitations as a comprehensive adversarial metric. First, ReRo and existing bounds [39], [62] assume attackers have no target-specific auxiliary knowledge, ignoring partial information such as demographic attributes or social media data—information that real-world attacks often exploit [9], [10], [72]. Second, ReRo is defined as a success probability; therefore, it considers an attack successful whenever the adversary correctly reconstructs the target, regardless of the role played by the data release in that success. For instance, an adversary may guess correctly based solely on public knowledge, independently of the dataset. Such a success does not constitute a privacy risk derived from participating in a data release, since it would have occurred even without participating. Nevertheless, ReRo would still count this as a privacy breach. As a consequence, this notion may penalize mechanisms for revealing global statistical information—which is, in fact, the primary goal of data release—and may incorrectly interpret success achieved through statistical imputation as participation risk [23], [140]. This incorrect assessment can lead to unnecessary utility loss when ReRo is used for noise calibration, as to compensate for this perceived risk we may add excessive noise, even though the actual privacy risk does not warrant such degradation.

Given these theoretical limitations, this thesis aims to empirically investigate the impact of potentially misleading ReRo assessments in realistic attack scenarios and to develop more precise metrics and bounds. In particular, we seek to establish a connection between DP and an attack metric that is sufficiently general to encompass data reconstruction attacks (DRA), while accurately measuring participation risk. A suitable metric should account for auxiliary knowledge, as well as imputation and prior distribution effects, thereby enabling a more principled and precise calibration of the privacy parameters.

Moreover, to better understand the attack-mitigation properties of DP and enable noise injection calibrated to a participant's true risk of information disclosure, we aim to establish bounds that connect DP mechanisms and their privacy parameters directly to the mitigation of real attacks. Specifically, we focus on providing: (i) a worst-case bound that is independent of the attacker's auxiliary knowledge, and (ii) an auxiliary-dependent bound enabling tighter analysis when the attacker's auxiliary information is known.

For the application of DP attack-mitigation bounds in noise calibration and auditing, the tightness of the bounds is crucial. Overestimating the risk can lead to unnecessary noise injection and inaccurate parameter estimation. To assess tightness, we construct and prove the optimal attack strategy for any reconstruction goal, auxiliary knowledge, and mechanism. This strategy also serves as a practical tool for DP auditing.

Moreover, to address several practical scenarios and broaden the applicability of our results, we aim to consider a wide spectrum of attackers and assumptions. In particular, we seek to derive bounds that are as tight as possible in fully white-box settings, where the attacker's knowledge and the DP mechanism is fully known. Additionally, we provide

fallback bounds for cases where full access to the mechanism is not available, hence the risk must be bounded solely according to the mechanism's privacy parameters.

Overall, this chapter aims to advance the understanding of how DP mitigates attacks in realistic scenarios. In particular, we investigate the extent to which mechanisms with the same ε can result in markedly different levels of adversarial advantage, highlighting that ε alone does not fully capture the effective privacy risk.

Finally, we investigate the applications of DP bounds for attack mitigation in both noise calibration and DP auditing. We analyze whether these bounds can reduce the noise required compared to existing ReRo-based methods and validate these improvements through extensive experimental evaluation. In particular, we aim to obtain a universally tight bound that implies an optimal calibration method: for a given mechanism, no greater utility can be achieved for a specified risk threshold than when using our bound. We further study how optimal attacks and bounds can enhance existing DP auditing frameworks [40], [41] by capturing all reconstruction risks and providing more accurate and actionable privacy assessments. Specifically, our goal is to obtain precise estimates and a data-agnostic method that does not require expending resources to first identify the worst-case pair of datasets.

Summarizing, this chapter focuses on addressing the following questions:

- Does ReRo and their bounds adequately capture reconstruction success arising from imputation and target-specific auxiliary knowledge in real scenarios?
- Can we define a consistent and unified reconstruction risk metric that naturally incorporates auxiliary knowledge?
- How do DP mechanisms limit attack performance for different attacker assumptions?
- What is the optimal attack strategy for any given reconstruction objective, mechanism, and prior distribution, and can its optimality be formally proven?
- Can we develop an auditing framework providing broader threat analysis, scalability and more accurate privacy budget estimation than existing auditing techniques?

Following the discussion in Section 2.3, we focus our analysis on bounded DP and consider, for any target record x, an informed adversary [39]. In this chapter, we operate under the independence assumption and adopt the most general DRA framework to provide a unified perspective on the attack pipeline.

5.1. Review of the Related Work

In this section, we review the relevant previous work on measuring the effective attack resilience of DP mechanisms for calibration and auditing, discussing novel insights and gaps that motivate our work.

Attack-based DP noise calibration. Several recent studies [56], [139], [141] demonstrate that calibrating DP noise based on resilience to specific attacks can significantly improve utility. Such approaches, however, primarily target MIAs, which leads to unnecessary utility degradation without offering meaningful privacy benefits when membership is public or considered non-sensitive [39].

Beyond MIAs, privacy concerns often involve AIA, where the adversary aims to infer sensitive attributes of individuals from released data [142], [143]. A common metric for evaluating such attacks is the attribute advantage [54]. Existing works that provide theoretical bounds for AIAs either analyze specific attack strategies [54] or adopt more general DRA frameworks [39], [57]. Within the latter, the notion of ReRo has emerged as the metric for measuring the risk of DRAs, under which attribute inference can be modeled as a special case [39]. Moreover, Balle et al.[39] and Hayes et al.[62] provide novel bounds relating DP parameters and ReRo enabling ReRo-based DP noise calibration.

A note on limitations of ReRo. Balle et al.'s work introducing ReRo and establishing its connection to DP was pioneering, as it provided a tangible risk assessment framework for DRAs under DP. Moreover, it broadened the perspective of the community by highlighting that privacy risks extend MIAs. While ReRo is appropriate in settings where we can confidently assume that the adversary's reconstruction capability derives entirely from the record participation, extending it to more general scenarios introduces significant limitations.

A general-purpose risk metric would be expected to cover all relevant attack scenarios. However, ReRo bounds do not formally account for the impact of target-specific auxiliary knowledge, hence excluding MIAs, AIAs and targeted DRAs as introduced in Section 2.3. Formally, the attack A considered in the original ReRo model [39] (see Definition 2.18 for details), only has access to the mechanism output $\mathcal{M}(D)$, i.e., $A\colon \Theta \to \mathcal{D}(\mathcal{X})$, implying that $\Pr(A(\mathcal{M}(D), a(x)) \in S) = \Pr(A(\mathcal{M}(D), a(x')) \in S)$ for any pair of possible targets x, x' and output set S. Under this assumption, the attacker A cannot adapt its strategy to a specific target x. This choice is reasonable for attacks that attempt to reconstruct a record without relying on auxiliary knowledge, such as the trajectory reconstruction studies discussed in Section 3.2. However, it fundamentally prevents assessing the risk of MIA and AIA, as they use full or partial knowledge of some target records. This is a relevant limitation since most real-world privacy attacks historically exploit publicly available information about the target [9], [10], [11]. Moreover, we show in Section 5.2 that several attacks leverage target-specific auxiliary knowledge, and their success highly depends on it.

All formal bounds connecting ReRo and DP were proven under this restrictive exclusion. The requirement that the attack depends only on $\mathcal{M}(D)$, ignoring target-specific information, is critical to establishing both Equations (2.3) and (2.6). This is not merely a theoretical limitation: we show in Section 5.5 that these bounds do not hold for attacks that exploit target-specific knowledge against well-known mechanisms such as DP-SGD.

A direct extension of ReRo to targeted attacks $A(\theta, a(x))$ may also lead to problematic assessments: Not only do the original bounds no longer hold, but the metric also collapses to a substantial overestimation of risk due to imputation and background knowledge.

For instance, the trivial MIA, $A(\theta, x) = x$, has success probability 1, which ReRo would interpret as a catastrophic privacy risk, even though no actual leakage occurs. This is not a negligible edge case; it has caused misleading overestimation of risk in black-box attacks on classification models [58], where much of the reported success arose from data imputation rather than exploiting the mechanism's output. Such overestimation obscures the true leakage and can lead to unnecessary utility loss when ReRo is used to calibrate noise in DP.

Even under the original assumption that the attacker has no target-specific knowledge, ReRo still overestimates risk, as we discussed in our preliminary work [57]. The mechanism output $\mathcal{M}(D)$ inherently reveals distributional information and population-level statistics, which are the primary goals of any learning process. This information can be used to perform imputation and infer attributes of individual records—even those not in D—with high accuracy, particularly when strong correlations exist (e.g., smoking correlating with cancer). In this case, the apparent attack success is driven by statistical inference rather than actual privacy violations, a phenomenon often referred to as a *privacy fallacy* [15], [23]. Indeed, several works establish that it is impossible to simultaneously provide utility and eliminate absolute information gain [15], [23].

We conclude that, while foundational, ReRo may be misleading as a general-purpose attack resilience metric, as it overlooks key statistical phenomena that distort privacy risk assessment, such as data imputation and targeted attacks. Both cases are very common and have an impact in practice (see Section 5.5), motivating the need for a novel framework to more accurately assess the risk of DP mechanisms with respect to attacks.

DP auditing. DP auditing [144] seeks to demonstrate tight estimates of the privacy budget, discover implementation flaws, and estimate empirical privacy. However, auditing in practice remains a significant challenge. For instance, implementation bugs or design flaws can severely degrade privacy guarantees in ways that are not immediately obvious. To address this, black-box discovery methods such as DP-Sniper [42] and Eureka [43] have been developed to detect DP violations by training classifiers to distinguish between mechanism outputs from "worst-case" neighboring inputs. However, DP-Sniper and its predecessor Eureka were both designed for *continuous or low-dimensional* mechanisms such as Laplace, Geometric, and Sparse Vector. Their methodology fundamentally relies on training machine-learning classifiers to distinguish between outputs of a mechanism under two adjacent inputs. This methodology implicitly assumes that the mechanism's output distribution lies in a low-dimensional, learnable representation. In contrast, in *frequency-oracle protocols for categorical domains* (e.g., GRR, SS, OUE), the input space has size k (often dozens, hundreds, or thousands of categories). For such mechanisms, the output distributions are not low-dimensional vectors but discrete randomized encodings whose structure is combinatorial rather than continuous [136]. Consequently, the learned classifiers fail to scale as the domain dimension grows.

Beyond identifying bugs, existing empirical privacy auditing approaches primarily focus on MIAs [56], [134], [145], [146], which limits their ability to detect broader forms of privacy leakage. Some auditing techniques extend beyond MIAs to consider AIAs, but these are restricted to specific contexts—such as Label DP [147] or synthetic data

generation [148]. In the LDP setting, the state-of-the-art framework LDP Auditor [41] relies specifically on perfect reconstruction without target-specific auxiliary knowledge for auditing.

Summarizing, despite its practical importance, no existing auditing framework incorporates auxiliary information or supports a DRA-based analysis and enables systematic evaluation across diverse DP mechanisms. This gap motivates the development of a general auditing methodology designed to capture realistic adversaries and to quantify broader classes of privacy risks.

5.2. Reconstruction Advantage

In this section, we introduce reconstruction advantage (RAD) as a novel, unifying metric for adversarial risk assessment. We first establish a worst-case bound on RAD that holds for any mechanism, data distribution, and auxiliary knowledge, ensuring robustness when the attacker's prior knowledge is unknown. We then refine this result by deriving a tighter bound under known auxiliary knowledge and prove its tightness by constructing the corresponding optimal attack that achieves it. Together, these results provide a noise calibration method to optimize utility for a given risk. We empirically validate the practical tightness of our bounds in Section 5.5.

In order to address ReRo's lack of accounting for the impact of target-specific auxiliary knowledge, we explicitly incorporate this concept into RAD. Formally, each record $x \in \mathcal{X}$ may be associated with target-specific auxiliary information $a(x) \in aux$. The auxiliary information can take different forms. For instance, in the classical AIA setting, where records are pairs $x = (z, y)$, one may define $a(x) = z$ and attempt to infer y. Alternatively, in the image reconstruction setting, the target may be the full record x, while $a(x)$ could correspond to a label such as "image of a person" or "image of an animal". The only structural assumption we impose is that the type of auxiliary information is consistent across all records: if $a(x)$ corresponds to a set of pixels, then for any other record x', $a(x')$ must also be a set of pixels (and not, for example, a semantic label). Having established this formalization, we are now in a position to introduce our metric.

Definition 5.1 (η-RAD)**.** Let π be a prior over $\mathcal{X}$, $\ell\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}_{\geq 0}$ an error function, and $a(x) \in aux$ the target-specific auxiliary information for each $x \in \mathcal{X}$. Given a mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{D}(\Theta)$, any dataset $D_- \in \mathcal{X}^{n-1}$ and any adversary $A : \Theta \times aux \to \mathcal{D}(\mathcal{X})$ we define the η-reconstruction advantage, η-RAD, as

$$\eta\text{-RAD} = \Pr_{\substack{X_1\sim\pi \\ \theta\sim\mathcal{M}(D_{X_1})}}[\ell(X_1, A(\theta, a(X_1))) \leq \eta] - \Pr_{\substack{X_0,X_1\sim\pi \\ \theta\sim\mathcal{M}(D_{X_0})}}[\ell(X_1, A(\theta, a(X_1)) \leq \eta].$$

RAD explicitly accounts for target-specific auxiliary knowledge, providing a generalization of the membership and attribute advantages to arbitrary reconstruction attacks. Importantly, RAD outputs values between -1 and $(1 - \kappa_\pi) \leq 1$ where $\kappa_\pi = \Pr_{X,X'\sim\pi}[X = X']$, i.e., the probability of resampling from the distribution π,

analogously to membership and attribute advantage (see Section 3.2). Intuitively, this reflects the fact that if dataset members are drawn from a finite universe, when we randomly sample a record from the universe to simulate non-members, there is a probability, κ_π, that it coincides with the record of the actual participant.

Intuitively, RAD measures the increase in the attacker's success probability that arises solely from the target's participation in the private learning process. In this way, RAD avoids the overestimation of risk that is inherent in ReRo. If RAD ≤ 0, participation carries no risk, since the attacker's probability of correctly reconstructing the record is no greater than if the individual had not participated. Larger values of RAD indicate higher participation risk. In the extreme case where RAD $= 1 - \kappa_\pi$, participation entails absolute risk: the attacker always succeeds in reconstructing the participant's record, while no sensitive information can be reconstructed from non-participants.

Previous bounds for ReRo assume that DRAs perform equally for every target. This assumption holds when the adversary has no target-specific auxiliary knowledge ($aux = \{\varnothing\}$), but breaks once aux is available: for instance, knowing that a target's surname is "Smith" might give less information than knowing that it is "Sainthorpe-Burton", as the latter is less frequent and hence carries more information. Such differences are not captured by ReRo, nor reflected in the proofs of the corresponding bounds [39], [57]. Hence, we provide the first theoretical bound that explicitly accounts for aux and covers any possible attack from MIAs to the most general DRAs:

Theorem 5.2 ((ε, δ)-DP implies η-RAD)**.** *Let $\pi, \ell, \eta \geq 0$ as in Def. 5.1, and $\kappa_\pi = \Pr_{X,X' \sim \pi}[X = X']$. If a mechanism $\mathcal{M} \colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ satisfies (ε, δ)-DP, then for any attack $A \colon \Theta \times aux \to \mathcal{D}(\mathcal{X})$, and database D_- we have*

$$\eta\text{-RAD} \leq \mathrm{TV}(\mathcal{M})(1 - \kappa_\pi) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}(1 - \kappa_\pi).$$

Proof. We use $\int f(x)\,\mathrm{d}\mu(x)$ as unified notation that represents either a sum (if μ is the counting measure) or an integral (if μ is the Lebesgue measure), aggregating both the discrete and continuous case in one single notation.

First, note that for every $x \in \mathcal{X}$ and target-specific knowledge $a(x)$, any attack admits the representation $A(\mathcal{M}(D), a(x)) \equiv \mathcal{A}_x(\mathcal{M}(D))$, verifying

$$p_{\mathcal{A}_x}(s \mid D) \equiv p_A(s \mid a(x), D) = \int_\Theta p_{\mathcal{M}}(\theta \mid D)\, p_A(s \mid \theta, a(x))\,\mathrm{d}\mu(\theta).$$

Note that the attack outputs values in $\mathcal{X}$. Therefore,

$$\begin{aligned}
\mathrm{TV}(\mathcal{A}_x(D), \mathcal{A}_x(D')) &\coloneqq \sup_{S \subseteq \mathcal{X}} \left| \Pr(\mathcal{A}_x(D) \in S) - \Pr(\mathcal{A}_x(D') \in S) \right| \\
&= \tfrac{1}{2} \int_{\mathcal{X}} \Big| p_A(s \mid \mathcal{M}(D), a(x)) - p_A(s \mid \mathcal{M}(D'), a(x)) \Big|\,\mathrm{d}\mu(s) \qquad (5.1) \\
&= \tfrac{1}{2} \int_{\mathcal{X}} \Big| \int_\Theta p_A(s \mid \theta, a(x)) \left(p_{\mathcal{M}}(\theta \mid D) - p_{\mathcal{M}}(\theta \mid D') \right) \mathrm{d}\mu(\theta) \Big|\,\mathrm{d}\mu(s) \\
&\leq \tfrac{1}{2} \int_{\mathcal{X}} \int_\Theta p_A(s \mid \theta, a(x)) \left| p_{\mathcal{M}}(\theta \mid D) - p_{\mathcal{M}}(\theta \mid D') \right| \mathrm{d}\mu(\theta) \Big|\,\mathrm{d}\mu(s) \qquad (5.2)
\end{aligned}$$

$$
\begin{aligned}
&= \tfrac{1}{2} \int_{\Theta} \left| p_{\mathcal{M}}(\theta \mid D) - p_{\mathcal{M}}(\theta \mid D') \right| \mathrm{d}\mu(\theta) \int_{\mathcal{X}} p_A(s \mid \theta, a(x)) \, \mathrm{d}\mu(s) \\
&= \tfrac{1}{2} \int_{\theta} \left| p_{\mathcal{M}}(\theta \mid D) - p_{\mathcal{M}}(\theta \mid D') \right| \mathrm{d}\mu(\theta) \\
&= \mathrm{TV}(\mathcal{M}(D), \mathcal{M}(D')), \qquad (5.3)
\end{aligned}
$$

where Equation (5.1) follows from [81, Proposition 4.2, p. 48] and Equation (5.2) from Minkowski's inequality.

Moreover, given any success set $S_\eta(x) = \{x' \in \mathcal{X} \colon \ell(x, x') \leq \eta\}$, and using the notation $A(\mathcal{M}(D), a(x)) \equiv \mathcal{A}_x(\mathcal{M}(D))$, we have

$$
\Pr_{\substack{X_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{X_0})}} [\ell(X_1, A(\theta, a(X_1))) \leq \eta] = \Pr_{X_1 \sim \pi} [\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)].
$$

Hence, applying Equation (5.3) and Definition 2.24 to RAD Definition 5.1 we obtain:

$$
\begin{aligned}
\eta\text{-RAD} &= \Pr_{X_1 \sim \pi} [\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr_{X_0, X_1 \sim \pi} [\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \\
&= \mathop{\mathbb{E}}_{X_0 \sim \pi} \left[\Pr_{X_1 \sim \pi} [\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr_{X_1 \sim \pi} [\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \right] \\
&= \mathop{\mathbb{E}}_{X_0, X_1 \sim \pi} \left[\mathbf{1}_{\{X_0 \neq X_1\}} \left(\Pr[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \right) \right] \\
&\overset{\text{Eq.5.3}}{\leq} \mathrm{TV}(\mathcal{M}) \mathop{\mathbb{E}}_{X_0, X_1 \sim \pi} \left[\mathbf{1}_{\{X_0 \neq X_1\}} \right].
\end{aligned}
$$

The result follows from the fact that $\mathbb{E}_{X_0, X_1 \sim \pi} \left[\mathbf{1}_{\{X_0 \neq X_1\}} \right] = 1 - \sum_z \pi_z^2$ for discrete variables and 1 for continuous ones. Finally, Equation (2.5) completes the proof. □

Note that in the discrete case, $\kappa_\pi = \sum_x \pi_x^2$, which is maximized when π is uniform over two possible records (e.g., $\pi = U\{X_0, X_1\}$). In the continuous case, the resampling probability is, by definition, zero. Consequently, the result simplifies to η-RAD $\leq \mathrm{TV}(\mathcal{M})$, unaffected by the prior distribution.

Theorem 5.2 is the first bound for RAD under the strongest threat model, where the attacker may leverage auxiliary knowledge. Particularly, this results states that if the mechanism is fully known, we can determine the attack mitigation it provides by its total variation. When only the DP parameters ε and δ are available, the bound quantifies how each parameter contributes to the attacker's advantage. Consequently, this theorem serves as a key tool for DP noise calibration, improving on ReRo by encompassing a broader spectrum of potential attackers.

Moreover, Theorem 5.2 allows upper bounding RAD under composition. As we discussed in Section 2.3.1.1, given $\mathrm{TV}(\mathcal{M}_i) = \Delta$, the T-adaptive composition satisfies $\mathrm{TV}(M) \leq (1 - (1 - \Delta)^T)$. Hence, η-RAD $\leq (1 - (1 - \Delta)^T)(1 - \kappa_\pi)$.

Theorem 5.2 does not depend on the attacker's auxiliary knowledge. Therefore, the same bound holds whether the attacker has no auxiliary information ($aux = \{\varnothing\}$) or complete knowledge of the record ($a(x) = x$), since the result is derived in a worst-case

manner. However, when the attacker's goal is to reconstruct an entire record (as in DRA) or infer parts of it (as in AIA), it is unreasonable to assume that the attacker already knows the full record ($a(x) = x$)—as assumed for MIA. Therefore, we next provide a tighter bound that explicitly incorporates the target-specific auxiliary knowledge.

In order to simultaneously cover continuous and discrete distributions, we use the Lebesgue–Stieltjes formulation (following the same strategy as in Chapter 4). That is, we state the theorem with respect to the prior probability measure P on $\mathcal{X}$ and let μ be a reference measure (i.e., the counting measure in the discrete case, and the Lebesgue measure in the continuous case). If P is absolutely continuous with respect to μ and $\pi = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ is the Radon–Nikodym derivative, then for every measurable set $B \subseteq \mathcal{X}$ we have:

$$P(B) = \int_B \mathrm{d}P = \int_B \pi_x \,\mathrm{d}\mu(x),$$

which reduces to $P(B) = \sum_{x \in B} \pi_x$ in the discrete case, and $P(B) = \int_B \pi_x \,\mathrm{d}x$ in the continuous case. Similarly, the randomize mechanism $\mathcal{M}$ generates a probability distribution,

$$\Pr[\mathcal{M}(D) \in S] = \int_S \mathrm{d}P_{\mathcal{M}(D)} = \int_S p_{\mathcal{M}}(\theta \mid D) \,\mathrm{d}\mu(\theta).$$

Given the established notation, we state the following result, which provides an upper bound on the RAD of any attacker with prior distribution P and auxiliary-information function $a : \mathcal{X} \to aux$, which maps each target x to the auxiliary knowledge $a(x)$ available to the attacker.

Theorem 5.3. *Given $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$, P the prior probability measure on $\mathcal{X}$ and $a\colon \mathcal{X} \to aux$ measurable, then for any attack $A\colon \Theta \times aux \to \mathcal{D}(\mathcal{X})$, we have*

$$\eta\text{-RAD} \leq \int_\Theta \int_{aux} \max_{x_\theta \in \mathcal{X}} \left(\int_{S^z_\eta(x_\theta)} w(\theta, x) \,\mathrm{d}P_z(x) \right) \mathrm{d}\nu(z) \,\mathrm{d}\mu(\theta),$$

where $w(x, \theta) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta)$, $S^z_\eta(x_\theta) = \{x\colon a(x) = z \wedge \ell(x_\theta, x) \leq \eta\}$, $\nu(z) = P \circ a^{-1}(z)$ and P_z the disintegration theorem measure. Additionally, μ denotes the counting (or Lebesgue) measure in the discrete (or continuous) case. The discrete case simplifies to

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in aux} \max_{x_\theta \in \mathcal{X}} \sum_{\substack{\ell(x, x_\theta) \leq \eta \\ a(x) = z}} w(\theta, x) \pi_x$$

by direct application of Remark 2.29.

Proof. Following the notation introduced in Theorem 5.2, we consider $A(\mathcal{M}(D), a(x)) \equiv \mathcal{A}_x(\mathcal{M}(D))$. Moreover, denoting $\Pr[A(\theta, z) \in S] \equiv \Pr_A[S \mid \theta, z]$, for any $z \in aux$ and $\theta \in \Theta$,

$$\Pr_A[S \mid \theta, z] = \int_S \mathrm{d}P_{A(\theta,z)} = \int_S p_A(x \mid \theta, z) \,\mathrm{d}\mu(x) = \int_{\mathcal{X}} \mathbf{1}_{\{x \in S\}} p_A(s \mid \theta, x) \,\mathrm{d}\mu(x), \quad (5.4)$$

for μ the reference measure as previously defined. First, we rewrite RAD definition as

$$
\begin{aligned}
\eta\text{-RAD} &= \Pr_{X_1\sim\pi}[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr_{X_0,X_1\sim\pi}[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \\
&= \mathop{\mathbb{E}}_{X_0\sim\pi}\Big[\Pr_{X_1\sim\pi}[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr_{X_1\sim\pi}[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]\Big] \\
&= \mathop{\mathbb{E}}_{X_0,X_1\sim\pi}\Big[\mathbf{1}_{\{X_0\neq X_1\}}\big(\Pr[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]\big)\Big] \\
&= \mathop{\mathbb{E}}_{X_0,X_1\sim\pi}\Big[\mathbf{1}_{\{X_0\neq X_1\}}\int_\Theta \Pr_A\left(S_\eta(X_1) \mid \theta, a(X_1)\right)\Big(p_{\mathcal{M}}(\theta \mid D_{X_1}) - p_{\mathcal{M}}(\theta \mid D_{X_0})\Big)\,\mathrm{d}\mu(\theta)\Big] \\
&= \mathop{\mathbb{E}}_{X_1\sim\pi}\Big[\int_\Theta \Pr_A\left(S_\eta(X_1) \mid \theta, a(X_1)\right) \mathop{\mathbb{E}}_{X_0\sim\pi}\Big[\mathbf{1}_{\{X_0\neq X_1\}}\big(p_{\mathcal{M}}(\theta \mid D_{X_1}) - p_{\mathcal{M}}(\theta \mid D_{X_0})\big)\Big]\,\mathrm{d}\mu(\theta)\Big] \\
&= \mathop{\mathbb{E}}_{X_1\sim\pi}\Big[\int_\Theta \Pr_A\left(S_\eta(X_1) \mid \theta, a(X_1)\right) \\
&\qquad\qquad \cdot \Big(p_{\mathcal{M}}(\theta \mid D_{X_1}) \mathop{\mathbb{E}}_{X_0\sim\pi}[\mathbf{1}_{\{X_0\neq X_1\}}] - \mathop{\mathbb{E}}_{X_0\sim\pi}[\mathbf{1}_{\{X_0\neq X_1\}} p_{\mathcal{M}}(\theta \mid D_{X_0})]\Big)\,\mathrm{d}\mu(\theta)\Big] \\
&= \mathop{\mathbb{E}}_{X_1\sim\pi}\Big[\int_\Theta \Pr_A\left(S_\eta(X_1) \mid \theta, a(X_1)\right) \underbrace{\left(p_{\mathcal{M}}(\theta \mid D_{X_1}) - p_{\mathcal{M}}(\theta)\right)}_{w(X_1,\theta)}\,\mathrm{d}\mu(\theta)\Big] \qquad (5.5) \\
&= \int_{\mathcal{X}}\int_\Theta \Pr_A\left(S_\eta(x_1) \mid \theta, a(x_1)\right) w(x_1,\theta)\,\mathrm{d}\mu(\theta)\,\mathrm{d}P(x),
\end{aligned}
$$

where Equation (5.5) follows trivially for the continuous case, since $\mathbb{E}_{X_0\sim\pi}\left[\mathbf{1}_{\{X_0\neq X_1\}}\right] = 1$, and for the discrete one, since

$$
\begin{aligned}
&p_{\mathcal{M}}(\theta \mid D_{X_1}) \mathop{\mathbb{E}}_{X_0\sim\pi}\left[\mathbf{1}_{\{X_0\neq X_1\}}\right] - \mathop{\mathbb{E}}_{X_0\sim\pi}\left[\mathbf{1}_{\{X_0\neq X_1\}} p_{\mathcal{M}}(\theta \mid D_{X_0})\right] \\
&= p_{\mathcal{M}}(\theta \mid D_{X_1})(1-\pi_1) - \mathop{\mathbb{E}}_{X_0\sim\pi}[p_{\mathcal{M}}(\theta \mid D_{X_0})] + p_{\mathcal{M}}(\theta \mid D_{X_1})\pi_1 \\
&= p_{\mathcal{M}}(\theta \mid D_{X_1}) - p_{\mathcal{M}}(\theta).
\end{aligned}
$$

Moreover, for all records, x_1, x_2, with the same auxiliary knowledge, i.e., $a(x_1) = a(x_2) = z$, and for any fixed output θ, we have that

$$
\Pr_A(S_\eta(x_1) \mid \theta, a(x_1)) = \Pr_A(S_\eta(x_1) \mid \theta, a(x_2)) = \Pr_A(S_\eta(x_1) \mid \theta, z).
$$

Hence, given $a^{-1}(z) = \{x \colon a(x) = z\}$ for all $z \in aux$, and $\nu(z) = P \circ a^{-1}(z)$, applying disintegration theorem (Cf. Theorem 2.28) there exists a unique measure P_z such that

$$
\begin{aligned}
\eta\text{-RAD} &= \int_{\mathcal{X}}\int_\Theta \Pr_A(S_\eta(x) \mid a(x), \theta)\, w(x,\theta)\ \mathrm{d}\mu(\theta)\,\mathrm{d}P(x) \\
&= \int_\Theta \int_{aux}\int_{a^{-1}(z)} \Pr_A(S_\eta(x) \mid z, \theta)\, w(x,\theta)\ \mathrm{d}P_z(x)\,\mathrm{d}\nu(z)\,\mathrm{d}\mu(\theta)
\end{aligned}
$$

Combining the previous equation with Equation (5.4) we obtain

$$
\begin{aligned}
\eta\text{-RAD} &= \int_\Theta \int_{aux}\int_{a^{-1}(z)} \left(\int_{\mathcal{X}} \mathbf{1}_{\{\ell(z,\tilde{x})\le\eta\}}\, p_A(\tilde{x} \mid z, \theta)\,\mathrm{d}\mu(\tilde{x})\right) w(x,\theta)\ \mathrm{d}P_z(x)\,\mathrm{d}\nu(z)\,\mathrm{d}\mu(\theta) \\
&= \int_\Theta \int_{aux}\int_{\mathcal{X}} p_A(\tilde{x} \mid z, \theta) \int_{a^{-1}(z)} \mathbf{1}_{\{\ell(x,\tilde{x})\le\eta\}}\, w(x,\theta)\ \mathrm{d}P_z(x)\,\mathrm{d}\mu(\tilde{x})\,\mathrm{d}\nu(z)\,\mathrm{d}\mu(\theta)
\end{aligned}
$$

$$\leq \int_\Theta \int_{aux} \int_{\mathcal{X}} p_A(\tilde{x} \mid z, \theta) \max_{x_\theta \in \mathcal{X}} \int_{a^{-1}(z)} \mathbf{1}_{\{\ell(x,x_\theta)\leq\eta\}}\, w(x,\theta)\ \mathrm{d}P_z(x)\, \mathrm{d}\mu(\tilde{x})\, \mathrm{d}\nu(z)\, \mathrm{d}\mu(\theta)$$

$$= \int_\Theta \int_{aux} \max_{x_\theta \in \mathcal{X}} \int_{a^{-1}(z)} \mathbf{1}_{\{\ell(x,x_\theta)\leq\eta\}} w(x,\theta)\, \mathrm{d}P_z(x) \left(\int_{\mathcal{X}} p_A(\tilde{x} \mid z, \theta)\, \mathrm{d}\mu(\tilde{x}) \right) \mathrm{d}\nu(z)\, \mathrm{d}\mu(\theta)$$

$$= \int_\Theta \int_{aux} \max_{x_\theta \in \mathcal{X}} \int_{a^{-1}(z)\cap\{x\colon \ell(x,x_\theta)\leq\eta\}} w(x,\theta)\, \mathrm{d}P_z(x)\, \mathrm{d}\nu(z)\, \mathrm{d}\mu(\theta)$$

$$\int_\Theta \int_{aux} \max_{x_\theta \in \mathcal{X}} \int_{S_\eta^z(x_\theta)} w(x,\theta)\, \mathrm{d}P_z(x)\, \mathrm{d}\nu(z)\, \mathrm{d}\mu(\theta)$$

where $S_\eta^z(x_\theta) = \{x\colon a(x) = z \wedge \ell(x_\theta, x) \leq \eta\}$. □

Theorem 5.3 bounds RAD when the specific mechanism, $\mathcal{M}$, and auxiliary knowledge, aux, are known. At the same time, it becomes more precise than our worst-case bound Theorem 5.2, as we illustrate in Figure 5.1. Moreover, although this bound is inherently complex due to its generality, it admits simpler characterizations for commonly studied threat models—such as MIAs and DRAs where no target-specific auxiliary knowledge is assumed.

In particular, in an MIA, where the attacker has full-knowledge about the record and just seek to infer participation (i.e. $a(x) = x$ for all records), then $aux = \mathcal{X}$, the push-forward metric simplifies to $\nu(z) = P \circ a^{-1}(z) = P(z)$, and

$$P_z(\mathcal{X}\backslash a^{-1}(z)) = P_z(\mathcal{X}\backslash\{z\}) = 0 \Rightarrow P_{z'}(\{z\}) = \mathbf{1}_{\{z'=z\}} \equiv \delta_{z'}(z).$$

Therefore satisfying that, for any measurable set $B \subseteq \mathcal{X} (\equiv aux)$,

$$\int_{aux} P_z(B)\, \mathrm{d}\nu(z) = \int_{aux} \mathbf{1}_{\{z\in B\}}\, \mathrm{d}P(z) = P(B).$$

Then, according to the disintegration theorem (see Chapter 2), $\nu(z) = P(z)$, and $P_z = \delta_z$. Moreover, since a is the identity function,

$$S_\eta^z(x_\theta) = \{x\colon a(x) = z \wedge \ell(x, x_\theta) \leq \eta\} = \begin{cases} \{z\} & \text{if } \ell(z, x_\theta) \leq \eta, \\ \varnothing & \text{otherwise.} \end{cases}$$

Hence, applying our Theorem 5.3,

$$\eta\text{-RAD} \leq \int_\Theta \int_{aux} \max_{x_\theta \in \mathcal{X}} \int_{S_\eta^z(x_\theta)} w(x,\theta)\, \mathrm{d}P_z(x)\, \mathrm{d}\nu(z)\, \mathrm{d}\mu(\theta) \tag{5.6}$$

$$= \int_\Theta \int_{aux} \max_{\substack{x_\theta \in \mathcal{X} \\ \ell(x_\theta,z)\leq\eta}} \int_{\{z\}} \mathbf{1}_{\{\ell(x_\theta,x)\leq\eta\}} w(\theta,x)\, \mathrm{d}\delta_z(x)\, \mathrm{d}P(z)\, \mathrm{d}\mu(\theta) \tag{5.7}$$

$$= \int_\Theta \int_{aux} \max_{\substack{x_\theta \in \mathcal{X} \\ \ell(x_\theta,z)\leq\eta}} w(\theta,z)\, \mathrm{d}P(z)\, \mathrm{d}\mu(\theta), \tag{5.8}$$

$$= \int_\Theta \int_{\{z\colon w(\theta,z)>0\}} w(\theta,z)\pi_z\, \mathrm{d}\mu(z)\, \mathrm{d}\mu(\theta), \tag{5.9}$$

since $z \in \mathcal{X}$, $\arg\max_{x_\theta} = S_\eta(z)$ if $w(\theta, z) > 0$ and $\arg\max_{x_\theta} = \mathcal{X} \backslash S_\eta(z)$ otherwise, avoiding negative values. For discrete variable previous formula simplifies to

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{\substack{x \in \mathcal{X} \\ w(\theta,x)>0}} w(\theta, x)\pi_x. \tag{5.10}$$

We can consider the other extreme, when $aux = \{\varnothing\}$. Here, $\varnothing$ is treated as the single element in aux. To avoid confusion with properties of the empty set, we instead denote $aux = \{a\}$, meaning that the auxiliary function is constant for every user, i.e., $a^{-1}(a) = \mathcal{X}$. In this case, there is no target-specific auxiliary knowledge.

Consequently, $\nu(a) = P(a^{-1}(a)) = P(\mathcal{X}) = 1$. Hence, ν is the Dirac measure δ_a. The first condition defining P_a according to Theorem 2.28 is

$$P_a(\mathcal{X} \backslash a^{-1}(a)) = P_a(\mathcal{X} \backslash \mathcal{X}) = P_a(\varnothing) = 0,$$

which is satisfied by any measure by definition. Hence, we look to the second defining condition, for any measurable set $B \subseteq \mathcal{X}$

$$P(B) = \int_{\{a\}} P_a(B)\, \mathrm{d}\delta_a(a) = P_a(B), \tag{5.11}$$

hence, $P_a = P$, obtaining:

$$\begin{aligned}
\eta\text{-RAD} &\leq \int_\Theta \int_{aux} \max_{x_\theta \in \mathcal{X}} \left(\int_{S_\eta^a(x_\theta)} w(\theta, x)\, \mathrm{d}P_z(x) \right) \mathrm{d}\nu(z)\, \mathrm{d}\mu(\theta) \\
&= \int_\Theta \int_{\{a\}} \max_{x_\theta \in \mathcal{X}} \left(\int_{S_\eta^a(x_\theta)} w(\theta, x)\, \mathrm{d}P_a(z) \right) \mathrm{d}\delta_a(z)\, \mathrm{d}\mu(\theta) \\
&= \int_\Theta \max_{x_\theta \in \mathcal{X}} \left(\int_{S_\eta^a(x_\theta)} w(\theta, x)\, \mathrm{d}P(x) \right) \mathrm{d}\mu(\theta) \\
&= \int_\Theta \max_{x_\theta \in \mathcal{X}} \left(\int_{S_\eta^a(x_\theta)} w(\theta, x)\pi_x\, \mathrm{d}\mu(x) \right) \mathrm{d}\mu(\theta).
\end{aligned}$$

Note that, $S_\eta^a(x_\theta) = \{x \in a^{-1}(a) \colon \ell(x_\theta, z) \leq \eta\} = \{x \in \mathcal{X} \colon \ell(x_\theta, z) \leq \eta\} = S_\eta(x_\theta)$. Particularly, it simplifies for the discrete case to:

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \max_{x' \in \mathcal{X}} \sum_{\ell(x', x) \leq \eta} w(\theta, x)\pi_x. \tag{5.12}$$

Moreover, if $\eta = 0$ (perfect reconstruction), such as any AIA setting and the original ReRo setting [62]), Theorem 5.3 formula simplifies to:

$$0\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{z \in aux} \max_{\substack{a(x)=z \\ w(x,\theta)>0}} w(x, \theta)\pi_x. \tag{5.13}$$

Importantly, 0-RAD is consistently zero for continuous random variables by definition.

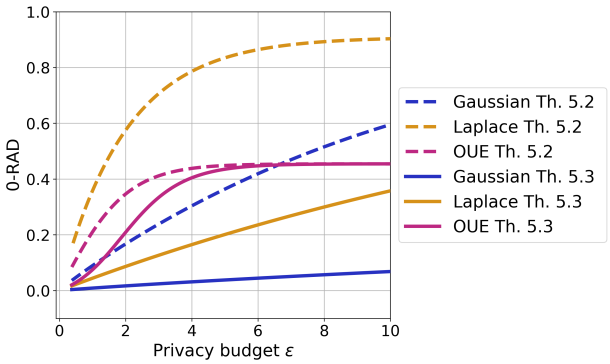


Figure 5.1.: Theorem 5.3 bound for different DP mechanisms with $|\mathcal{X}| = 11$, $aux = \{\varnothing\}$ and a uniform prior. Importantly, for the same ε, each mechanism offers different levels of attack mitigation, highlighting the need for RAD analysis as a complementary tool to traditional privacy parameters. Moreover, in all cases, we observe that the bound in Theorem 5.3 improves upon Theorem 5.2.

Finally, given $|\mathcal{X}| = m$ and $aux = \{\varnothing\}$, previous equation admits the simplification

$$0\text{-RAD} \leq \sum_{i=1}^{m} \left(\Pr_{\mathcal{M}}(\Theta_i \mid x_i) - \Pr_{\mathcal{M}}(\Theta_i) \right) \pi_i, \tag{5.14}$$

where $\Theta_1 = \{\theta \in \Theta \colon x_1 \in \arg\max_j w(\theta, x_j)\pi_j\}$ and for every $i \geq 1$, Θ_{i+1} is recursively defined as

$$\Theta_{i+1} = \{\theta \in \Theta \colon x_{i+1} \in \arg\max_j w(\theta, x_j)\pi_j\} \setminus \bigcup_{k=1}^{i} \Theta_k. \tag{5.15}$$

We illustrate the benefits of Theorem 5.3 for relevant DP mechanisms through the following examples and visualizations in Figures 5.1 and 5.6. To compute the RAD bounds for each mechanism, we directly apply the formula from Theorem 5.3 to the corresponding mechanism distribution. The full computational details are provided in Section A.3.

Example 5.4. The generalized randomized response mechanism (GRR) [149] is an LDP mechanism that outputs the true record x_1 with probability $p = \mathrm{e}^{\varepsilon}/(\mathrm{e}^{\varepsilon} + m - 1)$ and any other record $x_0 \neq x_1$ with probability $q = (\mathrm{e}^{\varepsilon} + m - 1)^{-1}$. Since, $p \geq q$ for all $\varepsilon \geq 0$,

$$w(\theta, x) = \begin{cases} (p - q)(1 - \pi_\theta) & \text{if } x = \theta \\ (q - p)\pi_\theta & \text{otherwise,} \end{cases} \tag{5.16}$$

and $w(x, \theta) > 0$ iff $x = \theta$. Hence, applying Theorem 5.3 for $a(x) = x$:

$$\eta\text{-RAD} \leq \sum_{\theta} (p - q)(1 - \pi_\theta)\pi_\theta = \frac{\mathrm{e}^{\varepsilon} - 1}{\mathrm{e}^{\varepsilon} + m - 1}(1 - \kappa_\pi) = \mathrm{TV}(1 - \kappa_\pi).$$

Hence, the advantage of an attacker only depends on the chosen ε, the total universe size $|\mathcal{X}| = m$ and the initial distribution. Particularly, given a maximum risk threshold RAD$\leq \gamma$, we can choose ε following:

$$\varepsilon = \ln \frac{1 + \gamma \frac{m-1}{1-\kappa_\pi}}{1 - \frac{\gamma}{1-\kappa_\pi}}.$$

For instance, to guarantee RAD below 0.1 on binary queries (with uniform prior) the user must set $\varepsilon = \ln(1.5) \approx 0.405$, while for the same RAD in a query with $m = 100$ possibilities must select $\varepsilon = \ln(12.2087) \approx 2.503$.

Now if we consider a reconstructions attack without target-specific auxiliary knowledge, i.e., $aux = \{\varnothing\}$, we obtain

$$\eta\text{-RAD} = (p - q)\left(1 - \sum_\theta \pi_\theta \inf_{\ell(x_\theta,\theta)\leq\eta} \Pr_{X\sim\pi}[\ell(X, x_\theta) \leq \eta]\right).$$

Hence, the advantage of such attacker is always less than one with full-knowledge. However, it is not much worse, since for instance considering $\eta = 0$ and a uniform distribution we get exactly the same formula.

Example 5.5. The optimal unary encoding (OUE) mechanism [150] maps each input $x \in \mathcal{X}$ to an m-dimensional one-hot binary vector and perturbs each bit independently. For each position $i \in [m]$, the obfuscated vector θ is sampled such that $\Pr[\theta_i = 1] = 1/2$ if $i = x$, and $\Pr[\theta_i = 1] = q = \frac{1}{e^\varepsilon+1}$ otherwise. Denoting $p = 1-q$, according to Theorem 5.3, we obtain that, for $a(x) = x$:

$$\eta\text{-RAD} \leq \frac{1}{2}\frac{e^\varepsilon - 1}{e^\varepsilon + 1}(1 - \kappa_\pi) = \text{TV(OUE)}(1 - \kappa_\pi).$$

First, note that for the same attack and prior distribution, OUE provides a different level of protection than GRR. In particular, while increasing ε in GRR always increases the attacker's advantage—approaching 1 as $\varepsilon \to \infty$—in the case of OUE, the attacker's advantage is upper bounded by 0.5, regardless of how large ε becomes. This illustrates that ε alone does not capture the full picture: mechanisms with the same ε can yield markedly different levels of attack mitigation.

If we consider $aux = \{\varnothing\}$, then the bound becomes:

$$0\text{-RAD} \leq \frac{p - q}{2p}\left(\sum_{i=1}^{m} p^{m-i}\pi_i(1 - \pi_i) - q\sum_{i=1}^{m} p^{m-i}\pi_i \sum_{z=1}^{i-1} \pi_z\right)$$

which in particular for $\pi = U[m]$:

$$0\text{-RAD} \leq \frac{(2p - 1)\left(1 - p^{m-1}\right)}{2m(1 - p)} = \frac{e^\varepsilon - 1}{2m}\left(1 - \left(\frac{e^\varepsilon}{1 + e^\varepsilon}\right)^{(m-1)}\right).$$

Note that when $\varepsilon \to \infty$ previous bound converges to $\frac{m-1}{2m}$, hence even if we keep reducing the noise (increasing ε), the attacker's advantage is limited. We plot this bound in Figure 5.1.

Example 5.6. In the subset selection mechanism (SS) [151] users report a subset $\theta \subseteq \mathcal{X} = \{x_1, \ldots, x_m\}$ containing their true value z with probability $p = \frac{\omega e^{\varepsilon}}{\omega e^{\varepsilon} + m - \omega}$, where $\omega = |\theta| = \max\left(1, \left\lfloor \frac{m}{e^{\varepsilon}+1} \right\rfloor\right)$. The subset is completed by sampling uniformly from $\mathcal{X} \setminus \{x\}$. According to Theorem 5.3 we obtain that for $\pi = U[m]$

$$0\text{-RAD} \leq \frac{pm - \omega}{m\omega}.$$

Once again, we obtain a direct formula to calibrate the mechanism parameters (in this case, p) to achieve a desired RAD. Furthermore, the protection offered by SS against reconstruction attacks lies between that of GRR, which provides weaker protection, and OUE, which provides stronger protection, as illustrated in Figure 5.7.

Example 5.7. The Laplace mechanism adds Laplace noise with scale $b = \Delta q/\varepsilon$ to the query value $q(D) \in \mathbb{R}$ [16]. If $\mathcal{X} = \{x_1, \ldots, x_m\}$ is uniformly distributed and $\Delta q = 1$ applying Theorem 5.3 we obtain

$$0\text{-RAD} \leq \frac{m-1}{m}\left(1 - e^{-\frac{\varepsilon}{2(m-1)}}\right).$$

First, we observe that the Laplace mechanism provides stronger protection against reconstruction attacks than OUE for small values of ε. For example, as shown in Figure 5.1, for all $\varepsilon \in [0, 14]$ the Laplace mechanism achieves lower RAD than OUE on a data domain with $|\mathcal{X}| = 11$.

Moreover, we derive a direct calibration method for the Laplace mechanism. As illustrated in Figure 5.2, calibrating ε according to the maximum admissible risk using our approach yields significantly higher accuracy compared to the state-of-the-art method based on ReRo.

Example 5.8. The Gaussian mechanism adds Gaussian noise $\mathcal{N}(0, \sigma)$ to the query value $q(D) \in \mathbb{R}$ [39]. Given Φ the CDF of the standard normal distribution, if $\mathcal{X} = \{x_1, \ldots, x_m\}$ is uniformly distributed and $\Delta q = 1$, applying Theorem 5.3 we obtain

$$0\text{-RAD} \leq \frac{m-1}{m}\left(2\Phi\Big(\frac{1}{2\sigma(m-1)}\Big) - 1\right).$$

We plot this bound in Figure 5.1 alongside the corresponding OUE and Laplace bounds under the same attack model. The comparison shows that the Gaussian mechanism provides substantially stronger protection against reconstruction attacks without auxiliary knowledge—for a universe of size 11 and a uniform prior—than both OUE and Laplace.

These examples highlight both the practical applicability of Theorem 5.3 for estimating reconstruction risk in real-world settings and the importance of conducting a dedicated RAD analysis of DP mechanisms. First, as illustrated in Figure 5.1, mechanisms with the same privacy parameters can provide substantially different levels of protection in terms of risk mitigation. This observation underscores the need for resilience analysis beyond the ε-based criterion. Moreover, these examples demonstrate that Theorem 5.3

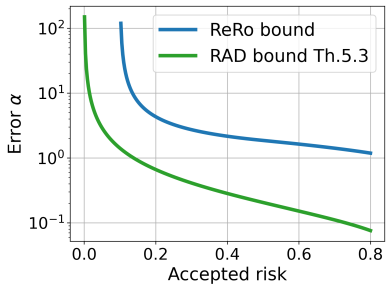


Figure 5.2.: Upper bound on the Laplace mechanism query error (utility) at 95% confidence when the noise is calibrated using ReRo vs. RAD. We see that for the same risk estimation, calibrating with using RAD improves utility.

Algorithm 1: Optimal Attack

Input : θ, η and $a(x) = z$
Output : $\tilde{x}$
Compute $a^{-1}(z) = \{x \colon a(x) = z\}$
for $x' \in \mathcal{X}$ **do**
  $\mathcal{W}_\eta^z(x') = \sum_{x \in a^{-1}(z) \colon \ell(x,x') \leq \eta} w(\theta, x)\pi_x$;
Select $\tilde{x} \in \arg\max_{x'} \mathcal{W}_\eta^z(x')$ (at random)

yields simple and explicit characterizations of the RAD for common DP mechanisms, directly relating their noise parameters to the resulting level of risk mitigation.

Moreover, in Figure 5.1 we see the improvement when we target specific auxiliary knowledge instead of using our worst-case bound (Theorem 5.2). Hence, Theorem 5.3 offers an improved noise calibration method to ensure protection against real attacks, when the auxiliary knowledge is well defined. For instance, when the entire record is considered private ($aux = \{\varnothing\}$); alternatively, when a specific attribute y is deemed sensitive, we consider all the remainder record public (we denote it as $a(x) = x \backslash y$).

Importantly, we illustrate in Figure 5.2 the utility gain of noise calibration using our RAD bounds compared to using the best existing ReRo bound [62], showing the benefit of our bounds for system design. Specifically, we consider $aux = \{\varnothing\}$—allowing comparison with [62]. We plot the upper bound on the Laplace mechanism's query error that can be guaranteed with 95% confidence, for $|\mathcal{X}| = 10$ and $\Delta = 1$, showing a substantial improvement in utility enabled by our RAD-based calibration.

Crucially, Theorem 5.3 is universally tight: for any mechanism and auxiliary knowledge, there exists an attack achieving the bound, so it cannot be further improved. We illustrate this by explicitly constructing such an attack in Algorithm 1, proving the existence of an optimal adversary for any auxiliary model.

The attack strategy is conceptually simple yet highly effective. We start with the case of $\eta = 0$, i.e., perfect reconstruction. In the fully informed setting—where the adversary knows the entire target record (as in an informed MIA)—the optimal strategy is to

declare the target a member whenever the mechanism's output provides any positive evidence of participation, that is, whenever

$$w(\theta, x) = p_{\mathcal{M}}(\theta \mid x) - p_{\mathcal{M}}(\theta) > 0.$$

Intuitively, if the observed output is more likely under the target record than under the prior distribution, the adversary should infer membership. If there is more than one candidate x sharing the same auxiliary knowledge, $a(x) = z$, (e.g., several users may shared a common attribute) the attacker can not optimize for all at the same time, therefore select $\tilde{x}$ such that it maximizes the posterior weight $w(\theta, x)\pi_x$, as long as it provides positive evidence. In the extreme case, when $aux = \{\varnothing\}$ (no auxiliary information), the attacker cannot narrow the candidate set and so the optimal reconstruction selects $x^* \in \arg\max_{x \in \mathcal{X}} w(x, \theta)\pi(x)$, i.e., any record that maximizes the posterior probability given θ. When the attacker does not require exact reconstruction but is satisfied with producing a candidate within a controlled error η of the true record, the optimal strategy retains a similar structure. However, rather than comparing records based on the posterior probability of a single output, the analysis evaluates the posterior mass of their associated success sets. The attacker then selects the record $\tilde{x}$ whose success set $S_\eta(\tilde{x})$ attains the largest posterior probability given the observed output.

This result is particularly relevant, as it implies that, for a given risk tolerance, the utility of a mechanism cannot exceed what our method achieves; in other words, our approach yields optimal noise calibration.

Corollary 5.9 (Attack Optimality)**.** *Given the conditions as in Theorem 5.3, Algorithm 1 achieves the highest attainable η-RAD.*

Proof. Following Algorithm 1, given θ, z, the attack always select (at random) an output from the set:

$$S_\theta^z = \arg\max_{\tilde{x} \in \mathcal{X}} \int_{\{x_1 \colon a^{-1}(z) \wedge \ell(\tilde{x}, x_1) \le \eta\}} w(x_1, \theta)\ \mathrm{d}P_z(x_1). \tag{5.17}$$

Hence, the attack A verifies

$$\Pr_A\big(A(\theta, z) \in S_\theta^z\big) = 1, \qquad p_A(\tilde{x} \mid \theta, z) = \frac{\mathbf{1}_{\{\tilde{x} \in S_\theta^z\}}}{\mu(S_\theta^z)},$$

and for all $\tilde{x} \in S_\theta^z$,

$$\int_{a^{-1}(z)} \mathbf{1}_{\{\ell(\tilde{x}, x_1) \le \eta\}}\, w(x_1, \theta)\ \mathrm{d}P_z(x_1) = \max_{s \in \mathcal{X}} \int_{S_\eta^z(s)} w(x_1, \theta)\ \mathrm{d}P_z(x_1) \equiv I_{z,\theta}.$$

Computing RAD according to the reformulation in Equation (5.5), we obtain

$$\begin{aligned}
\eta\text{-RAD}(A) &= \int_{\mathcal{X}} \int_{\Theta} \Pr_A\left(S_\eta(x_1) \mid \theta, a(x_1)\right) w(x_1, \theta)\ \mathrm{d}\mu(\theta)\, \mathrm{d}P(x_1) \\
&= \int_{\Theta} \int_{aux} \int_{a^{-1}(z)} \Pr_A\left(S_\eta(x_1) \mid \theta, z\right) w(x_1, \theta)\ \mathrm{d}P_z(x_1)\, \mathrm{d}\nu(z)\, \mathrm{d}\mu(\theta)
\end{aligned}$$

$$
\begin{aligned}
&= \int_\Theta \int_{aux} \int_{a^{-1}(z)} \left(\int_{S_\eta(x_1)} \frac{\mathbf{1}_{\{\tilde{x} \in S_\theta^z\}}}{\mu(S_\theta^z)} \, w(x_1, \theta) \, \mathrm{d}\mu(\tilde{x}) \right) \mathrm{d}P_z(x_1) \, \mathrm{d}\nu(z) \, \mathrm{d}\mu(\theta) \\
&= \int_\Theta \int_{aux} \int_{a^{-1}(z)} \left(\int_{\mathcal{X}} \mathbf{1}_{\{\ell(\tilde{x}, x_1) \leq \eta\}} \frac{\mathbf{1}_{\{\tilde{x} \in S_\theta^z\}}}{\mu(S_\theta^z)} \, w(x_1, \theta) \, \mathrm{d}\mu(\tilde{x}) \right) \mathrm{d}P_z(x_1) \, \mathrm{d}\nu(z) \, \mathrm{d}\mu(\theta) \\
&= \int_\Theta \int_{aux} \int_{\mathcal{X}} \frac{\mathbf{1}_{\{\tilde{x} \in S_\theta^z\}}}{\mu(S_\theta^z)} \left(\int_{a^{-1}(z)} \mathbf{1}_{\{\ell(\tilde{x}, x_1) \leq \eta\}} w(x_1, \theta) \, \mathrm{d}P_z(x_1) \right) \mathrm{d}\mu(\tilde{x}) \, \mathrm{d}\nu(z) \, \mathrm{d}\mu(\theta) \\
&= \int_\Theta \int_{aux} \int_{\mathcal{X}} \frac{\mathbf{1}_{\{\tilde{x} \in S_\theta^z\}}}{\mu(S_\theta^z)} \, I_{z,\theta} \, \mathrm{d}\mu(\tilde{x}) \, \mathrm{d}\nu(z) \, \mathrm{d}\mu(\theta) \\
&= \int_\Theta \int_{aux} I_{z,\theta} \, \mathrm{d}\nu(z) \, \mathrm{d}\mu(\theta) \left(\int_{\mathcal{X}} \frac{\mathbf{1}_{\{\tilde{x} \in S_\theta^z\}}}{\mu(S_\theta^z)} \, \mathrm{d}\mu(\tilde{x}) \right) \\
&= \int_\Theta \int_{aux} I_{z,\theta} \, \mathrm{d}\nu(z) \, \mathrm{d}\mu(\theta) \\
&= \int_\Theta \int_{aux} \max_{s \in \mathcal{X}} \int_{S_\eta^z(s)} w(x_1, \theta) \, \mathrm{d}P_z(x_1) \, \mathrm{d}\nu(z) \, \mathrm{d}\mu(\theta),
\end{aligned}
$$

which according to Theorem 5.3, coincides with the maximum attainable bound. □

Corollary 5.9 directly establishes that Theorem 5.3 is universally tight, i.e., for every mechanism, prior auxiliary knowledge and error threshold, Theorem 5.3 exactly determines the maximum achievable RAD. Moreover, Theorem 5.2 is tight, since there exists at least one mechanism (GRR, Example 5.4) for which Theorem 5.2 is achieved. We further validate that this is not an isolated case by empirically demonstrating tightness on additional mechanisms, such as DP-SGD (see Figure 5.4c).

Beyond the theoretical contribution, our results provide a practical tool: a general attack algorithm that practitioners can directly use to evaluate the privacy risks of their systems or the tightness of their bounds. As a concrete demonstration, we apply this attack in the context of LDP auditing (see Section 5.4) and to assess empirical risk and tightness of our bounds in (see Section 5.5). We also provide the application of our optimal attack in the specific case of DP-SGD:

Example 5.10 (Optimal Attack on DP-SGD)**.** Our analysis of DP-SGD is motivated by its central role in private learning: distributionally robust attacks were first introduced in this context [54], and DP-SGD remains the most widely used algorithm in practice [49]. In particular, we study the reconstruction setting considered by Hayes et al. [62], where the adversary attempts to reconstruct the target record x^* from a candidate set $\{x_1, \ldots, x_m\}$ with uniform prior using access to the privatized gradients $\{\bar{g}_1, \ldots, \bar{g}_T\}$ released during training, i.e., white-box setting. Note that in each DP-SGD iteration $\bar{g}_t$ is obtained as

$$
\bar{g}_t = \sum_x \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x)) + \mathcal{N}(0, c^2 \sigma^2 I),
$$

where σ is the noise scale, $\mathrm{clip}_C(\vec{v}) = \vec{v} \min(1, \frac{C}{\|\vec{v}\|_2})$ and θ_t the released weights in the previous iteration.

Given $\theta = (\theta_1, \dots, \theta_T)$, our optimal attack is determined by $\arg\max_{x:a(x)=z} w(\theta, x)$, and its sign, i.e., whether $w(\theta, x) > 0$ or not, for each candidate x and auxiliary knowledge z. Concretely, since the public dataset D_- is known, we can isolate the noisy contribution of the target's gradient at iteration t:

$$g_t = \bar{g}_t - \sum_{x \in D_-} \text{clip}_C(\nabla_\theta \ell(\theta_t, x)),$$

and simplify w maximization to

$$\arg\max_{x:a(x)=z} w(\theta, x) = \arg\max_{x:a(x)=z} \sum_t W(g_t, \text{clip}_C(\nabla_\theta \ell(\theta_t, x))) \tag{5.18}$$

where $W(u, v) = \langle u, v \rangle - \frac{1}{m} \sum_x \langle u, \text{clip}_C(\nabla_\theta \ell(\theta_t, x)) \rangle$, since W preserves the sign and $\arg\max$ of w. We present the pseudo-code of the optimal attack in Algorithm 2.

Indeed, given θ, x, under DP-SGD the privatized gradient at step t is

$$g_t \sim \mathcal{N}\Big(\mu_x,\, C^2\sigma^2 I\Big), \qquad \mu_x = \text{clip}_C(\nabla_\theta \ell(\theta_t, x)),$$

where C is the clipping parameter and I the identity function of dimension d, corresponding to the dimension of the gradients. Hence the likelihood is

$$p_{\mathcal{M}}(g_t \mid x) \;=\; \underbrace{\frac{1}{(2\pi C^2\sigma^2)^{d/2}}}_{A} \exp\Big(\underbrace{-\frac{1}{2C^2\sigma^2}}_{B} \|g_t - \mu_x\|^2\Big),$$

where both A, B are independent from x. Consequently,

$$w(g, x) = \prod_t p_{\mathcal{M}}(g_t \mid x) - \prod_t p_{\mathcal{M}}(g_t) > 0 \Leftrightarrow \tag{5.19}$$

$$A^T \left(\prod_t \mathrm{e}^{B\langle g_t, \mu_x \rangle} - \prod_t \frac{1}{m} \sum_i \mathrm{e}^{B\langle g_t, \mu_{x_i} \rangle} \right) > 0 \Leftrightarrow \tag{5.20}$$

$$\mathrm{e}^{B \sum_t \langle g_t, \mu_x \rangle} > \prod_t \frac{1}{m} \sum_i \mathrm{e}^{B\langle g_t, \mu_{x_i} \rangle} \Leftrightarrow \tag{5.21}$$

$$B \sum_t \langle g_t, \mu_x \rangle > \sum_t \ln \left(\frac{1}{m} \sum_i \mathrm{e}^{B\langle g_t, \mu_{x_i} \rangle} \right) \Leftrightarrow \tag{5.22}$$

$$B \sum_t \langle g_t, \mu_x \rangle > \sum_t \frac{1}{m} \sum_z \ln(\mathrm{e}^{B\langle g_t, \mu_{x_i} \rangle}) \Leftrightarrow \tag{5.23}$$

$$B \sum_t \langle g_t, \mu_x \rangle > \sum_t \frac{B}{m} \sum_i \langle g_t, \mu_{x_i} \rangle \Leftrightarrow \tag{5.24}$$

$$\sum_t \langle g_t, \mu_x \rangle - \sum_t \frac{1}{m} \sum_{x_i} \langle g_t, \mu_{x_i} \rangle > 0 \Leftrightarrow \tag{5.25}$$

$$\sum_t W(g_t, x) > 0. \tag{5.26}$$

Where Equation (5.23) follows from the application of Jensen's inequality to the logarithm. Moreover, $\arg\max_x w(g,x) = \arg\max_x \ln(p_{\mathcal{M}}(g,x)) = \arg\max_x \sum_t \ln p_{\mathcal{M}}(g_t,x)$, where

$$\ln p_{\mathcal{M}}(g_t \mid x_i) \propto -\tfrac{1}{2C^2\sigma^2} \left\| g_t - \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x_i)) \right\|^2.$$

Expanding the squared norm leads to

$$\|g_t\|^2 + \|\mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x_i))\|^2 - 2\langle g_t, \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x_i))\rangle.$$

The term $\|g_t\|^2$ is independent of x_i, and the term $\|\mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x_i))\|^2$ is bounded by C^2 (often nearly constant across candidates). Therefore, maximizing the log-likelihood is equivalent to maximizing

$$\langle g_t, \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x_i))\rangle.$$

Consequently, our optimal attack can be simplified by using $W(g_t, x)$ instead of $w(g_t, x)$.

When $aux = \{\varnothing\}$, our optimal attack coincides with the attack presented in [62]. Whereas they identified such an attack as the empirically best, we formally establish that this choice is indeed optimal. Moreover, we extend the optimal attack for any attacker that has target-specific auxiliary information. In particular, our optimal attack for attackers with $aux \neq \{\varnothing\}$ is empirically tested in Section 5.5, showing that previous bounds for ReRo indeed do not hold for attackers with target-specific auxiliary knowledge.

Algorithm 2: Optimal Attack for DP-SGD

Input : $\theta = (\theta_1, \ldots, \theta_T)$, $a(x) = z$ and $g = (g_1, \ldots, g_T)$
Output : $\tilde{x}$
for $x \colon a(x) = z$ **do**
  compute $\sum_t W(\bar{g}_t, \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x)))$;
Select $x^* = \arg\max_{x \colon a(x)=z} \sum_t W(\bar{g}_t, \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, x)))\pi_x$;
if $W(\bar{g}_t, x^*) > 0$ **then**
  $\tilde{x} = x^*$;
else
  $\tilde{x} \longleftarrow U[\mathcal{X} \backslash \{x \colon a(x) = z\}]$;

The bounds presented in this section offer concrete guidance for algorithm design. They can be directly leveraged for noise calibration, achieving rigorous privacy guarantees while maximizing utility. In particular, they induce a simple protocol for practitioners. First, one must specify which information is deemed private (e.g., the full record, a subset of attributes, or membership), which determines the choice of the auxiliary information aux and $a \colon \mathcal{X} \to aux$. Second, if prior knowledge about the distribution of $\mathcal{X}$ is available, it should be encoded in a distribution π. If this is not the case, however, one must resort to the worst-case prior; otherwise, the attacker's risk may be underestimated. This worst-case prior typically corresponds to $\pi_x = \pi_y = 1/2$ for the two records that are easiest to distinguish (see Examples 5.4 and 5.5 and Figure 5.6). Nevertheless, even

when the worst-case prior cannot be explicitly identified, the total variation bound given in Theorem 5.2 provides a safe upper bound for any choice of prior and aux.

Third, the resulting RAD of the mechanism can be computed using Theorem 5.3—an auxiliary-dependent bound proven to be universally tight, or upper-bounded by a worst-case guarantee when the nature of aux is unknown (Theorem 5.2). Finally, by inverting the corresponding bound, one can directly derive the noise-injection parameters that meet a prescribed risk level. Since our bounds are tight, this procedure yields mechanisms that are utility-optimal for any given risk acceptance.

Note that while the closed form of Theorem 5.3 is easy to derive for discrete data, this may not hold for continuous data, where the bound involves Lebesgue integrals. In such case, the bound can be evaluated numerically using a nested Monte Carlo procedure. Since numerical approximations introduce error, as a safer alternative, one may always use our closed-form upper bound in Theorem 5.2. However, this bound can be overly conservative when $aux = \{\varnothing\}$, motivating the tighter closed-form upper-bounds derived in the next section, which avoid numerical procedures even for continuous data.

5.3. η-RAD Upper Bounds under $aux = \{\varnothing\}$

Our bound in Theorem 5.3 is universally tight, but two limitations remain. First, it requires full knowledge of the mechanism, making it suitable for noise calibration; however, in DP auditing, we often have only query access (e.g., auditing external software) without insight into the internal protocol [152]. Second, the bound lacks a closed form hence may rely on numerical approximation, particularly for continuous data domains. Consequently, in this section we provide black-box bounds for the case $aux = \{\varnothing\}$, both because this is the standard assumption in prior DP auditing [41], [153] and data reconstruction studies [39], [62], and because it makes practical sense: for other auxiliary-information models, one can always rely on the closed-form bound provided by Theorem 5.2.

First, we present a general bound that applies to any reconstruction setting as long as no target-specific auxiliary knowledge is available. For this purpose, we introduce $\kappa^-_{\pi,\ell}(\eta)$ as the infimum counterpart of $\kappa^+_{\pi,\ell}(\eta)$, formally defined as

$$\kappa^-_{\pi,\ell}(\eta) = \inf_{x_0 \in \mathcal{X}} \Pr_{X\sim\pi}\left[\ell(X, x_0) \leq \eta\right], \tag{5.27}$$

representing the success probability of an oblivious attacker attempting to reconstruct the most difficult target only using π.

Theorem 5.11. *If a mechanism $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ satisfies f-DP, then for any attack with $aux = \{\varnothing\}$, $A\colon \Theta \to \mathcal{D}(\mathcal{X})$, it satisfies*

$$\eta\text{-RAD} \leq \max_{\alpha \in [\kappa^-_{\pi,\ell}(\eta), \kappa^+_{\pi,\ell}(\eta)]} 1 - f(\alpha) - \alpha.$$

If $\mathcal{X}$ is discrete, then it also holds

$$\eta\text{-RAD} \leq (1 - \kappa_\pi) \max_{\alpha \in [0, \frac{\kappa^+_{\pi,\ell}(\eta)}{1-\kappa_\pi}]} 1 - f(\alpha) - \alpha.$$

Proof. Kifer et al. [23, p.23] showed that for any $S \subseteq \Theta$, for any f-DP mechanism, and $x_0, x_1 \in \mathcal{X}$,

$$\Pr_{\mathcal{M}}(S \mid D_{x_1}) \leq 1 - f(\Pr_{\mathcal{M}}(S \mid D_{x_0}). \tag{5.28}$$

Moreover, since f is convex (see Section 2.3.1.1), applying Jensen's inequality:

$$f(\mathbb{E}_X[X]) \leq \mathbb{E}_X[f(X)] \Leftrightarrow -\mathbb{E}_X[f(X)] \leq -f(\mathbb{E}_X[X]). \tag{5.29}$$

Combining both Equation (5.28) and Equation (5.29) we obtain

$$\begin{aligned}
\eta\text{-RAD} &= \Pr_{X_1\sim\pi}[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr_{X_0,X_1\sim\pi}[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \\
&= \mathop{\mathbb{E}}_{X_0\sim\pi}\Big[\Pr_{X_1\sim\pi}[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr_{X_1\sim\pi}[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]\Big] \\
&= \mathop{\mathbb{E}}_{X_0,X_1\sim\pi}\Big[\Pr[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]\Big] \\
&\leq \mathop{\mathbb{E}}_{X_0,X_1\sim\pi}\Big[1 - f(\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]) - \Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]\Big] \\
&= 1 - \mathop{\mathbb{E}}_{X_1,X_0}[f\left(\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]\right)] - \mathop{\mathbb{E}}_{X_1,X_0}[\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]] \\
&\leq 1 - f\left(\mathop{\mathbb{E}}_{X_1,X_0}[\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]]\right) - \mathop{\mathbb{E}}_{X_1,X_0}[\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]],
\end{aligned}$$

where last inequality follows from Equation (5.29). Therefore, it suffices to determine the interval containing $\mathbb{E}_{X_1,X_0}[\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]$.

$$\begin{aligned}
&\mathbb{E}_{X_1,X_0\sim\pi}\left[\Pr_{X\sim\pi}[\mathcal{A}(D_{X_0}) \in S_\eta(X)]\right] \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \Pr[\mathcal{A}(D_{x_0}) \in S_\eta(x_1)]\pi_{x_0}\pi_{x_1}\,\mathrm{d}x_0\,\mathrm{d}x_1 \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{X}} p_{\mathcal{A}}[x \mid D_{x_0}]\mathbf{1}_{\{\ell(x,x_1)\leq\eta\}}\pi_{x_0}\pi_{x_1}\,\mathrm{d}x_0\,\mathrm{d}x_1\,\mathrm{d}x \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} p_{\mathcal{A}}[x \mid D_{x_0}]\left(\int_{\mathcal{X}} \mathbf{1}_{\{\ell(x,x_1)\leq\eta\}}\pi_{x_1}\,\mathrm{d}x_1\right)\pi_{x_0}\,\mathrm{d}x_0\,\mathrm{d}x \\
&\leq \kappa^+_{\pi,\ell}(\eta)\int_{\mathcal{X}}\int_{\mathcal{X}} p_{\mathcal{A}}[x \mid D_{x_0}]\pi_{x_0}\,\mathrm{d}x_0\,\mathrm{d}x = \kappa^+_{\pi,\ell}(\eta).
\end{aligned}$$

and analogous for $\kappa^-_{\pi,\ell}(\eta)$ since any attack output $x \in \mathcal{X}$ and hence it follows by definition. Note that last inequality assumes no auxiliary knowledge is available, therefore $p_{\mathcal{A}}[z \mid D_{x_0}, a(x_1)] = p_{\mathcal{A}}[z \mid D_{x_0}]$, hence it factors out of the integral with respect to x_1.

Now, we prove that, for discrete variables, the bound can be further improved. We follow the same notation as in Theorem 5.2, i.e.,

$$\mathop{\mathbb{E}}_{X_0,X_1\sim\pi}\left[\mathbf{1}_{\{X_0\neq X_1\}}\right] = 1 - \Pr_{X,X'\sim\pi}[X = X'] = \begin{cases} 1 & \text{if } \pi \text{ continuous,} \\ 1-\kappa_\pi & \text{if } \pi \text{ discrete.} \end{cases} \tag{5.30}$$

and $\sum_{x_1} \sum_{x_0 \neq x_1} \frac{\pi_0 \pi_1}{(1-\kappa_\pi)} = 1$. Now, combining Equation (5.28) and Equation (5.29) we obtain:

$$
\begin{aligned}
\eta\text{-RAD} &= \Pr_{X_1 \sim \pi}[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr_{X_0, X_1 \sim \pi}[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \\
&= \mathop{\mathbb{E}}_{X_0 \sim \pi} \Big[\Pr_{X_1 \sim \pi}[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr_{X_1 \sim \pi}[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \Big] \\
&= \mathop{\mathbb{E}}_{X_0, X_1 \sim \pi} \Big[\mathbf{1}_{\{X_0 \neq X_1\}} \big(\Pr[\mathcal{A}_{X_1}(D_{X_1}) \in S_\eta(X_1)] - \Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \big) \Big] \\
&\leq \mathop{\mathbb{E}}_{X_0, X_1 \sim \pi} \Big[\mathbf{1}_{\{X_0 \neq X_1\}} \big(1 - f(\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]) - \Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \big) \Big] \\
&= \mathop{\mathbb{E}}_{X_1, X_0 \sim \pi} [\mathbf{1}_{\{X_0 \neq X_1\}}] - \mathop{\mathbb{E}}_{X_1, X_0 \sim \pi} \Big[\mathbf{1}_{\{X_0 \neq X_1\}} f(\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]) \Big] \\
&\quad - \mathop{\mathbb{E}}_{X_1, X_0 \sim \pi} \Big[\mathbf{1}_{\{X_0 \neq X_1\}} \Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] \big) \Big] \\
&= (1 - \kappa_\pi) \Big(1 - \mathop{\mathbb{E}}_{X_1, X_0} [\mathbf{1}_{\{X_0 \neq X_1\}} \frac{f\left(\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]\right)}{(1 - \kappa_\pi)}] \\
&\qquad - \mathop{\mathbb{E}}_{X_1, X_0} [\mathbf{1}_{\{X_0 \neq X_1\}} \frac{\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)]}{(1 - \kappa_\pi)}] \Big) \\
&\leq (1 - \kappa_\pi) \Big(1 - f(\sum_{x_1} \sum_{x_0 \neq x_1} \Pr[\mathcal{A}_{x_1}(D_{x_0}) \in S_\eta(x_1)] \tfrac{\pi_0 \pi_1}{(1-\kappa_\pi)}) \\
&\qquad - \sum_{x_1} \sum_{x_0 \neq x_1} \Pr[\mathcal{A}_{x_1}(D_{x_0}) \in S_\eta(x_1)] \tfrac{\pi_0 \pi_1}{(1-\kappa_\pi)} \Big).
\end{aligned}
$$

Therefore, the proof follows from the following upper-bound:

$$
\begin{aligned}
&\sum_{x_1} \sum_{x_0 \neq x_1} \Pr[\mathcal{A}_{x_1}(D_{x_0}) \in S_\eta(x_1)] \tfrac{\pi_0 \pi_1}{(1-\kappa_\pi)} \\
&\leq \frac{1}{(1 - \kappa_\pi)} \mathop{\mathbb{E}}_{X_0, X_1} [\Pr[\mathcal{A}_{X_1}(D_{X_0}) \in S_\eta(X_1)] = \frac{\kappa^+}{(1 - \kappa_\pi)}.
\end{aligned}
$$

Concluding both bounds. □

If Theorem 5.3 cannot be computed in closed form, this result provides an upper bound for RAD when $aux = \{\varnothing\}$. It avoids numerical approximation errors and yields a tighter estimate than the conservative upper bound given in Theorem 5.2.

In the following example we see its practical application to Gaussian DP:

Example 5.12. We consider uniform prior and $\eta = 0$, hence $\kappa^+ = \frac{1}{m}$. Applying Theorem 5.11 we obtain

$$
0\text{-RAD} \leq \max_{\alpha \in [0, \frac{1}{m-1}]} 1 - f(\alpha) - \alpha \;=\; \max_{\alpha \in [0, \frac{1}{m-1}]} 1 - \Phi\Big(\Phi^{-1}(1 - \alpha) - \mu \Big) - \alpha \equiv \max_{\alpha \in [0, \frac{1}{m-1}]} g(\alpha),
$$

where Φ and φ denote respectively the CDF and PDF of the standard normal distribution. Using the chain rule and the following identity

$$
\frac{d}{d\alpha} \Phi^{-1}(1 - \alpha) = - \frac{1}{\varphi(\Phi^{-1}(1 - \alpha))},
$$

we obtain

$$\begin{aligned} g'(\alpha) &= -\varphi\Big(\Phi^{-1}(1-\alpha)-\mu\Big)\cdot\frac{d}{d\alpha}\Big[\Phi^{-1}(1-\alpha)-\mu\Big]-1 \\ &= \frac{\varphi(\Phi^{-1}(1-\alpha)-\mu)}{\varphi(\Phi^{-1}(1-\alpha))}-1. \end{aligned}$$

Moreover, the derivative can be rewritten in closed form. Recall that the standard normal density is

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}.$$

Therefore,

$$g'(\alpha)=\frac{\varphi(\Phi^{-1}(1-\alpha)-\mu)}{\varphi(\Phi^{-1}(1-\alpha))}-1 \tag{5.31}$$

$$=\frac{\exp\Big(-\frac{1}{2}(\Phi^{-1}(1-\alpha)-\mu)^2\Big)}{\exp\Big(-\frac{1}{2}(\Phi^{-1}(1-\alpha))^2\Big)}-1 \tag{5.32}$$

$$=\exp\left(\mu\,\Phi^{-1}(1-\alpha)-\frac{\mu^2}{2}\right)-1. \tag{5.33}$$

An interior maximizer satisfies $g'(\alpha)=0$, i.e.,

$$\mu\,\Phi^{-1}(1-\alpha)-\frac{\mu^2}{2}=0.$$

Because $\mu>0$, the unique solution is

$$\Phi^{-1}(1-\alpha)=\frac{\mu}{2}\Leftrightarrow\alpha=1-\Phi(\frac{\mu}{2}).$$

Moreover, since

$$g'(\alpha)=\exp\left(\mu\,\Phi^{-1}(1-\alpha)-\frac{\mu^2}{2}\right)-1,$$

we have $g'(\alpha)>0$ for $\alpha<1-\Phi(\mu/2)$ and $g'(\alpha)<0$ for $\alpha>1-\Phi(\mu/2)$. Hence, g increases up to $\alpha^\star$ and decreases thereafter, and the maximizer is unique.

It follows that the unconstrained maximizer is

$$\alpha^\star_{\text{free}}=1-\Phi\left(\frac{\mu}{2}\right).$$

Imposing the constraint $\alpha\le\frac{1}{m-1}$ yields

$$\alpha^\star=\min\left\{\frac{1}{m-1},\;1-\Phi\left(\frac{\mu}{2}\right)\right\}.$$

Consequently,

$$0\text{-RAD}\le\frac{m-1}{m}\left(1-\Phi\left(\Phi^{-1}(1-\alpha^\star)-\mu\right)-\alpha^\star\right).$$

We plot this bound for DP-SGD in Figures 5.4 and 5.5. While it is not perfectly tight, it provides a reliable approximation, avoiding the numerical computations required by the exact bound of Theorem 5.3.

Moreover, as a consequence of the previous result, we can obtain a bound of the RAD of any (ε, δ)-DP mechanism:

Proposition 5.13. *If a mechanism $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ satisfies (ε, δ)-DP, then for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{X})$, it satisfies*

$$\eta\text{-RAD} \leq \min\{\kappa^+_{\pi,\eta}(\mathrm{e}^\varepsilon - 1) + \delta, \tfrac{(1-\kappa^-_{\pi,\eta})(\mathrm{e}^\varepsilon-1)+\delta}{\mathrm{e}^\varepsilon}, \tfrac{e^\varepsilon-1+2\delta}{e^\varepsilon+1}(1-\kappa_\pi)\}.$$

Proof. Follows from combining previous theorem with [38] result that any (ε, δ)-DP mechanism is f-DP with, $f(\alpha) = \max\{1 - \delta - \mathrm{e}^\varepsilon\alpha, \frac{1-\delta-\alpha}{\mathrm{e}^\varepsilon}\}$, and analyze the different cases until we arrive to the bound. Formally, every (ε, δ)-DP mechanism verifies the that f-DP, with f

$$f(\alpha) = \max\Bigg\{\underbrace{1 - \delta - \mathrm{e}^\varepsilon\alpha}_{f_1(\alpha)}, \underbrace{\frac{1-\delta-\alpha}{\mathrm{e}^\varepsilon}}_{f_2(\alpha)}\Bigg\}. \tag{5.34}$$

On the other side, applying Theorem 5.11 we have

$$\eta\text{-RAD} \leq \max_{\alpha\in[\kappa^-,\kappa^+]} (1 - f(\alpha) - \alpha)\,. \tag{5.35}$$

Combining both equations we obtain,

$$\begin{aligned}
\eta\text{-RAD} &\leq \max_{\alpha\in[\kappa^-,\kappa^+]} (1 - f(\alpha) - \alpha) \\
&= \max_{\alpha\in[\kappa^-,\kappa^+]} 1 - \max\{f_1(\alpha), f_2(\alpha)\} - \alpha \\
&= \max_{\alpha\in[\kappa^-,\kappa^+]} (1 - \max\{f_1(\alpha) + \alpha, f_2(\alpha) + \alpha\}) \\
&= \max_{\alpha\in[\kappa^-,\kappa^+]} (\min\{1 - f_1(\alpha) - \alpha, 1 - f_2(\alpha) - \alpha\}) \\
&\leq \min\Bigg\{\max_{\alpha\in[\kappa^-,\kappa^+]} 1 - f_1(\alpha) - \alpha, \max_{\alpha\in[\kappa^-,\kappa^+]} 1 - f_2(\alpha) - \alpha\Bigg\}
\end{aligned}$$

Therefore, we analyze both maximums.

First, for f_1 we have:

$$1 - f_1(\alpha) - \alpha = \delta + \mathrm{e}^\varepsilon\alpha - \alpha \tag{5.36}$$

$$= \alpha(\mathrm{e}^\varepsilon - 1) + \delta \leq \kappa^+(\mathrm{e}^\varepsilon - 1) + \delta \tag{5.37}$$

Second, for f_2 we obtain:

$$1 - f_2(\alpha) - \alpha = 1 - \frac{1 - \delta - \alpha}{\mathrm{e}^\varepsilon} - \alpha \tag{5.38}$$

$$= 1 - \frac{1-\delta}{\mathrm{e}^\varepsilon} + \alpha(\mathrm{e}^{-\varepsilon} - 1) \leq 1 - \kappa^-(1 - \mathrm{e}^{-\varepsilon}) - \frac{1-\delta}{\mathrm{e}^\varepsilon} = \frac{(1-\kappa^-)(\mathrm{e}^\varepsilon - 1) + \delta}{\mathrm{e}^\varepsilon}. \tag{5.39}$$

Combined with the general bound Theorem 5.2 it follows the result. □

This bound enables to better interpret DP parameters in terms of reconstruction attacks without auxiliary knowledge.

Moreover, this bound remains informative even when the mechanism in use is completely unknown. For instance, consider auditing external software from a company that claims to provide (ε, δ)-DP but does not disclose the mechanism used. In such a black-box setting, where we are allowed to query the model but never know the underlying mechanism, our Proposition 5.13 still applies.

However, as discussed in Section 5.2, the actual privacy protection of a DP mechanism depends on its specific design and cannot be characterized solely by its privacy parameters. Consequently, this upper-bound should be used as a last-option estimate when the mechanism is unknown, rather than as a substitute for proper noise calibration in mechanism design.

Next, we focus on improving this black-box bound for perfect reconstruction, i.e., $\eta = 0$, in categorical data. This case is particularly relevant since many sensitive attributes, such as diseases, political opinions, or religious beliefs, are categorical and do not trivially support partial reconstruction, e.g., [154], [155]. For such settings, we derive more precise bounds. To do so, we first introduce the following auxiliary lemma:

Lemma 5.14. *Given $|\mathcal{X}| = m$ and $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ an (ε, δ)-DP mechanism, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{X})$ and $\gamma_x = \Pr_{\mathcal{M}}(\Theta_x \mid x) - \Pr_{\mathcal{M}}(\Theta_x)$, with Θ_x as in Equation (5.15), then*

$$\Gamma \coloneqq \sum_{x \in \mathcal{X}} \gamma_x \leq \frac{(m-1)(\mathrm{e}^{\varepsilon} - 1 + \delta m)}{\mathrm{e}^{\varepsilon} + m - 1}. \tag{5.40}$$

Proof. By definition $\Theta_x \cap \Theta_{x'} = \varnothing$. Besides, for all θ it exits at least one $x_\theta \in \arg\max_x p_{\mathcal{M}}(\theta \mid x)\pi_x$, and $\bigcup_{x \in \mathcal{X}} \Theta_x = \Theta$. Hence, $\{\Theta_x\}_{x \in \mathcal{X}}$ determines a partition in Θ. Therefore, by the law of total probability, for each x_0 we have

$$\sum_{x \in \mathcal{X}} \Pr_{\mathcal{M}}(\Theta_x \mid x_0) = \sum_{x \in \mathcal{X}} \int_{\Theta_x} p_{\mathcal{M}}(\theta \mid x_0)\, \mathrm{d}\mu(\theta) = \int_{\Theta} p_{\mathcal{M}}(\theta \mid x_0)\, \mathrm{d}\mu(\theta) = 1. \tag{5.41}$$

On the other hand, since $\mathcal{M}$ is (ε, δ)-DP, for every $x_1, x_0 \in \mathcal{X}$,

$$\Pr_{\mathcal{M}}(\Theta_1 \mid x_0) \geq \mathrm{e}^{-\varepsilon}(\Pr_{\mathcal{M}}(\Theta_1 \mid x_1) - \delta). \tag{5.42}$$

Substituting Equation (5.42) in Equation (5.41) we obtain, for all $i, j \in [m]$,

$$\Pr_{\mathcal{M}}(\Theta_i \mid x_i) + \mathrm{e}^{-\varepsilon} \sum_{i \neq j} \Pr_{\mathcal{M}}(\Theta_j \mid x_j) \leq 1 + \delta \mathrm{e}^{-\varepsilon}(m-1) \tag{5.43}$$

Summing the above inequality over all $i \in [m]$,

$$\sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid x_i) + (m-1)\mathrm{e}^{-\varepsilon} \sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid x_i) \leq m(1 + \delta \mathrm{e}^{-\varepsilon}(m-1)) \Leftrightarrow \tag{5.44}$$

$$\sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid x_i) \leq \frac{m(1+\delta \mathrm{e}^{-\varepsilon}(m-1))}{1+(m-1)\mathrm{e}^{-\varepsilon}} = \frac{m\mathrm{e}^{\varepsilon}+\delta m(m-1)}{\mathrm{e}^{\varepsilon}+(m-1)}. \tag{5.45}$$

Hence,

$$\begin{aligned}
\Gamma &= \sum_{x\in\mathcal{X}} \gamma_x && (5.46)\\
&= \sum_{x\in\mathcal{X}} \left(\Pr_{\mathcal{M}}(\Theta_x \mid x) - \Pr_{\mathcal{M}}(\Theta_x) \right) && (5.47)\\
&= \sum_{x\in\mathcal{X}} \Pr_{\mathcal{M}}(\Theta_x \mid x) - 1 && (5.48)\\
&\leq \frac{m\mathrm{e}^{\varepsilon}+\delta m(m-1)}{\mathrm{e}^{\varepsilon}+m-1} - 1 && (5.49)\\
&= \frac{(m-1)(\mathrm{e}^{\varepsilon}-1+\delta m)}{\mathrm{e}^{\varepsilon}+m-1}. && \square
\end{aligned}$$

Applying this lemma we obtain the following RAD bound:

Theorem 5.15 (0-RAD under (ε,δ)-DP)**.** *Given $|\mathcal{X}| = m$ with prior $\pi_1(1-\pi_1) \geq \cdots \geq \cdots \geq \pi_m(1-\pi_m)$ and $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{D}(\Theta)$ an (ε,δ)-DP mechanism, any attack $A\colon \Theta \to \mathcal{D}(\mathcal{X})$ verifies that*

$$0\text{-RAD} \leq \frac{e^{\varepsilon}-1+2\delta}{e^{\varepsilon}+1} K_\pi + R \max_{i>K} \pi_i$$

where $K_\pi = \sum_i^K (1-\pi_i)\pi_i$ and $K \in [m]$ is the largest index satisfying

$$R = (m-1)\frac{e^{\varepsilon}-1+m\delta}{e^{\varepsilon}+m-1} - (K - \sum_{i=1}^{K}\pi_i)\frac{e^{\varepsilon}-1+2\delta}{e^{\varepsilon}+1} \geq 0.$$

Proof. Since $|\mathcal{X}| = m$ and $aux = \{\varnothing\}$, Theorem 5.3 simplifies to Equation (5.14), hence

$$0\text{-RAD} \leq \sum_{i=1}^{m} \left(\Pr_{\mathcal{M}}(\Theta_i \mid x_i) - \Pr_{\mathcal{M}}(\Theta_i) \right) \pi_i \equiv \sum_{i=1}^{m} \gamma_i \, \pi_i. \tag{5.50}$$

For one side, we obtain that for all $i \in [m]$,

$$\begin{aligned}
\gamma_i = \Pr_{\mathcal{M}}(\Theta_i \mid x_i) - \Pr_{\mathcal{M}}(\Theta_i) &= \int_{\Theta_i} p_{\mathcal{M}}(\theta \mid x_i) - \sum_{j\in[m]} p_{\mathcal{M}}(\theta \mid x_j)\pi_j \, \mathrm{d}\mu(\theta) && (5.51)\\
&= \int_{\Theta_i} \sum_{j\in[m]} \left(p_{\mathcal{M}}(\theta \mid x_i) - p_{\mathcal{M}}(\theta \mid x_j) \right) \pi_j \, \mathrm{d}\mu(\theta) && (5.52)\\
&= \sum_{j\neq i} \left(\Pr_{\mathcal{M}}(\Theta_i \mid x_i) - \Pr_{\mathcal{M}}(\Theta_i \mid x_j) \right) \pi_j && (5.53)\\
&\leq \mathrm{TV}(\mathcal{M}) \sum_{j\neq i} \pi_j \leq \frac{e^{\varepsilon}-1+2\delta}{e^{\varepsilon}+1}(1-\pi_i). && (5.54)
\end{aligned}$$

If we simply apply this bound we recover Theorem 5.2 result:

$$0\text{-RAD} \leq \sum_{i=1}^{m} \gamma_i \, \pi_i \leq \sum_{i=1}^{m} \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}(1 - \pi_i)\pi_i = \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}(1 - \kappa_\pi).$$

However, due to Lemma 5.14, we know that this bound is loose, since in this case,

$$\Gamma = \sum_{i=1}^{m} \gamma_i = \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}(m-1) \geq \frac{e^{\varepsilon} - 1 + m\delta}{e^{\varepsilon} + m - 1}(m-1) = \Gamma_{\max}, \tag{5.55}$$

contradicting Lemma 5.14; therefore, it is impossible to achieve the local inequality $\gamma_i \leq \mathrm{TV}(\mathcal{M})(1 - \pi_i)$ simultaneously for all $i \in [m]$. In most cases, we can apply the local bound to a reduced set of indexes k, and the remainders must adjust so that the total sum $\sum_i \gamma_i = \Gamma$. Formally, at most, we can sum k summands such that,

$$\sum_{r=1}^{k} \gamma_{i_r} \leq \frac{e^{\varepsilon} - 1 + m\delta}{e^{\varepsilon} + m - 1}(m-1) \Leftrightarrow \tag{5.56}$$

$$\sum_{r=1}^{k} (1 - \pi_{i_r}) \leq (m-1)\frac{(\mathrm{e}^{\varepsilon} - 1 + m\delta)((\mathrm{e}^{\varepsilon} + 1)}{(\mathrm{e}^{\varepsilon} - 1 + 2\delta)((\mathrm{e}^{\varepsilon} - 1 + 2\delta))} \tag{5.57}$$

Hence, without loss of generality we order the indices so that

$$\pi_1(1 - \pi_1) \geq \pi_2(1 - \pi_2) \geq \cdots \geq \pi_m(1 - \pi_m).$$

obtaining,

$$0\text{-RAD} \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} \sum_{i=1}^{k_\pi} \pi_i(1 - \pi_i) + R \max_{r > k_\pi} \pi_r \tag{5.58}$$

with k_π the maximum index verifying:

$$\sum_{i=1}^{k_\pi} (1 - \pi_i) \leq (m-1)\frac{(\mathrm{e}^{\varepsilon} - 1 + m\delta)((\mathrm{e}^{\varepsilon} + 1)}{(\mathrm{e}^{\varepsilon} - 1 + 2\delta)((\mathrm{e}^{\varepsilon} - 1 + 2\delta))},$$

and R the reminder, i.e, K the biggest index such that

$$R = (m-1)\frac{e^{\varepsilon} - 1 + m\delta}{e^{\varepsilon} + m - 1} - (K - \sum_{i=1}^{K} \pi_i)\frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} \geq 0. \qquad \square$$

Note that in the extreme case where $\pi_1 = \pi_2 = \frac{1}{2}$ and $\pi_i = 0$ for all $i \neq 1, 2$, we recover exactly the same result as in Theorem 5.2.

Importantly, Theorem 5.15 is less applicable than Proposition 5.13, since it only applies for perfect reconstruction, $\eta = 0$, in categorical data. However, under these assumptions it offers a more accurate bound as we see in Figure 5.3.

This formulation enables the assessment of intermediate configurations of π. Notably, when $\pi = U[m]$ yields a marked improvement:

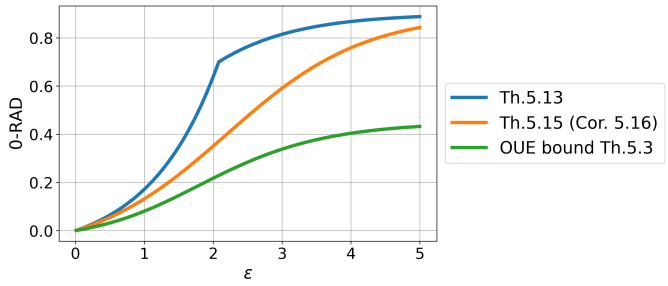


Figure 5.3.: Comparison of black-box bounds for 0-RAD without auxiliary knowledge, $\pi = U[10]$ and $\delta = 10^{-5}$. The bound given in Proposition 5.13 is more general and applies in any setting. In contrast, Theorem 5.15 is specific to categorical data but provides a tighter risk estimate when applicable. Finally, if the mechanism is known—here, OUE—it is always preferable to use the tighter bound provided by Theorem 5.3.

Corollary 5.16 (Black-box Uniform Prior)**.** *Let $\pi = \mathrm{U}[m]$ the uniform distribution over $\mathcal{X}$. If a mechanism $\mathcal{M}$ satisfies (ε, δ)-DP, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{X})$ it guarantees*

$$0\text{-RAD} \leq \frac{e^\varepsilon - 1 + \delta m}{e^\varepsilon + m - 1}\frac{m-1}{m}.$$

Proof. For every $K \in [m]$, $K_\pi = \sum_{i=1}^{K}(1-\pi_i)\pi_i = K\frac{m-1}{m^2}$ and $(K - \sum_{i=1}^{K}\pi_i) = K\frac{m-1}{m}$, therefore, denoting $A = \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}$ and applying Theorem 5.15 we get:

$$0\text{-RAD} \leq AK\frac{m-1}{m^2} + \frac{1}{m}(\Gamma - K\frac{m-1}{m}A) = \frac{1}{m}\Gamma = \frac{e^\varepsilon - 1 + \delta m}{e^\varepsilon + m - 1}\frac{m-1}{m}. \tag{5.59}$$

□

Remark on composition. Since our η-RAD bounds depend explicitly on the privacy parameters—namely ε, δ, and/or f—they can be directly recomputed under composition by first applying the corresponding composition results to obtain the composed privacy parameters (see Chapter 4), and then evaluating the bounds on these composed values. In the following example, we illustrate how to derive RAD composition bounds for the particular case of DP-SGD.

Example 5.17. Given a risk threshold, RAD $\leq \gamma$, we aim to calibrate the noise scale σ (i.e., the standard deviation of the Gaussian noise added to the gradients during training [49]) on a full-batch DP-SGD, for T steps to protect against the threat model considered by Hayes et al.[62], i.e., white-box access to private gradients, uniform prior over $|\mathcal{X}| = m$ and $\eta = 0$, hence $\kappa_- = \kappa_+ = 1/m$.

Each iteration of a full-batch DP-SGD performs a Gaussian mechanism on the gradient computation, hence, we discussed in Section 2.3.1.1, it verifies μ-GDP [38], with $\mu = 1/\sigma$.

The adaptive composition of T iterations of a μ-GDP mechanism is $(\mu\sqrt{T})$-GDP, as discussed in Section 6.3. Hence, a complete training of DP-SGD with T iterations, is

Notion	Assumptions	RAD bound	Composition	ReRo bound
Total variation	—	Theorem 5.2	✓	∄
f-DP	$aux = \{\varnothing\}$	Theorem 5.11	✓	Equation (2.6) [62]
$\mathcal{M}$	aux known	Theorem 5.3	×	∄
(ε, δ)-DP	—	Theorem 5.2	✓	∄
(ε, δ)-DP	$aux = \{\varnothing\}$	Proposition 5.13	✓	Equation (2.3) [39]
(ε, δ)-DP	$aux = \{\varnothing\}, \eta = 0$	Theorem 5.15	✓	Equation (2.3) [39]

Table 5.1.: Summary of RAD bounds applicability.

$(\sqrt{T}\sigma^{-1})$-GDP. Moreover, any μ-GDP mechanism has total variation $\mathrm{TV} \leq 2\Phi(\frac{\mu}{2}) - 1$ [67], hence DP-SGD after T iterations satisfies $\gamma \leq \frac{m-1}{m}(2\Phi(\frac{\sqrt{T}}{2\sigma}) - 1)$. Combining this composition result with our theorems we obtain direct calibration rules:

Without information about aux, we use Theorem 5.2. Obtaining,

$$\eta\text{-RAD} \leq \mathrm{TV}(\mathcal{M}) \left(1 - \frac{1}{m}) = \frac{m-1}{m}(2\Phi(\frac{\sqrt{T}}{2\sigma}) - 1\right).$$

We plot this bound for $T = 100$ in Figures 5.4b and 5.4c. We can then solve σ for any desired risk γ.

If we consider the whole records sensitive, $aux = \{\varnothing\}$, then we apply Theorem 5.11:

$$0\text{-RAD} \leq \frac{m-1}{m} \max_{\alpha \in [0, \frac{1}{m-1}]} \left(1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sqrt{T}}{\sigma}\right) - \alpha\right)$$

Hence, given $\alpha^* = \min\left\{\frac{1}{m-1},\ 1 - \Phi\left(\frac{\sqrt{T}}{2\sigma}\right)\right\}$ (see Example 5.12), the minimum σ to guarantee $0\text{-RAD} \leq \gamma$ is:

$$\sigma \ \geq \ \frac{\sqrt{T}}{\Phi^{-1}(1-\alpha^*) - \Phi^{-1}\left(1 - \frac{m}{m-1}\gamma - \alpha^*\right)}.$$

We plot this bound for the case of $T = 100$ in Figure 5.4a. A practitioner can then choose the minimum noise scale σ for any given risk threshold γ. For instance, given that a set of $m = 10$ individuals do not tolerate a risk bigger than 0.1, for a training of $T = 100$ iterations, one must add noise calibrated to $\sigma = 22$.

In summary, this section provides reasonable closed-form upper bounds (as we show in Section 5.5.3) for estimating RAD when Theorem 5.3 cannot be computed explicitly or $\mathcal{M}$ is unknown and $aux = \{\varnothing\}$, hence Theorem 5.2 would overestimate the risk. Importantly, these bounds offer composition results as we summarize in Table 5.1.

5.4. RAD for DP Auditing

DP auditing is crucial for assessing the tightness of DP mechanisms, establishing the practical impact of the mechanism parameters, and detecting implementation flaws in deployed DP mechanisms [42], [134], [144]. While previous DP auditing tools focus on solving specifically one of the aforementioned aspects, we propose a general-purpose DP auditing framework: RAD-based DP auditing.

RAD provides a unifying framework for analyzing adversarial risk under arbitrary threat models. Moreover, our bounds establish a tight and explicit connection between RAD and the standard privacy parameters. Taken together, these results yield a simple and principled approach to general-purpose DP auditing. Precision and tightness are especially critical in this context, since loose estimates may underestimate privacy risks or fail to detect bugs and implementation flaws.

The core idea of RAD-based auditing is straightforward: given a measured RAD value $\tilde{\gamma}$, we invert our theoretical bounds to estimate an empirical privacy budget. This empirical $\tilde{\varepsilon}$ reflects the observed privacy loss in practice, complementing theoretical worst-case values and providing a more realistic perspective on real-world risk. Formally, in previous sections, we provide bounding functions B such that $\mathrm{RAD}(\mathcal{M}) \leq B(\varepsilon, \delta)$ for any (ε, δ)-DP mechanism. Given a bound η-RAD $\leq B(\varepsilon, \delta)$, we compute RAD empirically obtaining γ, and estimate $\tilde{\varepsilon} \geq B^{-1}(\gamma, \delta)$.

The bound we employ depends on the specific setting. For instance, in a completely black-box scenario—where not even the mechanism used is known—for categorical data, in which we assume $\pi = U[m]$, the best bound is Corollary 5.16. Therefore, the DP auditing framework consists of running an attack, measuring its empirical RAD $\widetilde{\gamma}$, and deriving $\widetilde{\varepsilon}$ as follows:

$$\widetilde{\varepsilon} = \begin{cases} \ln\left(\frac{\widetilde{\gamma}\, m+1}{1-\widetilde{\gamma}\frac{m}{m-1}}\right) & \text{if the term can be evaluated,} \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{5.60}$$

However, if the mechanism $\mathcal{M}$ is known, we can use our improved bound from Theorem 5.3 (see Examples 5.4, A.17 and A.18).

Our auditing framework overcomes the fundamental scalability limitations of prior learning-based approaches such as DP-Sniper and Eureka [42], [43], enabling auditing in high-dimensional categorical LDP settings. Unlike these methods, our approach avoids costly hyperparameter tuning and the search for worst-case neighboring databases, and remains computationally feasible even when the input domain contains thousands of categories (see Section 5.5).

Despite the importance of LDP mechanisms [45], only one major work has so far focused on LDP auditing: LDP Auditor [41]. Applying our RAD-based DP auditing to LDP, we address key limitations of prior work. In contrast to LDP Auditor, which focuses exclusively on perfect reconstruction without target-specific auxiliary knowledge—excluding important use-cases such as AIAs—we allow auditing under broader threat models by leveraging optimal attacks (see Algorithm 1). Moreover, LDP Auditor uses

the Clopper–Pearson method to compute confidence intervals for the attacker's success probability. Since the upper bound of the interval must conservatively cover the true probability with high confidence, it systematically produces estimates that are higher than the actual value [41]. This intrinsic limitation is avoided in our approach, which does not rely on confidence intervals.

We investigate and empirically show the improvement in accuracy of our auditing approach in Section 5.5 (cf. Figure 5.8 for results), where we audit three main LDP mechanisms—GRR, SS and OUE—showing improved accuracy for all of them.

5.5. Experiments

In this section, we empirically examine the limitations of ReRo described in Section 5.1, focusing on how existing bounds fail to account for realistic attackers with target-specific auxiliary information. Moreover, we validate our theoretical bounds and our RAD-based DP auditing framework in real-world databases and DP mechanisms. Our experiments show that RAD accurately distinguishes privacy leakage from imputation, with tight bounds in practice, making it a reliable tool for interpretable noise calibration. RAD also enables auditing of LDP mechanisms, improving both scope and accuracy over the state-of-the-art [41].

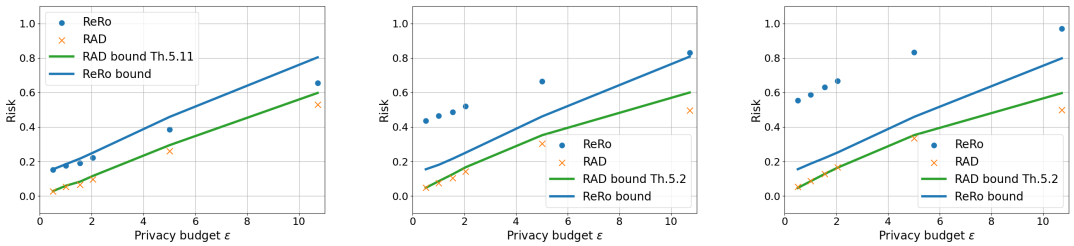
To ensure a fair comparison between ReRo and RAD in risk assessment, we emulate their experimental designs and dataset choices whenever possible. These design choices are particularly suitable for evaluating the tightness of our bounds, as the original works used them to assess the tightness of the ReRo bounds, reporting nearly tight results [62]. This suggests that these settings already serve as strong testbeds for tightness evaluation. Similarly, when comparing our RAD-based auditing framework to LDP Auditor, we adhere to their experimental design choices to ensure a coherent and consistent evaluation. We provide detailed descriptions of the datasets and experimental parameters in the following sections.

5.5.1. Database Description

We evaluate private learning, aggregation and LDP scenarios, using tailored datasets for each setting. The database selection is guided by their relevance in prior work and availability.

For DP-SGD, we use the same dataset as in ReRo [62] for consistency: MNIST [156], with 70 000 grayscale images of handwritten digits. We also replicate results on Fashion-MNIST [157] (Fashion), which similarly contains 70 000 grayscale images of clothes.

To evaluate the imputation attack [58], we use the Census and Texas-100X datasets in consistency with the original paper. The Census dataset [58] contains 1 676, records with 14 attributes, where race is treated as the sensitive attribute with eight categories. The Texas-100X dataset [58] comprises 925 128 patient records from 441 hospitals, including demographic and medical attributes, where ethnicity is treated as the sensitive attribute with two categories.



(a) DRA, $aux = \{\varnothing\}$. (b) DRA, $a(x) =$ image label. (c) MIA, $a(x) = x$.

Figure 5.4.: RAD vs. ReRo results for optimal attacks against DP-SGD on MNIST. Lines show theoretical bounds and markers of empirical risk as estimated by RAD/ReRo. Empirical results exceed the bounds estimated by ReRo, whereas our RAD bounds remain close to the true risk. Moreover, while ReRo sharply increases when auxiliary knowledge is available, RAD effectively discounts imputation effects.

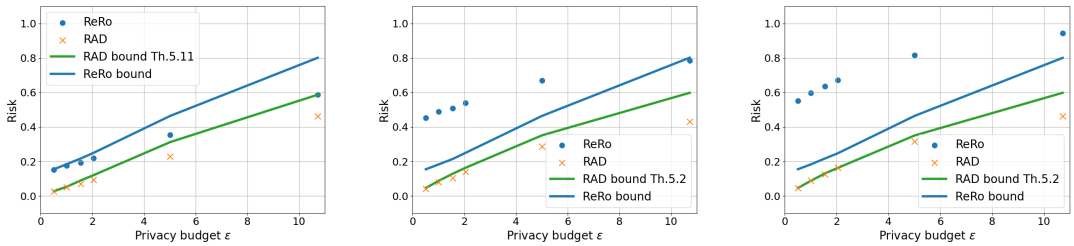
We evaluate aggregation in the Adult dataset [158], a census dataset commonly used in privacy-preserving aggregation [159]. It consists of 32 561, records with two numerical attributes, from which we select (working) hours-per-week following previous work [159], leading to the domain $\mathcal{X} = \{0, \ldots, 100\}$.

Finally, we evaluate our LDP auditing framework on location-reconstruction attacks using two real-world mobility datasets: the Porto dataset [88] and the Geolife dataset [125]. Both datasets are widely used in privacy and mobility research (e.g., [142], [160], [161]) and are publicly available. Each dataset consists of GPS coordinates, which we map to the OpenStreetMap (OSM) graph format [107] like prior work. The Porto dataset contains a total of 83, 409, 386 location reports that we map to the OSM roadgraph at Porto's city center (41.1475° N, 8.5870° W) with a 2.7 km radius, capturing the urban core of Porto. This radius leads to a universe size $|\mathcal{X}| = 3\,052$. The Geolife dataset contains a total of 24 876 978 locations that we mapped to an OSM graph centered near Tiananmen Square (39.9130° N, 116.3703° E) with a 5 km radius covering major central districts, leading to a universe of size $|\mathcal{X}| = 5\,356$.

5.5.2. Experiment Design

We investigate attacks on private learning (DP-SGD), aggregation queries (Laplace mechanism), and LDP protocols (GRR, OUE, SS) under varying auxiliary information settings to validate our bounds, compare RAD and ReRo, and evaluate the accuracy of our auditing framework.

To demonstrate that *ReRo overestimates risk*—and how RAD overcomes this limitation—we consider an attack that completely ignores the output of the private mechanism and relies solely on public information. This allows us to assess how ReRo behaves in a scenario where no private information is disclosed due to participation. To this end, we select the pure imputation attack [58]. This attack uses a public dataset D_- to train



(a) DRA, $aux = \{\varnothing\}$. (b) DRA, $a(x) =$ image label. (c) MIA, $a(x) = x$.

Figure 5.5.: RAD vs. ReRo results for optimal attacks against DP-SGD on Fashion. Lines show theoretical bounds and markers of empirical risk as estimated by RAD/ReRo. Both ReRo and RAD show a consistent behavior with respect to the MINST dataset.

a separate attack classifier A_I that, given the public attributes of a target, returns as label a prediction for the sensitive one. The adversary is given only the target public attribute $a(x)$ and outputs the prediction $\widetilde{s_z} = \arg\max_{s_i \in \Theta} \Pr_{\mathcal{I}}[s_i \mid a(x)]$, where the conditional distribution $\Pr[s_i \mid a(x)]$ is estimated by A_I, once the imputation model has been trained on D_-. This attack does not use any information from the target model $\mathcal{M}(D)$; therefore, adversarial success cannot be privacy leakage resulting from a user's participation in the training dataset of $\mathcal{M}(D)$. Following the original paper [58], we tested in both the Census and Texas datasets. We set $|D_-| = 49\,000$ and a universe $\mathcal{X}$ of $m = 1\,000$, randomly selected from the remaining data records consistent with [58]. We define the attack to be successful, $\ell(x, x') = 0$, if $a(x) = a(x')$, as is typical for AIAs.

We demonstrate how *RAD improves over ReRo* and establish the optimality of our bounds in both private learning and DP aggregation settings. In both cases, we evaluate tightness by testing our corresponding optimal attacks. To ensure a fair comparison, we emulate the original ReRo experimental setup for private learning, where the authors report their bounds to be nearly tight for $aux = \{\varnothing\}$. For DP aggregation, although no experimental results are reported in the original work, we adhere as closely as possible to the same parameter choices.

For private learning we run the attacks against DP-SDG on the MNIST and Fashion image datasets in three settings: $aux = z$ (a MIA), $aux = \{\varnothing\}$ (a DRA, replicating the setting in [62]), and $aux = a(x)$ (a DRA, where the adversary also knows the target image's label, i.e., which object is contained). To ensure a fair comparison with ReRo bounds, we select the parameters and thresholds exactly as specified in the original paper [62]. Namely, we declare an attack successful when $A(\theta, a(x)) = x$, that is, $\eta = 0$. We set $|D_-| = 999$ (and so the training set size is $|D_- \cup \{x\}| = 1\,000$) and train with full-batch DP-SGD for $T = 100$ steps. We set the clipping rate, i.e., the maximum norm we clip the real gradients to while training, $C = 0.1$ and $\delta = 10^{-5}$ and adjust the noise scale σ (see Example 5.17) for a given target ε. We set the uniform prior with size $|\mathcal{X}| = 8$ (disjoint from D_-), meaning that $\kappa^+_{\pi,0} = \kappa_\pi = 0.125$. Hence, we exactly replicate the original ReRo study [62] parameters.

For DP aggregation, we evaluate the optimal attack against the Laplace mechanism on sum queries using the "working-hours" attribute of Adult, employing truncation as a post-processing operation. Analogously to the private learning experiments, we set $|D| = 999$, $aux = \{\varnothing\}$ but in this case we evaluate the performance for $\eta \in \{0, 40, 80, 100\}$ to assess the impact or the error threshold on the risk estimation. Moreover, to understand the impact of the prior distribution on risk assessment, we compare three different distributions. As a baseline, we consider a uniform distribution. To simulate a realistic setting, we empirically estimate the distribution π from the original data, reflecting real-world frequencies (e.g., working 40 hours per week is *a priori* more likely than working 100 hours per week). Finally, we evaluate a fully skewed distribution with $\pi(100) = \pi(0) = 0.5$, representing a worst-case scenario in which the attacker's prior is concentrated on the two records that are easiest to distinguish in the dataset—analogous to the worst-case perspective in the original DP definition.

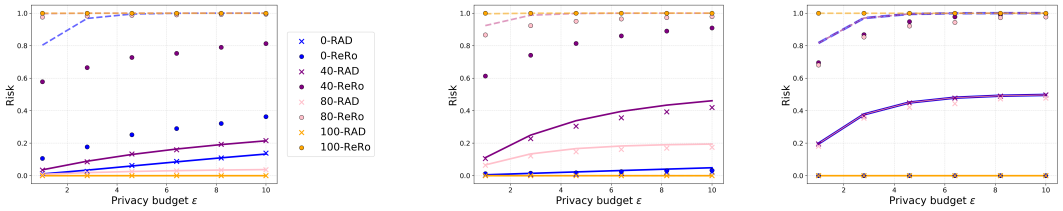
Finally, *we evaluate our RAD framework in LDP*, and we compare our auditing framework with the state-of-the-art tool LDP Auditor [41] for three relevant LDP mechanisms: GRR, OUE and SS [136], [162] .The results for LDP Auditor were obtained in collaboration with Héber H. Arcolezi, based on the implementation provided in Arcolezi and Gambs's public GitHub repository [163]. LDP Auditor estimates the empirical privacy budget in 10^6 runs.

We evaluate RAD based on our optimal attack (see Alg. 1) under a uniform prior and without auxiliary knowledge, allowing comparison with LDP Auditor. We then test our own LDP auditing framework: based on the obtained RAD value γ, we evaluate $B^{-1}(\gamma)$ for B following Theorem 5.3 and obtain an estimate of the empirical privacy budget. The precise $B(\varepsilon)$ for GRR, OUE and SS are shown in Examples 5.4 to 5.6 respectively. Since B^{-1} is not explicit for OUE, we approximate it numerically using the bisection method, which converges in $\mathcal{O}(\log(\tau^{-1}))$ iterations, where τ denotes the tolerance level [164]. We set $\tau = 10^{-6}$. Consistent with [41], we repeat the ε estimation five times and report the mean and standard deviation.

All experiments rely on empirical estimates of ReRo and RAD, i.e., estimates of a probability and a difference of probabilities, respectively. To obtain these estimates, we use Monte Carlo methods, approximating expected values by repeatedly sampling from the random process and computing the average. Following [62], ReRo is estimated by repeating J times the attack $A(\mathcal{M}(D_x), a(x))$ for each $x \in \mathcal{X}$ and computing the π-weighted average. The RAD correction term is estimated analogously by evaluating J times the attacks $A(\mathcal{M}(D_{x_0}), a(x_1))$ for each target–challenger pair $x_1, x_0 \in \mathcal{X}$ and averaging the results.

For MNIST, Fashion and Adult, we set $J = 1\,000$ (as in [62]). Note that in the LDP cases $D_- = \varnothing$, and we set $J = 10^6/m$ ensuring the total number of runs matches those 10^6 repetitions of LDP Auditor. Finally, for the imputation attack, we do not require a target model as it is target model-independent and set $J = 1$. We repeat the imputation attack with five different seeds and report the averaged ReRo and RAD scores.

We use Python and TensorFlow [165] to evaluate the attacks. For DP-SGD, we rely on a minimal implementation provided by Hayes et al. [62], which we extend to incorporate



(a) Adult distribution. (b) Uniform distribution. (c) $\pi(0) = \pi(100) = \frac{1}{2}$.

Figure 5.6.: Empirical risk of RAD (crosses) and ReRo (dots) for different error tolerances $\eta \in [0, 100]$ on the Adult dataset with truncated Laplace noise. Straight lines show theoretical RAD bounds, dashed lines ReRo bounds. While RAD bounds closely match the empirical risk, ReRo bounds consistently overestimate the risk. Moreover, ReRo increasingly overestimates the risk at larger error tolerances across all considered distributions.

Dataset	ReRo	RAD
Census	0.81	0
Texas	0.73	0

Table 5.2.: ReRo vs. RAD risk estimation for imputation attack. This type of attack does not access the dataset directly and therefore cannot induce any participation risk, which RAD correctly captures while ReRo significantly overestimates the risk.

RAD and target-specific auxiliary knowledge. For the imputation attack [58], we adapt the authors' public implementation [166].

5.5.3. Results

In this section, we present the RAD and ReRo empirical risk results on real attacks, along with their corresponding theoretical bounds. For both measures, the y-axis represents the estimated risk, with values close to one indicating high risk and values near zero indicating low risk.

5.5.3.1. RAD covers, but ReRo breaks for auxiliary knowledge

Figure 5.4 shows the results of ReRo and RAD risk estimation for our optimal attacks against DP-SGD on the MNIST dataset. Analogous results for the Fashion dataset are provided in Figure 5.5. We also include the corresponding theoretical bounds for ReRo and RAD for comparison. As expected, the existing ReRo bounds [62] correctly provide an upper limit on the empirically observed ReRo risk when the adversary has no prior knowledge of the target record ($aux = \{\varnothing\}$). Figure 5.4a). However, when the adversary has prior knowledge of the victim record (Figures 5.4b and 5.4c), ReRo estimates exceed

the values predicted by their theoretical bounds—which are meant to be upper bounds and, therefore, should never be surpassed by the true risk. In contrast, our RAD bounds consistently upper-limit the empirically estimated RAD risks across all tested attacks.

This supports our expectation that the ReRo bounds only hold under the assumption that the adversary has no auxiliary knowledge about the victim ($aux = \{\varnothing\}$), but fail to correctly estimate privacy risks when target-specific auxiliary knowledge exists.

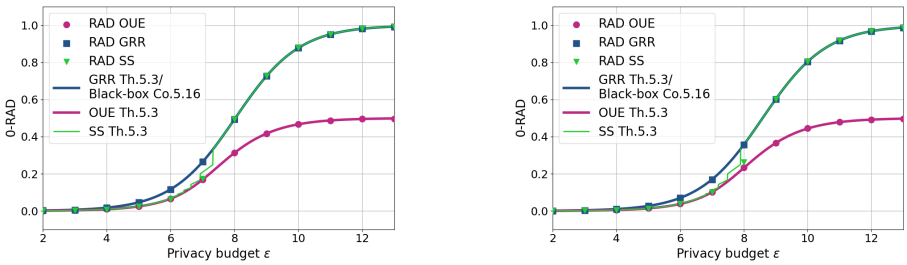
We can also observe that our bounds for RAD overcome this estimation error: they hold for any auxiliary knowledge and are nearly tight. In particular, Figures 5.4b and 5.4c show that the tightness of our worst-case bound Theorem 5.2 is not an isolated feature of GRR, but a reliable property that also applies to other widely used mechanisms, such as DP-SGD. Finally, Figure 5.4a shows that our closed-form bound Theorem 5.11 offers a reasonable upper-bound also when Theorem 5.3 needs to be numerically approximated (as is the case, for instance, with DP-SGD).

5.5.3.2. Leakage vs. Imputation

Table 5.2 compares the risk estimates of RAD and ReRo for the imputation attack. This attack is not based on any information leakage from the mechanism and ignores any output in the process. RAD in this case does estimate the privacy risk to be 0, whereas ReRo reports notably higher values (0.81 for Census and 0.73 for Texas). This underlines how RAD is the more reliable measure of actual privacy risks: RAD shows the absence of leakage when the attack's success relies solely on imputation, whereas ReRo suggests serious disclosures (or: attack potential), effectively overestimating the privacy risk. This result suggests that RAD is a safer choice for risk estimation, as it allows practitioners to measure the true risk of data disclosure without being affected by data imputation.

This tendency of ReRo to overestimate risk is not confined to this setting. In our optimal attacks on DP-SGD (Figure 5.4), ReRo consistently overestimates leakage across all investigated cases, with the effect becoming more pronounced as more auxiliary information is incorporated. Membership inference ($a(x) = z$) provides the clearest example, where ReRo reports risk values exceeding 0.6 even for privacy budgets $\varepsilon \leq 4$, which are commonly considered to offer strong privacy guarantees [133]. This behavior aligns with expectations, as ReRo cannot discount auxiliary information; consequently, greater attacker knowledge leads to larger overestimation.

Similarly, Figure 5.6 shows that ReRo fails to capture the effect of the success threshold η. As η increases, an oblivious attacker's success probability rises, but ReRo cannot account for this since it depends only on success probability and thus converges to 1 for all ε. This results in substantial overestimation: for $\eta = 100$, a trivial setting where any guess is correct, ReRo reports maximal risk despite the mechanism providing no advantage. In contrast, RAD properly discounts this effect, showing that increasing η boosts advantage only up to a point (here, $\eta = 40$), after which the advantage decreases as success becomes nearly granted.



(a) Porto (b) Geolife

Figure 5.7.: RAD results for LDP mechanisms. Lines show theoretical bounds and markers empirical RAD. First, we note that our bounds are perfectly tight for all tested mechanisms and datasets. Adittionally, we see that OUE offers the higher protection among the LDP mechanisms even for the same ε choices.

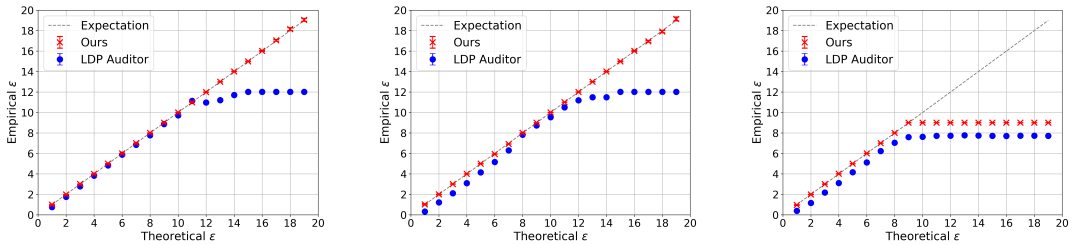
5.5.3.3. Bound tightness

Figure 5.6 shows the results of RAD and ReRo for our optimal attack against Laplace mechanism on Adult including their corresponding theoretical bounds. Figures 5.7a and 5.7b shows the analogous for LDP mechanisms, GRR, OUE and SS, on the Porto and Geolife datasets. On the x-axis, we see ε, and on the y-axis, the exact estimated risk for such ε selection. Note that for LDP, RAD and ReRo results coincide, since the attack relies solely on the released output (with no auxiliary information or imputation effects). Moreover, the prior-based chance level under the uniform prior is negligible for $|\mathcal{X}| = 3,052$. We therefore report only RAD to avoid redundancy.

We observe that our bounds (cf. Theorem 5.3) are tight for every prior π and capture even subtle differences between mechanisms. In particular, the RAD estimates for GRR perfectly match our perfect-reconstruction black-box bound (Theorem 5.15), confirming its tightness.

Moreover, Figure 5.6 clearly illustrates the impact of the data distribution: the skewed distribution (Figure 5.6c) constitutes the worst case, while the empirical distribution represents the best case. This highlights that knowledge of the data distribution can substantially improve utility; in the absence of such knowledge, the only safe choice is calibration with respect to the worst case.

Finally, these results provide concrete evidence for the importance of attack-based noise calibration. For identical values of ε, OUE offers significantly stronger protection against DRAs than GRR and SS. Hence, ε alone does not capture the full privacy picture, and RAD is essential for understanding the actual privacy implications of a mechanism for users.



(a) (GRR) (b) (SS) (c) (OUE)

Figure 5.8.: LDP Audit results from RAD-based auditing and LDP Auditor [41] on Porto dataset. Values along the diagonal indicate perfect accuracy; below it, privacy is overestimated; above it, underestimated.

5.5.3.4. Auditing Local DP with RAD

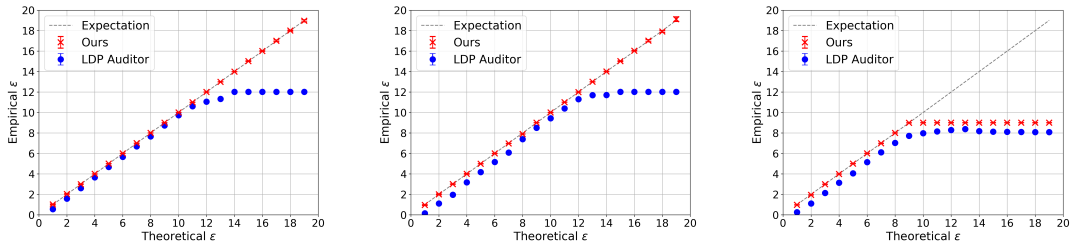
Figures 5.8 and 5.9 show the results from our LDP auditing experiments using the Porto and Geolife datasets. They compare the accuracy of predicting the actual ε using our RAD-based auditing versus LDP auditor. The closer the empirical ε is to the theoretical value (diagonal line), the more accurate the auditing tool. Additionally, smaller standard deviations indicate greater stability of the method.

For all tested mechanisms, our auditing approach improves over LDP Auditor for all ε values. In particular, we see that the highest ε LDP Auditor manages to estimate for both GRR and SS are capped around $\widetilde{\varepsilon} \approx 12.25$, hence preventing auditing of deployments with higher values. This limitation was already acknowledged by the authors of LDP Auditor, as it stems from the intrinsic shortcomings of the Clopper-Pearson method underlying their approach [41]. In contrast, the tightness of our RAD bound enables our auditing approach to accurately estimate empirical privacy budgets for the whole range, without such a limitation. Notably, for GRR and SS, our DP auditing yields near-perfect estimates for all epsilon values. For the OUE mechanism, our approach also outperforms LDP Auditor, however, the estimation accuracy declines at $\varepsilon \leq 9$. Note that this is an inherent limitation of OUE auditing as already mentioned in [41]: as we prove in Example A.17, 0-RAD converges to $\frac{m-1}{2m}$ when ε tends to infinity. Overall, these results support that the universal tightness of our theoretical bound Theorem 5.3 enables precise and reliable auditing based on DRAs.

5.6. Conclusion

In this chapter, we formally and empirically investigate the reconstruction risk that users incur when their data are processed by DP mechanisms.

Our results reveal that the current state-of-the-art risk metric, ReRo [39], tends to overestimate the actual leakage of DP mechanisms when the attacker leverages prior knowledge or public statistics, which can lead to excessive utility loss if used for noise calibration. Furthermore, we demonstrate that under realistic attacks—where the attacker



(a) (GRR) (b) (SS) (c) (OUE)

Figure 5.9.: LDP Audit results from RAD-based auditing and LDP Auditor [41] on Geolife dataset. Values along the diagonal indicate perfect accuracy; below it, privacy is overestimated; above it, underestimated.

exploits target-specific public knowledge—existing ReRo bounds, which were originally proven only for $aux = \{\varnothing\}$, can be violated.

To address these limitations, we first introduce η-RAD, a novel metric that captures the general reconstruction threat and accurately quantifies the privacy risk imposed by any specific mechanism. This metric explicitly accounts for target-specific knowledge and the effects of imputation and prior information on the risk. Our theoretical and empirical evaluation—across different data domains, DP mechanisms, and levels of attacker knowledge—demonstrates the robustness of RAD as a risk measure and its ability to correctly account for the leakage arising from participation in the dataset.

More importantly, we advance the understanding and practical interpretation of DP guarantees by proving tight bounds that connect DP mechanisms with their risk, using RAD. Offering new insights and clarity beyond existing analyses, we establish (i) universally tight bounds when the attacker's knowledge is specified, along with optimal strategies achieving them, (ii) closed-form bounds that remain valid regardless of auxiliary knowledge, and (iii) black-box upper bounds enabling risk estimation based on the privacy parameters, applicable in settings where the whole record is considered secret.

We provide a principled analysis of the tightness of our bounds by formalizing the optimal attack strategy for any reconstruction goal. Specifically, given an attack goal, an error threshold, and the auxiliary knowledge of the attacker, we provide a general algorithm that the attacker can follow to maximize RAD. We prove that this attack strategy is indeed optimal and further validate these findings empirically by implementing the attack on real datasets.

Our theoretical and empirical evaluation—across private learning, DP aggregation, and LDP settings—demonstrates not only the robustness of RAD as a risk measure, but also the significant impact of our bounds on improving DP noise calibration (resulting in better utility) and auditing (broadening scope and enhancing accuracy). In particular, we present a RAD-based auditing framework that is data-agnostic, meaning it does not require artificially approximating worst-case dataset pairs. This increases efficiency, especially for high-dimensional categorical data, while providing better accuracy than existing auditing tools of comparable scope.

Overall, our work demonstrates that privacy risk depends on the mechanism's structure, not just its nominal privacy parameters, and provides both fundamental insight and practical tools for privacy risk assessment and calibration—enabling notable utility gains without increasing the effective privacy risk.

6. Balancing Privacy and Utility in Correlated Data

This chapter is based on the contributions:

- Martin Lange*, **Patricia Guerra-Balboa***, Javier Parra-Arnau, and Thorsten Strufe. "Balancing Privacy and Utility in Correlated Data: A Study of Bayesian Differential Privacy". In: Proceedings of the VLDB Endowment, 2025, DOI: 0.14778/3749646.3749679.

In previous chapters, we focused on addressing DP challenges arising from complex data while adhering to the original assumptions of the DP model. Under this model, the adversary is assumed to be fully informed. In particular, in the bounded DP setting, the attacker is assumed to know all database entries except that of the target individual. For general granularities, this assumption is extended to all information in the database except what is deemed indistinguishable under the chosen granularity relation.

However, under this assumption the protection guarantees are limited to statistically independent data records, i.e., DP mechanisms underestimate participation risk when the underlying data are correlated (see Sections 2.3 and 3.5).

The limitations of DP for protecting correlated data have been theoretically exposed [29], [30], [32], [36] and empirically confirmed with attacks on real databases [33]. This is a significant issue, since correlations among data records are common in real-world databases, such as those induced by friendships in social networks [34], genetic similarities among family members [35] or physical limitations in trajectories (see Chapter 3).

As a response to the limitations of DP in the presence of correlation, several DP-based notions have been proposed to specifically address this challenge [30], [68], [167], [168], [169]. Among them, *Bayesian Differential Privacy* (BDP) [37] stands out for its simplicity and generality: it provides stronger privacy guarantees than DP and supports arbitrary correlation structures—capabilities that are not generally achievable within the Pufferfish framework. BDP also underlies extensions such as prior DP [168] and correlated DP for location data [169].

While DP assumes the adversary knows all records except the target, BDP considers arbitrary priors, including those where unknown records are correlated. It ensures bounded changes in output distributions even when the target record is part of a correlated subset. When data is independent, BDP and DP coincide. Under correlation, however, BDP quantifies worst-case leakage by integrating the mechanism's output with the data distribution via Bayes' rule, capturing adversarial advantages that DP

overlooks. Hence, BDP mitigates correlation-driven reconstruction attacks that breach DP's guarantees as empirically shown in [69].

While BDP provides a robust framework for assessing privacy leakage under data dependencies, its practical applicability remains uncertain. The few mechanisms that satisfy this notion [37], [69] are limited to specific correlation models, such as Gaussian Markov random fields—a subclass of multivariate Gaussian distributions forming a Markov random field where missing edges correspond to zeros in the inverse covariance matrix [64]—and binary-state Markov chains with a symmetric transition matrix. Given the scarcity of mechanisms and their applicability restrictions, it remains unclear whether BDP can provide utility in real-world applications. Moreover, the only solution for Gaussian Markov fields reported poor utility, since noise addition scales linearly with the number of records in the database and their only mitigation is to weaken BDP privacy by incorporating assumptions about the adversary [37]. Moreover, prior impossibility results [29], [68] show that strong utility under BDP without distributional assumptions is fundamentally limited.

In summary, DP privacy leakage estimation does not provide sufficient protection under data dependencies, and there is a need for improved utility with the robust BDP framework. Motivated by this issue, this chapter examines BDP's utility from both theoretical and practical perspectives, analyzing its limitations and proposing new strategies to reduce utility loss while maintaining BDP privacy guarantees.

Our investigation focuses on understanding how DP leakage can be translated into BDP leakage across different data contexts. This understanding provides a systematic way to construct BDP mechanisms by appropriately adjusting the parameters of existing DP mechanisms. Furthermore, we analyze the accuracy of BDP mechanisms obtained through this calibration methodology.

In particular, we investigate theoretical bounds on the accuracy of BDP mechanisms and derive specific utility guarantees under certain correlation models. Specifically, we analyze the impact of limiting the number of correlated records and study the applicability of BDP to both discrete and continuous correlation models. For each case, we focus our analysis on a representative distribution that has a significant impact on data analysis. Specifically, for the discrete case, we analyze data following a Markov chain and, for continuous data, we analyze multivariate Gaussian correlation. We focus on these two particular correlation models following previous work in BDP [37], [168] and due to their relevance in many real-world applications such as medical [170], location [100], or activity data [171] analyses.

Finally, we complement our formal analysis with an empirical evaluation on real-world data containing Gaussian and Markov correlations. This allows us to validate the formal utility analysis of BDP mechanisms in practical applications and to gain insight into how our theoretical results translate into practice. In summary, in light of the fundamental question of whether it is possible to obtain accurate global information while protecting privacy in the presence of correlations—and given the existing impossibility results that generally suggest otherwise—this chapter investigates the following questions:

- Is it possible to translate DP leakage into BDP leakage in a way that allows obtaining accurate BDP mechanisms by recalibrating the noise injection of existing DP mechanisms?

- Can useful analyses under correlation be performed by making realistic and sensible distributional assumptions, either through high-level restrictions, such as bounding the number of correlated records, or through stronger assumptions, such as specifying particular distributions?

6.1. Related Work

The challenge of designing privacy mechanisms that remain robust under arbitrary correlations has been a central concern in the development of privacy frameworks. Foundational work by Kifer and Machanavajjhala [29] introduced free-lunch privacy, the first formalism to consider the impact of correlations on privacy guarantees. Their no-free-lunch theorem shows that, under arbitrary data distributions, achievable utility is fundamentally constrained. However, they express utility in terms of discriminants—an abstraction that is neither intuitive nor translatable into practical utility metrics.

The existing strategy applicable to obtain BDP [172] requires noise calibration based on the Wasserstein distance. However, it requires computing the Wasserstein distance between the conditional output distributions corresponding to all pairs of sensitive values. This is computationally intractable [172], [173] in the general case. While a closed-form mechanism is derived for specific Markov chain models, it relies on a weakened notion that assumes limited adversarial background knowledge, and therefore cannot be meaningfully compared to BDP.

The only concrete evidence of the potential applicability of pure BDP in practice has been provided in the context of Gaussian and Markov correlation models. In their foundational work, Yang et al. [37] proposed adapting the Laplace mechanism to defend against correlated leakage in Gaussian Markov Random Fields. They also established preliminary theoretical connections between DP and BDP in this setting. Despite these important contributions, the proposed mechanisms' privacy guarantees degrade linearly with the number of correlated records, resulting in excessive noise that renders the mechanism impractical. Although the authors suggest mitigating this by limiting the adversary's knowledge, such a compromise weakens the privacy model and undermines the core guarantees of BDP. Moreover, the proposed mechanisms have not been evaluated in real-world scenarios, leaving their practical effectiveness uncertain.

A more recent effort by Chakrabarti et al. [69] proposes an adaptation of the randomized response to BDP over binary Markov chains. However, this mechanism is extremely constrained: it only applies to lazy, binary, stationary Markov chains and does not provide any general bounds relating DP and BDP leakage. Moreover, the only closed-form expressions for mechanism parameters holds under the restrictive assumption of a symmetric transition matrix limiting its usability even further.

In response to these limitations, several relaxed privacy notions have been proposed to strike a better balance between privacy and utility. Mutual Information Privacy (MI DP) [174] and its extension [173], for example, can be viewed as a relaxation, offering a framework where traditional mechanisms like Laplace and Gaussian can be recalibrated to account for correlation. These methods yield promising theoretical utility guarantees. However, MI guarantees are weaker, in particular, MI characterizes average-case privacy leakage rather than worst-case guarantees, and therefore cannot substitute the BDP framework when worst-case guarantees are desired.

In conclusion, while previous work highlights the limitations of DP protection and the need for BDP as a privacy standard, the challenge of providing utility with BDP protection remains unresolved. Moreover, the relationship between DP and BDP is not yet fully understood.

6.2. Limited Number of Correlated Variables

Given that the actual data distribution is often unclear or hard to estimate [175], it would be great to be able to protect against any potential correlation in the data. Formally, to protect against potential correlations without making distributional assumptions a mechanism must satisfy BDP with respect to all possible correlation distributions Π, a condition we call protection under *arbitrary correlation*. However, Kifer and Machanavajjhala showed that under this assumption BDP collapses to free-lunch privacy [29], [37][1]. Free-lunch privacy is a granularity notion that considers all pairs of datasets neighbors. That implies that all dataset pairs are at distance ε, forcing all query outputs $f(D)$ and $f(D')$ to be ε-indistinguishable [176]—intuitively implying a complete loss of utility.

To better understand the underlying utility issue, we formalize this limitation using the standard (α, β)-accuracy metric, providing a concrete, interpretable, and widely adopted measure of utility loss that enables clearer reasoning and meaningful comparisons across different randomized mechanisms.

Proposition 6.1. *Let $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}$ be an ε-BDP mechanism protecting against arbitrary correlation. Let $0 \leq \beta < \frac{1}{e^{\varepsilon}+1}$ be a real number and let $f : \mathcal{X}^n \to \mathbb{R}$ be a deterministic function. If $\mathcal{M}$ is (α, β)-accurate w.r.t. f, then*

$$\alpha > \frac{1}{2} \max_{D,D'} |f(D) - f(D')|.$$

Proof. First, note that any BDP mechanism protecting against arbitrary correlation is free-lunch [68]. Therefore for every $S \subseteq \mathbb{R}$, and for every pair $D, D \in \mathcal{X}^n$,

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S]. \tag{6.1}$$

[1]The original proof is stated within the Pufferfish privacy framework, which generalizes Bayesian differential privacy (BDP) to arbitrary forms of prior knowledge and to secrets beyond the protection of a single record.

We use this property and proceed by *reductio ad absurdum*. We assume that $\mathcal{M}$ fulfills an (α, β)-accuracy with respect to f with $\alpha \leq \frac{1}{2} \max_{D,D'} |f(D) - f(D')|$ and $\beta < \frac{1}{e^{\varepsilon}+1}$, i.e., for all D, D'

$$\Pr\left[|f(D) - \mathcal{M}(D)| \geq \frac{1}{2}|f(D) - f(D')|\right] \leq \frac{1}{\mathrm{e}^{\varepsilon} + 1},$$

and derive a contradiction for D':

$$\begin{aligned}
\Pr\left[|f(D') - \mathcal{M}(D')| \geq \alpha\right] &= \Pr[\mathcal{M}(D') \in \mathbb{R} \setminus (f(D') - \alpha, f(D') + \alpha)] \\
&\geq \Pr[\mathcal{M}(D') \in (f(D) - \alpha, f(D) + \alpha)] && (6.2) \\
&\geq e^{-\varepsilon} \Pr[\mathcal{M}(D) \in (f(D) - \alpha, f(D) + \alpha)] && (6.3) \\
&= e^{-\varepsilon} (1 - \Pr[\mathcal{M}(D) \in \mathbb{R} \setminus (f(D) - \alpha, f(D) + \alpha)]) \\
&= e^{-\varepsilon} (1 - \Pr[|f(D) - \mathcal{M}(D)| \geq \alpha]) \\
&\geq e^{-\varepsilon} (1 - \beta) && (6.4) \\
&> e^{-\varepsilon} (1 - \frac{1}{\mathrm{e}^{\varepsilon} + 1}) && (6.5) \\
&= \frac{1}{e^{\varepsilon} + 1} \\
&> \beta,
\end{aligned}$$

where Eq. 6.2 follows because the probability of a subset is always smaller or equal to the probability of the greater set. Here, this subset relationship $(f(D) - \alpha, f(D) + \alpha) \subseteq \mathbb{R} \setminus (f(D') - \alpha, f(D') + \alpha)$ holds because α is smaller or equal to half the distance between $f(D)$ and $f(D')$. This probability can once again be limited in Eq. 6.3 applying Eq. 6.1.

Finally, we use that $\mathcal{M}$ is (α, β)-accurate in Eq. 6.4 and subsequently replace β by its upper bound in Eq. 6.5.

Overall, it follows that $\Pr[|f(D') - \mathcal{M}(D')| \geq \alpha]$ is strictly greater than β. This is in contradiction to (α, β)-accuracy. □

Specifically, the result from Proposition 6.1 indicates that for theoretically relevant privacy levels $\varepsilon \in (0, 4)$ [133], the only confidence interval where we can reliably estimate the actual value of our query function f, with standard confidence levels (e.g., between 90% and 99%), includes almost all possible query values. For instance, consider a free-lunch algorithm used to compute $f(D)$, where f counts the number of infections in a database of n individuals. If the algorithm outputs $\frac{n}{2}$, it suggests that half of the population is infected. However, with a 90% confidence interval, we cannot tell whether there is no infection at all, or whether the entire population is infected.

While designing accurate BDP mechanisms is infeasible under arbitrary correlation—potentially involving all records—it is often reasonable in practice to assume that only subsets of records are correlated. For instance, in the context of genomic data, an individual's genome is strongly correlated with that of their relatives, but not with the entire population [35]. Hence, we assume that only m of n records in the database are correlated with each other, formally:

Definition 6.2. We say the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ has *at most* $m \leq n$ *correlated random variables* if there exist disjoint sets of indices $C_1, \ldots, C_r$ that make up $[n] = \bigcup_{l=1}^{r} C_l$ so that each set C_l has maximum cardinality $m \geq |C_l|$ for any $l \in [r]$, and for any $l \in [r]$, the random variables $\{X_j \mid j \in C_l\}$ are independent of the remaining random variables $\{X_j \mid j \in [n] \setminus C_l\}$.

This definition considers multiple groups of up to m correlated records as long as they do not "overlap", i.e., the records in one group are independent of the records in the other groups. Otherwise, we do not make any further assumptions about the distribution of the data. This allows us to find acceptable utility guarantees in Corollary 6.5 as long as m is sufficiently small.

6.2.1. Relationship between DP and BDP

We begin by introducing and proving a general bound on the BDP leakage of an ε-DP mechanism. Specifically, we show that if an ε-DP mechanism operates on data drawn from a distribution involving at most m correlated random variables, then it satisfies $m\varepsilon$-BDP. The practice of scaling the DP leakage by the number of correlated records to estimate worst-case leakage under correlation has been used in prior work [30], [177], but to our knowledge, this approach had not been formally shown to satisfy the BDP definition. We further prove that this bound is tight.

Theorem 6.3 (The General Bound)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a random vector with at most* $m \leq n$ *correlated random variables. Then, any* ε*-DP mechanism with input data drawn from* $\mathbf{X}$ *is* $m\varepsilon$*-BDP.*

Proof. Since BDP considers the supremum of the leakage over all possible adversaries, we prove that this bound holds for any adversary (K, i) with $i \in [n]$, $K \subseteq [n] \setminus \{i\}$ and $k = |K|$; hence, it also holds for the supremum.

Since $\{C_j\}_{j \in r}$ is a partition of $[n]$, there exists an $l \in [r]$ so that the target index $i \in C_l$. Thus, C_l contains the index i and all indices of random variables potentially correlated with X_i. Let $\tilde{C} := C_l \setminus K$ be the indices of random variables correlated with X_i that are not already included in K.

Then, we first show that the adversary-specific BDPL can be upper bounded as follows:

$$\mathrm{BDPL}_{(K,i)} = \sup_{S, \mathbf{x}_K, x_i, x_i'} \ln \frac{\Pr[Y \in S \mid \mathbf{x}_K, x_i]}{\Pr[Y \in S \mid \mathbf{x}_K, x_i']} \leq \sup_{S, \mathbf{x}_K, \mathbf{x}_{\tilde{C}}, \mathbf{x}'_{\tilde{C}}} \ln \frac{\Pr[Y \in S \mid \mathbf{x}_K, \mathbf{x}_{\tilde{C}}]}{\Pr[Y \in S \mid \mathbf{x}_K, \mathbf{x}'_{\tilde{C}}]}. \quad (6.6)$$

Assume that for all $\mathbf{x}_{\tilde{C}} \in \mathcal{X}^{|\tilde{C}|}$, we have

$$\Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i] > \Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, \mathbf{X}_{\tilde{C}} = \mathbf{x}_{\tilde{C}}]. \quad (6.7)$$

Now, we bring this to a contradiction, thereby proving the opposite of Eq. 6.7. Let index set $\tilde{C}_{-i} := \tilde{C} \setminus \{i\}$ include all indices of $\tilde{C}$ except for i. Then, we have

$$\Pr[Y \in S \mid \mathbf{x}_K, x_i] = \int_{\mathcal{X}^{|\tilde{C}_{-i}|}} \Pr[Y \in S | \mathbf{x}_K, x_i, \mathbf{x}_{\tilde{C}_{-i}}] \, p_{\mathbf{X}_{\tilde{C}_{-i}}}(\mathbf{x}_{\tilde{C}_{-i}} | \mathbf{x}_K, x_i) \, \mathrm{d}\mathbf{x}_{\tilde{C}_{-i}}$$

$$= \int_{\mathcal{X}^{|\tilde{C}_{-i}|}} \Pr[Y \in S \mid \mathbf{x}_K, (x_i, \mathbf{x}_{\tilde{C}_{-i}})]\; p_{\mathbf{X}_{\tilde{C}_{-i}}}(\mathbf{x}_{\tilde{C}_{-i}} \mid \mathbf{x}_K, x_i)\, \mathrm{d}\mathbf{x}_{\tilde{C}_{-i}} \tag{6.8}$$

$$< \int_{\mathcal{X}^{|\tilde{C}_{-i}|}} \Pr[Y \in S \mid \mathbf{x}_K, x_i]\; p_{\mathbf{X}_{\tilde{C}_{-i}}}(\mathbf{x}_{\tilde{C}_{-i}} \mid \mathbf{x}_K, x_i)\, \mathrm{d}\mathbf{x}_{\tilde{C}_{-i}} \tag{6.9}$$

$$= \Pr[Y \in S \mid \mathbf{x}_K, x_i] \int_{\mathcal{X}^{|\tilde{C}_{-i}|}} p_{\mathbf{X}_{\tilde{C}_{-i}}}(\mathbf{x}_{\tilde{C}_{-i}} \mid \mathbf{x}_K, x_i)\, \mathrm{d}\mathbf{x}_{\tilde{C}_{-i}} \tag{6.10}$$

$$= \Pr[Y \in S \mid \mathbf{x}_K, x_i], \tag{6.11}$$

where the random variable X_i is already included in the condition, so only indices $\tilde{C}_{-i}$ need to be added. In Eq. 6.8, we combine the two conditions $X_i = x_i$ and $\mathbf{X}_{\tilde{C}_{-i}} = \mathbf{x}_{\tilde{C}_{-i}}$ into one condition $\mathbf{X}_{\tilde{C}} = (x_i, \mathbf{x}_{\tilde{C}_{-i}})$. Notice that this is the same condition, just stated differently. Then, we use Eq. 6.7, which applies to all $\mathbf{x}_{\tilde{C}} \in \mathcal{X}^{|\tilde{C}|}$, in Eq. 6.9. Now, the first probability can be pulled out of the integral in Eq. 6.10 as it no longer depends on the value $\mathbf{x}_{\tilde{C}_{-i}}$. The final Eq. 6.11 follows because a probability density integrated over its entire domain is always one.

Note that if the variables are discrete the integral must be changed by a sum and the result follows analogously.

We have shown that the initial probability is strictly smaller than itself—a contradiction. Thus, the opposite of our assumption in Eq. 6.7 must be true and there must exist a vector $\mathbf{x}_{\tilde{C}}$ so that

$$\Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i] \leq \Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, \mathbf{X}_{\tilde{C}} = \mathbf{x}_{\tilde{C}}]. \tag{6.12}$$

Analogously, we can show that there exists a vector $\mathbf{x}'_{\tilde{C}} \in \mathcal{X}^{|\tilde{C}|}$ verifying

$$\Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x'_i] \geq \Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, \mathbf{X}_{\tilde{C}} = \mathbf{x}_{\tilde{C}}]. \tag{6.13}$$

Thus, with Eq. 6.12 and Eq. 6.13 we have Eq. 6.6. For any values of $\mathbf{x}_K$ and x_i, x'_i included in the left supremum, we can always find values $\mathbf{x}_{\tilde{C}}$ and $\mathbf{x}'_{\tilde{C}}$ so that the ratio becomes greater or equal, and these values $\mathbf{x}_{\tilde{C}}, \mathbf{x}'_{\tilde{C}}$ are included in the supremum on the right-hand side.

Let the set $U = [n] \setminus (K \cup \tilde{C})$, with $u = |U|$, include all remaining indices. Since by hypotheses $|\tilde{C}| \leq |C_l| \leq m$, for any known values $\mathbf{x}_K \in \mathcal{X}^k$, the correlated values $\mathbf{x}_{\tilde{C}} \in \mathcal{X}^{|\tilde{C}|}$ and $\mathbf{x}'_{\tilde{C}} \in \mathcal{X}^{|\tilde{C}|}$. Applying Eq. 6.6 we have

$$\begin{aligned}
\Pr_{\mathcal{M}}[Y \in S \mid \mathbf{x}_K, \mathbf{x}_{\tilde{C}}] &= \int_{\mathcal{X}^u} \Pr_{\mathcal{M}}[Y \in S \mid \mathbf{x}_K, \mathbf{x}_{\tilde{C}}, \mathbf{x}_U]\; p_{\mathbf{X}_U}(\mathbf{x}_U \mid \mathbf{x}_K, \mathbf{x}_{\tilde{C}})\, \mathrm{d}\mathbf{x}_U \\
&\leq \int_{\mathcal{X}^u} \mathrm{e}^{m\varepsilon} \Pr_{\mathcal{M}}[Y \in S | \mathbf{x}_K, \mathbf{x}'_{\tilde{C}}, \mathbf{x}_U]\; p_{\mathbf{X}_U}(\mathbf{x}_U | \mathbf{x}_K, \mathbf{x}_{\tilde{C}})\, \mathrm{d}\mathbf{x}_U \\
&= e^{m\varepsilon} \int_{\mathcal{X}^u} \Pr_{\mathcal{M}}[Y \in S \mid \mathbf{x}_K, \mathbf{x}'_{\tilde{C}}, \mathbf{x}_U]\; p_{\mathbf{X}_U}(\mathbf{x}_U \mid \mathbf{x}_K)\, \mathrm{d}\mathbf{x}_U \\
&= e^{m\varepsilon} \int_{\mathcal{X}^u} \Pr[Y \in S | \mathbf{x}_K, \mathbf{x}'_{\tilde{C}}, \mathbf{x}_U]\; p_{\mathbf{X}_U}(\mathbf{x}_U | \mathbf{x}_K, \mathbf{x}'_{\tilde{C}})\, \mathrm{d}\mathbf{x}_U \\
&= e^{m\varepsilon} \Pr[Y \in S \mid \mathbf{x}_K, \mathbf{x}'_{\tilde{C}}].
\end{aligned}$$

Combining both inequalities we obtain the result. □

This bound may appear overly pessimistic, as it seems to assume perfect correlation—where a change in one record fully determines the changes in all others.

However, as we show in the following example, the bound remains tight even when this extreme case is excluded. Specifically, we provide a counterexample in which there exists a family of probability distributions $\{\Pi_r\}_{r\in\mathbb{N}}$ such that for every r the variables do not deterministically determine one another while for any $\delta > 0$ there exist one r verifying that BDPL $> 2\varepsilon - \delta$. This confirms both the tightness of our result and that the bound cannot be improved, even excluding the edge case of perfect correlation.

	$X_1 = 0$	$X_1 = 1$	Total
$X_2 = 0$	$\frac{1}{r^2}$	$\frac{r-1}{r^2}$	$\frac{1}{r}$
$X_2 = 1$	$\frac{1}{r^3}$	$\frac{r^3-r^2-1}{r^3}$	$\frac{r-1}{r}$
Total	$\frac{1+r}{r^3}$	$\frac{r^3-r-1}{r^3}$	1

Table 6.1.: Probability distribution of Example 6.4

Example 6.4. For all $r \in \mathbb{N}$ with $r \geq 2$, Table 6.1 shows a valid probability distribution Π_r for $\mathbf{X} = (X_1, X_2)$.

Moreover, given any $r > 2$, there is not perfect dependency, i.e., neither X_1 is fully determined by X_2 nor vice versa (see Section A.4).

Moreover, if $\mathcal{M}$ is ε-DP, then there are two neighboring databases $D, D' \in \{0,1\}^2$ for which the privacy loss reaches ε; otherwise, ε is not tight for $\mathcal{M}$, and a smaller value could be used with the same reasoning. Without loss of generality, we assume they differ in the first coordinate, otherwise by inverting Table 6.1 we get the same result and we assume that it exits $S \subset \Theta$ such that

$$\Pr[A(0,0) \in S] = \mathrm{e}^{\varepsilon} \Pr[A(0,1) \in S] = e^{\varepsilon} \Pr[A(1,0) \in S] = \mathrm{e}^{2\varepsilon} \Pr[A(1,1) \in S],$$

as it is the case, for instance, for the GRR mechanism with $S = \{(1,1)\}$ (see Example 5.4). Then, computing the BDPL we obtain:

$$\begin{aligned}
\mathrm{e}^{\mathrm{BDPL}} &\geq \frac{\Pr[Y \in S \mid X_1 = 0]}{\Pr[Y \in S \mid X_1 = 1]} \\
&= \frac{\sum_{x_2\in\{0,1\}} \Pr[Y \in S \mid X_1 = 0, X_2 = x_2] \Pr[X_2 = x_2 \mid X_1 = 0]}{\sum_{x_2\in\{0,1\}} \Pr[Y \in S \mid X_1 = 1, X_2 = x_2] \Pr[X_2 = x_2 \mid X_1 = 1]} \\
&= \frac{e^{2\varepsilon} \Pr[\mathcal{M}(1,1) \in S] \frac{r}{r+1} + e^{\varepsilon} \Pr[\mathcal{M}(1,1) \in S] \frac{1}{r+1}}{e^{\varepsilon} \Pr[\mathcal{M}(1,1) \in S] \frac{r^2-r}{r^3-r-1} + \Pr[\mathcal{M}(1,1) \in S] \frac{r^3-r^2-1}{r^3-r-1}} \\
&= \frac{e^{2\varepsilon} \frac{r}{r+1} + e^{\varepsilon} \frac{1}{r+1}}{e^{\varepsilon} \frac{r^2-r}{r^3-r-1} + \frac{r^3-r^2-1}{r^3-r-1}},
\end{aligned}$$

for all $r > 2$. Since

$$\lim_{r\to\infty} \frac{e^{2\varepsilon}\,\frac{r}{r+1} + e^{\varepsilon}\,\frac{1}{r+1}}{e^{\varepsilon}\,\frac{r^2-r}{r^3-r-1} + \frac{r^3-r^2-1}{r^3-r-1}} = \mathrm{e}^{2\varepsilon},$$

taking the limit when r tends to infinity we have BDPL $\geq 2\varepsilon$. Since the general upper bound of the BDPL of an ε-DP mechanism is 2ε we have BDPL $= 2\varepsilon$. Therefore, taking arbitrary large r, we obtain not perfectly determined variables, and the BDPL arbitrary close to 2ε.

Example 6.4 shows that, without additional hypotheses, Theorem 6.3 is tight, even excluding perfect correlation.

6.2.2. Accuracy

Theorem 6.3 enables to use $(\frac{\varepsilon}{m})$-DP mechanisms as ε-BDP mechanisms. However, reducing ε in a DP mechanism often has a negative impact on utility (see Section 2.1). In particular, we investigate the impact on the accuracy of the Laplace mechanism. As a consequence of Theorem 6.3 and Proposition 2.10 we obtain the following result:

Corollary 6.5. *Let $\mathcal{M}_{\varepsilon,f}$ be the Laplace ε-DP mechanism that approximates the query $f\colon \mathcal{X}^n \to \mathbb{R}$ with input described by the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with at most $m \leq n$ correlated random variables. Then, if $\mathcal{M}_{\varepsilon,f}$ is (α,β)-accurate w.r.t. f, there exists an ε-BDP mechanism $\widetilde{\mathcal{M}}$ whose input is drawn from $\mathbf{X}$ and that is $(m\alpha,\beta)$-accurate w.r.t. f.*

Proof. From Proposition 2.10, we know that the (α,β)-accuracy of $\mathcal{M}_{\varepsilon,f}$ with respect to f is

$$\alpha = \ln\left(\frac{1}{\beta}\right) \cdot \frac{\Delta f}{\varepsilon}$$

for any $\beta \in (0,1]$ because $\mathcal{M}_{\varepsilon,f}$ uses the Laplace mechanism.

The idea is to also use the Laplace mechanism for $\widetilde{\mathcal{M}}$, but to use an adjusted DP privacy parameter ε' so that $\widetilde{\mathcal{M}}$ is ε-BDP. We will see that this results in $(m\alpha,\beta)$-accuracy. With the general bound from Theorem 6.3, we know that the Laplace mechanism $\tilde{\mathcal{M}} = \mathcal{M}_{\varepsilon',f}$ is $m\varepsilon'$-BDP. Thus, we have

$$m\varepsilon' = \varepsilon \;\Leftrightarrow\; \varepsilon' = \frac{1}{m}\varepsilon.$$

Now, we can calculate the accuracy of mechanism $\widetilde{\mathcal{M}}$ because it also uses the Laplace mechanism and we now know the used DP privacy leakage $\varepsilon' = \varepsilon/m$. Mechanism $\widetilde{\mathcal{M}}$ is (α',β)-accurate with

$$\alpha' = \ln\left(\frac{1}{\beta}\right) \cdot \frac{\Delta f}{\varepsilon'} = \ln\left(\frac{1}{\beta}\right) \cdot \frac{m\Delta f}{\varepsilon} = m\alpha.$$

Thus, the mechanism $\widetilde{\mathcal{M}}$ is $(m\alpha,\beta)$-accurate. □

This result shows that the error α of the Laplace mechanism increases proportionally with the number of correlated records when moving from ε-DP to ε-BDP, and while making no assumption about the distribution of the records. This may be acceptable when the number of correlated records m is small. For example, if $m = 2$, the error α doubles when transitioning from DP to BDP. If the DP mechanism's error is small, this increase may be acceptable. However, utility sharply decreases as m grows.

Since we have shown that our bound on BDPL is tight under the assumption of arbitrary correlation, the utility bound cannot be improved. This motivates the next two sections, where we investigate whether additional assumptions on the correlation model can lead to tighter bounds, enabling reduced noise and improved utility while still protecting against correlation attacks.

6.3. Multivariate Gaussian Correlation

A wide variety of phenomena are effectively modeled using a Gaussian distribution [178]. For example, physiological measures such as height and weight are correlated among family members, and the joint distribution of height and weight in a large population is well fit by a bivariate Gaussian distribution [170]. Consequently, we explore the applicability of BDP to multivariate Gaussian data.

When we are dealing with a database of n records, and each record is drawn from a Gaussian distribution, we can model the joint distribution of all records as a multivariate Gaussian distribution. This model also captures linear correlation between records [179].

Definition 6.6 (Multivariate Gaussian Distribution [179])**.** Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector, let vector $\mu \in \mathbb{R}^n$ be real and let matrix $\Sigma \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. We say $\mathbf{X}$ follows the *multivariate Gaussian distribution with mean* μ *and covariance* Σ if the probability density of $\mathbf{X}$ for any point $\mathbf{x} \in \mathbb{R}^n$ is

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\Pi)^n |\Sigma|}} \exp(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)),$$

where $|\Sigma|$ is the determinant of Σ. We write $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$.

We establish a relationship between DP and BDP for data drawn from a multivariate Gaussian distribution, based on the maximum Pearson correlation coefficient, which is calculated directly from the covariance matrix [179]. This provides a new, tighter upper bound for the BDPL that improves upon the specific Gaussian bound given in [37] and upon the general bound $m\varepsilon$ for any correlation model.

However, our bound applies only to a specific class of mechanisms: those that satisfy both DP and metric privacy under the ℓ_1 metric. We show in Section 6.3.2 that the clipped Laplace mechanism meets these criteria and develop a practical application in Section 6.5.1. To establish this result, we first connect metric privacy with an analogous form of BDP, termed Bayesian metric privacy, which we define below.

6.3.1. Relationship between Metric Privacy and Bayesian Metric Privacy

Unbounded continuous data domains, such as $\mathbb{R}^n$, usually produce challenges on DP application due to infinite sensitivities [180]. In the context of BDP, Yang et al. [37] defined a relaxation to work in those domains: If the data domain is equivalent to the real numbers (i.e., $\mathcal{X}^n = \mathbb{R}^n$), they defined a modified leakage, BDPL($\mathcal{M}; M$), where they only take into account the leakage between points with a distance smaller than $M \in \mathbb{R}$, i.e.,

$$\sup_{|x_i - x_i'| \leq M, \mathbf{x}_K, S} \ln \frac{\Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i]}{\Pr[Y \in S \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i']}.$$

Applying this BDP relaxation leaves indistinguishability between records at distances greater than M entirely uncontrolled. While this may increase applicability, it reduces privacy and limits insights into the impact of correlation.

However, metric privacy—or d-privacy when the metric is explicit—provides a solution to quantify privacy leakage as the distance $d(D, D')$ for each pair of databases D, D' when the maximum privacy leakage cannot be bounded [47].

To extend this advantage to correlated settings, we introduce the first metric version of BDP: Bayesian metric privacy. We define it following the same intuition as metric privacy; namely, the indistinguishability between two records x, x' depends on the distance $d(x, x')$ between them. Note that the change from databases to records is necessary because BDP does not apply to neighboring databases, but to target records, as we describe in Section 2.3.2.1. In this way, we can work with $\mathbb{R}^n$ as the data domain without losing information about the privacy leakage.

Definition 6.7 (Target Dependent Leakage)**.** Given a randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{D}(\Theta)$, $\mathbf{X}$ the input random vector following the distribution Π, the targeted record index $i \in [n]$, and the known record indices $K \subseteq [n] \backslash \{i\}$, the *adversary-specific target dependent* BDPL of $\mathcal{M}$ w.r.t. adversary (K, i) for any target values $x, x' \in \mathcal{X}$ is[2]

$$\text{BDPL}_{(K,i)}(x, x') = \sup_{\mathbf{x}_K, s} \ln \frac{p_Y(s \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x)}{p_Y(s \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x')}.$$

Given that we understand the leakage for each pair of data records we can simply define Bayesian metric privacy analogously to the original metric privacy notion:

Definition 6.8 (Bayesian Metric Privacy)**.** Let d be a (pseudo)metric on $\mathcal{X}^2$. A mechanism $\mathcal{M}$ is *Bayesian d-private* if for all $x, x' \in \mathcal{X}$,

$$\text{BDPL}(x, x') = \sup_{i,K} \text{BDPL}_{(K,i)}(x, x') \leq d(x, x'),$$

where the supremum is taken over all the possible sets of indices $i \in [n]$ and $K \subseteq [n] \setminus \{i\}$. BDPL$(x, x')$ is called *target dependent* BDPL.

[2] Analogously to BDPL Definition 2.26, if both the numerator and denominator are zero, we conventionally set $\text{BDPL}_{(K,i)}(x, x') = 0$.

The only difference between BDP and Bayesian metric privacy is that Bayesian metric privacy does not take the supremum over all pairs $x, x' \in \mathcal{X}$. Moreover, both notions are equivalent when we define $d(x, x') = \varepsilon$ for $x \neq x'$ and $d(x, x') = 0$ otherwise.

Now we can prove the relation between a d-private and a Bayesian d-private mechanism when the data distribution is a multivariate Gaussian. Particularly, we focus on the ℓ_1 distance due to its direct application to the Gaussian case. We formalize the conditions needed to obtain our bound:

Definition 6.9. For $\rho \in [0, 1]$ and $n \in \mathbb{N}$, we call the matrix $\Sigma_\rho \in \mathbb{R}^{n \times n}$ a *limited covariance matrix* if

- the matrix Σ_ρ is symmetric and positive definite,
- the diagonal of Σ_ρ is constant, i.e., there is a variance $\sigma^2 > 0$ so that $\Sigma_{\rho,ii} = \sigma^2$ for all $i \in [n]$ and,
- any pairwise correlation is limited by ρ. That is, for all $i \neq j$ we have $|\Sigma_{\rho,ij}| \leq \rho\sigma^2$.

The first condition is required to be a valid covariance matrix for a Gaussian distribution (see Definition 6.2). The second condition ensures that no records have a deviating variance, i.e., all records are drawn from the same one-dimensional distribution. The final condition imposes that the maximum Pearson correlation coefficient between any two random variables X_i and X_j is bounded by ρ. If we limit ρ to be small enough, we get a novel bound on the BDPL (see Theorem 6.11). However, before we can prove Theorem 6.11, we require the following lemma that establishes the maximum BDPL of a single adversary.

Lemma 6.10. *Let $\mathcal{M}$, with data domain $\mathbb{R}^n$, be an $(\varepsilon\ell_1)$-private mechanism with $\varepsilon > 0$, whose input data are drawn from a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. Here, $m \geq 2$ denotes the maximum number of correlated variables.*

Let $K = \{m-k, \ldots, m-1\}$ be the set of known indices for the adversary H correlated with the target X_m, with $k \leq m-2$, $T = K \cup \{m\}$ and U the set of unknown records correlated with X_m. If the principal submatrix Σ_T spanning $k+1$ rows and columns is invertible, then the adversary-specific target dependent BDPL *of $\mathcal{M}$ for any target values $x_m, x'_m \in \mathbb{R}$ is bounded by*

$$\mathrm{BDPL}_{(H,m)}(x_m, x'_m) \leq \varepsilon |x_m - x'_m| \left(\|\Sigma_{U;T} \Sigma_T^{-1} \mathbf{e}_{k+1}\|_1 + 1 \right)$$

where $\mathbf{e}_{k+1} \equiv (0, \ldots, 0, 1)^\top \in \mathbb{R}^{k+1}$ and the notation of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is reordered as

$$\mu = \begin{pmatrix} \mu_U \\ \mu_T \\ \mu_S \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_U & \Sigma_{U;T} & \mathbf{0} \\ \Sigma_{U;T}^\top & \Sigma_T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_S \end{pmatrix}. \tag{6.14}$$

Proof. The proof is derived from the multivariate Gaussian distribution properties. We introduce the following notation:

- From the set of unknown records V, U are correlated with X_m and W are independent.
- From the set of known records H, K are correlated with X_m and L are independent. $T = K \cup \{m\}$ and $R = H \cup \{m\}$.

Note that, by definition of the covariance matrix, $\Sigma_{ij} = 0$ for all pairs of independent variables $X_i \perp X_j$, which leads to the zero submatrices in Eq. 6.10.

First, we prove that for any $\mathbf{x}_V \in \mathbb{R}^v$ and $\mathbf{x}_T, \mathbf{x}'_T \in \mathbb{R}^{n-v}$, there exists a $\gamma \in \mathbb{R}^u$, such that

$$p_{\mathbf{X}_V}(\mathbf{x}_V \mid \mathbf{X}_T = \mathbf{x}_T) \equiv p_{\mathbf{X}_V}(\mathbf{x}_U, \mathbf{x}_W \mid \mathbf{X}_T = \mathbf{x}_T) = p_{\mathbf{X}_U}(\mathbf{x}_U + \gamma, \mathbf{x}_W \mid \mathbf{X}_T = \mathbf{x}'_T). \quad (6.15)$$

Then, the combination of this property with the $\varepsilon\ell_1$ condition gives the result.

We prove Equation (6.15) by using that the conditional distribution of a Gaussian distribution also follows the Gaussian distribution [181], i.e., $\mathbf{X}_U \mid \mathbf{X}_T = \mathbf{x}_T \sim \mathcal{N}(\hat{\mu}_U, \hat{\Sigma}_U)$ with

$$\hat{\mu}_U = \mu_U + \Sigma_{U;T}\Sigma_T^{-1}(\mathbf{x}_T - \mu_T), \quad \hat{\Sigma}_U = \Sigma_U - \Sigma_{U;T}\Sigma_T^{-1}\Sigma_{U;T}^\top.$$

While the conditional mean $\hat{\mu}_U$ depends on the specific value $\mathbf{x}_T$, the conditional covariance $\hat{\Sigma}_U$ remains fixed. Therefore, the two distributions (conditioned on $\mathbf{x}_T$ and $\mathbf{x}'_T \in \mathbb{R}^{k+1}$ respectively) only differ by a translation, i.e., for any $\mathbf{x}_U \in \mathbb{R}^u$ we have

$$\begin{aligned}
&p_{\mathbf{X}_U}(\mathbf{x}_U + \Sigma_{U;T}\Sigma_T^{-1}(\mathbf{x}_T - \mathbf{x}'_T) \mid \mathbf{X}_T = \mathbf{x}_T) \\
&\quad = \frac{\exp\Big(-\frac{1}{2}(\mathbf{x}_U + \Sigma_{U;T}\Sigma_T^{-1}(\mathbf{x}_T - \mathbf{x}'_T) - \hat{\mu}_U)^\top \hat{\Sigma}_U^{-1}(\mathbf{x}_U + \Sigma_{U;T}\Sigma_T^{-1}(\mathbf{x}_T - \mathbf{x}'_T) - \hat{\mu}_U)\Big)}{\sqrt{(2\Pi)^u |\hat{\Sigma}_U|}} \\
&\quad = \frac{\exp\Big(-\frac{1}{2}(\mathbf{x}_U - \hat{\mu}'_U)^\top \hat{\Sigma}_U'^{-1}(\mathbf{x}_U - \hat{\mu}'_U)\Big)}{\sqrt{(2\Pi)^u |\hat{\Sigma}'_U|}} \\
&\quad = p_{\mathbf{X}_U}(\mathbf{x}_U \mid \mathbf{X}_T = \mathbf{x}'_T).
\end{aligned}$$

Therefore, we have shown that the condition of the probability density $p_{\mathbf{X}_U}$ can simply be changed by additively shifting the input for the density, where the shift is $\gamma = \Sigma_{U;T}\Sigma_T^{-1}(\mathbf{x}'_T - \mu_T)$.

Second, we can use the fact that mechanism $\mathcal{M}$ is $(\varepsilon\ell_1)$-private, therefore, for all $\mathbf{x}_n, \mathbf{x}'_n \in \mathbb{R}^n$,

$$p_Y(s \mid \mathbf{X}_n = \mathbf{x}_n) \leq \exp(\varepsilon \|(\mathbf{x}_n, \mathbf{x}'_n)\|_1)\, p_Y(s \mid \mathbf{X}_n = \mathbf{x}'_n). \quad (6.16)$$

Now, applying Equation (6.15) and Equation (6.16) to the density function computation we can bound the adversary-specific target dependent BDPL.

Let Y be the random variable that represents the output of mechanism $\mathcal{M}$. Let random vector $\mathbf{X} = (X_1, \ldots, X_n)$ follow the multivariate Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$. To simplify notation, we combine the random vector $\mathbf{X}_H$ and random variable X_m into one target-inclusive vector $\mathbf{X}_R$ with $\mathbf{x}_R = (\mathbf{x}_H, x_m)^\top$ and $\mathbf{x}'_H = (\mathbf{x}_K, x'_m)^\top$. Considering the definition of the adversary-specific target dependent BDPL of (H, m) we have

$$\mathrm{BDPL}_{(H,m)}(x_m, x'_m) = \sup_{\mathbf{x}_H, s} \ln \frac{p_Y(s \mid \mathbf{X}_H = \mathbf{x}_H, X_m = x_m)}{p_Y(s \mid \mathbf{X}_H = \mathbf{x}_H, X_m = x'_m)} \equiv \sup_{\mathbf{x}_R, \mathbf{x}'_R, s} \ln \frac{p_Y(s \mid \mathbf{X}_R = \mathbf{x}_R)}{p_Y(s \mid \mathbf{X}_R = \mathbf{x}'_R)}$$

with the supremum taken over $s \in \mathbb{R}$ and known values $\mathbf{x}_H \in \mathbb{R}^h$.

We calculate the adversary-specific target dependent BDPL by rewriting the density $p_Y(s \mid \mathbf{x}_H, x_m)$ in terms of $p_Y(s \mid \mathbf{x}_H, x'_m)$:

$$\begin{aligned}
p_Y(s \mid \mathbf{X}_H = x_H, X_n = x_n) &\equiv p_Y(s \mid \mathbf{x}_R) \\
&= \int_{\mathbb{R}^v} p_Y(s \mid \mathbf{x}_V, \mathbf{x}_R)\, p_{\mathbf{X}_V}(\mathbf{x}_V \mid \mathbf{x}_R)\, \mathrm{d}\mathbf{x}_V \\
&= \int_{\mathbb{R}^v} p_Y(s \mid \mathbf{x}_U, \mathbf{x}_W, \mathbf{x}_R)\, p_{\mathbf{X}_W}(\mathbf{x}_W \mid \mathbf{x}_L)\, p_{\mathbf{X}_U}(\mathbf{x}_U \mid \mathbf{x}_T)\, \mathrm{d}\mathbf{x}_V && (6.17) \\
&= \int_{\mathbb{R}^v} p_Y(s \mid \tilde{\mathbf{x}}_U + \gamma, \mathbf{x}_W, \mathbf{x}_R)\, p_{\mathbf{X}_W}(\mathbf{x}_W \mid \mathbf{x}_L)\, p_{\mathbf{X}_U}(\tilde{\mathbf{x}_U} \mid \mathbf{x'}_T)\, \mathrm{d}\mathbf{x}_V && (6.18) \\
&\leq \mathrm{e}^{((\|\gamma\|_1 + |x_m - x'_m|)\varepsilon)} \int_{\mathbb{R}^v} p_Y(s \mid \tilde{\mathbf{x}}_U, \mathbf{x}_W \mathbf{x}'_R)\, p_{\mathbf{X}_W}(\mathbf{x}_W \mid \mathbf{x}_L)\, p_{\mathbf{X}_U}(\tilde{\mathbf{x}}_U \mid \mathbf{x'}_T)\, \mathrm{d}\tilde{\mathbf{x}}_V && (6.19) \\
&= \mathrm{e}^{((\|\gamma\|_1 + |x_m - x'_m|)\varepsilon)}\, p_Y(s \mid \mathbf{X}_H = x_H, X_m = x'_m)
\end{aligned}$$

Eq. 6.17 is obtained applying that $\mathbf{X_W} \perp \mathbf{X}_T$ and $\mathbf{X_U} \perp \mathbf{X}_L$. Then we substitute $\mathbf{x}_U$ for $\tilde{\mathbf{x}}_U + \gamma$, using the change of variable theorem for multiple integrals [182, p. 310]. The substitution is linear so the domain $\mathbb{R}^v$ over which we integrate does not change, and the determinant of the Jacobian of the substitution is simply 1. Hence, combining the change of variable with Eq. 6.15 we obtain Eq. 6.17. Finally, we use the inequality from Eq. 6.16 to derive Eq. 6.19, since $\|(\mathbf{x}_U, \mathbf{x}_T) - (\mathbf{x}_U + \gamma, \mathbf{x}'_T)\|_1 = \|\gamma\|_1 + |x_m - x'_m|$.

Now, we can formulate the upper bound of the adversary-specific target dependent BDPL for (H, m) for all $x_m, x'_m \in \mathbb{R}$:

$$\begin{aligned}
\mathrm{BDPL}_{(H,m)}(x_m, x'_m) &\leq (\|\gamma\|_1 + |x_m - x'_m|)\varepsilon \\
&= (\|\Sigma_{U;T}\Sigma_T^{-1}(\mathbf{x}_T - \mathbf{x}'_T)\|_1 + |x_m - x'_m|)\varepsilon \\
&= (\|\Sigma_{U;T}\Sigma_T^{-1}((\mathbf{x}_K, x_m)^\top - (\mathbf{x}_K, x'_m)^\top)\|_1 + |x_m - x'_m|)\varepsilon \\
&= (\|\Sigma_{U;T}\Sigma_T^{-1}(\mathbf{0}, x_m - x'_m)^\top\|_1 + |x_m - x'_m|)\varepsilon \\
&= (\|\Sigma_{U;T}\Sigma_T^{-1}\mathbf{e}_{k+1}\|_1 + 1)|x_m - x'_m|\varepsilon. \qquad \square
\end{aligned}$$

Applying previous lemma, when we limit ρ to be small enough—specifically $\rho(m-2) < 1$—we get the following bound:

Theorem 6.11. *Let $\mathcal{M}$, with data domain $\mathbb{R}^n$, be an $(\varepsilon\ell_1)$-private mechanism, where $\varepsilon > 0$, with input data drawn from a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma_\rho)$, with*

mean $\mu \in \mathbb{R}^n$ and limited covariance matrix $\Sigma_\rho \in \mathbb{R}^{n\times n}$ (Def. 6.9). Given a limited number of correlated records, $m \leq n$, such that $\rho(m-2) < 1$ is the maximum correlation coefficient. Then, for any $x, x' \in \mathbb{R}$ we have

$$\mathrm{BDPL}(x, x') \leq \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right)|x' - x|.$$

Proof. To prove an upper bound for the target dependent BDPL, we must bound the adversary-specific target dependent BDPL of every possible adversary (H, i) with $i \in [n]$ and $H \subseteq [n] \setminus \{i\}$. The remaining indices besides H and i make up the unknown indices $V = [n] \setminus \{H, i\}$. We differentiate between two cases.

Case 1: There are no unknown indices correlated with the target $i \in [n]$, i.e., we have $U = \varnothing \subseteq V$. Therefore, we can calculate the target dependent BDPL of this adversary by using the fact that $\mathcal{M}$ is $(\varepsilon\ell_1)$-private. We following the same notation as in previous lemma: Set of unknown records V, U are correlated with X_i and W are independent. Set of known records H, K are correlated with X_i and L are independent, so in particular $i \notin L$, $T = K \cup \{i\}$ and $R = H \cup \{i\}$. Hence, we obtain:

$$\begin{aligned}
p_Y(s \mid \mathbf{X}_H = x_H, X_i = x_i) &\equiv p_Y(s \mid \mathbf{x}_R) \\
&= \int_{\mathbb{R}^v} p_Y(s \mid \mathbf{x}_V, \mathbf{x}_R)\, p_{\mathbf{X}_V}(\mathbf{x}_V \mid \mathbf{x}_R)\, \mathrm{d}\mathbf{x}_V \\
&= \int_{\mathbb{R}^v} p_Y(s \mid \mathbf{x}_V, \mathbf{x}_R)\, p_{\mathbf{X}_V}(\mathbf{x}_V \mid \mathbf{x}_L)\, \mathrm{d}\mathbf{x}_V \\
&\leq \int_{\mathbb{R}^u} \mathrm{e}^{\varepsilon|x_i - x_i'|} p_Y(s \mid \mathbf{x}_V, \mathbf{x'}_R)\, p_{\mathbf{X}_V}(\mathbf{x}_V \mid \mathbf{x}_L)\, \mathrm{d}\mathbf{x}_V \\
&= \mathrm{e}^{\varepsilon|x_i - x_i'|} p_Y(s \mid \mathbf{X}_H = x_H, X_i = x_i').
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathrm{BDPL}_{(H,i)}(x_i, x_i') &= \sup_{\mathbf{x}_H, s} \ln \frac{p_Y(s \mid \mathbf{X}_H = \mathbf{x}_H, X_i = x_i)}{p_Y(s \mid \mathbf{X}_H = \mathbf{x}_H, X_i = x_i')} \\
&= \ln e^{\varepsilon|x_i - x_i'|} = \varepsilon|x_i - x_i'| \\
&\leq \left(\frac{m^2}{4(\frac{1}{\rho} - m + 3)} + 1\right)\varepsilon|x_i - x_i'|.
\end{aligned}$$

Case 2: There is at least one unknown record correlated with the target, i.e., $U \neq \varnothing$.

Let $k = |K|$ be the number of known records correlated with the target and $u = |U|$ the number of unknown ones. Without loss of generality we have $K = \{m-k, \ldots, m-1\}$ and $i = m$. Otherwise, we simply reorder the components of the random vector $\mathbf{X}$ so that the statements about K and i apply.

Choose $\Sigma_U \in \mathbb{R}^{u\times u}$, $\Sigma_{U;T} \in \mathbb{R}^{u\times(k+1)}$ and $\Sigma_T \in \mathbb{R}^{(k+1)\times(k+1)}$ so that the following holds:

$$\Sigma_\rho = \begin{pmatrix} \Sigma_U & \Sigma_{U;T} & \mathbf{0} \\ \Sigma_{U;T}^\top & \Sigma_T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_S \end{pmatrix}$$

In order to use Lemma 6.10, we require the principal submatrix Σ_T spanning the last $k+1$ rows and columns to be invertible. We separate in two subcases:

Case 2.1 ($m = 2$). In such case, given than $U \neq \varnothing$ we have that $k = 0$ and

$$\Sigma_\rho = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \mathbf{0} \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_S \end{pmatrix}$$

Therefore, for all $\sigma_2 \neq 0$ we have that $\gamma = \frac{\rho\sigma_1\sigma_2}{\sigma_2^2} = \frac{\rho\sigma_1}{\sigma_2} \leq \rho$. Applying Lemma 6.10 we obtain

$$\mathrm{BDPL}_{(H,i)}(x_i, x_i') \leq (\|\gamma\|_1 + 1)\varepsilon|x_i' - x_i| = (\rho + 1)\varepsilon|x_i' - x_i|. \tag{6.20}$$

Case 2.2 ($m > 2$). We proceed finding a strictly positive lower bound for the eigenvalues of Σ_T. Conveniently, we can later use this fact to bound the entries of Σ_T^{-1}. We denote the individual cells of Σ_T as a_{jl} for all $j, l \in [k+1]$.

According to the Gershgorin circle theorem [183], every eigenvalue of a real matrix such as Σ_T is contained in a closed disc on the complex number plane with center a_{jj} and radius $\sum_{l\neq j} |a_{jl}|$ for $j \in [k+1]$. Since the eigenvalues of a symmetric matrix must be real [184, p. 1], we are only concerned with the real part of this disc, i.e., the interval $[a_{jj} - \sum_{l\neq j} |a_{jl}|, a_{jj} + \sum_{l\neq j} |a_{jl}|]$. We can construct a lower bound of the smallest eigenvalue λ_- of Σ_T by finding the lowest border of these intervals.

$$\begin{aligned}
\lambda_- &\geq \min_j a_{jj} - \sum_{l\neq j} |a_{jl}| \\
&\geq \min_j \sigma^2 - \sum_{l\neq j} |\rho\sigma^2| && (6.21) \\
&= \sigma^2 - k\rho\sigma^2 \\
&= (1 - k\rho)\sigma^2 && (6.22) \\
&> (1 - (m-2)\,\frac{1}{m-2})\sigma^2 = 0 && (6.23)
\end{aligned}$$

In Eq. 6.21, we use the fact that every random variable has the same variance σ^2 and the correlation between any two random variables in $\mathbf{X}$ is in the interval $[-\rho, \rho]$. Then we show in Eq. 6.23 that the bound is positive since k must be $m-2$ or smaller because there is one targeted record and $U \neq \varnothing$ and the maximum correlation ρ is bounded with $\rho < \frac{1}{m-2}$. Thus, we have shown that each eigenvalue of Σ_T is strictly positive and the matrix is therefore invertible.

Now, by direct application of Lemma 6.10 we have an upper bound of the adversary-specific target dependent BDPL for any $x_i, x_i' \in \mathbb{R}$ with

$$\mathrm{BDPL}_{(H,i)}(x_i, x_i') \leq (\|\Sigma_{U;T}\Sigma_T^{-1}\mathbf{e}_{k+1}\|_1 + 1)\varepsilon|x_i' - x_i|.$$

This bound depends on the adversary-specific matrices $\Sigma_{U;T}$ and Σ_T. Our goal is to find a bound for the total target dependent BDPL, irrespective of the specific adversary.

We denote the individual cells of $\Sigma_{U;T}$ as b_{jl} for all $j \in [u]$, $l \in [k+1]$ and the cells of Σ_T^{-1} as α_{jl} for all $j, l \in [k+1]$.

Since $\Sigma_{U;T}$ only contains covariances between random variables from U and K or n and the covariance is bounded by $\rho\sigma^2$, for all indices $j, l \in [u]$ we obtain that

$$|b_{jl}| \leq \rho\sigma^2.$$

Now, we bound the entries of the inverse matrix Σ_T^{-1}. The 2-norm of any symmetric matrix $A \in \mathbb{R}^{m \times m}$ is defined as

$$\|A\|_2 := \sup_{\mathbf{x} \in \mathbb{R}^m} \{\|A\mathbf{x}\|_2 \colon \|\mathbf{x}\|_2 = 1\}.$$

The 2-norm of A is equal to its maximum singular value (i.e., the largest absolute eigenvalue for a symmetric matrix) [185, p. 47]. Thus, we can use the 2-norm to bound the entries of a symmetric matrix by its largest absolute eigenvalue $|\lambda_+|$. Let $\mathbf{e}_j \in \mathbb{R}^m$ be the vector with 1 at position j and 0 elsewhere. For every entry a_{ij} in A, we have

$$|a_{ij}| = \sqrt{a_{ij}^2} \leq \sqrt{\sum_{k \in [m]} a_{kj}^2} = \|A\mathbf{e}_j\|_2 \leq \|A\|_2 = |\lambda_+|.$$

Additionally, the eigenvalues of an inverse A^{-1} can be determined knowing the eigenvalues of A. Let $\lambda \in \mathbb{R}$ be any eigenvalue of A; λ cannot be zero because A is invertible, consequently,

$$\begin{aligned} & Ax = \lambda x \\ \Leftrightarrow \quad & A^{-1}Ax = \lambda A^{-1}x \\ \Leftrightarrow \quad & x = \lambda A^{-1}x \\ \Leftrightarrow \quad & \frac{1}{\lambda}x = A^{-1}x \end{aligned}$$

Thus, the eigenvalues of A^{-1} are the inverses of the eigenvalues of A. Putting it all together, the entries of Σ_T^{-1} are smaller or equal to the inverse of the smallest absolute eigenvalue of Σ_T. In Eq. 6.22, we have shown that all eigenvalues of Σ_T are positive and larger than $(1 - k\rho)\sigma^2$. Therefore, the smallest *absolute* eigenvalue of Σ_T, denoted λ_-, is also larger than this bound. Now, these two facts (the entries of a matrix can be bounded by the largest absolute eigenvalue, and the eigenvalues of A^{-1} are the inverses of the eigenvalues of A) are brought together to bound the entries of Σ_T^{-1}:

$$\alpha_{jl} \leq \frac{1}{\lambda_-} \leq \frac{1}{(1 - k\rho)\sigma^2}.$$

With the bounds for the entries of $\Sigma_{U;T}$ and Σ_T^{-1} in hand, we can find a general upper bound for the adversary-specific target dependent BDPL for any $x_i, x_i' \in \mathbb{R}$.

$$\mathrm{BDPL}_{(H,i)}(x_i, x_i') \leq (\|\Sigma_{U;T}\Sigma_T^{-1}\mathbf{e}_{k+1}\|_1 + 1)|x_i' - x_i|\varepsilon \tag{6.24}$$

$$= (\|\Sigma_{U;T}(\alpha_{1,k+1}, \ldots, \alpha_{k+1,k+1})^\top\|_1 + 1)|x_i' - x_i|\varepsilon \tag{6.25}$$

$$= \left(\|(\sum_{l=1}^{k+1} \alpha_{l,k+1}\, b_{1,l}, \ldots, \sum_{l=1}^{k+1} \alpha_{l,k+1}\, b_{u,l})^\top\|_1 + 1\right)|x_i' - x_i|\varepsilon \tag{6.26}$$

$$= \left(\sum_{j=1}^{u} |\sum_{l=1}^{k+1} \alpha_{l,k+1}\, b_{j,l}| + 1\right)|x_i' - x_i|\varepsilon \tag{6.27}$$

$$\leq \left(\sum_{j=1}^{u} |\sum_{l=1}^{k+1} \frac{\rho\sigma^2}{(1-k\rho)\sigma^2}| + 1\right)|x_i' - x_i|\varepsilon \tag{6.28}$$

$$= \left(\frac{u(k+1)\rho}{1-k\rho} + 1\right)|x_i' - x_i|\varepsilon \tag{6.29}$$

$$= \left(\frac{u(k+1)}{\frac{1}{\rho} - k} + 1\right)|x_i' - x_i|\varepsilon$$

$$\leq \left(\frac{(\frac{m}{2})^2}{\frac{1}{\rho} - m + 2} + 1\right)|x_i' - x_i|\varepsilon \tag{6.30}$$

$$= \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right)|x_i' - x_i|.\varepsilon$$

We first use the inequality from Lemma 6.10 in Eq. 6.24. Then, we multiply Σ_T^{-1} and $\mathbf{e}_{k+1}$; only the last column of Σ_T^{-1} remains in Eq. 6.25. The result is multiplied with $\Sigma_{U;T}$ in Eq. 6.26. Afterwards, in Eq. 6.27 we apply the ℓ_1-distance to the remaining vector. The bounds for the entries α and b are used in Eq. 6.28. Keep in mind that the denominator is always positive because $k \leq m-2$ and $\rho < \frac{1}{m-2}$. Then the two sums can be simplified by multiplying by their cardinality in Eq. 6.29 since the entries of the sum no longer depend on j or l. Finally, we bound the numerator and denominator in Eq. 6.30: The numerator $u(k+1)$ is smaller or equal to $(m/2)^2$ because u and $k+1$ are positive and together form $m = u+k+1$. Thus, their product is maximal if they meet at the exact midpoint to m. As mentioned previously, the denominator is always positive. It therefore becomes minimal (and the entire expression maximal) if k becomes maximal. This is the case for $k = m-2$.

In both cases we were able to bound adversary-specific target dependent BDPL as required. Hence, the proof is complete. □

Theorem 6.11 provides a concrete formula for the increase in privacy leakage due to linear correlations of a multivariate Gaussian model relative to independent data. Higher Pearson coefficients lead to greater leakage. Additionally, we can extend this result to derive a relation between DP and BDP.

6.3.2. Relationship between DP and BDP

Observe that any d-private mechanism is an ε-DP mechanism with $\varepsilon = \sup_{D \sim D'} d(D, D')$. Similarly, any Bayesian d-private mechanism is an ε-BDP mechanism considering $\varepsilon =$

$\sup_{x,x'} d(x, x')$. By leveraging these relationships between privacy notions we can establish a connection between DP and BDP. However, since this supremum may be unbounded, it can lead to undesirable privacy guarantees. To manage this relationship effectively we apply clipping techniques, resulting in Theorem 6.14, which enables the construction of BDP mechanisms from DP mechanisms. Formally, clipping is defined as:

Definition 6.12. For any interval $I = [a, b] \subset \mathbb{R}$, we define the *clipping function* $c_I : \mathbb{R}^n \to \mathbb{R}^n$, which, for all $D \in \mathbb{R}^n$ and all $i \in [n]$, outputs

$$c_I(D)_i = \max(a, \min(b, D_i)).$$

Let $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}$ be a mechanism. We define its *clipped version* $\mathcal{M}_I : \mathbb{R}^n \to \mathbb{R}$ as $\mathcal{M}_I = \mathcal{M} \circ c_I$.

Due to the data domain reduction derived from the clipping pre-processing, we can bound the DP leakage of $(\varepsilon\ell_1)$-private mechanisms.

Lemma 6.13. *If $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}$ is $(\varepsilon\ell_1)$-private, then its clipped version $\mathcal{M}_I$ is $(\varepsilon\ell_1)$-private and $(M\varepsilon)$-DP with $M = |b - a|$.*

Proof. We begin by showing that $\mathcal{M}_I$ is $(\varepsilon\ell_1)$-private. Let $D_1, D_2 \in \mathbb{R}^n$ be arbitrary and $S \subseteq \mathbb{R}$ be any measurable set. We have

$$\begin{aligned}
\Pr[\mathcal{M}_I(D_1) \in S] &= \Pr[\mathcal{M}(c_I(D_1)) \in S] \\
&\leq e^{\varepsilon \|c_I(D_1) - c_I(D_2)\|_1} \Pr[\mathcal{M}(c_I(D_2)) \in S] && (6.31)\\
&\leq e^{\varepsilon \|D_1 - D_2\|_1} \Pr[\mathcal{M}(c_I(D_2)) \in S] && (6.32)\\
&= e^{\varepsilon \|D_1 - D_2\|_1} \Pr[\mathcal{M}_I(D_2) \in S]. && (6.33)
\end{aligned}$$

In Eq. 6.31 we use that $\mathcal{M}$ is $(\varepsilon\ell_1)$-private. Then, we apply the fact that the ℓ_1-distance of two clipped data sets is smaller or equal to the ℓ_1-distance of the original data sets in Eq. 6.32. Finally, Eq. 6.33 shows that $\mathcal{M}_I$ is $(\varepsilon\ell_1)$-private.

Now, we will show that $\mathcal{M}_I$ is also DP. Let $D, D' \in \mathbb{R}^n$ be arbitrary neighboring data sets, i.e., there only exists a single index $i \in [n]$ with $D_i \neq D'_i$. For any measurable set $S \subseteq \Theta$ we have

$$\begin{aligned}
\Pr[\mathcal{M}_I(D) \in S] &= \Pr[\mathcal{M}(c_I(D)) \in S] \\
&\leq e^{\varepsilon \|c_I(D) - c_I(D')\|_1} \Pr[\mathcal{M}(c_I(D')) \in S] && (6.34)\\
&= e^{\varepsilon \sum_{j \in [n]} |\max(a, \min(b, D_j)) - \max(a, \min(b, D'_j))|} \Pr[\mathcal{M}(c_I(D')) \in S] && (6.35)\\
&= e^{\varepsilon |\max(a, \min(b, D_i)) - \max(a, \min(b, D'_i))|} \Pr[\mathcal{M}(c_I(D')) \in S] && (6.36)\\
&\leq e^{\varepsilon(b-a)} \Pr[\mathcal{M}(c_I(D')) \in S] = e^{\varepsilon(b-a)} \Pr[\mathcal{M}_I(D') \in S].
\end{aligned}$$

Once again, we use that $\mathcal{M}$ is $(\varepsilon\ell_1)$-private in Eq. 6.34. We expand the definition of the ℓ_1-distance and of the clipping function c_I in Eq. 6.35. Then, we use for Eq. 6.36 that D and D' only differ for index i. Finally, we can bound the difference between the two entries because they are clipped to the interval $[a, b]$. We have shown that $\mathcal{M}_I$ is $(b - a)\varepsilon$-DP. □

With Lemma 6.13 and Theorem 6.11, we can directly show that this class of DP-mechanisms has a limited BDPL.

Theorem 6.14 (The Gaussian Bound)**.** *Let $\mathcal{M}_I$, with data domain $\mathbb{R}^n$, be the clipped version of an $(\varepsilon\ell_1)$-private mechanism $\mathcal{M}$, where $\varepsilon > 0$, with input data drawn from a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma_\rho)$, with mean $\mu \in \mathbb{R}^n$. Given a maximum of $m \leq n$ correlated variables, and a limited covariance matrix $\Sigma_\rho \in \mathbb{R}^{n\times n}$ (Def. 6.9), such that $\rho(m-2) < 1$ is the maximum correlation coefficient. Then, the clipped mechanism $\mathcal{M}_I$ is*

$$\left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right) M\varepsilon\text{-BDP.}$$

where M is the diameter of the interval I.

Proof. We know that $\mathcal{M}_I$ is a $(\varepsilon\ell_1)$-private query because of Lemma 6.13. Thus, we can apply Theorem 6.11 to find a universal bound of the target dependent BDPL in this situation. Therefore, the idea is to show how the BDPL is bounded by the target dependent BDPL, and to then apply Theorem 6.11.

$$\begin{aligned}
\mathrm{BDPL} &\coloneqq \sup_{K,i} \mathrm{BDPL}_{(K,i)} \\
&= \sup_{K,i} \left(\sup_{\mathbf{x}_K, s, |x_i - x_i'| \leq M} \ln \frac{p_Y(s \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i)}{p_Y(s \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i')} \right) && (6.37) \\
&= \sup_{K,i} \left(\sup_{|x_i - x_i'| \leq M} \mathrm{BDPL}_{(K,i)}(x_i, x_i') \right) && (6.38) \\
&= \sup_{|x_i - x_i'| \leq M} \left(\sup_{K,i} \mathrm{BDPL}_{(K,i)}(x_i, x_i') \right) && (6.39) \\
&= \sup_{|x_i - x_i'| \leq M} \mathrm{BDPL}(x_i, x_i') && (6.40) \\
&\leq \sup_{|x_i - x_i'| \leq M} \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \right) |x_i' - x_i| \varepsilon && (6.41) \\
&= \sup_{|x_i - x_i'| \leq M} \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \right) |x_i - x_i'| \varepsilon \\
&= \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \right) M\varepsilon.
\end{aligned}$$

In Eq. 6.37 we expand the definition of BDPL. Then, in Eq. 6.38 we use that the adversary-specific *target dependent* BDPL is defined equivalently, except it does not take the supremum over x_i, x_i'. We switch the order of the suprema in Eq. 6.39 to subsequently plug in the definition of the general target dependent BDPL. Then, we use Theorem 6.11 to derive Eq. 6.41. Finally, the last supremum can be resolved by writing out $d(x_i, x_i')$ and bounding it with $M\varepsilon$. It follows that clipped mechanism $\mathcal{M}$ is BDP since the BDPL is limited. □

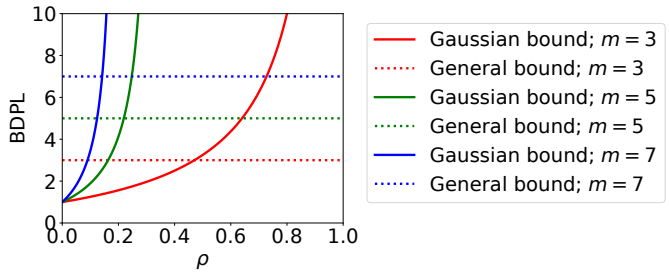


Figure 6.1.: Gaussian-specific bound compared to the general bound, which coincides with the state-of-the-art bound for Gaussian models [37].

Theorem 6.14 allows us to systematically build a BDP mechanism by recalibrating the noise of a DP mechanism when $\rho(m-2) < 1$. For instance, given the clipped Laplace mechanism $\mathcal{M}_I$ that adds noise to a data point $x \in \mathbb{R}$ following $\text{Lap}(\frac{M}{\tau})$, where

$$\tau = \varepsilon \frac{4(\frac{1}{\rho} - m + 2)}{m^2 + 4(\frac{1}{\rho} - m + 2)} \tag{6.42}$$

we obtain an ε-BDP mechanism. Moreover,

$$\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1 \leq m \text{ if and only if } \rho \leq \frac{m-1}{\frac{5}{4}m^2 - 3m + 2}. \tag{6.43}$$

Hence, the Gaussian bound improves on the general bound if ρ is on the order of $\rho \approx \frac{1}{m}$ (see Figure 6.1). The higher the number of correlated records m, the better the relative improvement of the Gaussian specific bound compared to the general bound for small correlation. Importantly, Yang et al. [37] establish a bound for Gaussian Markov random fields. They establish that a clipped $\mathcal{M}_{\varepsilon,f}$ satisfies $(nM\varepsilon)$-BDP, which coincides with the general bound when all records are correlated. Theorem 6.14 applies to this particular case since a Gaussian Markov random field is an example of Gaussian Multivariate distribution. Moreover, our bound improves over theirs in the same cases it improves over the general bound.

6.3.3. Accuracy

When the Pearson correlation is bounded as specified in Equation (6.43), we showed that a larger ε' than $\frac{\varepsilon}{m}$ is sufficient to guarantee ε-BDP via an ε'-DP mechanism. Since a larger privacy budget generally correlates with improved utility, we can therefore anticipate enhanced utility results. In particular, we express the accuracy improvement of the Laplace mechanism when it is calibrated to protect data drawn from a multivariate Gaussian distribution. As a consequence of our Theorem 6.14 and Proposition 2.10 from [16] we obtain the following result:

Corollary 6.15. *Let $\mathcal{M}_{\varepsilon,f_I}$ be the clipped Laplace ε-DP mechanism that approximates the query f_I as in Definition 6.12 with input data drawn from a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma_\rho)$ with mean $\mu \in \mathbb{R}^n$ and limited covariance $\Sigma_\rho \in \mathbb{R}^{n\times n}$ with a maximum number of correlated variables $m \leq n$ such that $\rho(m-2) < 1$. Then, if the Laplace mechanism $\mathcal{M}_{\varepsilon,f_I}$ is (α,β)-accurate w.r.t. f_I, there exists an ε-BDP mechanism $\widetilde{\mathcal{M}}$ that is $(h\alpha,\beta)$-accurate w.r.t. f_I with*

$$h = \frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1.$$

Proof. The idea of this proof is to construct mechanism $\widetilde{\mathcal{M}}$ with the Laplace mechanism as well, but to choose a carefully selected privacy leakage $\varepsilon' < \varepsilon$ so that mechanism $\widetilde{\mathcal{M}}$ is (1) ε-BDP and (2) $(h\alpha,\beta)$-accurate.

First, we determine the accuracy of mechanism $\mathcal{M}_{\varepsilon,f_I}$. According to Proposition 2.10, the (α,β)-accuracy of the Laplace mechanism for a given probability $\beta \in (0,1]$ and privacy parameter ε is

$$\alpha = \ln\left(\frac{1}{\beta}\right)\frac{\Delta f_I}{\varepsilon}.$$

So this is the (α,β)-accuracy of $\mathcal{M}_{\varepsilon,f_I}$. We have to show that there exists an ε-BDP mechanism $\widetilde{\mathcal{M}}$ which is $(h\alpha,\beta)$-accurate. We choose $\widetilde{\mathcal{M}}$ as the Laplace mechanism applied to f_I with an adjusted privacy parameter $\varepsilon' > 0$. Thus, $\widetilde{\mathcal{M}}$ will be ε'-DP. Observe that $\widetilde{\mathcal{M}}$ is d-private with $d(D,D') = \frac{\varepsilon'}{M}\|D - D'\|_1$. Therefore, we can use Theorem 6.14 to show that $\widetilde{\mathcal{M}}$ is BDP. We must choose ε' in a way that ensures that the BDPL is limited to ε, so that we have ε-BDP. With Theorem 6.14, $\widetilde{\mathcal{M}}$ is

$$\left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right)\varepsilon'\text{-BDP}.$$

Therefore, to achieve ε-BDP, we must have

$$\left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right)\varepsilon' = \varepsilon \quad \Leftrightarrow \quad \varepsilon' = \varepsilon\left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right)^{-1}.$$

Now, we can calculate the accuracy of $\widetilde{\mathcal{M}}$ because it also uses the Laplace mechanism. Then, we find an upper bound for this accuracy. Mechanism $\widetilde{\mathcal{M}}$ is (α',β)-accurate, with

$$\begin{aligned}\alpha' &= \ln(\frac{1}{\beta})\frac{\Delta f_I}{\varepsilon'} = \ln(\frac{1}{\beta})\frac{\Delta f_I}{\varepsilon}(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1)\\ &= \alpha\left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right)\\ &= \alpha h\end{aligned}$$

Finally, we find that $\widetilde{\mathcal{M}}$ is $(h\alpha,\beta)$-accurate. □

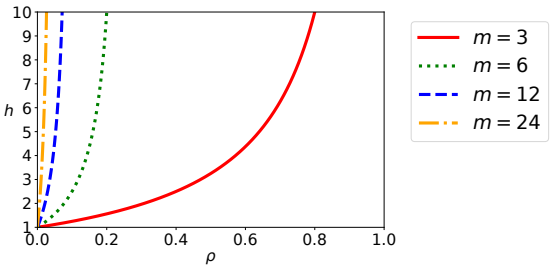


Figure 6.2.: Relative accuracy of an ε-BDP mechanism to an ε-DP mechanism for a Multivariate Gaussian distribution.

The statement of Corollary 6.15 is visualized in Figure 6.2. This figure shows that in order to provide similar utility to DP, ρ must be small. The larger the number of correlated records m, the smaller ρ has to be to provide similar utility. The results in this section enable the protection of weakly correlated data drawn from a multivariate Gaussian distribution. Furthermore, a comparison of the accuracy achieved by our method versus the state-of-the-art bound from [37] and the general BDP bound is presented in Figure 6.6, demonstrating a consistent improvement enabled by our approach.

6.4. Markov Chain Correlation Model

In streaming processes or time series data, states at successive time steps are often correlated, meaning that the state at a given time step depends on the state at the previous one. For example, a user's location at time step t is correlated with their previous location at $t-1$ (see Section 3.1). This dependency pattern is commonly modeled using Markov chains [186]. Consequently, in this section we investigate the impact of correlations following a Markov model on the privacy leakage and utility of BDP mechanisms.

Particularly, we prove Theorem 6.20, a new bound on the BDPL of any ε-DP mechanism when data is correlated corresponding to a Markov chain. Additionally, we use our results to elaborate on the utility gain compared to protecting against arbitrary correlation.

For the remainder of this work, we adopt the definition of a Markov chain from [186], which specifically refers to finite, time-homogeneous Markov chains, i.e., those with finite state spaces and time-invariant transition probabilities. Formally,

Definition 6.16 (Markov Chain [186])**.** Let $\mathcal{S}$ be a finite set of possible states of size $s \in \mathbb{N}$ and let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector. We say $\mathbf{X}$ is a *Markov chain* with transition matrix $P \in \mathbb{R}^{s \times s}$ and initial distribution $w \in \mathbb{R}^s$ if all of the following holds.

1. For all states $x, y \in \mathcal{S}$ and all indices $i \in [n-1]$ we have $\Pr[X_{i+1} = x | X_i = y] = P_{y,x}$.
2. For all states $x \in \mathcal{S}$ we have $\Pr[X_1 = x] = w_x$.

3. The Markov property: For all indices $i \in [n-1]$ and for all states $x_1, \ldots, x_i, x_{i+1} \in \mathcal{S}$ we have

$$\begin{aligned}&\Pr[X_{i+1} = x_{i+1} \mid X_1 = x_1, \ldots, X_i = x_i] \\ &\quad = \Pr[X_{i+1} = x_{i+1} \mid X_i = x_i].\end{aligned}$$

Note that the Markov property from Definition 6.16 holds not only when the full history is known, but also when only the partial history is known as we show in the following remark.

Remark 6.17. Given an index $i \in [n-1]$, and a set $A \subseteq \{1, \ldots, i-1\}$ containing only indices smaller than i. Then, for any states $x, y \in \mathcal{S}$ and any state tuple $\mathbf{x}_A \in \mathcal{S}^{|A|}$,

$$\begin{aligned}\Pr[x_{i+1} \mid x_i, \mathbf{x}_A] &= \sum_{\mathbf{x}_B \in \mathcal{S}^b} \Pr[x_{i+1} | x_i, \mathbf{x}_A, \mathbf{x}_B] \Pr[\mathbf{x}_B | x_i, \mathbf{x}_A] \\ &\overset{(*)}{=} \sum_{\mathbf{x}_B \in \mathcal{S}^b} \Pr[x_{i+1} \mid x_i] \Pr[\mathbf{x}_B \mid x_i, \mathbf{x}_A] \\ &= \Pr[x_{i+1} \mid x_i] \sum_{\mathbf{x}_B \in \mathcal{S}^b} \Pr[\mathbf{x}_B \mid x_i, \mathbf{x}_A] = \Pr[x_{i+1} \mid x_i],\end{aligned}$$

where $B = [n] \setminus (A \cup \{i\})$ is the set of remaining indices, and $b = |B|$. We use the law of total probability to introduce all remaining indices in B. Then, in $(*)$, we apply the Markov property. Finally, $\Pr[x_{i+1} \mid x_i]$ can be factored out of the sum because it no longer depends on $\mathbf{x}_B$. The remaining sum adds up to 1, so we are left with only the condition of the direct predecessor.

6.4.1. Relationship between DP and BDP

In this section, we show that it is possible to obtain a bound on the BDPL of any DP mechanism based on the maximum ratio between the largest and smallest transition probabilities in the Markov chain. The intuition is that if all transition probabilities are similar, changing the random variable X_i from state x_i to state x_i' will have minimal impact on the subsequent time steps of the Markov chain. However, if the transition probabilities differ significantly, this change could have a large effect over many time steps. To prove our main result Theorem 6.20 we need the auxiliary Lemmas 6.18 and 6.19.

Lemma 6.18 (Generalized Markov Property)**.** *Given* $\mathbf{X}$ *a Markov chain, for all sets of indices* $A \in [n] \setminus \{i\}$*:*

$$\Pr[x_i \mid \mathbf{x}_A] = \Pr[x_i \mid x_\ell, x_r]$$

where ℓ, r *are the nearest indices to* i *in* A *both left and right., i.e.,* $\ell, r \in A$ *with* $\ell < i$ *and* $i < r$ *so that for all indices* $j \in A$ *we have* $j < \ell$ *or* $r < j$*. Notably, if all indices in* A *are smaller than* i *we only need to consider* ℓ *and if all are above we only need to consider* r*.*

Proof. We derive this statement directly from probability rules and the Markov property. Let $i, j \in [n]$ be indices and $A' \subseteq [n] \setminus \{i, j\}$ be a set of indices so that there exists an index $\ell \in A'$, $|A'| = a$ that is "in between" i and j, i.e., we have $i > \ell > j$ or $i < \ell < j$. If for all states $x_i, x_j \in \mathcal{S}$ and for all state tuples $\mathbf{x}_{A'} \in \mathcal{S}^a$ we have

$$\Pr[x_i \mid \mathbf{x}_{A'}, x_j] = \Pr[x_i \mid \mathbf{x}_{A'}], \tag{6.44}$$

then it follows

$$\begin{aligned}\Pr[x_i \mid \mathbf{x}_A] &= \Pr[x_i \mid x_{i_1}, \dots x_\ell, x_r, \dots, x_{i_a}] \\ &= \Pr[x_i \mid x_{i_1}, \dots x_\ell, x_r] &(6.45)\\ &= \Pr[x_i \mid x_\ell, x_r]. &(6.46)\end{aligned}$$

Where Eq. 6.45 holds because for all $i_j > r$, there exists $r \in A'$ such that $i < r < i_j$; therefore we can apply Equation (6.44). Analogously Eq. 6.46 holds because for all $i_j < \ell$, there exists $\ell \in \{\ell, r\}$ such that $i_j < \ell < i$.

Consequently, for the rest of the proof we focus on proving Equation (6.44).

We proceed separately for the case in which the "irrelevant" index j is smaller or left, i.e., there exists $\ell \in A$ such that $j < \ell < i$ and for the case in which j is above or right, i.e., exist $r \in A$ such that $i < r < j$.

Case 1: First, we show that Equation (6.44) holds if $A_1 \subseteq \{1, \dots, i-1\}$ and $\ell \in A_1$ such that $j < \ell$ where $\ell \in A_1$ and $j < \ell < i$ so that no other index in A_1 lies between i and l. This means that the set of indices between ℓ and i—defined as $B = \{\ell+1, \dots, i-1\}$—is disjoint with A_1.

If B is empty, then Eq. 6.44 follows immediately from Remark 6.17 because index ℓ is the direct predecessor of index i, i.e., $\ell = i - 1$.

If $B \neq \varnothing$, using the law of total probability, we have

$$\begin{aligned}\Pr[x_i|\mathbf{x}_{A_1}, x_j] &= \sum_{\mathbf{x}_B \in \mathcal{S}^b} \Pr[x_i|\mathbf{x}_{A_1}, x_j, \mathbf{x}_B] \Pr[\mathbf{x}_B|\mathbf{x}_{A_1}, x_j] \\ &= \sum_{\mathbf{x}_B \in \mathcal{S}^b} \Pr[x_i|\mathbf{x}_{A_1}, \mathbf{x}_B] \Pr[\mathbf{x}_B|\mathbf{x}_{A_1}] = \Pr[x_i|\mathbf{x}_{A_1}], &(6.47)\end{aligned}$$

where Equation (6.47) follows by applying Remark 6.17, since

$$\begin{aligned}\Pr[\mathbf{x}_B|\mathbf{x}_{A_1}, x_j] &= \Pr[x_{\ell+1}, \dots, x_{i-1}|\mathbf{x}_{A_1}, x_j] \\ &= \Pr[x_{\ell+1}, \dots, x_{i-2}|\mathbf{x}_{A_1}, x_j] \Pr[x_{i-1}|\mathbf{x}_{A_1}, x_j, x_{\ell+1}, \dots, x_{i-2}] &(6.48)\\ &= \dots \\ &= \Pr[x_{\ell+1}|\mathbf{x}_{A_1}, x_j] \prod_{k=l+2}^{i-1} \Pr[x_k|\mathbf{x}_{A_1}, x_j, x_{\ell+1}, \dots, x_{k-1}] &(6.49)\\ &= \Pr[x_{\ell+1}|\mathbf{x}_{A_1}] \prod_{k=l+2}^{i-1} \Pr[x_k|\mathbf{x}_{A_1}, x_{\ell+1}, \dots, x_{k-1}] = \Pr[\mathbf{x}_B|\mathbf{x}_{A_1}] &(6.50)\end{aligned}$$

In Eq. 6.48, we rewrite the joint probability of $\mathbf{X}_B$ as the two parts X_{i-1} and $\mathbf{X}_{B\setminus\{i-1\}}$. This uses the fact that a joint probability can be rewritten using $\Pr[A \cap B \mid C] = \Pr[A \mid C]\ \Pr[B \mid C, A]$. Here, event A corresponds to conditions $X_{\ell+1} = x_{\ell+1}, \ldots, X_{i-2} = x_{i-2}$, event B to condition $X_{i-1} = x_{i-1}$ and event C to conditions $\mathbf{X}_{A_1} = \mathbf{x}_{A_1}, X_j = x_j$. This step is repeatedly used to fully split $\mathbf{X}_B$ into its components and derive Eq. 6.49. Then, we use the Remark 6.17 for Eq. 6.50: Random variable $X_{\ell+1}$ is conditionally independent of X_j given the direct predecessor X_ℓ with index $\ell \in A$. Similarly, random variable X_k is conditionally independent of X_j, given the direct predecessor X_{k-1}.

Note that, this result can be extended by induction to say that given any set of indices $C \subseteq [\ell+1, \ldots, n]$, if $A_1 \subseteq \{1, \ldots, i-1\}$, ℓ the biggest index A_1 and $j < \ell$, then

$$\Pr[x_c|\mathbf{x}_{A_1}, x_j] = \Pr[x_c|\mathbf{x}_{A_1}] \tag{6.51}$$

For $|C| = 1$, we have just proven Equation (6.51). Now we assume it true for all $|C| \leq n-1$ and we prove it for $|C| = n$:

$$\begin{aligned}
\Pr[x_c|\mathbf{x}_{A_1}, x_j] &= \Pr[x_{c_1}, \ldots, x_{c_n} \mid \mathbf{x}_{A_1}, x_j] \\
&= \Pr[x_{c_n}|\mathbf{x}_{A_1}, x_j, \mathbf{x}_{C\setminus\{c_n\}}] \Pr[\mathbf{x}_{C\setminus\{c_n\}} \mid \mathbf{x}_{A_1}, x_j] \\
&= \Pr[x_{c_n}|\mathbf{x}_{A_1}, \mathbf{x}_{C\setminus\{c_n\}}] \Pr[\mathbf{x}_{C\setminus\{c_n\}} \mid \mathbf{x}_{A_1}] \\
&= \Pr[x_c|\mathbf{x}_{A_1}]
\end{aligned}$$

where the last equality follows directly from the induction hypothesis since $|C \setminus \{c_n\}| = n-1$.

Now, we derive that Equation (6.44) also holds for an arbitrary set of indices A, not necessarily all smaller than i, i.e., $A \subseteq [n]\setminus\{i\}$. We can partition A into the indices before i and after i, i.e., $A = A_1 \cup A_2$ where $A_1 = \{i_j \in A \colon i_j < i\}$ and $A_2 = \{i_j \in A \colon i_j > i\}$. Then, we have

$$\begin{aligned}
\Pr[x_i|\mathbf{x}_A, x_j] &= \Pr[x_i|\mathbf{x}_{A_1}, \mathbf{x}_{A_2}, x_j] \coloneqq \frac{\Pr[x_i, \mathbf{x}_{A_1}, \mathbf{x}_{A_2}, x_j]}{\Pr[\mathbf{x}_{A_1}, \mathbf{x}_{A_2}, x_j]} \\
&= \frac{\Pr[x_i, \mathbf{x}_{A_1}, x_j] \Pr[\mathbf{x}_{A_2} \mid x_i, \mathbf{x}_{A_1}, x_j]}{\Pr[\mathbf{x}_{A_1}, x_j] \Pr[\mathbf{x}_{A_2} \mid \mathbf{x}_{A_1}, x_j]} \\
&= \Pr[x_i|\mathbf{x}_{A_1}, x_j] \frac{\Pr[\mathbf{x}_{A_2} \mid x_i, \mathbf{x}_{A_1}, x_j]}{\Pr[\mathbf{x}_{A_2} \mid \mathbf{x}_{A_1}, x_j]} \\
&= \Pr[x_i|\mathbf{x}_{A_1}] \frac{\Pr[\mathbf{x}_{A_2} \mid x_i, \mathbf{x}_{A_1}]}{\Pr[\mathbf{x}_{A_2} \mid \mathbf{x}_{A_1}]} \qquad (6.52) \\
&= \Pr[x_i|\mathbf{x}_{A_1}, \mathbf{x}_{A_2}] = \Pr[x_i|\mathbf{x}_A, x_j]
\end{aligned}$$

where Equation (6.52) follows from Equation (6.51).

Case 2: There exists an index $r \in A$ with $i < l < r$. Here, Equation (6.44) is obtained by applying Bayes' rule to reduce the problem to the already proven first case:

$$\Pr[x_i \mid \mathbf{x}_A, x_j] = \frac{\Pr[x_j \mid \mathbf{x}_A, x_i]\ \Pr[x_i \mid \mathbf{x}_A]}{\Pr[x_j \mid \mathbf{x}_A]} \tag{6.53}$$

$$= \frac{\Pr[x_j \mid \mathbf{x}_A] \Pr[x_i \mid \mathbf{x}_A]}{\Pr[x_j \mid \mathbf{x}_A]} \tag{6.54}$$

$$= \Pr[x_i \mid \mathbf{x}_A] \tag{6.55}$$

We use Bayes' theorem in Eq. 6.53. The first probability of the numerator is now in the situation of Case 1, since it is the probability of random variable X_j, conditioned on $\mathbf{X}_A$ and X_i with $i < r < j$ and $r \in A$. Equation (6.44) has already been proven for that case, so we apply it to derive Eq. 6.54. Finally, Eq. 6.55 follows directly by simplifying. □

Lemma 6.19. *Let random vector $\mathbf{X} = (X_1, \ldots, X_n)$ be a Markov chain with transition probabilities $P \in \mathbb{R}^{s \times s}$ and initial distribution $w \in \mathbb{R}^s$ with the following properties:*

(H1) Every cell of P is positive, i.e., for all $k, l \in \mathcal{S}$ we have $P_{k,l} > 0$.

(H2) Vector w is an eigenvector of P to the eigenvalue 1, i.e., $wP = w$.

Let $i \in [n]$ be the target index and let sets $U, K \subseteq [n] \backslash \{i\}$ be disjoint, with $[n] = U \cup K \cup \{i\}$ and at least one index in U. Then, for any unknown states $\mathbf{x}_U \in \mathcal{S}^u$, known states $\mathbf{x}_K \in \mathcal{S}^k$ and target states $x_i, x_i' \in \mathcal{S}$ we have

$$\frac{\Pr[\mathbf{X}_U = \mathbf{x}_U \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i]}{\Pr[\mathbf{X}_U = \mathbf{x}_U \mid \mathbf{X}_K = \mathbf{x}_K, X_i = x_i']} \leq \left(\frac{\max_{kl} P_{kl}}{\min_{kl} P_{kl}}\right)^4 \equiv \gamma^4.$$

Proof. First, combining Lemma 6.18 and Bayes' rule we obtain that, for all $A \in [n] \backslash \{i\}$:

$$\Pr[x_i \mid \mathbf{x}_A] = \frac{\Pr[x_r \mid x_i, x_\ell] \Pr[x_i \mid x_\ell]}{\Pr[x_r \mid x_\ell]} = \frac{\Pr[x_r \mid x_i] \Pr[x_i \mid x_\ell]}{\Pr[x_r \mid x_\ell]} \tag{6.56}$$

where 6.56 follows by application of Lemma 6.18 since i is closer to r than ℓ, i.e., $\ell < i < r$.

Second, we use (H1) and (H2) to prove that given $\mathbf{X}$ a Markov chain, for all indices i, j, not necessarily consecutive, and for all $x_i, x_i', y_j, y_j' \in \mathcal{S}$,

$$\frac{\Pr[X_i = x_i]}{\Pr[X_i = x_i']} \leq \gamma, \frac{\Pr[X_i = x_i | X_j = y_j]}{\Pr[X_i = x_i | X_j = y_j']} \leq \gamma \text{ and } \frac{\Pr[X_i = x_i | X_j = y_j]}{\Pr[X_i = x_i' | X_j = y_j]} \leq \gamma \tag{6.57}$$

We begin by proving $\frac{\Pr[x_i]}{\Pr[x_i']} \leq \gamma$. First, we show that w—as an eigenvector of P—not only contains the prior probabilities of X_1, but of any random variable X_i for $i \in [n]$. This means that the equality $\Pr[X_i = x_i] = w_{x_i}$ holds for any state $x_i \in \mathcal{S}$ because w is the equilibrium distribution. We proceed by induction, therefore we assume it true for $i - 1$ and prove it for i:

$$\Pr[X_i = x_i] = \sum_{y \in \mathcal{S}} \Pr[X_i = x_i \mid X_{i-1} = y] \Pr[X_{i-1} = y] \tag{6.58}$$

$$= \sum_{y \in \mathcal{S}} P_{y,x_i} w_y \tag{6.59}$$

$$= (wP)_{x_i} = w_{x_i} \tag{6.60}$$

We apply the law of total probability in Eq. 6.58. Then, we use the transition matrix P and the induction hypothesis ($\Pr[X_{i-1} = y] = w_y$) to replace the probabilities with entries of P and w in Eq. 6.59. Then, we rewrite the sum as a matrix-vector product in Eq. 6.60. Finally, we take advantage of the fact that w is an eigenvector of P to complete the result. With the basis $\Pr[X_1 = x] = w_x$ (which is the definition of w, see Definition 6.16) and this derivation, we prove by induction that the entries of w are equal to the prior probabilities of any random variable X_i.

Now we can bound the entries of w, thereby also bounding the probabilities $\Pr[X_i = x]$. We prove the upper bound $w_x \leq \max_{k,l \in \mathcal{S}} P_{k,l}$ by contradiction; the lower bound $w_x \geq \min_{k,l \in \mathcal{S}} P_{k,l}$ follows analogously. Assume that there exists a state $y \in \mathcal{S}$ so that its prior probability in w is greater than any transition probability, i.e., $w_y > \max_{k,l \in \mathcal{S}} P_{k,l}$. This leads to a contradiction as follows.

$$
\begin{aligned}
w_y &= (wP)_y && (6.61)\\
&= \sum_{k \in \mathcal{S}} P_{k,y}\, w_k && (6.62)\\
\Leftrightarrow (1 - P_{y,y}) w_y &= \sum_{k \neq y} P_{k,y}\, w_k \\
\Leftrightarrow 1 - P_{y,y} &= \sum_{k \neq y} \frac{P_{k,y}}{w_y}\, w_k \\
&< \sum_{k \neq y} w_k && (6.63)\\
&= 1 - w_y && (6.64)\\
\Leftrightarrow 1 + w_y &< 1 + P_{y,y} \\
\Leftrightarrow w_y &< P_{y,y}
\end{aligned}
$$

We use that w is an eigenvector in Eq. 6.61 and subsequently rewrite the matrix-vector multiplication. In Eq. 6.63, we apply the assumption that w_y is greater than any transition probability, so $P_{k,y}/w_y$ must be strictly smaller than one. Equation (6.64) follows because the entries of w must sum to one as w is a probability distribution (see Definition 6.16). Finally, we arrive at the statement $w_y < P_{y,y}$ which is contradictory to our assumption that w_y is bigger than every $P_{k,y}$. Thus, this assumption was false and probability w_y must be smaller or equal to $\max_{k,l \in \mathcal{S}} P_{kl}$ for any state $y \in \mathcal{S}$.

Second, we prove that $\frac{\Pr[x_i|y_j]}{\Pr[x_i|y'_j]} = \frac{P_{y_j,x_i}}{P_{y'_j,x_i}} \leq \gamma$. Let indices $i, j \in [n]$ such that $j < i$ and let states $x_i, y_j, y'_j \in \mathcal{S}$. If random variables X_i and X_j are direct neighbors, i.e., $i = j+1$, then the probabilities are transition probabilities from matrix P and the bound follows trivially. In the other case (i.e., $j + 1 < i$), we have

$$
\begin{aligned}
\Pr[X_i = x_i \mid X_j = y_j] &= \sum_{y_{j+1} \in \mathcal{S}} \Pr[x_i \mid y_j, X_{j+1} = y_{j+1}] \Pr[X_{j+1} = y_{j+1} \mid y_j] \\
&= \sum_{y_{j+1} \in \mathcal{S}} \Pr[x_i \mid y_{j+1}] \Pr[y_{j+1} \mid y_j]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{y_{j+1}\in\mathcal{S}} \Pr[x_i \mid y_{j+1}]\ \Pr[y_{j+1} \mid y'_j]\ \frac{\Pr[y_{j+1} \mid y_j]}{\Pr[y_{j+1} \mid y'_j]} \\
&= \sum_{y_{j+1}\in\mathcal{S}} \Pr[x_i \mid y'_j, y_{j+1}]\ \Pr[y_{j+1} \mid y'_j]\ \frac{P_{y_j,y_{j+1}}}{P_{y'_j,y_{j+1}}} \qquad (6.65)\\
&\leq \sum_{y_{j+1}\in\mathcal{S}} \Pr[x_i \mid y'_j, y_{j+1}]\ \Pr[y_{j+1} \mid y'_j]\,\gamma \qquad (6.66)\\
&\leq \gamma \Pr[X_i = x \mid X_j = y'] \,.
\end{aligned}
$$

We use Lemma 6.18 to remove $X_j = y_j$ from the condition of the first probability in Eq. 6.65 because $j+1$ is closer to i. If $i < j$ then the results follow from applying Bayes' rule and the previous case.

Finally, we prove the last inequality from Equation (6.57), in a similar fashion to the one before. Given $x_i, x'_i, y_j \in \mathcal{S}$. If random variables X_i and X_j are direct neighbors (i.e., $j = i-1$), the ratio can once again be bounded straightforwardly as it only contains probabilities from P.

Otherwise for $j < i-1$, we have

$$
\begin{aligned}
\Pr[X_i = x_i | X_j = y_j] &\sum_{x_{i-1}\in\mathcal{S}} \Pr[x_i | y_j, x_{i-1}]\ \Pr[x_{i-1} | y_j] \\
&= \sum_{x_{i-1}\in\mathcal{S}} \Pr[x_i | x_{i-1}]\ \Pr[x_{i-1} \mid y_j] \qquad (6.67)\\
&= \sum_{x_{i-1}\in\mathcal{S}} \Pr[x'_i \mid x_{i-1}]\ \frac{\Pr[x_i \mid x_{i-1}]}{\Pr[x'_i | x_{i-1}]}\ \Pr[x_{i-1} | y_j] \\
&= \sum_{x_{i-1}\in\mathcal{S}} \Pr[x'_i \mid x_{i-1}]\ \Pr[x_{i-1} | y_j]\ \frac{P_{x_{i-1},x_i}}{P_{x_{i-1},x'_i}} \\
&\leq \sum_{x_{i-1}\in\mathcal{S}} \Pr[x'_i \mid x_{i-1}, y_j]\ \Pr[x_{i-1} | y_j]\,\gamma \qquad (6.68)\\
&= \gamma \Pr[X_i = x' | X_j = y] \,.
\end{aligned}
$$

Lemma 6.18 is used to drop $X_j = y$ from the condition in Eq. 6.67 and add it again in Equation (6.68). If $i < j$ then the results follow from applying the Bayes rule and the previous case.

Combining Equation (6.56) and Equation (6.57) we obtain that

$$
\frac{\Pr[x_i \mid \mathbf{x}_A]}{\Pr[x'_i \mid \mathbf{x}_A]} = \frac{\Pr[x_r \mid x_i]\ \Pr[x_i \mid x_\ell]}{\Pr[x_r \mid x'_i]\ \Pr[x'_i \mid x_\ell]} \leq \gamma^2 \qquad (6.69)
$$

Note that if A only contains indices smaller than i, then previous equation gets simplified to

$$
\frac{\Pr[x_i \mid \mathbf{x}_A]}{\Pr[x'_i \mid \mathbf{x}_A]} = \frac{\Pr[x_i \mid x_\ell]}{\Pr[x'_i \mid x_\ell]} \leq \gamma \leq \gamma^2
$$

and the analogous holds if A only contains indices bigger than i.

Finally, combining Bayes' rule with Equation (6.69) applied to $A = K$ and $A = [n]\backslash\{i\}$, we obtain the result:

$$\frac{\Pr[\mathbf{x}_U|\mathbf{x}_K, x_i]}{\Pr[\mathbf{x}_U|\mathbf{x}_K, x_i']} = \frac{\Pr[x_i|\mathbf{x}_K, \mathbf{x}_U]\ \Pr[x_i'|\mathbf{x}_K]}{\Pr[x_i'|\mathbf{x}_K, \mathbf{x}_U]\ \Pr[x_i|\mathbf{x}_K]} = \frac{\Pr[x_i|\mathbf{x}_{-i}]\ \Pr[x_i'|\mathbf{x}_K]}{\Pr[x_i'|\mathbf{x}_{-i}]\ \Pr[x_i|\mathbf{x}_K]} \leq \gamma^2\gamma^2 = \gamma^4$$

Note that if $K = \varnothing$ the previous expression gets simplified to

$$\frac{\Pr[\mathbf{x}_U|x_i]}{\Pr[\mathbf{x}_U|x_i']} = \frac{\Pr[x_i|\mathbf{x}_U]\ \Pr[x_i']}{\Pr[x_i'|\mathbf{x}_U]\ \Pr[x_i]} \leq \gamma^3 \leq \gamma^4. \qquad \square$$

Finally, combining previous lemmas we obtain our novel bound:

Theorem 6.20 (The Markov Chain Bound)**.** *Let $s \in \mathbb{N}$ be the number of states. Let $\mathcal{M} : \mathcal{S}^n \to \mathcal{D}(\Theta)$ be an ε-DP mechanism. Let the databases follow a Markov chain with transition matrix $P \in \mathbb{R}^{s\times s}$ and initial distribution $w \in \mathbb{R}^s$ with the following properties:*

(H1) For all $x, y \in \mathcal{S}$ we have $P_{x,y} > 0$ and,

(H2) $wP = w$.

Then, $\mathcal{M}$ is an $(\varepsilon + 4\ln\gamma)$-BDP mechanism where

$$\gamma \coloneqq \frac{\max_{x,y\in\mathcal{S}} P_{xy}}{\min_{x,y\in\mathcal{S}} P_{xy}}.$$

Proof. If there are no unknown indices $U = \varnothing$ the adversary knows every index $K = [n] \setminus \{i\}$ except the target index and the adversary-specific $\mathrm{BDPL}_{(K,i)}$ becomes the same as the DP privacy leakage [37]. Thus, since $\varepsilon \leq \varepsilon + 4\gamma$ for all $\gamma \geq 1$, the inequality is trivially satisfied.

We denote by $u = |U|$ the size of unknown indices. If there is at least one unknown index for the adversary, i.e., $u \geq 1$ then we have

$$\begin{aligned}
\Pr_{\mathcal{M}}[Y \in S|\mathbf{x}_K, x_i] &= \sum_{\mathbf{x}_U\in\mathcal{S}^u} \Pr_{\mathcal{M}}[Y \in S|\mathbf{x}_K, x_i, \mathbf{x}_U]\ \Pr_{\Pi}[\mathbf{x}_U|\mathbf{x}_K, x_i] \\
&= \sum_{\mathbf{x}_U\in\mathcal{S}^u} \Pr[Y \in S|\mathbf{x}_K, x_i, \mathbf{x}_U]\ \Pr[\mathbf{x}_U|\mathbf{x}_K, x_i']\ \frac{\Pr[\mathbf{x}_U|\mathbf{x}_K, x_i]}{\Pr[\mathbf{x}_U|\mathbf{x}_K, x_i']} \\
&\overset{\text{(Lemma 6.19)}}{\leq} \sum_{\mathbf{x}_U\in\mathcal{S}^u} \Pr[Y \in S|\mathbf{x}_K, x_i, \mathbf{x}_U]\ \Pr[\mathbf{x}_U|\mathbf{x}_K, x_i']\ \left(\frac{\max_{kl} P_{kl}}{\min_{kl} P_{kl}}\right)^4 \\
&= \left(\frac{\max_{kl} P_{kl}}{\min_{kl} P_{kl}}\right)^4 \sum_{\mathbf{x}_U\in\mathcal{S}^u} \Pr[Y \in S|\mathbf{x}_K, x_i, \mathbf{x}_U]\ \Pr[\mathbf{x}_U|\mathbf{x}_K, x_i'] \\
&\leq \left(\frac{\max_{kl} P_{kl}}{\min_{kl} P_{kl}}\right)^4 e^{\varepsilon} \sum_{\mathbf{x}_U\in\mathcal{S}^u} \Pr[Y \in S|\mathbf{x}_K, x_i', \mathbf{x}_U]\ \Pr[\mathbf{x}_U|\mathbf{x}_K, x_i'] \\
&= \left(\frac{\max_{kl} P_{kl}}{\min_{kl} P_{kl}}\right)^4 e^{\varepsilon}\ \Pr[Y \in S|\mathbf{x}_K, x_i']
\end{aligned}$$

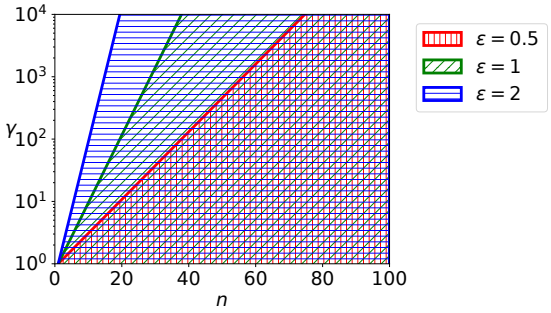


Figure 6.3.: Comparison of Markov-specific bound to general bound. The Markov-specific bound improves upon the general bound for values of n and γ in the respective shaded area.

Therefore, for every possible adversary with $u \geq 1$ we have that $\mathrm{BDPL}_{(K,i)} \leq \varepsilon + 4 \ln \gamma$. Since we have bounded the $\mathrm{BDPL}_{(K,i)}$ of every possible adversary (K, i), we also bound the total BDPL. □

(H1) states that all entries in the transition matrix are strictly positive, while (H2) requires that the initial distribution is a *stationary distribution*, meaning the distribution over states w_t (without considering the previous one) remains constant at each time—a common modeling assumption in various data mining tasks such as weather forecasting [187] or electricity consumption [188]. Notably, condition (H1) implies that the chain is both irreducible and aperiodic, which in turn guarantees the existence of a unique stationary distribution w [81], thereby satisfying (H2). Furthermore, for any initial distribution w', the distribution at time t converges geometrically fast to w as t increases [81]. Consequently, even when the initial distribution is not stationary, it asymptotically approaches the stationary distribution, effectively satisfying (H2) after discarding a sufficient initial portion of the process.

While prior work provides a mechanism for BDP protection of lazy binary Markov chains with a symmetric transition matrix [69], we present the first direct and general relationship between DP and BDP leakage for arbitrary Markov chains, including non-binary ones. When compared this novel bound with the general one we obtain that for any $\varepsilon > 0$, and maximum transition probability ratio $\gamma \geq 1$ we have

$$\varepsilon + 4 \ln \gamma < n\varepsilon \quad \text{if and only if } \gamma < \exp\left(\frac{n-1}{4}\varepsilon\right). \tag{6.70}$$

Therefore, the Markov chain bound outperforms the general bound in most cases. For instance, with an ε-DP mechanism where $\varepsilon = 0.5$ and a database size of $n = 80$, it remains tighter even when the largest transition probability is $10,000$ times the smallest. For the same $\varepsilon = 0.5$, the Markov bound only becomes looser than the general one when the number of correlated records is small, e.g., $n = 20$, and the transition probability ratio γ is as high as 100, which still represents a significant disparity (See Figure 6.3).

Moreover, Theorem 6.20 enables the systematic design of BDP mechanisms by adjusting the noise of an existing DP mechanism. Noise calibration depends only on the maximum ratio between the Markov transition probabilities of the model, γ, and the adjusted mechanism must be calibrated to $\varepsilon' = \varepsilon - 4\ln(\gamma)$. Note that the best BDPL privacy achievable using Theorem 6.20 is $\varepsilon = 4\ln(\gamma)$, since $\varepsilon' \geq 0$. Consequently, the minimum achievable ε is fundamentally constrained by the structure of the underlying Markov model—specifically, by the maximum transition ratio γ. We illustrate how the transition matrix changes the minimum ε in theoretical settings in Figure 6.5, and in real-world data in Section 6.5.

6.4.2. Accuracy

The Markov chain bound enables us to derive improved utility guarantees for the Laplace mechanism when γ is sufficiently small.

Corollary 6.21. *Let $\mathcal{M}_{\varepsilon,f}$ be the ε-Laplace mechanism that approximates the query $f : \mathcal{S}^n \to \mathbb{R}$ and inputs a database drawn from a Markov chain satisfying (H1) and (H2). If $\mathcal{M}_{\varepsilon,f}$ is (α,β)-accurate w.r.t. f and $\varepsilon \geq 4\ln(\gamma)$ then, there exists an ε-BDP mechanism $\widetilde{\mathcal{M}}$ that is $(h\alpha,\beta)$-accurate w.r.t. f with*

$$h = \frac{\varepsilon}{\varepsilon - 4\ln(\gamma)}.$$

Proof. The idea of this proof is to construct mechanism $\widetilde{\mathcal{M}}$ with the Laplace mechanism as well, but to choose a carefully selected privacy leakage $\varepsilon' < \varepsilon$ so that mechanism $\widetilde{\mathcal{M}}$ is (1) ε-BDP and (2) $(h\alpha,\beta)$-accurate.

First, we determine the accuracy of mechanism $\mathcal{M}_{\varepsilon,f_I}$. With Proposition 2.10, we know that the (α,β)-accuracy of the Laplace mechanism for a given probability $\beta \in (0,1]$ and privacy parameter ε is

$$\alpha = \ln\left(\frac{1}{\beta}\right)\frac{\Delta f}{\varepsilon}.$$

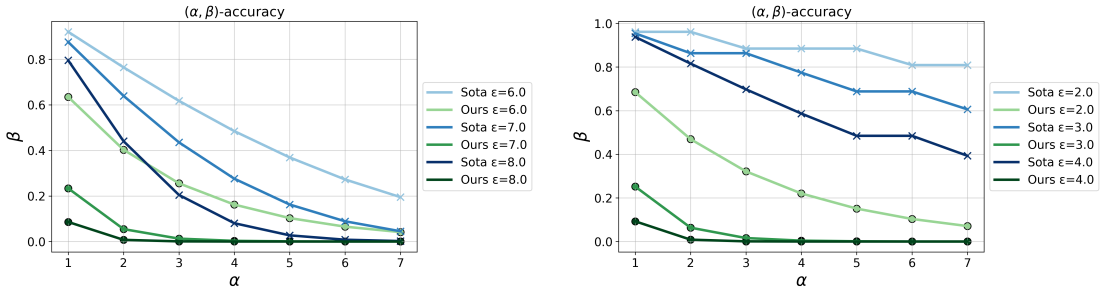
So this is the (α,β)-accuracy of $\mathcal{M}_{\varepsilon,f}$.

We have to show that there exists an ε-BDP mechanism $\widetilde{\mathcal{M}}$ which is $(h\alpha,\beta)$-accurate. We choose $\widetilde{\mathcal{M}}$ as the Laplace mechanism applied to f with an adjusted privacy parameter $\varepsilon' > 0$. Thus, $\widetilde{\mathcal{M}}$ will be ε'-DP. Therefore, we can use Theorem 6.20 to show that $\widetilde{\mathcal{M}}$ is BDP. We must choose ε' in a way that ensures that the BDPL is limited to ε, so that we have ε-BDP. With Theorem 6.20, $\widetilde{\mathcal{M}}$ is

$$(\varepsilon' + 4\ln\gamma)\text{-BDP}.$$

Therefore, to achieve ε-BDP, we must have

$$\varepsilon' + 4\ln\gamma = \varepsilon \Leftrightarrow \varepsilon' = \varepsilon - 4\ln\gamma.$$



(a) Self-transition probability $P_{ss} = 0.8$.

(b) Self-transition probability $P_{ss} = 0.6$.

Figure 6.4.: (α, β)-accuracy comparison of our mechanism vs. the state-of-the-art approach [69] for $n = 500$. Our mechanism based on Theorem 6.20 (in green) consistently achieves lower error than the state of the art (in blue) for the same privacy and confidence levels.

Now, we can calculate the accuracy of $\widetilde{\mathcal{M}}$ because it also uses the Laplace mechanism. Then, we find an upper bound for this accuracy. Mechanism $\widetilde{\mathcal{M}}$ is (α', β)-accurate, with

$$\alpha' = \ln(\frac{1}{\beta})\frac{\Delta f}{\varepsilon'} = \ln(\frac{1}{\beta})\frac{\Delta f}{\varepsilon - 4\ln\gamma} = \alpha h.$$

□

6.4.2.1. Comparison with the State of the Art

Chakrabarti et al. [69] propose a BDP adaptation of the randomized response mechanism for symmetric, lazy stationary Markov chains with binary states, i.e., a Markov chain with $s \in \{0, 1\}$, $w = (0.5, 0.5)$ and symmetric transition matrix

$$P = \begin{pmatrix} 1-r & r \\ r & 1-r \end{pmatrix}$$

with a constant self-transition probability $P_{ss} = r \in (0, 0.5)$ for all $s \in \{0, 1\}$, indicating the laziness, i.e., it is more likely to remain in the same state that a change. They prove that the adapted randomized response such that

$$\Pr_{\mathcal{M}}(Y_i \mid X_i) = \begin{cases} 1-\rho & \text{if } Y_i = X_i \\ \rho & \text{otherwise,} \end{cases}$$

where

$$\rho \geq \frac{4 + r(re^{\varepsilon} - 2) - \sqrt{r^2 e^{\varepsilon}(4 + r(re^{\varepsilon} - 4))}}{8 + 2r(re^{\varepsilon} + r - 4)}$$

fulfills ε-BDP. While they do not give any utility estimate or experiment, we compute the (α, β)-accuracy of this mechanism. Moreover, we show that it provides worse accuracy than our Laplace-based mechanism.

The target query in a randomized mechanism is the true number of positive answers, denoted by n_1, to a given question (i.e., $s = 1$). Rather than using the noisy total count directly, we can compute an unbiased estimator of n_1 for randomized response with parameter $p = 1 - \rho$ [189]. Specifically, given y_i, the noisy response of user i, we compute

$$\hat{n}_1 = \frac{\sum_{i=1}^{n} y_i - n(1-p)}{2p-1}.$$

Additionally, considering $Z = \sum_{i=1}^{n} Y_i$, the random variable Z can be expressed as the convolution of two binomial distributions: $Z = Z_0 + Z_1$. Here, Z_0 represents the number of reported 1s originating from individuals with $X_i = 0$ who lied, and Z_1 represents the number of correctly reported 1s where $X_i = 1$ was preserved. Formally,

$$Z_0 \sim \mathrm{Bin}(N - n_1, \rho),\ Z_1 \sim \mathrm{Bin}(n_1, 1-\rho).$$

Hence, by definition of (α, β)-accuracy we obtain:

$$\begin{aligned}\Pr\left[|\hat{n}_1 - n_1| \geq \alpha\right] &= \Pr\left[|n_1 - \frac{(Z - n(1-p))}{(2p-1)}| \geq \alpha\right] \\ &= \Pr\left[|n_1(2p-1) + n(1-p) - Z| \geq \alpha(2p-1)\right]\end{aligned}$$

Therefore, the probability of interest can be decomposed as:

$$\Pr\left[\,|Z - (n_1(2p-1) + n(1-p))| \geq t\,\right] = \Pr\left[Z \leq \mu - t\right] + \Pr\left[Z \geq \mu + t\right],$$

where $t = \alpha(2p-1)$ and $\mu = n_1(2p-1) + n(1-p)$. Since Z is a discrete random variable, for all $t \neq 0$, this can be written as:

$$\sum_{k=0}^{\lfloor \mu - t \rfloor} \Pr[Z = k] + \sum_{k=\lceil \mu + t \rceil}^{n} \Pr[Z = k].$$

Since Z is the convolution of two binomial random variables,

$$\Pr(Z = k) = \sum_{i=0}^{k} \Pr(Z_1 = i) \cdot \Pr(Z_0 = k - i),$$

therefore, the full expression becomes:

$$\begin{aligned}\beta &= \Pr\left[\,|Z - \mu| \geq t\,\right] \\ &= \sum_{k=0}^{\lfloor \mu - t \rfloor} \sum_{i=0}^{k} \binom{n_1}{i} (1-\rho)^k \rho^{\mu - k} \binom{n-\mu}{k-i} \rho^{k-i} (1-\rho)^{n-\mu-k+i} \\ &+ \sum_{k=\lceil \mu + t \rceil}^{n} \sum_{i=0}^{k} \binom{n_1}{i} (1-\rho)^k \rho^{n_1 - k} \binom{n-n_1}{k-i} \rho^{k-i} (1-\rho)^{n-n_1-k+i}\end{aligned}$$

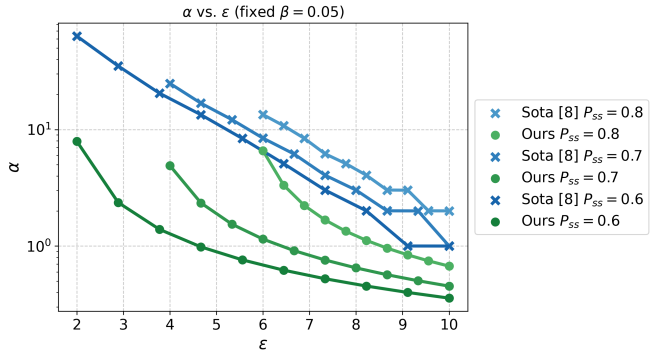


Figure 6.5.: Accuracy of our mechanism vs. the one proposed in [69] for $n = 700$ and various self-transition probabilities P_{ss}. At 95% confidence, confidence, our mechanism (in green) consistently ensures lower error than the state of the art (in blue) for all values of ε.

$$= \sum_{k=0}^{\lfloor \mu - t \rfloor} \sum_{i=0}^{k} \binom{n_1}{i} \binom{n-n_1}{k-i} (1-\rho)^{n-n_1+i} \rho^{n_1-i} \tag{6.71}$$

$$+ \sum_{k=\lceil \mu + t \rceil}^{n} \sum_{i=0}^{k} \binom{n_1}{i} \binom{n-n_1}{k-i} (1-\rho)^{n-n_1+i} \rho^{n_1-i} \tag{6.72}$$

where $t = \alpha(2p-1)$ and $\mu = n_1(2p-1) + n(1-p)$.

At the same time, since $\Delta f = 1$ for binary counting queries, we have that the (α, β)-accuracy of our Laplace-based mechanism is:

$$\beta = \mathrm{e}^{-\alpha(\varepsilon - 4\ln(\gamma))},$$

where in this case $\gamma = \frac{1-P_{ss}}{P_{ss}}$.

We compare both accuracies in Figure 6.4, showing that for all values, our mechanism has a better accuracy, i.e., lower β for the same γ. Additionally, we provide the variation of α respect to different ε values with fixed $\beta = 0.05$, i.e. 95% confidence in Figure 6.5. Note that, since Eq. 6.72 is not invertible, to obtain Figure 6.5, we numerically approximate α using the bisection method.

It is important to note that while their mechanism supports arbitrary BDPL, ours applies only for $\varepsilon \geq 4\ln(\gamma)$. However, our approach generalizes to arbitrary Markov chains, whereas theirs is limited to lazy, symmetric binary models. In the intersection of both applicability domains, our use of Laplace-based recalibration yields improved utility.

In conclusion, the Markov-specific bound significantly improves upon the general bound in most cases (as shown in Figure 6.3), especially for large numbers of correlated records, where the general bound increases drastically while our bound remains stable. Moreover, our result enables improved utility (Figure 6.5) compared to prior work [69]. Its advantage is most notable when the number of correlated records is large, as it remains independent of dataset size—unlike the general bound, which grows linearly. However, this comes at

the cost of a minimum privacy level determined by the data distribution, a limitation absent in the general bound and Chakrabarti et al.'s approach [69].

6.5. Utility Experiments

Theoretical bounds on privacy and utility do not always translate directly to practical implementations. For instance, while it may be theoretically feasible to achieve a given (α, β)-accuracy, designing or implementing a mechanism that attains this in practice can be challenging. In this section, we use our theoretical results to construct a BDP mechanism and empirically evaluate its utility on real-world databases that follow either multivariate Gaussian correlations or Markov chains. Our goal is to show that the theoretical utility gains predicted under specific correlation structures are indeed achievable in practice as well as measure the improvement over previous approaches.

We calibrate the Laplace mechanism using Theorem 6.14 and Theorem 6.20 to derive BDP mechanisms. We then run these BDP mechanisms on the selected databases and compare the utility results with those of BDP mechanisms designed to protect against arbitrary correlation, in order to assess the improvements offered by the correlation-specific approach. Moreover, we also plot, when applicable, the accuracy results of the state-of-the-art solutions for Gaussian BDP [37]. Unfortunately, none of the evaluated datasets meet the strict assumptions needed to apply the only prior mechanism for Markov models [69]. Finally, we plot the utility of the classical DP Laplace mechanism as a baseline, representing the best-case utility achievable ignoring correlation.

6.5.1. Databases

We use four real-world databases, two for each correlation model. Additionally, we use a synthetic dataset to test scalability for Gaussian correlations. The selection criteria are public availability, quality of the databases, and the fulfillment of the theoretical assumptions, namely, following the correlation model and fulfilling the extra hypotheses of the corresponding theorem in each case, regarding the Pearson correlation coefficient and the transition matrix.

6.5.1.1. Multivariate Gaussian

We use two datasets that align well with our modeling framework: the Galton Height Data [190], a historical dataset originally compiled to study the correlation between parents' and children's heights, and the FamilyIQ dataset [191], which includes IQ scores of gifted children and their parents.

The Galton Height Data—considered a classical example of linear correlation modeling, where regression and correlation are interpreted within the framework of a multivariate Gaussian distribution [192]—is especially well known in statistical analysis for introducing the very concept of regression [170]. In contrast, several studies provide evidence that IQ scores in the general population are standardized to follow a multivariate Gaussian distribution, where non-zero correlations are observed only among close relatives [193].

Database	n	m	Parameters	Sensitivity
Galton	897	3	$\rho = 0.275$	$\Delta q = 254\,\text{cm}$
FamilyIQ	868	2	$\rho = 0.4483$	$\Delta q = 120$
SyntheticIQ	20 000	2	$\rho = 0.45$	$\Delta q = 120$
Activity	17 568	n	$\gamma = 7.54$	$\Delta q = 1$
Activity Single Day	288	n	$\gamma = 7.54$	$\Delta q = 1$
Electricity	731	n	70 kWh, $\gamma = 3.29$ 80 kWh, $\gamma = 4.49$ 90 kWh, $\gamma = 8.43$	$\Delta q = 1$

Table 6.2.: Data description. Here, m denotes the max number of correlated records and n the total amount.

These properties make both datasets well-suited for evaluating the practical transferability of our Gaussian-based bounds. Additionally, we generate the dataset SyntheticIQ to test the scalability of our approach. Following the findings among several populations summarized in [193], we generate data following a Gaussian distribution with $\mu = 100$, $\sigma^2 = 15$ and $\rho = 0.45$ for parent-child.

To ensure bounded sensitivities, all records are clipped to the range of 0 cm to 254 cm (0 to 100 inches) for Galton, and from 40 to 160 for IQ datasets as summarized in Table 6.2.

All explored datasets fulfill the conditions of our Theorem 6.14: In Galton Data the Pearson correlation coefficient is $\rho = 0.275$, satisfying the condition $\rho = 0.275 < 1 = \frac{1}{m-2}$, hence our bounded-correlation assumptions hold. For $m = 2$, the condition trivially holds for all ρ values, so in particular for FamilyIQ and SyntheticIQ.

6.5.1.2. Markov Model

We study two use cases—human activity and electricity consumption—well-suited for Markov modeling. Human activity representations such as "inactive" versus "active" are modeled by Markov chains [194]. Similarly, electricity usage patterns, particularly transitions between periods of high and low consumption, have been effectively modeled using Markov processes [188], [195], [196]. We select a representative database for each domain to evaluate our framework. For human activity, we use Activity Data [197], which contains the time series of step counts recorded every 5 minutes from a personal activity monitoring device worn by a single individual during October and November 2012. This allows us to extract the "active" state if any steps are recorded and the "inactive" when the user does not move. Besides, to assess the data size impact, we split Activity Data into 61 unique subdatabases, each corresponding to the activity states of a single day.

For electricity usage, we use the Electricity Dataset [198], which captures a single residence electricity usage in Canada from 2012 to 2014. Modeling home energy consumption

is essential for studying demand-response, and particularly for detecting on-peak periods (high consumption) and off-peak periods (low consumption) [188]. Hence, we classify each hour as low or high consumption depending on whether the usage falls below or exceeds a fixed threshold of 80 kWh—the central value of the range. Additionally, we study different threshold values, 70 and 90 kWh, to assess their impact on utility. In all cases, we evaluate event-level local privacy guarantees, assuming no trusted curator and focusing on user-side privacy protection [16]. The technical details of the three datasets are summarized in Table 6.2.

In order to fulfill the conditions of Theorem 6.20 we require the transition probabilities of the Markov chain to be positive. We calculate them empirically and receive the following transition matrices for Activity and for Electricity $70, 80, 90$ kWh in this order:

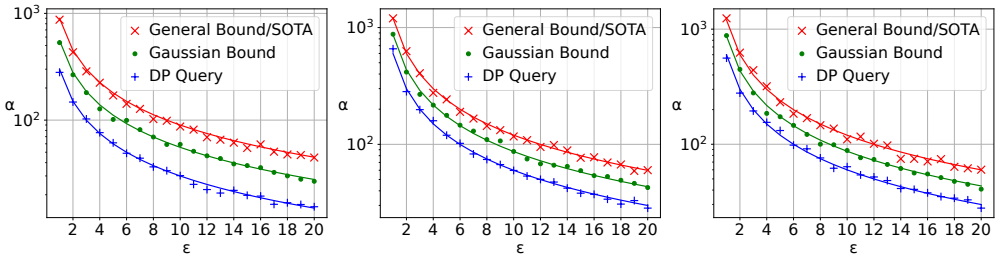
$$\begin{pmatrix} 0.882 & 0.117 \\ 0.305 & 0.695 \end{pmatrix}, \begin{pmatrix} 0.445 & 0.555 \\ 0.149 & 0.850 \end{pmatrix}, \begin{pmatrix} 0.818 & 0.182 \\ 0.371 & 0.629 \end{pmatrix}, \begin{pmatrix} 0.894 & 0.106 \\ 0.478 & 0.522 \end{pmatrix},$$

representing $P_{00}, P_{01}, P_{10}, P_{11}$ with $s = 0$ inactive/low consumption and $s = 1$ active/high consumption. Our theoretical results also require w to be a stationary distribution. While w can not be empirically computed since we only have one initial state, both Markov chains are irreducible, since both states are reachable from each other, aperiodic, since $P_{ss} \neq 0$ for both $s \in \{0, 1\}$, and $P_{st} > 0$ hence there exists a stationary initial distribution [199]. Therefore, we conclude that the databases fulfill the conditions for testing our results.

6.5.2. Target Queries and Utility Metrics

We focus our utility study on two commonly used queries: sum and counting queries. These queries are relevant to our use cases because they allow aggregating numerical data—such as the average height or IQ of a population—or counting events, like the number of daily energy peaks or the amount of daily activity. Formally, given a database $D = (x_1, x_2, \ldots, x_n)$, where each x_i represents a numerical value, a sum query is defined as: $q_S(D) = \sum_{i=1}^{n} x_i$. In the case of the Gaussian data, each x_i corresponds to an individual's height or IQ. If each record is binary, i.e., $x_i \in \{0, 1\}$, as is the case for the activity and electricity datasets, $q_S(D)$ is called a counting query since it outputs the count of states with the attribute 1.

Our theoretical results are expressed in terms of (α, β)-accuracy. To evaluate empirical utility, we use the upper bound of a $(1 - \beta)$ confidence interval for the absolute query error, which serves as a practical counterpart. Specifically, we report the upper limit of a 95% confidence interval (i.e., $\beta = 0.05$), a standard choice in practice [200]. A smaller upper bound indicates higher utility. When this bound is close to the theoretical error α, it demonstrates a strong alignment between empirical and theoretical results, highlighting their practical applicability. To facilitate comparison with our theoretical results, we plot the theoretical error tolerance α for each mechanism, derived from Proposition 2.10 for the baseline DP mechanism and Corollary 6.5, Corollary 6.15, and Corollary 6.21 for the general bound, the Gaussian bound and the Markov chain bound respectively. Additionally, we report the mean absolute percentage error (MAPE) to estimate the expected relative error for a single execution.



(a) Galton, $n = 897$ $m = 3$ (b) FamilyIQ, $n = 868$, $m = 2$. (c) SyntheticIQ, $n = 20000$, $m = 2$.

Figure 6.6.: Gaussian data results. Lines show theoretical error at $\beta = 5\%$ and markers indicate empirical 95% upper bounds. Both align well, as expected. Our Gaussian bound shows significant improvements over prior work, being approximately 50% closer to utility lower bound corresponding to the DP mechanism (insecure under this model).

6.5.3. Mechanism and Experiment Design

In order to provide BDP mechanisms that approximate the target queries presented in Section 6.5.2, we use the Laplace mechanism with the noise calibrated through Theorem 6.3 for the DP baseline, Theorem 6.14 for Gaussian data and Theorem 6.20 for Markov data. In this section, we refer to the DP privacy leakage by τ, to avoid confusion with the actual maximum BDPL denoted by ε.

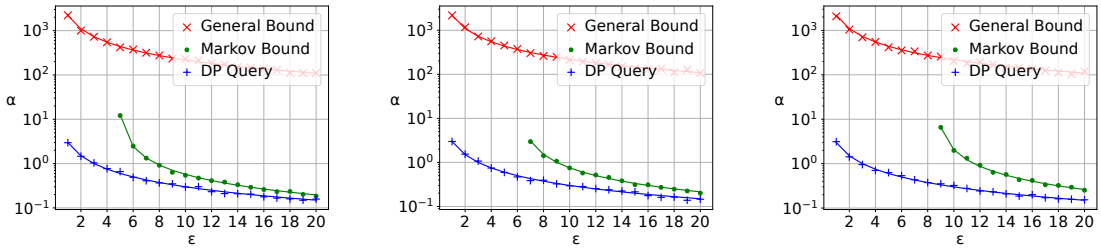
6.5.3.1. Gaussian Data

As explained in Section 6.5.1, we assume that the data is drawn from a multivariate Gaussian distribution with maximum number of correlated variables m respectively. Both the general bound and state of the art [37] indicate that for the Laplace mechanism $\mathcal{M}_{\tau,f}$, we have $\varepsilon = m\tau$, i.e., $\varepsilon = 3\tau$ for Galton and $\varepsilon = 2\tau$ for IQ datasets. Alternatively, according to the Pearson coefficients described in Table 6.2, Theorem 6.14 tells us that $\mathcal{M}_{\tau,f}$ is ε-BDP, with $\varepsilon \approx 1.853\tau, 1.45\tau$ for Galton and IQ datasets respectively. Consequently, we fix BDPL values $\varepsilon \in (0, 20]$ and compute the corresponding τ using Eq. 6.42 for the Gaussian-specific correlation approach and $\tau = \frac{\varepsilon}{3}$ for the general correlation and state of the art. For $\varepsilon \in (0, 5)$, we ensure strong theoretical privacy guarantees, while also considering the higher range $\varepsilon \in [5, 20]$, which has shown empirical resilience to certain privacy attacks [201], [202].

6.5.3.2. Markov Data

As discussed in Section 6.5.1 we assume that the data follows a Markov chain. According to the γ values summarized in Table 6.2, Theorem 6.20 tells us that the Laplace mechanism $\mathcal{M}_{\tau,f}$ applied to a counting query f is ε-BDP, with

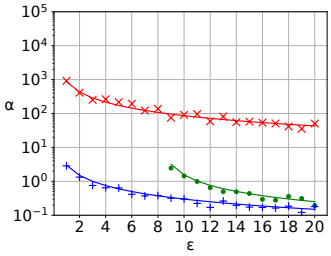
$$\varepsilon_A = \tau + 8.05,\ \varepsilon_{E,70} \approx \tau + 4.7,\ \varepsilon_{E,80} \approx \tau + 6.03,\ \varepsilon_{E,90} \approx \tau + 8.54. \tag{6.73}$$



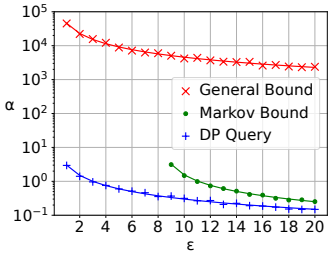
(a) Electricity 70 kWh, $n = 731$.

(b) Electricity 80 kWh, $n = 731$.

(c) Electricity 90 kWh, $n = 731$.



(d) Activity Single Day, $n = 288$.



(e) Activity, $n = 17\,568$.

Figure 6.7.: Markov Data Results. Lines show theoretical error at $\beta = 5\%$ and markers indicate empirical 95% upper bounds. Both align well, as expected. We observe a drastic utility improvement for our Markov bound compared to the general bound. This is expected, since the general bound escalates with the number of correlated records, whereas the Markov bound remains stable.

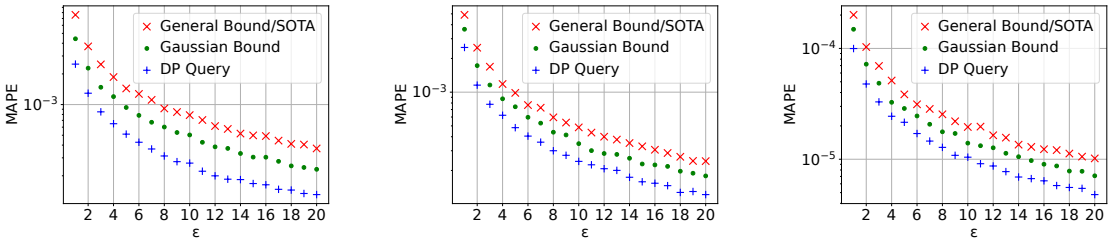
In comparison, with the general bound we have $\varepsilon = n\tau$ for mechanism $\mathcal{M}_{\tau,f}$. Similar to Gaussian data, we apply the Laplace mechanism to compute the sum query of each subgroup with BDPL values $\varepsilon \in (0, 20]$ and compute the corresponding τ using Eq. 6.73 for the Markov-specific mechanism and taking $\tau = \frac{\varepsilon}{n}$ for the general correlation approach. However, none of the datasets provide a symmetric transition matrix, which means that the proposal in [69] is not applicable, making an empirical comparison impossible.

Note that, while ε-BDP can be provided for all values using the general bound and state of the art [69], Eq.6.73 only allows for $\varepsilon \geq 8.05, 6.9, 4.7$ and 8.45 for Activity and Electricity data respectively, since τ must be positive (See Section 6.4).

In both Markov and Gaussian experiments, to calculate empirical confidence intervals, we execute the mechanism for each dataset 1000 times. Since Activity Single Day provides 56 unique datasets, we average the result over them.

6.5.4. Results and Discussion

Figure 6.6 presents the results for the Gaussian models, including our Gaussian-specific bound, the state-of-the-art bound from [37] (which coincides with the general bound), and the DP Laplace mechanism for sum queries. We plot the DP mechanism as the



(a) Galton data, $n = 897$ (b) FamiliyIQ, $n = 868$ (c) SyntheticIQ, $n = 20\,000$

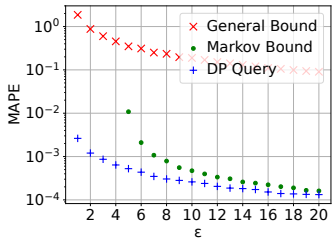
Figure 6.8.: MAPE of private sum queries on data correlated according to a Gaussian mutivariate distribution. Our Gaussian bound shows significant improvements over prior work, being approximately 50% closer to utility lower bound corresponding to the DP mechanism (insecure under this model).

baseline for the best possible utility; however, it is important to note that DP does not offer meaningful protection in this experiment, given correlation. Among the correlation-protecting mechanisms, those that use the Gaussian bound consistently outperform the state-of-the-art mechanism [37] for all ε in all datasets. Note that we plot all results on a logarithmic scale. This makes it harder to visually see the substantial reduction of error achieved by our mechanisms—particularly for small values of ε. For instance, for $\varepsilon = 1$ the error is reduced by more than 400 units for both IQ datasets and 200 inches for the Galton. Note that the Galton height data uses imperial units (inches), thus the errors are also interpreted in inches.

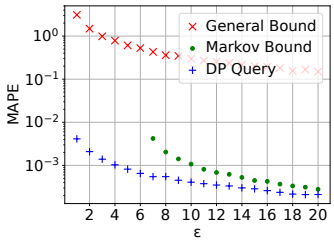
The results for Markov chains are shown in Figure 6.7. Again, we use the DP mechanism as the baseline for the best possible utility, not as a comparable protective mechanism. For BDP mechanisms, we observe that the different Markov models tested lead to varying minimum achievable BDPL levels, as determined by our Markov-based bound: Electricity 70 kWh yields the most favorable case with a minimum $\varepsilon = 4.9$, while 90 kWh imposes the weakest bound with a minimum $\varepsilon = 8.45$. In contrast, the general bound supports all $\varepsilon > 0$. Hence, if we want a lower ε than that allowed by the Markov bound, we should fall back to the general bound.

In all cases where the Markov chain bound is applicable, mechanisms using it significantly outperform those relying on the general bound. While the error of mechanisms based on the general bound increases sharply, the error of both the Markov chain–based mechanism and the standard DP mechanism remains stable. The larger n, the larger the improvement of our approach with respect to the general bound. For the largest dataset—Activity—the general bound results in a 10^5 times larger error than that of our proposed Markov chain bound. This is because the general bound scales with the size of the database n, while the Markov bound is independent of n, highlighting the huge benefit of using our novel bound for large datasets.

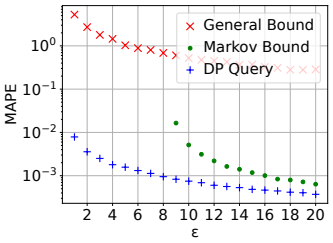
The results demonstrate that BDP mechanisms calibrated with our newly proven Gaussian and Markov chain bounds outperform prior BDP mechanisms and mechanisms calibrated with the general bound in terms of utility on real-world data. Moreover, the



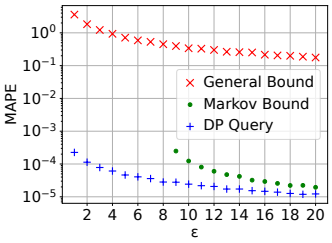
(a) Electricity 70 kWh, $n = 731$



(b) Electricity 80 kWh, $n = 731$



(c) Electricity 90 kWh, $n = 731$



(d) Activity, $n = 17\,568$

Figure 6.9.: MAPE results for databases following a Markov distribution. We observe a drastic utility improvement for our Markov bound compared to the general bound. This is expected, since the general bound escalates with the number of correlated records, whereas the Markov bound remains stable.

empirical errors from our experiments closely align with our theoretical utility results, validating the practical applicability of our theorems.

We extend this study with the analysis of the relative error. Figure 6.8 show the MAPE for Galton and IQ datasets (Gaussian correlation model) and Figure 6.9 Activity and Electricity data (Markov chain model). As expected, the DP query has the lowest MAPE however it does not offer protection against correlation. When we offer BDP protections we see that the correlation-specific BDP mechanisms (i.e., using the Gaussian bound or the Markov chain bound) outperform the BDP mechanism protecting against arbitrary correlation with the general bound following the same trend as for the (α, β)-accuracy. The benefit is particularly prominent for data following a Markov chain where the MAPE of the general bound reaches values above 100%, resulting in an error as large as the ground truth itself. In comparison, the Markov chain bound achieves errors below 10% for single day activity data, and below 0.1% for Activity and Electricity datasets.

We acknowledge certain limitations when extrapolating our results. The validity of our experimental findings is constrained by the specific databases used. While the Galton height data serves as a well-known example of record correlation, it reflects only one of many possible correlation patterns. Similarly, most practical applications of a Markov chain would involve more than two states, introducing complexity beyond the binary-state model used in our study. Nevertheless, our results provide valuable insight into the

practical applicability of our theorems and indicate their potential for real-world scenarios. Furthermore, these experiments demonstrate that achieving meaningful utility while protecting against correlation is feasible in practice.

6.6. Conclusion

In this chapter, we explored the utility of BDP mechanisms. We addressed prior limitations by analyzing broader correlation models and providing a detailed study of privacy-utility trade-offs, supported by theoretical results and empirical evidence.

Specifically, we established new connections between DP and BDP mechanisms and demonstrated how they can be leveraged for privacy protection under correlation. We proved that any ε-DP mechanism satisfies $m\varepsilon$-BDP, where m is the size of the correlated group, and showed this bound is tight. We then improved upon it by considering multivariate Gaussian and Markov models, deriving novel bounds on BDP leakage that provide stronger utility guarantees than the state-of-the-art approaches under the same privacy constraints. The advantage of our correlation-specific bounds is particularly evident under Markov-modeled correlations. While mechanisms based on the general bound exhibit high sensitivity to the number of correlated records, our Markov-based bound remains robust and stable regardless of the dataset size.

In particular, our results make it possible to overcome the limitations of event-level privacy in streaming settings that arise due to temporal correlations (cf. Section 3.3). Our BDP mechanism for Markov chains explicitly accounts for these dependencies and provides meaningful privacy guarantees for time series released in a streaming fashion, without incurring the substantial utility losses typically associated with user-level protection.

While it remains a futile attempt to apply BDP without assuming a specific correlation model, both our theoretical and experimental results demonstrate that it is possible to achieve better utility without weakening the adversary model in practical scenarios: (a) when the number of correlated records is small, (b) when the data follows a weakly correlated Gaussian model, or (c) when the data is a time series following a Markov chain with sufficiently similar transition probabilities. Note that, while this distributional assumption can be restrictive, they serves as a first step towards understanding whether actual BDP guarantees and utility are achievable. Moreover, we present several real-world datasets that satisfy this assumption.

Overall, our results Theorems 6.3, 6.14 and 6.20 advance the theoretical and practical understanding of BDP, enabling the reuse of DP mechanisms in correlated settings. This opens future directions for deriving correlation-specific bounds to design more accurate BDP mechanisms protecting against real-world correlation-based attacks.

7. Conclusion

In this thesis, we examined several challenges and limitations of DP in complex data, a domain that has become increasingly important in private data analysis. Trajectories, time series, graphs, and other high-dimensional structured objects exhibit rich semantics, strong internal correlations, and multiple different sensitive aspects to be protected. These characteristics fundamentally challenge the assumptions underlying classical DP and significantly complicate both the interpretation and the applicability of DP mechanisms.

Given that complex data structures are inherently more difficult to protect with existing DP techniques, our research is driven by two central questions. First, we seek to understand precisely where and why these methods fail. Second, we aim to determine how these limitations can be addressed in a principled and rigorous manner.

To address the first question, this thesis adopts a systematic approach. We conducted a comprehensive literature analysis on the DP publication of trajectory data, chosen as a representative and motivating instance of complex data. We classified DP mechanisms for trajectory data protection into four main categories and formally analyzed their privacy.

Our analysis revealed widespread formal and conceptual flaws in the literature. Some mechanisms ignore temporal information, leaving them vulnerable to time-based attacks, while others disregard correlations intrinsic to trajectories, enabling correlation-based inference attacks. We also identified incorrect adaptations of standard mechanisms, such as the Laplace mechanism, whose naïve adaptations to complex data domains often fail to provide DP guarantees. Beyond these methodological shortcomings, existing approaches are largely restricted to narrow application scenarios and simplified toy universes, highlighting that the private release of human mobility data remains an open and fertile research area.

From this systematic analysis, we identify and address four core obstacles that hinder the applicability of DP to complex data: (i) The formal flaws and bugs when trying to adapt existing DP mechanisms to more complex systems, (ii) the proliferation of granularities and neighborhood definitions and the challenge of extending DP properties, such as composition, to them, (iii) the lack of interpretability of privacy parameters with respect to realistic attacks, and (iv) the presence of correlations.

Building on this diagnosis, the thesis advances the state of the art by systematically addressing each of these fundamental limitations that complex data introduce for the adoption of DP.

First, we address the problem of composition across multiple granularities and data domains. We introduce a unified composition framework based on metric privacy, which simultaneously accommodates all granularities and data domains. In this framework, composition is determined solely by the preprocessing functions and the induced metrics,

removing the need to distinguish between sequential and parallel composition. This perspective enables tighter privacy bounds by explicitly accounting for preprocessing effects and applies uniformly across diverse data domains and levels of granularity. Importantly, it also allows for intermediate and previously unexplored composition regimes, which are particularly relevant for complex data pipelines.

Within this unified framework, we identify the precise conditions under which optimal privacy loss bounds can be achieved for partitioned databases. This addresses a notable gap in the literature, where practitioners are often forced to rely on sequential composition and consequently incur significant utility loss, as parallel composition does not readily generalize to these settings.

We extend our general composition framework to metric Gaussian DP, bringing the best of both notions by leveraging metric privacy for tighter composition analysis while retaining the interpretability in terms of attacks and the strong compositional properties of GDP. Notably, in the metric Gaussian setting, our partitioning theorems yield tighter composition bounds than the classical approach of taking the maximum of the individual metric guarantees, leading to more accurate privacy estimates for complex workflows.

Second, we address the gap between formal DP guarantees and protection against concrete attacks. Our analysis shows that commonly used evaluation metrics, such as ReRo, are inconsistent in practice and can overestimate the true privacy risk, as they do not discount attacks that merely reconstruct records based on prior knowledge. To remedy this, we introduce the RAD metric, prove its theoretical consistency, and show its empirical superiority across a wide range of scenarios, including private learning, DP aggregation, and LDP. Crucially, RAD reveals that privacy risk depends on the structure of the mechanism rather than solely on its nominal DP parameters. This insight is especially important for complex data, where different mechanisms with the same ε can provide drastically different protection levels.

We further derive tight bounds linking DP parameters to attack success under varying adversarial knowledge, extending the analysis beyond MIAs to a broader class of attacks relevant in complex domains. By constructing the optimal attack for any given mechanism, auxiliary knowledge, and prior distribution, we not only prove the tightness of our formal bounds but also provide a powerful auditing tool for DP mechanisms. Together with our impossibility results, our auditing framework provides a practical solution to detect DP implementation and formal flaws and ensure that the privacy guarantees advertised to individuals are preserved.

This attack-centric perspective significantly improves the interpretability and calibration of privacy guarantees in complex data systems, enabling substantial utility gains without increasing effective privacy risk.

Finally, we confront the issue of correlations. After revisiting impossibility results that rule out meaningful guarantees under arbitrary correlation models, the thesis focuses on structured and practically relevant correlation families. In particular, we derive new leakage bounds in BDP for certain multivariate Gaussian models and Markov chains. These correlation-specific bounds allow standard DP mechanisms to be reused in correlated settings with substantially improved utility compared to state-of-the-art

approaches. The benefits are especially pronounced for Markov-modeled time series, where the proposed bounds remain stable regardless of dataset size. Our results open the path for meaningful privacy guarantees for temporally correlated data releases without resorting to user-level protection and its associated utility loss.

Taken together, this thesis not only identifies several fundamental limitations of DP adoption in complex data but also advances concrete theoretical and practical solutions to each of them. It provides tools to detect formal errors and flawed adaptations of DP mechanisms, improves the interpretability of privacy guarantees through attack-aware risk measures, and introduces new BDP mechanisms and bounds capable of protecting against correlation-based attacks with competitive utility. By systematically addressing composition, interpretability, and correlation, this work contributes to bridging the gap between DP theory and its reliable application in real-world systems operating on complex data domains.

Limitations and open challenges. While this thesis provides several advances, the application of DP to complex data domains is far from complete. We therefore conclude by highlighting key limitations of our work that naturally open promising directions for future research towards fully realizing the applicability of DP in complex systems.

Regarding composition, we extend our framework to GDP. However, as discussed in Chapter 5, not all mechanisms admit a meaningful characterization of attack resilience within the GDP framework. This limitation motivates extending our composition results to the more general framework of f-DP. Such an extension is non-trivial, as the group privacy property in f-DP fundamentally relies on composing a trade-off function a natural number of times, corresponding to discrete group sizes. In contrast, metric privacy and complex data settings require reasoning over real-valued distances. Bridging this gap would require extending group privacy to continuous metrics, and tools from functional analysis—such as fractional function iteration via Schröder's or Abel's equations—appear to offer a promising mathematical foundation for generalizing the metric privacy framework beyond GDP to arbitrary f-DP.

Importantly, the composability of our universally tight bound for RAD analysis (Theorem 5.3) remains an open question. Since composability is crucial for the adoption of DP, this represents a major limitation that must be further investigated. Note that the solution we currently propose for composition is to use upper bounds, which comes at the cost of losing tightness. Moreover, while the RAD framework and the associated bounds substantially improve the interpretability of DP guarantees and their relation to concrete attacks, our results still rely on the assumption of independent data. This assumption is often violated in complex data domains, where correlations are intrinsic and frequently exploited by adversaries. Extending the RAD framework to BDP and to correlation-aware attack models therefore represents a particularly important and challenging direction for future work.

Finally, the BDP results presented in this thesis rely on specific distributional assumptions. While we provide real-world datasets for both the Gaussian and Markov cases, these assumptions can be restrictive and may limit the applicability of our results to general data analysis tasks. The key contribution of this work is to demonstrate that

there are real scenarios where it is possible to achieve BDP guarantees while retaining utility. However, this is only a first step and does not solve the broader problem of privacy under correlated data.

Moreover, our Markov bound offers a significant noise reduction compared to the general bound, but it has a lower limit on the minimum BDP leakage it can provide, which is a major limitation—especially for distributions where this minimum leakage is nearly ten, as observed in the Activity database. Improving this bound to allow arbitrarily low leakage is a promising direction for future work. Particularly, all our BDP mechanisms are obtained by re-calibrating the noise of classical DP mechanisms to account for correlations. Although this approach enables the reuse of well-understood mechanisms, it remains an open question whether privacy mechanisms designed specifically for BDP—rather than derived from DP mechanisms—could achieve superior utility. Exploring the possible design of ad hoc BDP mechanisms instead of calibrating existing DP ones, as well as extending our correlation-specific bounds to additional data distributions and dependency structures, constitutes a natural next step towards robust and practical privacy guarantees for complex data.

Bibliography

[1] D. Blazquez and J. Domenech, "Big data sources and methods for social and economic analyses," *Technological Forecasting and Social Change*, vol. 130, pp. 99–113, 2018. DOI: 10.1016/j.techfore.2017.07.027.

[2] X. Kong, M. Li, K. Ma, K. Tian, M. Wang, Z. Ning, and F. Xia, "Big trajectory data: A survey of applications and services," *IEEE Access*, vol. 6, pp. 58 295–58 306, 2018. DOI: 10.1109/ACCESS.2018.2873779.

[3] S. V. G. Subrahmanya, D. K. Shetty, V. Patil, B. Z. Hameed, R. Paul, K. Smriti, N. Naik, and B. K. Somani, "The role of data science in healthcare advancements: Applications, benefits, and future prospects," *Irish Journal of Medical Science*, vol. 191, no. 4, pp. 1473–1483, 2022. DOI: 10.1007/s11845-021-02730-z.

[4] A. J. Blumberg and P. Eckersley, "On locational privacy, and how to avoid losing it forever," *Electronic Frontier Foundation*, 2009. [Online]. Available: https://www.eff.org/files/eff-locational-privacy.pdf.

[5] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Res.*, vol. 2, no. 2, pp. 59–64, 2015. DOI: 10.1016/j.bdr.2015.01.006.

[6] B. Tarnoff, *Big data for the people: It's time to take it back from our tech overlords*, Accessed: 10 Feb 2026, 2018. [Online]. Available: https://www.theguardian.com/technology/2018/mar/14/tech-big-data-capitalism-give-wealth-back-to-people.

[7] S. Ovide, "Just collect less data, period.," *The New York Times*, 2020, accessed on 2021-01-18. [Online]. Available: https://www.nytimes.com/2020/07/15/technology/just-collect-less-data-period.html.

[8] European Parliament and Council of the EU, *Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR) (Text with EEA relevance)*, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[9] L. Sweeney, "Simple demographics often identify people uniquely," *Carnegie Mellon University, Data Privacy*, vol. 671, no. 3, pp. 1–34, 2000. DOI: 10.1184/R1/6625769.

[10] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *IEEE Symposium on Security and Privacy (S&P)*, 2008, pp. 111–125. DOI: 10.1109/SP.2008.33.

[11] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, pp. 536–539, 2015. DOI: 10.1126/science.1256297.

[12] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, "Identification via location-profiling in GSM networks," in *ACM Workshop on Privacy in the Electronic Society (WPES)*, 2008, pp. 23–32. DOI: 10.1145/1456403.1456409.

[13] C. Deusser, S. Passmann, and T. Strufe, "Browsing unicity: On the limits of anonymizing web tracking data," in *IEEE Symposium on Security and Privacy (S&P)*, 2020, pp. 777–790. DOI: 10.1109/SP40000.2020.00018.

[14] J. Todt, S. Hanisch, and T. Strufe, "Fantômas: Understanding face anonymization reversibility," *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, vol. 2024, no. 4, pp. 24–43, 2024. DOI: 10.56553/popets-2024-0105.

[15] C. Dwork, "Differential privacy," in *Automata, Languages and Programming (ICALP)*, vol. 4052, 2006, pp. 1–12. DOI: 10.1007/11787006_1.

[16] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy* (Found. Trends Theor. Comput. Sci.). Now Publishers, 2014. DOI: 10.1561/0400000042.

[17] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf, *Statistical Disclosure Control*. John Wiley & Sons, 2012. DOI: 10.1002/9781118348239.

[18] D. Desfontaines and B. Pejó, "SoK: Differential privacies," *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, vol. 2020, no. 2, pp. 288–313, 2020. DOI: 10.2478/popets-2020-0028.

[19] F. McSherry, "Privacy integrated queries," in *ACM SIGMOD International Conference on Management of Data*, 2009, pp. 19–30. DOI: 10.1145/1559845.1559850.

[20] Differential Privacy Team, Apple, *Learning with privacy at scale*, 2017. [Online]. Available: https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf.

[21] R. Rogers, S. Subramaniam, S. Peng, D. Durfee, S. Lee, S. K. Kancha, S. Sahay, and P. Ahammad, "Linkedin's audience engagements api: A privacy preserving data analytics system at scale," *Journal of Privacy and Confidentiality*, vol. 11, no. 3, 2021. DOI: DOI:10.29012/jpc.782.

[22] U.S. Census Bureau, *Understanding differential privacy*, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/differential-privacy.html, Accessed: 2026-01-25, 2021.

[23] D. Kifer, J. M. Abowd, R. Ashmead, R. Cumings-Menon, P. Leclerc, A. Machanavajjhala, W. Sexton, and P. Zhuravlev, *Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 census*, 2022. arXiv: 2209.03310.

[24] J. M. Abowd, "The U.S. census bureau adopts differential privacy," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2018, p. 2867. DOI: 10.1145/3219819.3226070.

[25] S. Guha, P. Kidwell, R. P. Hafen, and W. S. Cleveland, "Visualization databases for the analysis of large complex datasets," in *Artificial Intelligence and Statistics*, 2009, pp. 193–200. [Online]. Available: https://proceedings.mlr.press/v5/guha09a.html.

[26] D. Desfontaines, *The privacy loss random variable*, Blog post, 2020. [Online]. Available: https://differentialprivacy.org/privacy-loss-rv/.

[27] N. Li, M. Lyu, D. Su, and W. Yang, *Differential Privacy: From Theory to Practice*. Springer International Publishing, 2017. DOI: 10.1007/978-3-031-02350-7.

[28] L. Franceschi-Bicchierai, *Redditor cracks anonymous data trove to pinpoint Muslim cab drivers*, Jan. 28, 2015. [Online]. Available: https://mashable.com/archive/redditor-muslim-cab-drivers.

[29] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *ACM SIGMOD International Conference on Management of Data*, Jun. 12, 2011, pp. 193–204. DOI: 10.1145/1989323.1989345.

[30] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnerable: Differential privacy under dependent tuples," in *Network and Distributed System Security Symposium (NDSS)*, 2016. DOI: 10.14722/ndss.2016.23279.

[31] H. Wang, Z. Xu, S. Jia, Y. Xia, and X. Zhang, "Why current differential privacy schemes are inapplicable for correlated data publishing?" In *The Web Conference (WWW)*, vol. 24, 2021, pp. 1–23. DOI: 10.1007/s11280-020-00825-8.

[32] À. Miranda-Pascual, P. Guerra-Balboa, J. Parra-Arnau, J. Forné, and T. Strufe, "SoK: Differentially private publication of trajectory data," *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, vol. 2023, no. 2, pp. 496–516, 2023. DOI: doi.org/10.56553/popets-2023-0065.

[33] T. Humphries, S. Oya, L. Tulloch, M. Rafuse, I. Goldberg, U. Hengartner, and F. Kerschbaum, "Investigating membership inference attacks under data dependencies," in *IEEE Computer Security Foundations Symposium (CSF)*, 2023, pp. 473–488. DOI: 10.1109/CSF57540.2023.00013.

[34] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, 2003, pp. 556–559. DOI: 10.1145/956863.956972.

[35] N. Almadhoun, E. Ayday, and Ö. Ulusoy, "Differential privacy under dependent tuples—the case of genomic privacy," *Bioinformatics*, vol. 36, pp. 1696–1703, 2020. DOI: 10.1093/bioinformatics/btz837.

[36] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, "Quantifying differential privacy under temporal correlations," in *IEEE International Conference on Data Engineering (ICDE)*, 2017, pp. 821–832. DOI: 10.1109/ICDE.2017.132.

[37] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *ACM SIGMOD International Conference on Management of Data*, 2015, pp. 747–762. DOI: 10.1145/2723372.2747643.

[38] J. Dong, A. Roth, and W. J. Su, *Gaussian differential privacy*, 2019. arXiv: 1905.02383.

[39] B. Balle, G. Cherubin, and J. Hayes, "Reconstructing training data with informed adversaries," in *IEEE Symposium on Security and Privacy (S&P)*, 2022, pp. 1138–1156. DOI: 10.1109/SP46214.2022.9833677.

[40] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2018, pp. 475–489. DOI: 10.1145/3243734.3243818.

[41] H. H. Arcolezi and S. Gambs, "Revealing the true cost of locally differentially private protocols: An auditing perspective," *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, vol. 2024, pp. 123–141, 2024. DOI: 10.56553/popets-2024-0110.

[42] B. Bichsel, S. Steffen, I. Bogunovic, and M. Vechev, "DP-Sniper: Black-box discovery of differential privacy violations using classifiers," in *IEEE Symposium on Security and Privacy (S&P)*, 2021, pp. 391–409. DOI: 10.1109/SP40001.2021.00081.

[43] Y. Lu, M. Magdon-Ismail, Y. Wei, and V. Zikas, "Eureka: A general framework for black-box differential privacy estimators," in *Symposium on Security and Privacy (SP)*, 2024, pp. 913–931. DOI: 10.1109/SP54263.2024.00166.

[44] S. Meiser, *Approximate and probabilistic differential privacy definitions*, Cryptology ePrint Archive, Paper 2018/277, 2018. [Online]. Available: https://eprint.iacr.org/2018/277.

[45] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2014, pp. 1054–1067. DOI: 10.1145/2660267.2660348.

[46] M. Hay, C. Li, G. Miklau, and D. Jensen, "Accurate estimation of the degree distribution of private networks," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, Dec. 2009, pp. 169–178. DOI: 10.1109/ICDM.2009.11.

[47] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, 2013, pp. 82–102. DOI: doi.org/10.1007/978-3-642-39077-7_5.

[48] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to noninteractive database privacy," *Journal of the ACM*, vol. 60, no. 2, pp. 1–25, 2013. DOI: 10.1145/2450142.2450148.

[49] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318. DOI: 10.1145/2976749.2978318.

[50] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 394–403. [Online]. Available: https://proceedings.mlr.press/v80/balle18a.html.

[51] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Adv. Cryptology – Annual Int. Conf. Theory Appl. Cryptogr. Techniques (EUROCRYPT)*, 2006, pp. 486–503. DOI: 10.1007/11761679_29.

[52] B. Fung, k. Wang, R. Chen, and P. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys*, vol. 42, 2010. DOI: 10.1145/1749603.1749605.

[53] F. Jin, W. Hua, M. Francia, P. Chao, M. E. Orlowska, and X. Zhou, "A survey and experimental study on privacy-preserving trajectory data publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 5577–5596, 2023. DOI: 10.1109/TKDE.2022.3174204.

[54] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *IEEE Computer Security Foundations Symposium (CSF)*, 2018, pp. 268–282. DOI: 10.1109/CSF.2018.00027.

[55] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 634–646, 2018. DOI: 10.1145/3243734.3243855.

[56] D. Bernau, G. Eibl, P. W. Grassal, H. Keller, and F. Kerschbaum, "Quantifying identifiability to choose and audit ϵ in differentially private deep learning," *Proceedings of the VLDB Endowment*, vol. 14, no. 13, pp. 3335–3347, 2021. [Online]. Available: https://dl.acm.org/doi/10.14778/3484224.3484231.

[57] P. Guerra-Balboa, A. Sauer, and T. Strufe, "Analysis and measurement of attack resilience of differential privacy," in *ACM Workshop on Privacy in the Electronic Society (WPES)*, 2024, pp. 155–171. DOI: 10.1145/3689943.3695046.

[58] B. Jayaraman and D. Evans, "Are attribute inference attacks just imputation?" In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022, pp. 1569–1582, ISBN: 9781450394505. DOI: 10.1145/3548606.3560663.

[59] Ú. Erlingsson, I. Mironov, A. Raghunathan, and S. Song, *That which we call private*, 2020. arXiv: 1908.03566.

[60] M. Nasr, S. Songi, A. Thakurta, N. Papernot, and N. Carlin, "Adversary instantiation: Lower bounds for differentially private machine learning," in *IEEE Symposium on Security and Privacy (S&P)*, 2021, pp. 866–882. DOI: 10.1109/SP40001.2021.00069.

[61] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys*, vol. 54, no. 11s, 2022. DOI: 10.1145/3523273.

[62] J. Hayes, B. Balle, and S. Mahloujifar, "Bounding training data reconstruction in DP-SGD," in *Conference on Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 78 696–78 722. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/f8928b073ccbec15d35f2a9d39430bfd-Paper-Conference.pdf.

[63] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. Springer, 2005. DOI: 10.1007/0-387-27605-X_6.

[64] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*. New York, USA: Chapman and Hall/CRC, 2005. DOI: 10.1201/9780203492024.

[65] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *Journal of the Royal Statistical Society*, vol. 84, no. 1, pp. 3–37, 2022. DOI: 10.1111/rssb.12454.

[66] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *International Conference on Machine Learning (ICML)*, vol. 37, 2015, pp. 1376–1385. [Online]. Available: https://proceedings.mlr.press/v37/kairouz15.html.

[67] E. Ghazi and I. Issa, "Total variation meets differential privacy," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, pp. 207–220, 2024. DOI: 10.1109/JSAIT.2024.3384083.

[68] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, 3:1–3:36, 2014. DOI: 10.1145/2514689.

[69] D. Chakrabarti, J. Gao, A. Saraf, G. Schoenebeck, and F.-Y. Yu, *Optimal local bayesian differential privacy over markov chains*, 2022. arXiv: 2206.11402.

[70] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–21, 2010. DOI: 10.1126/science.1177170.

[71] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2011, pp. 1082–1090. DOI: 10.1145/2020408.2020579.

[72] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Nature Scientific Reports*, vol. 3, 2013. DOI: 10.1038/srep01376.

[73] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Personal and Ubiquitous Computing*, vol. 18, pp. 163–175, 2012. DOI: 10.1007/s00779-012-0633-z.

[74] E. Buchholz, A. Abuadbba, S. Wang, S. Nepal, and S. S. Kanhere, "Reconstruction attack on differential private trajectory protection mechanisms," in *Proceedings of the 38th Annual Computer Security Applications Conference (ACSAC'22)*, 2022, pp. 279–292. DOI: 10.1145/3564625.3564628.

[75] V. Primault, A. Boutet, S. B. Mokhtar, and L. Brunie, "The long road to computational location privacy: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2772–2793, 2019. DOI: 10.1109/COMST.2018.2873950.

[76] M. Fiore, P. Katsikouli, E. Zavou, M. Cunche, F. Fessant, D. Le Hello, U. Aïvodji, B. Olivier, T. Quertier, and R. Stanica, "Privacy in trajectory micro-data publishing: A survey," *Transactions on Data Privacy*, vol. 3, 2020. [Online]. Available: https://inria.hal.science/hal-02968279.

[77] T. T. Portela, F. Vicenzi, and V. Bogorny, "Trajectory data privacy: Research challenges and opportunities," in *Brazilian Symposium on Geoinformatics (GEOINFO)*, vol. 20, 2019, pp. 99–110. [Online]. Available: http://mtc-m16d.sid.inpe.br/col/sid.inpe.br/mtc-m16d/2019/11.27.13.45/doc/99-110.pdf.

[78] H. Jiang, J. Li, P. Zhao, F. Zeng, Z. Xiao, and A. Iyengar, "Location privacy-preserving mechanisms in location-based services: A comprehensive survey," *ACM Computing Surveys*, vol. 54, no. 1, 2021. DOI: 10.1145/3423165.

[79] E. Buchholz, A. Abuadbba, S. Wang, S. Nepal, and S. S. Kanhere, "Sok: Can trajectory generation combine privacy and utility?" *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, vol. 2024, pp. 75–93, 2024. DOI: doi.org/10.56553/popets-2024-0068.

[80] E. Buchholz, N. Fernandes, D. D. Nguyen, A. Abuadbba, S. Nepal, and S. S. Kanhere, *What is the cost of differential privacy for deep learning-based trajectory generation?* 2025. arXiv: 2506.09312.

[81] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*. American Mathematical Soc., 2017. DOI: 10.1007/s00283-018-9839-x.

[82] R. Mello, V. Bogorny, L. Alvares, L. Santana, C. Ferrero, A. A. Frozza, G. Schreiner, and C. Renso, "MASTER: A multiple aspect view on trajectories," *Transactions in GIS*, vol. 23, no. 4, 2019. DOI: 10.1111/tgis.12526.

[83] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2012, pp. 638–649. DOI: 10.1145/2382196.2382263.

[84] R. Chen, B. C. M. Fung, and B. C. Desai, *Differentially private trajectory data publication*, 2011. arXiv: 1112.2020.

[85] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava, "DPT: Differentially private trajectory synthesis using hierarchical reference systems," *Proceedings of the VLDB Endowment*, vol. 8, no. 11, pp. 1154–1165, 2015. DOI: 10.14778/2809974.2809978.

[86] M. E. Gursoy, L. Liu, S. Truex, and L. Yu, "Differentially private and utility preserving publication of trajectory data," *IEEE Transactions on Mobile Computing*, vol. 18, no. 10, pp. 2315–2329, 2018. DOI: 10.1109/TMC.2018.2874008.

[87] T. Brinkhoff, "A framework for generating network-based moving objects," *GeoInformatica*, vol. 6, no. 2, pp. 153–180, 2002. DOI: 10.1023/A:1015231126594.

[88] M. O'Connell, M. Moreira, and W. Kan, *ECML/PKDD 15: Taxi trajectory prediction (I)*, 2015. [Online]. Available: https://kaggle.com/competitions/pkdd-15-predict-taxi-service-trajectory-i.

[89] Y. Cao and M. Yoshikawa, "Differentially private real-time data release over infinite trajectory streams," in *IEEE International Conference on Mobile Data Management (MDM)*, 2015, pp. 68–73. DOI: 10.1109/MDM.2015.15.

[90] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proceedings of the VLDB Endowment*, Trondheim, Norway, 2005, pp. 901–909. [Online]. Available: https://vldb.org/conf/2005/papers/p901-aggarwal.pdf.

[91] A. Hern, "Fitness tracking app Strava gives away location of secret US army bases," *The Guardian*, Jan. 28, 2018. Accessed: Feb. 23, 2022. [Online]. Available: https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases.

[92] D. Gritten, "Strava app flaw revealed runs of Israeli officials at secret bases," *BBC*, 2022. [Online]. Available: https://www.bbc.com/news/world-middle-east-61879383.

[93] A. Haeberlen, B. C. Pierce, and A. Narayan, "Differential privacy under fire," in *USENIX Security Symposium*, 2011, p. 33. [Online]. Available: https://dl.acm.org/doi/10.5555/2028067.2028100.

[94] J. Trotter, "Public NYC taxicab database lets you see how celebrities tip," *Gawker*, 2014. [Online]. Available: https://www.gawkerarchives.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546.

[95] A. Tockar, *Riding with the stars: Passenger privacy in the NYC taxicab dataset*, 2014. [Online]. Available: https://agkn.wordpress.com/author/atockar/.

[96] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro, "Knock knock, who's there? membership inference on aggregate location data," *Network and Distributed System Security Symposium (NDSS)*, 2018. DOI: 10.14722/ndss.2018.23183.

[97] V. Guan, F. Guépin, A.-M. Cretu, and Y.-A. de Montjoye, "A zero auxiliary knowledge membership inference attack on aggregate location data," *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, vol. 4, pp. 80–101, 2024. DOI: 10.56553/popets-2024-0108.

[98] K. Sui, Y. Zhao, D. Liu, M. Ma, L. Xu, L. Zimu, and D. Pei, "Your trajectory privacy can be breached even if you walk in groups," in *IEEE/ACM International Symposium on Quality of Service (IWQoS)*, 2016, pp. 1–6. DOI: 10.1109/IWQoS.2016.7590444.

[99] C. Zhou, D. Frankowski, P. Ludford Finnerty, S. Shekhar, and L. Terveen, "Discovering personal gazetteers: An interactive clustering approach," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2004, pp. 266–273. DOI: 10.1145/1032222.1032261.

[100] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Show me how you move and I will tell you who you are," in *ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS (SPRINGL)*, vol. 4, 2010, pp. 34–41. DOI: 10.1145/1868470.1868479.

[101] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data," in *The Web Conference (WWW)*, 2017, pp. 1241–1250. DOI: 10.1145/3038912.3052620.

[102] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2003, pp. 31–42. DOI: 10.1145/1066116.1189037.

[103] L. Rossi, J. Walker, and M. Musolesi, "Spatio-temporal techniques for user identification by means of GPS mobility data," *EPJ Data Sci.*, vol. 4, no. 1, pp. 1–16, 2015. DOI: 10.1140/epjds/s13688-015-0049-x.

[104] J. Freudiger, R. Shokri, and J.-P. Hubaux, "Evaluating the privacy risk of location-based services," in *International Conference on Financial Cryptography and Data Security (FC)*, vol. 7035, 2011. DOI: 10.1007/978-3-642-27576-0_3.

[105] B. Zan, Z. Sun, M. Gruteser, and X. Ban, "Linking anonymous location traces through driving characteristics," in *ACM Conference on Data and Application Security and Privacy (CODASPY)*, 2013, pp. 293–300. DOI: 10.1145/2435349.2435391.

[106] S. Escher, M. Sontowski, K. Berling, S. Köpsell, and T. Strufe, "How well can your car be tracked: Analysis of the european C-ITS pseudonym scheme," in *IEEE Vehicular Technology Conference (VTC-Spring)*, 2021, pp. 1–6. DOI: 10.1109/VTC2021-Spring51267.2021.9449078.

[107] OpenStreetMap contributors, *Planet dump retrieved from https://planet.osm.org*, 2017. [Online]. Available: https://www.openstreetmap.org.

[108] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias, "Differentially private event sequences over infinite streams," *Proceedings of the VLDB Endowment*, vol. 7, no. 12, pp. 1155–1166, 2014. DOI: 10.14778/2732977.2732989.

[109] H. Asi, J. Duchi, and O. Javidbakht, "Element level differential privacy: The right granularity of privacy," in *Proceedings of the AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI)*, 2022. [Online]. Available: https://aaai-ppai22.github.io/files/11.pdf.

[110] N. Wang and M. S. Kankanhalli, "Protecting sensitive place visits in privacy-preserving trajectory publishing," *Computers & Security*, vol. 97, no. C, 2020. DOI: 10.1016/j.cose.2020.101949.

[111] F. Deldar and M. Abadi, "A differentially private location generalization approach to guarantee non-uniform privacy in moving objects databases," *Knowledge-Based Systems*, vol. 225, p. 107 084, 2021. DOI: 10.1016/j.knosys.2021.107084.

[112] X. Zhao, D. Pi, and J. Chen, "Novel trajectory privacy-preserving method based on prefix tree using differential privacy," *Knowledge-Based Systems*, vol. 198, p. 105 940, 2020. DOI: 10.1016/j.knosys.2020.105940.

[113] X. Zhao, Y. Dong, and D. Pi, "Novel trajectory data publishing method under differential privacy," *Expert Systems with Applications*, vol. 138, p. 112 791, 2019. DOI: 10.1016/j.eswa.2019.07.008.

[114] S. Yuan, D. Pi, X. Zhao, and M. Xu, "Differential privacy trajectory data protection scheme based on R-tree," *Expert Systems with Applications*, vol. 182, p. 115 215, 2021. DOI: 10.1016/j.eswa.2021.115215.

[115] J. Zhao, J. Mei, S. Matwin, Y. Su, and Y. Yang, "Risk-aware individual trajectory data publishing with differential privacy," *IEEE Access*, vol. 9, pp. 7421–7438, 2020. DOI: 0.1109/ACCESS.2020.3048394.

[116] S. Cai, X. Lyu, X. Li, D. Ban, and T. Zeng, "A trajectory released scheme for the internet of vehicles based on differential privacy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 16 534–16 547, 2022. DOI: 10.1109/TITS.2021.3130978.

[117] J. Hua, Y. Gao, and S. Zhong, "Differentially private publication of general time-serial trajectory data," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2015, pp. 549–557. DOI: 10.1109/INFOCOM.2015.7218422.

[118] S. Chen, A. Fu, J. Shen, S. Yu, H. Wang, and H. Sun, "RNN-DP: A new differential privacy scheme base on recurrent neural network for dynamic trajectory privacy protection," *Journal of Network and Computer Applications*, vol. 168, p. 102 736, 2020. DOI: 10.1016/j.jnca.2020.102736.

[119] M. Li, L. Zhu, Z. Zhang, and R. Xu, "Achieving differential privacy of trajectory data publishing in participatory sensing," *Information Sciences*, vol. 400, pp. 1–13, 2017. DOI: 10.1016/j.ins.2017.03.015.

[120] Q. Han, Z. Xiong, and K. Zhang, "Research on trajectory data releasing method via differential privacy based on spatial partition," *Security and communication networks*, 2018. DOI: 10.1155/2018/4248092.

[121] X. Zhao, D. Pi, and J. Chen, "Novel trajectory privacy-preserving method based on clustering using differential privacy," *Expert Systems with Applications*, vol. 149, p. 113 241, 2020. DOI: 10.1016/j.eswa.2020.113241..

[122] D. Shao, K. Jiang, T. Kister, S. Bressan, and K.-L. Tan, "Publishing trajectory with differential privacy: A priori vs. a posteriori sampling mechanisms," in *International Conference on Database and Expert Systems Applications (DEXA)*, vol. 8055, 2013, pp. 357–365. DOI: doi.org/10.1007/978-3-642-40285-2_31.

[123] B. Liu, S. Xie, H. Wang, Y. Hong, X. Ban, and M. Mohammady, "VTDP: Privately sanitizing fine-grained vehicle trajectory data with boosted utility," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 6, pp. 2643–2657, 2021. DOI: 10.1109/TDSC.2019.2960336.

[124] T. Cunningham, G. Cormode, H. Ferhatosmanoglu, and D. Srivastava, "Real-world trajectory sharing with local differential privacy," *Proceedings of the VLDB Endowment*, vol. 14, no. 11, pp. 2283–2295, 2021. DOI: doi:10.14778/3476249.3476280.

[125] Y. Zheng, H. Fu, X. Xie, W.-Y. Ma, and Q. Li, *Geolife GPS trajectory dataset - user guide*, 2011. [Online]. Available: https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/.

[126] H. Wang, H. Hong, L. Xiong, Z. Qin, and Y. Hong, "L-srr: Local differential privacy for location-based services with staircase randomized response," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022, pp. 2809–2823. DOI: 10.1145/3548606.3560636.

[127] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data – anonymisation groundhog day," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1451–1468. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.

[128] P. Nanayakkara, M. A. Smart, R. Cummings, G. Kaptchuk, and E. M. Redmiles, "What are the chances? explaining the epsilon parameter in differential privacy," in *USENIX Security Symposium*, 2023. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/nanayakkara.

[129] F. Galli, S. Biswas, K. Jung, T. Cucinotta, and C. Palamidessi, "Group privacy for personalized federated learning," in *Proc. Int. Conf. Inform. Syst. Security Priv. (ICISSP)*, pp. 252–263. DOI: 10.5220/0011885000003405.

[130] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003. DOI: 10.1137/S003614450342480.

[131] D. Romanini, S. Lehmann, and M. Kivelä, "Privacy and uniqueness of neighborhoods in social networks," *Scientific reports*, vol. 11, no. 1, p. 20 104, 2021. DOI: 10.1038/s41598-021-94283-5.

[132] D. Garcia, "Leaking privacy and shadow profiles in online social networks," *Science advances*, vol. 3, no. 8, e1701172, 2017. DOI: 10.1126/sciadv.1701172.

[133] J. Lee and C. Clifton, "How much is enough? choosing ε for differential privacy," in *Proceedings of the 14th International Conference on Information Security (ISC)*, 2011, pp. 325–340. DOI: 10.1007/978-3-642-24861-0_22.

[134] M. Jagielski, J. Ullman, and A. Oprea, "Auditing differentially private machine learning: How private is private sgd?" In *Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 22 205–22 216. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf.

[135] F. Tramer, A. Terzis, T. Steinke, S. Song, M. Jagielski, and N. Carlini, *Debugging differential privacy: A case study for privacy auditing*, 2022. arXiv: 2202.12219.

[136] H. H. Arcolezi, S. Gambs, J.-F. Couchot, and C. Palamidessi, "On the risks of collecting multidimensional data under local differential privacy," *Proceedings of the VLDB Endowment*, vol. 16, no. 5, pp. 1126–1139, 2023. DOI: 10.14778/3579075.3579086.

[137] G. Cormode, S. Gade, S. Maddock, and E. Ullah, "Synthetic tabular data: Methods, attacks and defenses," *Proceedings of the VLDB Endowment*, vol. 18, no. 12, pp. 5448–5450, 2025. DOI: 10.14778/3750601.3750692.

[138] C. Carey, T. Dick, A. Epasto, A. Javanmard, J. Karlin, S. Kumar, A. Muñoz Medina, V. Mirrokni, G. H. Nunes, S. Vassilvitskii, et al., "Measuring re-identification risk," in *IEEE International Conference on Mobile Data Management (MDM)*, vol. 1, 2023, pp. 1–26. DOI: 10.1145/3589294.

[139] B. Kulynych, J. F. Gomez, G. Kaissis, F. du Pin Calmon, and C. Troncoso, "Attack-aware noise calibration for differential privacy," in *Conference on Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024, pp. 134 868–134 901. DOI: 10.52202/079017-4286.

[140] M. Bun, D. Desfontaines, C. Dwork, M. Naor, K. Nissim, A. Roth, A. Smith, T. Steinke, J. Ullman, and S. Vadhan, *Statistical inference is not a privacy violation*, DifferentialPrivacy.org, Jun. 2021. [Online]. Available: https://differentialprivacy.org/inference-is-not-a-privacy-violation/.

[141] K. Chatzikokolakis, G. Cherubin, C. Palamidessi, and C. Troncoso, "Bayes security: A not so average metric," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2023, pp. 388–406. DOI: 10.1109/CSF57540.2023.00011.

[142] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro, "What does the crowd say about you? evaluating aggregation-based location privacy," *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, vol. 4, pp. 156–176, 2017. DOI: 10.1515/popets-2017-0043.

[143] B. Jayaraman, "Analyzing the leaky cauldron: Inference attacks on machine learning," Ph.D. dissertation, University of Virginia, Dec. 2022. [Online]. Available: https://libraetd.lib.virginia.edu/public_view/1r66j21378.

[144] M. S. M. S. Annamalai, B. Balle, J. Hayes, G. Kaissis, and E. D. Cristofaro, *The hitchhiker's guide to efficient, end-to-end, and tight dp auditing*, 2025. arXiv: 2506.16666.

[145] M. S. M. S. Annamalai and E. De Cristofaro, "Nearly tight black-box auditing of differentially private machine learning," in *Conference on Neural Information Processing Systems (NeurIPS)*, vol. 37, 2025, pp. 131 482–131 502. DOI: 10.52202/079017-4179.

[146] T. Steinke, M. Nasr, and M. Jagielski, "Privacy auditing with one (1) training run," in *Conference on Neural Information Processing Systems (NeurIPS)*, (New Orleans, LA, USA), 2023.

[147] M. Malek, I. Mironov, K. Prasad, I. Shilov, and F. Tramèr, "Antipodes of label differential privacy: PATE and ALIBI," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. DOI: 10.52202/079017-4179.

[148] F. Houssiau, J. Jordon, S. N. Cohen, O. Daniel, A. Elliott, J. Geddes, C. Mole, C. Rangel-Smith, and L. Szpruch, *Tapas: A toolbox for adversarial privacy auditing of synthetic data*, 2022. arXiv: 2211.06550.

[149] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *International Conference on Machine Learning (ICML)*, 2016, pp. 2436–2444. [Online]. Available: http://proceedings.mlr.press/v48/kairouz16.pdf.

[150] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *USENIX Security Symposium*, Vancouver, BC, Canada, 2017, pp. 729–745. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao.

[151] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under local differential privacy," in *IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 759–763. DOI: 10.1109/ISIT.2017.8006630.

[152] D. Gorla, L. Jalouzot, F. Granese, C. Palamidessi, and P. Piantanida, "On estimating the strength of differentially private mechanisms in a black-box setting," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 5, pp. 5494–5507, 2025. DOI: 10.1109/TDSC.2025.3568160.

[153] S. Mahloujifar, L. Melis, and K. Chaudhuri, *Auditing f-differential privacy in one run*, 2025. [Online]. Available: https://proceedings.mlr.press/v267/mahloujifar25a.html.

[154] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677.

[155] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *USENIX Security Symposium*, 2014, pp. 17–32.

[156] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[157] H. Xiao, K. Rasul, and R. Vollgraf, *Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms*, 2017. arXiv: 1708.07747.

[158] B. Becker and R. Kohavi, *Adult dataset*, UCI Machine Learning Repository, 1996. DOI: 10.24432/C5XW20.

[159] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based k-anonymity," *Proceedings of the VLDB Endowment*, vol. 23, no. 5, pp. 771–794, 2014. DOI: 10.1007/s00778-014-0351-4.

[160] S. Lestyán, G. Ács, and G. Biczók, "In search of lost utility: Private location data," *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS)*, vol. 3, pp. 354–372, 2022. DOI: 10.56553/popets-2022-0076.

[161] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2025, pp. 1298–1309. DOI: 10.1145/2810103.2813640.

[162] M. E. Gursoy, L. Liu, K.-H. Chow, S. Truex, and W. Wei, "An adversarial approach to protocol analysis and selection in local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1785–1799, 2022. DOI: 10.1109/TIFS.2022.3170242.

[163] H. H. Arcolezi, *LDP-Audit GitHub repository*, https://github.com/hharcolezi/ldp-audit, 2024.

[164] T. Sauer, *Numerical Analysis*, 2nd ed. USA: Pearson Education, Inc, 2012, ISBN: 0-321-78367-0. [Online]. Available: https://digitallibrary.srisathyasaicollege.in/bitstream/123456789/6844/1/Sauer%20-%20Numerical%20Analysis%202e.pdf.

[165] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. [Online]. Available: https://www.tensorflow.org/.

[166] B. Jayaraman, *EvaluatingDPML Github repository*, 2022. [Online]. Available: https://github.com/bargavj/EvaluatingDPML.

[167] X. He, A. Machanavajjhala, and B. Ding, "Blowfish privacy: Tuning privacy-utility trade-offs using policies," in *ACM SIGMOD International Conference on Management of Data*, 2014, pp. 1447–1458. DOI: 10.1145/2588555.2588581.

[168] Y. Li, X. Ren, S. Yang, and X. Yang, "Impact of prior knowledge and data correlation on privacy leakage: A unified analysis," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2342–2357, 2019. DOI: 10.1109/TIFS.2019.2895970.

[169] K. M. Chong and A. Malip, "May the privacy be with us: Correlated differential privacy in location data for ITS," *Computer Networks*, vol. 241, p. 110 214, 2024. DOI: 10.1016/j.comnet.2024.110214.

[170] J. Brainard and D. E. Burmaster, "Bivariate distributions for height and weight of men and women in the United States," *Risk Analysis*, vol. 12, no. 2, pp. 267–275, 1992. DOI: 10.1111/j.1539-6924.1992.tb00674.x.

[171] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-markov model," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 838–845. DOI: 10.1109/CVPR.2005.61.

[172] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish Privacy Mechanisms for Correlated Data," in *ACM SIGMOD International Conference on Management of Data*, New York, USA, 2017, pp. 1291–1306. DOI: 10.1145/3035918.3064025.

[173] T. Nuradha and Z. Goldfeld, "Pufferfish privacy: An information-theoretic study," *IEEE Transactions on Information Theory*, vol. 69, no. 11, pp. 7336–7356, 2023. DOI: 10.1109/TIT.2023.3296288. [Online]. Available: https://ieeexplore.ieee.org/document/10185108/.

[174] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 43–54. DOI: 10.1145/2976749.2978308.

[175] M. Sunnåker and J. Stelling, "Model extension and model selection," in *Uncertainty in Biology: A Computational Modeling Approach*, vol. 17, 2015, pp. 213–241. DOI: 10.1007/978-3-319-21296-8_9.

[176] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, 2006, pp. 265–284. DOI: 10.1007/11681878_14.

[177] R. Chen, B. C. M. Fung, P. S. Yu, and B. C. Desai, "Correlated network data publication via differential privacy," *Proceedings of the VLDB Endowment*, vol. 23, no. 4, pp. 653–676, 2014. DOI: 10.1007/s00778-013-0344-8.

[178] V. M. Panaretos, *Statistics for Mathematicians*. Switzerland: Springer International Publishing, 2016. DOI: 10.1007/978-3-319-28341-8.

[179] J. Shao, *Mathematical Statistics*. Springer, 2003. DOI: 10.1007/b97553.

[180] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2013, pp. 901–914. DOI: 10.1145/2508859.2516735.

[181] K. B. Petersen, M. S. Pedersen, et al., "The matrix cookbook," *Technical University of Denmark*, 2012. [Online]. Available: http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html.

[182] J. Shurman, *Calculus and Analysis in Euclidean Space*. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-49314-5.

[183] S. A. Gershgorin, "Über die abgrenzung der eigenwerte einer matrix," *Izvestija Rossijskoj akademii nauk. Serija matematičeskaja*, vol. 1, no. 6, pp. 749–754, 1931.

[184] T. Hawkins, "Cauchy and the spectral theory of matrices," *Historia Mathematica*, vol. 2, no. 1, pp. 1–29, 1975. DOI: 10.1016/0315-0860(75)90032-4.

[185] J. A. Trop, "Topics in sparse approximation," PhD Thesis, 2004. [Online]. Available: http://hdl.handle.net/2152/1272.

[186] E. Behrends, *Introduction to Markov Chains*. Wiesbaden, Germany: Vieweg+Teubner Verlag, 2000. DOI: 10.1007/978-3-322-90157-6.

[187] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences (Fourth Edition)*. Elsevier, 2020. DOI: 10.1016/C2017-0-03921-6.

[188] O. Ardakanian, S. Keshav, and C. Rosenberg, "Markovian models for home electricity consumption," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Green Networking*, 2011, pp. 31–36. DOI: 10.1145/2018536.2018544.

[189] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, Mar. 1965. DOI: 10.1080/01621459.1965.10480775.

[190] F. Galton, *Galton height data*, Harvard Dataverse, 2017. DOI: 10.7910/DVN/T0HSJ1.

[191] F. A. Graybill and H. K. Iyer, 1994. [Online]. Available: https://www.kaggle.com/datasets/jacopoferretti/child-vs-mother-iq/data?select=gifted.csv.

[192] Z. C. Luo, K. Albertsson-Wikland, and J. Karlberg, "Target height as predicted by parental heights in a population-based study," *Pediatric Research*, vol. 44(4), pp. 563–571, 1998. DOI: 10.1203/00006450-199810000-00016.

[193] R. Plomin, J. C. DeFries, V. S. Knopik, and J. M. Neiderhiser, *Behavioral Genetics: A Primer*, Sixth edition. New York: Worth Publishers, 2013, 503 pp., ISBN: 978-1-4292-4215-8.

[194] Q. Huang, D. Cohen, S. Komarzynski, X.-M. Li, P. Innominato, F. Lévi, and B. Finkenstädt, "Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data," *Journal of The Royal Society Interface*, vol. 15, no. 139, p. 20 170 885, 2018. DOI: 10.1098/rsif.2017.0885.

[195] J. Munkhammar, D. van der Meer, and J. Widén, "Very short term load forecasting of residential electricity consumption using the Markov-chain mixture distribution (MCM) model," *Applied Energy*, vol. 282, p. 116 180, 2021. DOI: 10.1016/j.apenergy.2020.116180.

[196] H. Dalkani, M. Mojarad, and H. Arfaeinia, "Modelling electricity consumption forecasting using the Markov process and hybrid features selection," *International Journal of Intelligent Systems and Applications*, vol. 10, no. 5, p. 14, 2021. DOI: 10.5815/ijisa.2021.05.02.

[197] S. Malik, *Activity Data*, 2020. [Online]. Available: https://www.kaggle.com/datasets/shambhavimalik/activity-data/data.

[198] S. Makonin, B. Ellert, I. V. Bajić, and F. Popowich, "Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014," *Scientific data*, vol. 3, no. 1, pp. 1–12, 2016. DOI: 10.1038/sdata.2016.37.

[199] W.-K. Ching and M. K. Ng, *Markov Chains: Models, Algorithms and Applications*. Boston, MA, USA: Springer, 2006. DOI: 10.1007/0-387-29337-X_7.

[200] D. K. Lee, "Alternatives to P value: Confidence interval and effect size," *Korean Journal of Anesthesiology*, vol. 69, no. 6, pp. 555–562, 2016. DOI: 10.4097/kjae.2016.69.6.555.

[201] J. Near and D. Darais, *Differential privacy: Future work & open challenges*, 2022. [Online]. Available: https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-future-work-open-challenges.

[202] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *IEEE Symposium on Security and Privacy (S&P)*, 2022, pp. 1897–1914. DOI: 10.1109/SP46214.2022.9833649.

[203] D. Burago, Y. Burago, and S. Ivanov, *A Course in Metric Geometry*. American Mathematical Society, 2001, vol. 33. DOI: 10.1090/gsm/033.

[204] K. Siegrist, *Probability, Statistics, and Stochastic Processes*. John Wiley & Sons, Ltd, 2012, ISBN: 9781118231296. DOI: 10.1002/9781118231296.

A. Additional Proofs and Remarks

In this chapter, we provide the omitted lemmas, proofs, and additional details needed to complete the results presented in this thesis.

A.1. Additional Details for Chapter 2

We begin elaborating on metric privacy properties and its relation with granularities.

Proposition A.1. *Let $\mathbb{D}$ be a database class and $\mathcal{G}$ a granularity notion over $\mathbb{D}$. Then the canonical metric $d_{\mathbb{D}}^{\mathcal{G}}$ is a well-defined extended metric.*

Proof. The canonical metric $d_{\mathbb{D}}^{\mathcal{G}}\colon \mathbb{D}^2 \to [0,\infty]$ is defined as the minimum number of neighboring-changes in $\mathbb{D}$ one needs to perform to obtain D from D' (with $d_{\mathbb{D}}^{\mathcal{G}}(D,D') = \infty$ if it is not possible). More formally, we define a *relational chain between elements* $D, D' \in \mathbb{D}$ as an ordered finite sequence of $D_i \in \mathbb{D}$ such that $D_0 \sim_{\mathcal{G}} D_1 \sim_{\mathcal{G}} \cdots \sim_{\mathcal{G}} D_n$ with $D = D_0$ and $D' = D_n$, and define $d_{\mathbb{D}}^{\mathcal{G}}(D,D')$ as the minimum length of any relation chain connecting D and D' (with $d_{\mathbb{D}}^{\mathcal{G}}(D,D') = \infty$ if no chain exists).

We need to prove that $d_{\mathbb{D}}^{\mathcal{G}}$ is a well-defined extended metric. By construction, the image of $d_{\mathbb{D}}^{\mathcal{G}}$ is $[0,\infty]$, and $d_{\mathbb{D}}^{\mathcal{G}}(D,D') = 0$ if and only if $D = D'$. Symmetry also follows from the fact that $\sim_{\mathcal{G}}$ is a symmetric relation, i.e., any chain from D to D' can also be seen as a chain from D' to D.

Finally, concatenating the chains gives us the triangle inequality. Let $D, D', D'' \in \mathbb{D}$ such that $d_{\mathbb{D}}^{\mathcal{G}}(D,D') = m$ and $d_{\mathbb{D}}^{\mathcal{G}}(D',D'') = n$. The triangle inequality holds if $n = \infty$ or $m = \infty$, so suppose $n, m < \infty$. Then, by definition, there exists a relational chain of length n connecting D and D', and a relational chain of length m connecting D' and D''. Joining the chains at D' gives us a relational chain of length $n+m$. By definition of $d_{\mathbb{D}}^{\mathcal{G}}$, we obtain the triangle inequality $d_{\mathbb{D}}^{\mathcal{G}}(D,D'') \leq n + m = d_{\mathbb{D}}^{\mathcal{G}}(D,D') + d_{\mathbb{D}}^{\mathcal{G}}(D',D'')$.

In conclusion, $d_{\mathbb{D}}^{\mathcal{G}}$ is a extended metric and $(\mathbb{D}, d_{\mathbb{D}}^{\mathcal{G}})$ is a metric space. □

Proposition A.2. *Let $\mathcal{G}$ be a granularity notion over the database class $\mathbb{D}$. Then, a mechanism $\mathcal{M}$ with domain $\mathbb{D}$ is $\varepsilon d_{\mathbb{D}}^{\mathcal{G}}$-private if and only if it is $\mathcal{G}$ ε-DP.*

Proof. First, we see that $\varepsilon d_{\mathbb{D}}^{\mathcal{G}}$-privacy implies $\mathcal{G}$ ε-DP. Suppose that $\mathcal{M}\colon \mathbb{D} \to \mathbb{D}(\Theta)$ is $\varepsilon d_{\mathbb{D}}^{\mathcal{G}}$-private. Then, for any $\mathcal{G}$-neighboring databases $D, D' \in \mathbb{D}$ and any measurable $S \subseteq \Theta$, we have that

$$\Pr[\mathcal{M}(D) \in S] \leq \mathrm{e}^{\varepsilon d_{\mathbb{D}}^{\mathcal{G}}(D,D')} \Pr[\mathcal{M}(D') \in S].$$

By construction of the canonical metric, $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = 1$ since D and D' are $\mathcal{G}$-neighboring, and therefore $\mathcal{M}$ is $\mathcal{G}$ ε-DP.

Now we prove the other implication. Suppose $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$ is $\mathcal{G}$ ε-DP. We want to see that, for all $D, D' \in \mathbb{D}$ and all measurable $S \subseteq \mathcal{S}$,

$$\Pr[\mathcal{M}(D) \in S] \leq \mathrm{e}^{\varepsilon d_{\mathbb{D}}^{\mathcal{G}}(D,D')} \Pr[\mathcal{M}(D') \in S].$$

The result clearly holds if $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = \infty$, so suppose $d_{\mathbb{D}}^{\mathcal{G}}(D, D') = n < \infty$. Since the distance is finite, there exists $D_0, \ldots, D_n \in \mathbb{D}$, such that $D = D_0$, $D' = D_n$ and

$$D_0 \sim_{\mathcal{G}} D_1 \sim_{\mathcal{G}} \cdots \sim_{\mathcal{G}} D_{n-1} \sim_{\mathcal{G}} D_n.$$

Since D_{i-1} and D_i are $\mathcal{G}$-neighboring, for all measurable $S \subseteq \mathcal{S}$ and $i \in [n]$ we have that

$$\Pr[\mathcal{M}(D_{i-1}) \in S] \leq \mathrm{e}^{\varepsilon} \Pr[\mathcal{M}(D_i) \in S],$$

and, by applying the inequalities in order, we obtain

$$\begin{aligned}
\Pr[\mathcal{M}(D) \in S\} &\leq \mathrm{e}^{\varepsilon} \Pr\{\mathcal{M}(D_1) \in S] \\
&\leq \mathrm{e}^{2\varepsilon} \Pr\{\mathcal{M}(D_2) \in S] \\
&\leq \cdots \\
&\leq \mathrm{e}^{n\varepsilon} \Pr\{\mathcal{M}(D') \in S] \\
&= \mathrm{e}^{\varepsilon d_{\mathbb{D}}^{\mathcal{G}}(D,D')} \Pr\{\mathcal{M}(D') \in S].
\end{aligned}$$

In conclusion, $\mathcal{M}$ is $\varepsilon d_{\mathbb{D}}^{\mathcal{G}}$-private. □

Furthermore, generalizing sensitivity to metric spaces allows us to quantify how preprocessing functions impact privacy guarantees:

Proposition A.3 (Preprocessing [47])**.** *Let $(\mathbb{D}_1, d_1)$ and $(\mathbb{D}_2, d_2)$ be two privacy spaces and let f be a deterministic map with sensitivity $\Delta f < \infty$ with respect to d_1 and d_2, and let $\mathcal{M}\colon \mathbb{D}_2 \to \mathcal{D}(\Theta)$ be a d_2-private mechanism. Then $\mathcal{M} \circ f$ satisfies $(\Delta f) d_1$-privacy.*

Remark A.4. The reciprocal of Proposition A.3 is not true. For example, consider ε-LDP over with domain $\mathbb{R}$, i.e., $(\mathbb{R}, d^{\mathcal{L}})$, such that

$$d^{\mathcal{L}} = \begin{cases} \varepsilon & \text{if } x \neq y \\ 0 & \text{otherwise.} \end{cases}$$

Take $\mathcal{M}\colon \mathbb{R} \to \mathcal{D}(\mathbb{R})$ such as $\mathcal{M}(x) = x + Z$ where $Z \sim \mathrm{Lap}(\frac{1}{\varepsilon})$. We can easily verify that this mechanism is not ε-LDP by selecting two numbers $|x - y| > 1$. Moreover, the sensitivity of the identity map over the real numbers is infinite, so there is not $\varepsilon > 0$ such that $\mathcal{M}$ is ε-LDP.

However, if we take $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x) = \frac{1}{1+\mathrm{e}^x}$, then $\Delta f = |f(x) - f(y)| \leq 1$. Therefore, $\mathcal{M} \circ f$ corresponds to a local Laplace mechanism, and is ε-LDP.

In conclusion, there exist $\mathcal{M}$ and f such that $\mathcal{M} \circ f$ is $(\varepsilon \Delta f) d_{\mathbb{R}}^{\mathcal{L}}$-private but $\mathcal{M}$ is not $\varepsilon' d_{\mathbb{R}}^{\mathcal{L}}$-private for any $\varepsilon' > 0$.

The privacy implications of the data domain. To understand the real privacy implications of a $d_{\mathbb{D}}$-privacy, we need to look not only at the distance but also at the domain of definition.

The domain, $\mathbb{D}$, encodes what information we consider public knowledge and what we want to protect up to $d_{\mathbb{D}}$. The larger the domain, the greater the privacy, but it also comes with the cost of greater sensitivities and harder-to-achieve privacy protection. The distance, $d_{\mathbb{D}}$, encodes how hard it is to distinguish any pair of databases, and therefore what information we are protecting.

Additionally, it is important to select domains with compatible metrics. For example, information is completely unprotected if $d_{\mathbb{D}}(D, D') = \infty$. Therefore, *connected* privacy spaces (i.e., $d_{\mathbb{D}}(D, D') < \infty$ for all $D, D' \in \mathbb{D}$) are preferable because the change across connected components is not guaranteed to be protected. For example, when $\mathbb{D}$ is totally disconnected, we can end up with nonsensical privacy guarantees like in the following example.

Example A.5. Consider $\mathcal{X}^n \coloneqq \{D \in \mathbb{D}_{\mathcal{X}} \mid |D| = n\}$, the class of databases of size n with elements drawn from $\mathcal{X}$, and choose the unbounded granularity notion. It is clear that unbounded-neighboring databases always differ by one element. Therefore, there are no unbounded-neighboring databases in $\mathcal{X}^n$ (i.e., the privacy space is totally disconnected).

This privacy space would imply, by *reductio ad absurdum*, that *any* mechanism is unbounded ε-DP for all $\varepsilon \geq 0$ since for all the neighbors (none) the definition holds. In particular, the identity mechanism (such that $\mathcal{M}(D) = D$) defined over $\mathcal{X}^n$ (which does not provide any protection) is unbounded 0-DP.

Note that choosing $\mathbb{D}_{\mathcal{X}}$ as the domain does not lead to the same problem, but as we mentioned before, relaxing the domain so that it is defined for subsets $\mathbb{D} \subset \mathbb{D}_{\mathcal{X}}$ and other database types is usually more convenient, coherent, and necessary. Following the same line, the bounded granularity defines a connected privacy space over $\mathcal{X}^n$, but defines a disconnected one over $\mathbb{D}_{\mathcal{X}}$.

Besides, note that the restriction of $d^{\mathcal{G}}_{\mathbb{D}}$ to the subclass $\mathbb{D}' \subseteq \mathbb{D}$ is not always $d^{\mathcal{G}}_{\mathbb{D}'}$ as we precise in the following remark:

Remark A.6. The *induced metric* of $d\colon \mathbb{D}^2 \to [0, \infty]$ to a subclass $\mathbb{D}' \subseteq \mathbb{D}$ is defined as the metric $d|_{\mathbb{D}'}$ such that $d|_{\mathbb{D}'}(D, D') = d(D, D')$ for all $D, D' \in \mathbb{D}'$.

Note that the induced metric of $d^{\mathcal{G}}_{\mathbb{D}}$ to the subclass $\mathbb{D}' \subseteq \mathbb{D}$ is not $d^{\mathcal{G}}_{\mathbb{D}'}$. Mathematically speaking, the $d^{\mathcal{G}}_{\mathbb{D}}$ is a *intrinsic metric* [203], i.e., defined as the infimum of the lengths of all paths from the first database to the second. However, the induced metric to $\mathbb{D}'$ is not necessarily the intrinsic metric over $\mathbb{D}'$ [203]. Therefore, the distance between two databases in $\mathbb{D}' \subseteq \mathbb{D}$ can be different over $\mathbb{D}'$ and $\mathbb{D}$.

As an example, the privacy space $(\mathbb{D}_{\mathcal{X}}, d^{\mathcal{U}}_{\mathbb{D}_{\mathcal{X}}})$ with the unbounded metric $d^{\mathcal{U}}_{\mathbb{D}_{\mathcal{X}}}(D, D') = |D \triangle D'|$. However, note that $d^{\mathcal{U}}_{\mathbb{D}}(D, D') \neq |D \triangle D'|$ in general for $\mathbb{D} \subseteq \mathbb{D}_{\mathcal{X}}$, e.g., in the class of databases of size n, $\mathcal{X}^n$. Therefore, there exist $\mathbb{D} \subseteq \mathbb{D}_{\mathcal{X}}$ such that $d^{\mathcal{U}}_{\mathbb{D}} \neq d^{\triangle}_{\mathbb{D}}$, even though $d^{\mathcal{U}}_{\mathbb{D}_{\mathcal{X}}} = d^{\triangle}_{\mathbb{D}_{\mathcal{X}}}$.

Relationship between Metrics. The notion of $d_{\mathbb{D}}$-privacy allows us to compare the privacy level between metrics over the same domain, which also helps to extend composability notions proven for one to others. Consider two metrics, d_1 and d_2, over $\mathbb{D}$ such that $d_1 \leq d_2$ (pointwise). In this case, we can say that d_1 offers more protection than d_2 because any mechanism $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$ that satisfies d_1-privacy also satisfies d_2-privacy [47]. This can be extended to compared the privacy offered by different granularities:

Proposition A.7 (Relation between granularities)**.** *Let $d_{\mathbb{D}}^{\mathcal{G}_1}$ and $d_{\mathbb{D}}^{\mathcal{G}_2}$ be two canonical metrics of granularities $\mathcal{G}_1$ and $\mathcal{G}_2$, such that*

$$k = d_{\mathbb{D}}(\mathcal{G}_1, \mathcal{G}_2) \coloneqq \max_{\substack{D,D' \in \mathbb{D} \\ D \sim_{\mathcal{G}_2} D'}} d_{\mathbb{D}}^{\mathcal{G}_1}(D, D') < \infty.$$

Then, $d_{\mathbb{D}}^{\mathcal{G}_1} \leq k d_{\mathbb{D}}^{\mathcal{G}_2}$.

Proof. We need to see that $d_{\mathbb{D}}^{\mathcal{G}_1}(D, D') \leq k d_{\mathbb{D}}^{\mathcal{G}_2}(D, D')$ for all $D, D' \in \mathbb{D}$. If $d_{\mathbb{D}}^{\mathcal{G}_2}(D, D') = \infty$, then the result holds, so we consider $d_{\mathbb{D}}^{\mathcal{G}_2}(D, D') = n < \infty$.

Since the distance is finite, there exists $D_0, \ldots, D_n \in \mathbb{D}$, such that $D = D_0$, $D' = D_n$ and

$$D_0 \sim_{\mathcal{G}_2} D_1 \sim_{\mathcal{G}_2} \cdots \sim_{\mathcal{G}_2} D_{n-1} \sim_{\mathcal{G}_2} D_n.$$

Since D_{i-1} and D_i are $\mathcal{G}_2$-neighboring, $d_{\mathbb{D}}^{\mathcal{G}_1}(D_{i-1}, D_i) \leq \mathrm{dist}_{\mathbb{D}}(\mathcal{G}_1, \mathcal{G}_2) = k$. Therefore, applying the triangle inequality with $d_{\mathbb{D}}^{\mathcal{G}_1}$ over the chain, we obtain

$$d_{\mathbb{D}}^{\mathcal{G}_1}(D, D') \leq \sum_{i=1}^{n} d_{\mathbb{D}}^{\mathcal{G}_1}(D_{i-1}, D_i) \leq \sum_{i=1}^{n} k = kn = k d_{\mathbb{D}}^{\mathcal{G}_2}(D, D'). \qquad \square$$

Therefore, if $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$ is $\mathcal{G}_1$ ε-DP, then $\mathcal{M}$ is $\mathcal{G}_2$ $k\varepsilon$-DP. This fact allows us to compare different granularity notions over the same domain, e.g., all information protected by $\mathcal{G}_1$ must also be protected by $\mathcal{G}_2$, while not necessarily the other way around.

From this result, we can deduce the well-known fact that unbounded ε-DP implies bounded 2ε-DP in $\mathbb{D}_{\mathcal{X}}$ [27] since $d_{\mathbb{D}_{\mathcal{X}}}(\mathcal{U}, \mathcal{B}) = 2$. However, $d_{\mathbb{D}_{\mathcal{X}}}(\mathcal{B}, \mathcal{U}) = \infty$ because $d_{\mathbb{D}_{\mathcal{X}}}^{\mathcal{B}}(D, D') = \infty$ for all $D \sim_{\mathcal{U}} D'$. Once again we can not forget the data domain: the privacy-level comparison between two granularity notions directly depends on which class we compare them in. While this result holds in $\mathbb{D}_{\mathcal{X}}$, we saw in Example A.5 that this is not the case for all database classes. Another example, the *free-lunch* granularity notion $\mathcal{FL}$ [29] is defined such that all pairs of databases are free-lunch neighboring, i.e., $d_{\mathbb{D}}^{\mathcal{FL}}(D, D') = 1$ for all $D \neq D'$. Therefore, $d_{\mathbb{D}}^{\mathcal{FL}} \leq d_{\mathbb{D}}^{\mathcal{G}}$ verifies for any canonical metric $d_{\mathbb{D}}^{\mathcal{G}}$, and thus free-lunch DP implies all others.

We conclude that metric privacy offers a consistent generalization of pure DP that encodes any possible granularity definition. This motivates our use of metric privacy as unifying privacy framework for composition in arbitrary data domains in Chapter 4.

Utility of DP mechanisms. Finally, we conclude with a self-contained proof of the accuracy of the Gaussian mechanism.

Proposition A.8 ((α, β)-accuracy of the Gaussian mechanism)**.** *Let $f : \mathbb{D} \to \mathbb{R}$ be a real-valued query and consider the Gaussian mechanism*

$$\mathcal{M}(D) = f(D) + Z, \qquad Z \sim \mathcal{N}(0, \sigma^2).$$

Then, for any $\beta \in (0, 1)$, the mechanism is (α, β)-accurate with

$$\alpha = \sigma \sqrt{2 \ln\left(\frac{2}{\beta}\right)}.$$

Proof. By definition, the mechanism is (α, β)-accurate if

$$\Pr\left[\, |\mathcal{M}(D) - f(D)| > \alpha \,\right] \leq \beta.$$

Since $\mathcal{M}(D) - f(D) = Z$, it suffices to bound

$$\Pr\left[\, |Z| > \alpha \,\right], \qquad Z \sim \mathcal{N}(0, \sigma^2).$$

Let $X = Z/\sigma \sim \mathcal{N}(0, 1)$. Then

$$\Pr\left[\, |Z| > \alpha \,\right] = \Pr\left[\, |X| > \alpha/\sigma \,\right] = 2 \Pr\left[\, X > \alpha/\sigma \,\right].$$

Using the standard Gaussian tail bound

$$\Pr[X > t] \leq \exp\left(-\frac{t^2}{2}\right) \quad \text{for all } t > 0,$$

we obtain

$$\Pr\left[\, |Z| > \alpha \,\right] \leq 2 \exp\left(-\frac{\alpha^2}{2\sigma^2}\right).$$

Setting the right-hand side equal to β and solving for α yields

$$\alpha = \sigma \sqrt{2 \ln\left(\frac{2}{\beta}\right)}.$$

Therefore,

$$\Pr\left[\, |\mathcal{M}(D) - f(D)| \leq \alpha \,\right] \geq 1 - \beta,$$

which proves that the Gaussian mechanism is (α, β)-accurate. □

A.2. Additional Details for Chapter 4

For the proofs on the adaptive composition, we need some basic probability results [204], which we will recompile in the following remark.

Remark A.9. Let $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$. As we mentioned earlier, $\mathcal{M}(D)$ for all $D \in \mathbb{D}$ are random elements (e.g., random variables, continuous or discrete; random vectors; random matrices). Note that for every $\mathcal{M}(D)$ and measurable set $S \subseteq \Theta$, $P_{\mathcal{M}(D)}(S) = \Pr[\mathcal{M}(D) \in S]$ defines a measure. This can also be defined with an integral, i.e.,

$$P_{\mathcal{M}(D)}(S) = \Pr[\mathcal{M}(D) \in S] = \int_S \mathrm{d}P_{\mathcal{M}(D)},$$

known as the *Lebesgue–Stieltjes integral*. It can be evaluated over any integrable (in the Lebesgue–Stieltjes sense) function $g\colon \mathcal{S} \to \mathbb{R}$ as $\int_S g \,\mathrm{d}P_{\mathcal{M}(D)}$. This is also denoted as $\int_S g(s) \,\mathrm{d}P_{\mathcal{M}(D)}(s)$, or as $\int_S g \,\mathrm{d}F_{\mathcal{M}(D)} = \int_S g \,\mathrm{d}F_{\mathcal{M}(D)}(s)$ with $F_{\mathcal{M}(D)}$ the distribution function of $\mathcal{M}(D)$. We will use the Lebesgue–Stieltjes integral because it allows us to generalize our results to any random element, such as discrete, continuous, and mixed random variables or random vectors. Specifically, the integral can be written as

$$\int_S g \,\mathrm{d}P_{\mathcal{M}(D)} = \sum_{s \in S} g(s) \Pr[\mathcal{M}(D) = s]$$

if $\mathcal{M}(D)$ is a discrete random variable, and as

$$\int_S g \,\mathrm{d}P_{\mathcal{M}(D)} = \int_S g(s)\, f_{\mathcal{M}(D)}(s) \,\mathrm{d}s$$

if $\mathcal{M}(D)$ is a continuous random variable with density function $f_{\mathcal{M}(D)}$.

Some of the well-known properties of the integrals that we will use in the proofs are linearity: for any integrable functions $f, g\colon \Theta \to \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$,

$$\int_S (\alpha f + \beta g) \,\mathrm{d}P_{\mathcal{M}(D)} = \alpha \int_S f \,\mathrm{d}P_{\mathcal{M}(D)} + \beta \int_S g \,\mathrm{d}P_{\mathcal{M}(D)},$$

and order: for any integrable functions $f, g\colon \mathcal{S} \to \mathbb{R}$ such that $f \leq g$,

$$\int_S f \,\mathrm{d}P_{\mathcal{M}(D)} \leq \int_S g \,\mathrm{d}P_{\mathcal{M}(D)}.$$

Additionally, from the probability properties, we have that

$$\int_\Theta \mathrm{d}P_{\mathcal{M}(D)} = \Pr[\mathcal{M}(D) \in \Theta] = 1,$$

and the *law of total probability*: for any event A,

$$\Pr[A] = \int_\Theta \Pr[A \mid \mathcal{M}(D) = s] \,\mathrm{d}P_{\mathcal{M}(D)}(s).$$

The last result that we will use concerns the sum of measures. Given two measures, μ and ν, over the same measure space and $a, b \geq 0$, we obtain that $a\mu + b\nu$ is also a measure over the same space. Extending to any $a, b \in \mathbb{R}$ gives us that $a\mu + b\nu$ is a signed measure. In either case, we have that

$$\int_S g \, \mathrm{d}(a\mu + b\nu) = a \int_S g \, \mathrm{d}\mu + b \int_S g \, \mathrm{d}\nu$$

for all measurable S and integrable g.

Lemma A.10. *Let $\mathcal{M} \colon \mathbb{D} \to \mathcal{D}(\Theta)$ be a d-private mechanism. Then,*

$$\int_S g \, \mathrm{d}P_{\mathcal{M}(D)} \leq \mathrm{e}^{d(D,D')} \int_S g \, \mathrm{d}P_{\mathcal{M}(D')}$$

for any integrable function $g \colon \Theta \to [0, 1]$.

Proof. Fix $D, D' \in \mathbb{D}$. The result is clear if $d(D, D') = \infty$, so we see the finite case.

Define the signed measure $\alpha \coloneqq P_{\mathcal{M}(D)} - \mathrm{e}^{d(D,D')} P_{\mathcal{M}(D')}$. Note that $\alpha \leq 0$ because $\mathcal{M}$ is d-private. Then, we have (see Remark A.9) that

$$\int_S g \, \mathrm{d}\alpha = \int_S g \, \mathrm{d}P_{\mathcal{M}(D)} - \mathrm{e}^{d(D,D')} \int_S g \, \mathrm{d}P_{\mathcal{M}(D')},$$

for all measurable $S \subseteq \Theta$ and any integrable function $g \colon \mathcal{S} \to [0, 1]$. Since $g \leq 1$, we have that

$$\int_S g \, \mathrm{d}\alpha \leq \int_S \mathrm{d}\alpha = \alpha(S) \leq 0,$$

and therefore

$$\int_S g \, \mathrm{d}P_{\mathcal{M}(D)} \leq \mathrm{e}^{d(D,D')} \int_S g \, \mathrm{d}P_{\mathcal{M}(D')}. \qquad \square$$

Additionally, we analyze different examples of partitioning functions and their corresponding behavior with respect to various metrics.

Proposition A.11. *Let $\mathbb{D} \subseteq \mathbb{D}_{\mathcal{X}}$. Any k-partitioning function p of Example 4.5 is $d_{\mathbb{D}}^{\triangle}$-commutative.*

Proof. Consider a partition of Example 4.5 and fix $D, D' \in \mathbb{D}$. Since $d^{\triangle}(D, D') = |D \triangle D'|$, we need to prove that

$$\sum_{i=1}^{k} |p_i(D) \triangle p_i(D')| = \left| \Big(\bigcup_{i=1}^{k} p_i(D) \Big) \triangle \Big(\bigcup_{i=1}^{k} p_i(D') \Big) \right| \leq |D \triangle D'|. \tag{A.1}$$

We prove first the equality, which corresponds to seeing that

$$\bigcup_{i=1}^{k} (p_i(D) \triangle p_i(D')) = \Big(\bigcup_{i=1}^{k} p_i(D) \Big) \triangle \Big(\bigcup_{i=1}^{k} p_i(D') \Big).$$

Since $p_i(D)$ and $p_j(D')$ are always disjoint for all $D, D' \in \mathbb{D}$ and $i \neq j$, it is sufficient to see that

$$(A \triangle A') \cup (B \triangle B') = (A \cup B) \triangle (A' \cup B')$$

for any arbitrary multisets such that $A \cap (B \cup B') = \varnothing$ and $A' \cap (B \cup B') = \varnothing$ (this case then extends by induction to k pairs of disjoint multisets). We denote by $\mathcal{A}$, $\mathcal{A}'$, $\mathcal{B}$ and $\mathcal{B}'$ the underlying set of the multisets of A, A', B and B', respectively. We use the multiset notation $A \coloneqq \langle \mathcal{A}, m_A \rangle$ where $m_A(a)$ corresponds to the multiplicity of $a \in \mathcal{A}$ in A. Under this notation we have for arbitrary multisets A and B that

(i) $A \cup B = \langle \mathcal{A} \cup \mathcal{B}, \max\{m_A, m_B\} \rangle$.

(ii) $A \cup B = \langle \mathcal{A} \cup \mathcal{B}, m_A + m_B \rangle$ when $A \cap B = \varnothing$.

(iii) $A \cap B = \langle \mathcal{A} \cap \mathcal{B}, \min\{m_A, m_B\} \rangle$.

(iv) $A \backslash B = \langle \mathcal{A}, m_A - m_B \rangle$ if $B \subseteq A$.

(v) $A \triangle B = \langle \mathcal{A} \cup \mathcal{B}, |m_A - m_B| \rangle$ using (i), (iii) and (iv).

Therefore, we have that

$$\begin{aligned}(A \cup B) \triangle (A' \cup B') &\overset{\text{(ii)}}{=} \langle \mathcal{A} \cup \mathcal{B}, m_A + m_B \rangle \triangle \langle \mathcal{A}' \cup \mathcal{B}', m_{A'} + m_{B'} \rangle \\ &\overset{\text{(v)}}{=} \langle \mathcal{A} \cup \mathcal{B} \cup \mathcal{A}' \cup \mathcal{B}', |m_A + m_B - m_{A'} - m_{B'}| \rangle,\end{aligned}$$

and

$$\begin{aligned}(A \triangle A') \cup (B \triangle B') &\overset{\text{(v)}}{=} \langle \mathcal{A} \cup \mathcal{A}', |m_A - m_{A'}| \rangle \cup \langle \mathcal{B} \cup \mathcal{B}', |m_B - m_{B'}| \rangle \\ &\overset{\text{(ii)}}{=} \langle \mathcal{A} \cup \mathcal{A}' \cup \mathcal{B} \cup \mathcal{B}', |m_A - m_{A'}| + |m_B - m_{B'}| \rangle.\end{aligned}$$

Since $A \cap (B \cup B') = \varnothing$ and $A' \cap (B \cup B') = \varnothing$, we obtain that $|m_A + m_B - m_{A'} - m_{B'}| = |m_A - m_{A'}| + |m_B - m_{B'}|$. Therefore,

$$(A \triangle A') \cup (B \triangle B') = (A \cup B) \triangle (A' \cup B'),$$

and by induction we obtain the equality of Equation (A.1).

The inequality in Equation (A.1) follows from the equality we just proved. Take $A = \bigcup_{i=1}^{k} p_i(D)$, $A' = \bigcup_{i=1}^{k} p_i(D')$, $B = D \backslash A$ and $B' = D' \backslash A'$, that obviously verify $A \cap B = \varnothing$ and $A' \cap B' = \varnothing$. Therefore,

$$\begin{aligned}\sum_{i=1}^{k} |p_i(D) \triangle p_i(D')| + |B \triangle B'| &= \left| \left(\bigcup_{i=1}^{k} p_i(D) \right) \triangle \left(\bigcup_{i=1}^{k} p_i(D') \right) \right| + |B \triangle B'| \\ &= |(A \triangle A') \cup (B \triangle B')| \\ &= |(A \cup B) \triangle (A' \cup B')| \\ &= |D \triangle D'|.\end{aligned}$$

Then, we obtain the inequality since $|B \triangle B'| \geq 0$. □

Proposition A.12. *If p is a k-partitioning function, $k > 1$, of Example 4.5, then p is not $d^{\mathcal{B}}_{\mathbb{D}_\mathcal{X}}$-compatible.*

Proof. Since the definition of compatible partition applies to any pair of neighboring databases, we just need to prove that there exists a pair of bounded-neighboring $D, D' \in \mathbb{D}_\mathcal{X}$ such that the condition of $d^{\mathcal{B}}_{\mathbb{D}_\mathcal{X}}$-compatibility is not satisfied.

In particular, we take $x_j \in \mathcal{X}_j$ and $x_i \in \mathcal{X}_i$ with $i \neq j$, and we build the two following bounded-neighboring databases: $D = \{x_i, x_i, x_i\}$ and $D' = \{x_i, x_i, x_j\}$. Then, $p_i(D) = D \neq D\backslash\{x_i\} = p_i(D')$ and $p_j(D) = \varnothing \neq \{x_j\} = p_j(D')$. Therefore there exist more than one $r \in [k]$ (particularly two: i, j), such that $p_r(D) \neq p_r(D')$, and thus $p_r(D) \not\sim_\mathcal{B} p_r(D')$. Therefore, p is not $d^{\mathcal{B}}_{\mathbb{D}_\mathcal{X}}$-compatible. □

Finally, we present the following lemma needed to prove Corollary A.14.

Lemma A.13. *Let $A, B \in \mathbb{D}_\mathcal{X}$ such that $|A| \leq |B|$ and $d^{\triangle}_{\mathbb{D}_\mathcal{X}}(A, B) = n$. Let $k = |B| - |A|$. Then, for any $\{x_i\}_{i\in[k]} \in \mathbb{D}_\mathcal{X}$, $C = A + \{x_i\}_{i\in[k]}$ verifies $d^{\mathcal{B}}_{\mathbb{D}_\mathcal{X}}(C, B) \leq n$ (where $+$ denotes the sum of multisets).*

Proof. Take $A, B \in \mathbb{D}_\mathcal{X}$ such that $r = |A| \leq |B| = s$ and $d^{\triangle}_{\mathbb{D}_\mathcal{X}}(A, B) = |A \triangle B| = n < \infty$. Observe that if $A \cap B = \{b_1, \ldots, b_l\}$ with $0 \leq l \leq r$, then we can express A and B as

$$B = \{\overbrace{b_1, \ldots, b_l}^{A\cap B}, \overbrace{b_{l+1}, \ldots, b_r, b_{r+1}, \ldots, b_s}^{B\backslash(A\cap B)}\},$$
$$A = \{\underbrace{b_1, \ldots, b_l}_{A\cap B}, \underbrace{a_{l+1}, \ldots, a_r}_{A\backslash(A\cap B)}\}.$$

In this case, note that

$$A \triangle B = (A\backslash(A \cap B)) \cup (B\backslash(A \cap B)) = \{b_{l+1}, \ldots, b_r, b_{r+1}, \ldots, b_s, a_{l+1}, \ldots, a_r\},$$

which has size n by hypothesis.

Consider the case where $|A| = |B|$. Then $|A\backslash(A \cap B)| = |B\backslash(A \cap B)|$ and $n = d^{\triangle}_{\mathbb{D}_\mathcal{X}}(A, B) = 2|A\backslash(A \cap B)|$ is even. In particular, $A \triangle B$ has the same number of elements of A and B, and therefore we can obtain B from A in $\frac{n}{2}$ bounded changes ($a_i \to b_i$ for $i \in \{l+1, \ldots, r\}$). That is, $d^{\mathcal{B}}_{\mathbb{D}_\mathcal{X}}(A, B) = \frac{n}{2}$. Therefore, if $|A| = |B|$, the statement verifies taking $A = C$ (since $k = 0$).

Suppose now $|A| < |B|$ (where $k \coloneqq |B| - |A|$) and define $C = A + \{x_i\}_{i\in[k]}$ for arbitrary $x_i \in \mathcal{X}$, i.e.,

$$C = \{\underbrace{b_1, \ldots, b_l}_{A\cap B}, \underbrace{a_{l+1}, \ldots, a_r}_{A\backslash(A\cap B)}, x_1, \ldots, x_k\}.$$

In particular, $|B| = |C|$, so we can apply the previous case. Thus, we obtain that $m \coloneqq d^{\triangle}_{\mathbb{D}_\mathcal{X}}(C, B)$ is even and $d^{\mathcal{B}}_{\mathbb{D}_\mathcal{X}}(C, B) = \frac{m}{2}$. Furthermore,

$$C \triangle B \subseteq (A \triangle B) + \{x_i\}_{i\in[k]},$$

so

$$m = |C \triangle B| \leq |A \triangle B| + k = n + k.$$

Note that $k \leq n$ since

$$\begin{aligned} n - k &= |A \triangle B| - (|B| - |A|) = |A \backslash (A \cap B)| + |B \backslash (A \cap B)| - |B| + |A| \\ &= |A| - |A \cap B| + |B| - |A \cap B| - |B| + |A| = 2|A| - 2|A \cap B| \geq 0. \end{aligned}$$

Thus, $m \leq n + k \leq 2n$. In conclusion, $d^{\mathcal{B}}_{\mathbb{D}_{\mathcal{X}}}(C, B) \leq n$. Since the proof does not depend on the choice of x_i, the proof is complete. □

Applying Proposition A.16 to bounded DP in $\mathbb{D}_{\mathcal{X}}$ we obtain the following result:

Corollary A.14. *Let $\mathbb{D}_{\mathcal{X}}$ be a database universe, $\mathcal{Y} \subsetneq \mathcal{X}$ and $f \colon \mathbb{D}_{\mathcal{X}} \to \mathbb{D}_{\mathcal{Y}}$ such that $f(D) = D \cap \mathcal{Y}$. Let $\mathcal{M} \colon \mathbb{D}_{\mathcal{X}} \to \mathcal{D}(\Theta)$ be a $d^{\mathcal{B}}_{\mathbb{D}_{\mathcal{X}}}$-private mechanism that is f-dependent. Then, $\mathcal{M}$ is $d_{\mathbb{D}_{\mathcal{X}}}$-private* with*

$$d_{\mathbb{D}_{\mathcal{X}}}(D, D') \coloneqq \min\{d^{\mathcal{B}}_{\mathbb{D}_{\mathcal{X}}}(D, D'), |f(D) \triangle f(D')|\} \leq \min\{d^{\mathcal{B}}_{\mathbb{D}_{\mathcal{X}}}(D, D'), d^{\mathcal{U}}_{\mathbb{D}_{\mathcal{X}}}(D, D')\}.$$

Proof. For all $D, D' \in \mathbb{D}_{\mathcal{X}}$, we need to prove that

$$\min_{\substack{\tilde{D}, \tilde{D}' \in \mathcal{D} \\ f(\tilde{D}) = f(D) \\ f(\tilde{D}') = f(D')}} d^{\mathcal{B}}_{\mathbb{D}_{\mathcal{X}}}(\tilde{D}, \tilde{D}') = \min\{d^{\mathcal{B}}_{\mathbb{D}_{\mathcal{X}}}, |f(D) \triangle f(D')|\}.$$

We have that $d^{\mathcal{B},f}_{\mathbb{D}_{\mathcal{X}}} \leq d^{\mathcal{B}}_{\mathbb{D}_{\mathcal{X}}}$ by definition of minimum privacy. Therefore, we just need to prove that $d^{\mathcal{B},f}_{\mathbb{D}_{\mathcal{X}}} \leq |f(D) \triangle f(D')|$ and we obtain the result.

First, note that since $f(D) = D \cap \mathcal{Y}$, $f(f(D)) = f(D)$ for all $D \in \mathbb{D}_{\mathcal{X}}$. Suppose without lost of generality that $|f(D)| \leq |f(D')|$, and let $k = |f(D')| - |f(D)|$. We take $x \in \mathcal{X} \backslash \mathcal{Y}$ and define $C \coloneqq f(D) + \{x, \overset{(k)}{\ldots}, x\}$. We see it verifies $f(D) = f(C)$. Then, $d^{\mathcal{B},f}_{\mathbb{D}_{\mathcal{X}}}(D, D') \leq d^{\mathcal{B}}_{\mathbb{D}_{\mathcal{X}}}(C, D') \leq |f(D) \triangle f(D')|$ by the definition of minimum privacy and Lemma A.13. □

Remark A.15. Let f be a deterministic map with domain $\mathbb{D}$, and let $\mathcal{M}$ with domain $\mathbb{D}$ be an f-dependent mechanism. If $f(D) = f(\tilde{D})$ for some $D, \tilde{D} \in \mathbb{D}$, then $\mathcal{M}(D)$ and $\mathcal{M}(\tilde{D})$ are equal random elements, i.e., $\Pr[\mathcal{M}(D) \in S\} = \Pr[\mathcal{M}(\tilde{D}) \in S\}$ for all measurable $S \subseteq \Theta$.

This is because, by definition of f-dependency, there exists a mechanism $\mathcal{M}^*$ such that $\mathcal{M} = \mathcal{M}^* \circ f$. Therefore

$$\Pr[\mathcal{M}(D) \in S] = \Pr[\mathcal{M}^*(f(D)) \in S] = \Pr[\mathcal{M}^*(f(\tilde{D})) \in S] = \Pr[\mathcal{M}(\tilde{D}) \in S]$$

for all measurable $S \subseteq \Theta$.

Under this definition, we arrive at the following result:

Proposition A.16 (Minimum privacy)**.** *Let $(\mathbb{D}, d_{\mathbb{D}})$ be a privacy space, let f be a deterministic map with domain $\mathbb{D}$, and let $\mathcal{M}\colon \mathbb{D} \to \mathcal{D}(\Theta)$ be a $d_{\mathbb{D}}$-private mechanism. If $\mathcal{M}$ is f-dependent, then $\mathcal{M}$ is $d_{\mathbb{D}}^f$-private* with*

$$d_{\mathbb{D}}^f(D, D') \coloneqq \min_{\substack{\tilde{D},\tilde{D}' \in \mathbb{D} \\ f(\tilde{D})=f(D) \\ f(\tilde{D}')=f(D')}} d_{\mathbb{D}}(\tilde{D}, \tilde{D}').$$

Proof. We fix $D, D' \in \mathbb{D}$ and choose $\tilde{D}, \tilde{D}' \in \mathbb{D}$ such that $f(D) = f(\tilde{D})$, $f(D') = f(\tilde{D}')$, and $d_{\mathbb{D}}(\tilde{D}, \tilde{D}')$ is minimum. In this case, $d_{\mathbb{D}}(\tilde{D}, \tilde{D}') = d_{\mathbb{D}}^f(D, D')$. Then, by definition of $d_{\mathbb{D}}$-privacy and Remark A.15,

$$\Pr[\mathcal{M}(D) \in S] = \Pr[\mathcal{M}(\tilde{D}) \in S] \leq \mathrm{e}^{d_{\mathbb{D}}(\tilde{D},\tilde{D}')} \Pr[\mathcal{M}(\tilde{D}') \in S] = \mathrm{e}^{d_{\mathbb{D}}^f(D,D')} \Pr[\mathcal{M}(D') \in S]$$

Since this holds for all $D, D' \in \mathbb{D}$ for all measurable $S \subseteq \Theta$, $\mathcal{M}$ is $d_{\mathbb{D}}^f$-private*. □

A.3. Additional Details for Chapter 5

Here, we provide the detailed computation of RAD for Examples 5.5 to 5.8 using Theorem 5.3.

Example A.17. In the optimal unary encoding (OUE) mechanism [150] each user encodes its input $x \in \mathcal{X}$ as a one-hot m-dimensional binary vector and perturbs each bit independently. For each position $i \in [m]$, the obfuscated vector θ is sampled such that $\Pr[\theta_i = 1] = 1/2$ if $i = x$, and $q = \frac{1}{\mathrm{e}^{\varepsilon}+1}$ otherwise. Denoting $p = 1 - q$ and $k_\theta = \#\{\theta_i = 1\}$, for every θ such that $k_\theta \geq 1$, we have that

$$\Pr(\theta \mid x) = \begin{cases} P \equiv \frac{1}{2} q^{k_\theta - 1} p^{m - k_\theta} & \text{if } \theta_x = 1 \\ Q \equiv \frac{1}{2} q^{k} p^{m - k_\theta - 1} & \text{if } \theta_x \neq 1 \end{cases} \tag{A.2}$$

and $\Pr(\vec{0} \mid x) = \frac{1}{2} p^{m-1}$. Hence, $\Pr(\vec{0}) = \frac{1}{2} p^{m-1}$ and $w(\vec{0}, z) = 0$ for all x. For all $\theta \neq \vec{0}$ we obtain,

$$p(\theta) = \frac{1}{2} q^{k_\theta - 1} p^{m - k_\theta} \underbrace{\left(\sum_{x\colon \theta_x = 1} \pi_x \right)}_{S_\theta} + \frac{1}{2} q^{k} p^{m - k_\theta - 1} \left(1 - \sum_{x\colon \theta_x = 1} \pi_x\right)$$

Note that $P - Q = \frac{p-q}{2}(q^{k_\theta - 1} p^{m - k_\theta - 1}) \geq 0$. Consequently,

$$w(\theta, x) = \begin{cases} (P - Q)\,(1 - S_\theta) \geq 0 & \text{if } \theta_x = 1 \\ (Q - P) S_\theta \ \leq 0 & \text{otherwise.} \end{cases}$$

Applying Theorem 5.3 for $a(x) = x$ we obtain,

$$\begin{aligned}
\eta\text{-RAD} &\leq \sum_{\theta} \sum_{x:\theta_x=1} (P-Q)\,(1 - \sum_{\theta_x=1} \pi_x)\pi_x \\
&= \frac{p-q}{2} \sum_{\theta} \sum_{x:\theta_x=1} (q^{k_\theta-1}p^{m-k_\theta-1})\,(1 - \sum_{\theta_x=1} \pi_x)\pi_x \\
&= \frac{p-q}{2p} \sum_{k=1}^{m} q^{k-1}p^{m-k} \sum_{\theta:k_\theta=k} \sum_{\theta_x=1} \pi_x\Big(1 - \sum_{\theta_x=1} \pi_x\Big) \\
&= \frac{p-q}{2p} \sum_{k=1}^{m} q^{k-1}p^{m-k} \binom{m-2}{k-1}(1-\kappa_\pi) \\
&= (1-\kappa_\pi)\frac{p-q}{2p} \sum_{k=1}^{m} \binom{m-2}{k-1} q^{k-1}p^{m-k} \\
&= (1-\kappa_\pi)\frac{p-q}{2p} \sum_{r=0}^{m-2} \binom{m-2}{r} q^{r}p^{m-1-r} \quad \text{(index change } r = k-1\text{)} \\
&= (1-\kappa_\pi)\frac{p-q}{2p}\, p^{m-1} \sum_{r=0}^{m-2} \binom{m-2}{r} (q/p)^{r} \\
&= (1-\kappa_\pi)\frac{p-q}{2p}\, p^{m-1}(1+q/p)^{m-2} \quad \text{(binomial identity)} \\
&= (1-\kappa_\pi)\frac{p-q}{2p}\, p^{m-1}(1/p)^{m-2} \quad \text{(since } p+q=1\text{)} \\
&= (1-\kappa_\pi)\frac{p-q}{2} \\
&= \frac{1}{2}\frac{e^\varepsilon-1}{e^\varepsilon+1}(1-\kappa_\pi) = \mathrm{TV}(\mathcal{M})\,(1-\kappa_\pi).
\end{aligned}$$

When $aux = \{\varnothing\}$, we have

$$\max_{x\in\mathcal{X}} w(\theta,x)\pi_x = (P-Q)(1 - \sum_{\theta_x=1} \pi_x) \max_{\theta_x=1} \pi_x.$$

Hence,

$$\begin{aligned}
\eta\text{-RAD} &\leq \sum_{\theta} (P-Q)\,(1 - \sum_{\theta_x=1} \pi_x) \max_{\theta_x=1} \pi_x \\
&= \frac{p-q}{2} \sum_{\theta} (q^{k_\theta-1}p^{m-k_\theta-1})\,(1 - \sum_{\theta_x=1} \pi_x) \max_{\theta_x=1} \pi_x
\end{aligned}$$

We order $\pi_1 \leq \cdots \leq \pi_m$. Then,

$$\Theta_i = \{\theta\colon \max_x w(\theta,x)\pi_x = w(\theta,x_i)\} = \{\theta\colon \theta_i = 1 \wedge \theta_j = 0 \text{ for all } j > i\},$$

and we can rewrite

$$\eta\text{-RAD} \leq \frac{p-q}{2p}\sum_{i=1}^{m}\pi_i \underbrace{\sum_{\theta\in\Theta_i} q^{k_\theta-1}p^{m-k_\theta}(1-\sum_{\theta_x=1}\pi_x)}_{A_i}$$

For every $k=1,\ldots,i$, there are $\binom{i-1}{k-1}$ vectors $\theta\in\Theta_i$ such that $k_\theta=k$. Moreover, the sum $\sum_{x\in S}\pi_x$ over all sets S of size k containing i:

$$\sum_{\substack{S\subseteq\{1,\ldots,i\}\\ i\in S,|S|=k}}\sum_{z\in S}\pi_x = \pi_i\binom{i-1}{k-1}+\sum_{z=1}^{i-1}\pi_x\binom{i-2}{k-2}.$$

Hence,

$$\begin{aligned}
A_i =& \sum_{k=1}^{i} q^{k-1}p^{m-k}\left[\binom{i-1}{k-1}(1-\pi_i)-\binom{i-2}{k-2}\sum_{z=1}^{i-1}\pi_x\right]\\
=&(1-\pi_i)\sum_{k=1}^{i}\binom{i-1}{k-1}q^{k-1}p^{m-k}-\left(\sum_{z=1}^{i-1}\pi_x\right)\sum_{k=1}^{i}\binom{i-2}{k-2}q^{k-1}p^{m-k}\\
=&(1-\pi_i)p^{m-i}\sum_{r=0}^{i-1}\binom{i-1}{r}q^{r}p^{i-1-r}-\left(\sum_{z=1}^{i-1}\pi_x\right)qp^{m-i}\sum_{j=0}^{i-2}\binom{i-2}{j}q^{j}p^{(i-2-j)}\\
=&(1-\pi_i)p^{m-i}+qp^{m-i}\left(\sum_{z=1}^{i-1}\pi_x\right).
\end{aligned}$$

since for $k=1$, $\binom{i-2}{k-1}=0$, so we can start in 2, i.e., $k=j+2$. Hence,

$$A_i = p^{m-i}\Big[(1-\pi_i)-q\sum_{x=1}^{i-1}\pi_x\Big], \qquad i=1,\ldots,m.$$

so

$$\eta\text{-RAD} \leq \frac{p-q}{2p}\left(\sum_{i=1}^{m}p^{m-i}\pi_i(1-\pi_i)-q\sum_{i=1}^{m}p^{m-i}\pi_i\sum_{z=1}^{i-1}\pi_x\right)$$

For instance, $\pi_i=\frac{1}{m}$

$$\begin{aligned}
\eta\text{-RAD} &\leq \frac{p-q}{2p}\sum_{i=1}^{m}p^{m-i}\pi_i(1-\pi_i)-q\sum_{i=1}^{m}p^{m-i}\pi_i\sum_{z=1}^{i-1}\pi_x\\
&= \frac{p-q}{2mp}\left(\frac{m-1}{m}\sum_{i=1}^{m}p^{m-i}-q\frac{1}{m}\sum_{i=1}^{m}p^{m-i}(1-i)\right)\\
&= \frac{p-q}{2p}\left[\frac{1}{m}\Big(1-\frac{1}{m}\Big)\frac{p^m-1}{p-1}-q\cdot\frac{1}{m^2}\cdot\frac{p^m-1-m(p-1)}{(p-1)^2}\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{p=1-q}{=} \frac{1-2q}{2(1-q)}\left[-\frac{m-1}{m^2}\cdot\frac{(1-q)^m-1}{q}-\frac{1}{m^2}\cdot\frac{(1-q)^m-1+mq}{q}\right]\\
&=\frac{1-2q}{2(1-q)}\left(-\frac{1}{mq}((1-q)^m-1+q)\right)\\
&=\frac{2q-1}{2(1-q)}\cdot\frac{(1-q)^m-1+q}{mq}\\
&\overset{q=1-p}{=}\frac{1-2p}{2(1-p)p}\cdot\frac{p^m-p}{m}=\frac{1-2p}{2(1-p)}\cdot\frac{p^{m-1}-1}{m}\\
&=\frac{(2p-1)(1-p^{m-1})}{2m(1-p)}.
\end{aligned}
$$

Note that

$$
\lim_{\varepsilon\to\infty}\frac{\mathrm{e}^\varepsilon-1}{2m}\left(1-\left(\frac{\mathrm{e}^\varepsilon}{1+\mathrm{e}^\varepsilon}\right)^{(m-1)}\right)=\frac{m-1}{2m},
$$

hence even if we keep reducing the noise (increasing ε), the attacker's advantage is limited.

Example A.18 (Subset Selection, $aux=\{\varnothing\}$)**.** In the subset selection mechanism (SS) [151] users report a subset $\theta\subseteq\mathcal{X}=\{x_1,\dots,x_m\}$ containing their true value x with probability $p=\frac{\omega\mathrm{e}^\varepsilon}{\omega\mathrm{e}^\varepsilon+m-\omega}$, where $\omega=|\theta|=\max\left(1,\left\lfloor\frac{m}{\mathrm{e}^\varepsilon+1}\right\rfloor\right)$. The subset is completed by sampling uniformly from $\mathcal{X}\setminus\{x\}$.

Note that, given $A=\binom{m-1}{\omega-1}$ and $B=\binom{m-1}{\omega}$,

$$
\Pr_{\mathcal{M}}(\theta\mid x)=\begin{cases}\frac{p}{A} & \text{if } x\in\theta\\ \frac{1-p}{B} & \text{if } x\notin\theta\end{cases}\tag{A.3}
$$

Since $|\Theta|=\binom{m}{\omega}$ we have that, according to Equation (5.14), for $\pi=U[m]$,

$$
\text{0-RAD}\le\frac{1}{m}(\sum_{\theta\in\Theta}\max_x p_{\mathcal{M}}(\theta\mid x)-1)\tag{A.4}
$$

$$
=\frac{1}{m}\binom{m}{\omega}\frac{p}{A}=\frac{m}{m\omega}p-\frac{1}{m}=\frac{pm-\omega}{m\omega}.\tag{A.5}
$$

Example A.19 (Gaussian mechanism and $aux=\{\varnothing\}$)**.** The Gaussian mechanism adds Gaussian noise $\mathcal{N}(0,\sigma)$ the query value $q(D)\in\mathbb{R}$ [50]. If $\mathcal{X}=\{x_1,\dots,x_m\}$ is uniformly distributed and $\Delta q=1$,

$$
x\in\arg\max_j w(\theta,x_j)\pi_j\Leftrightarrow x\in\arg\max_j w(\theta,x_j).
$$

Hence, applying Equation (5.14) we obtain

$$
\text{0-RAD}\le\frac{1}{m}\sum_{i=1}^m\left(\Pr_{\mathcal{M}}(\Theta_i\mid x_i)-\Pr_{\mathcal{M}}(\Theta_i)\right)=\frac{1}{m}\sum_{i=1}^m\left(\Pr_{\mathcal{M}}(\Theta_i\mid x_i)-1\right)\tag{A.6}
$$

Note that for each x, $\Pr_{\mathcal{M}}(\theta \mid x) = \Pr_{\mathcal{M}}(\theta \mid q(D_x))$. Since D_- is fixed, $q(D_x)$ is completely determined by x, hence we use the abuse of notation $q(D_x) \equiv x$. We want to compute $\Pr_{\mathcal{M}}(\Theta_i \mid x_i)$ for $i \in [m]$. Without loss of generality we re-order $x_1 < x_2 < \cdots < x_n$, and define the gaps $\Delta_i := x_{i+1} - x_i$. For fixed θ, the maximizing density corresponds to the x_i closest to θ. Thus $\mathbb{R}$ is partitioned into Voronoi intervals:

$$\Theta_1 = \big(-\infty, \tfrac{x_1+x_2}{2}\big], \tag{A.7}$$

$$\Theta_i = \big[\tfrac{x_{i-1}+x_i}{2}, \tfrac{x_i+x_{i+1}}{2}\big], \quad 2 \le i \le n-1, \tag{A.8}$$

$$\Theta_n = \big[\tfrac{x_{n-1}+x_n}{2}, \infty\big). \tag{A.9}$$

On Θ_i, the maximizer is x_i. Let Φ denote the standard normal CDF and φ its density function. Then, for $i = 1$

$$\Pr_{\mathcal{M}}(\Theta_1 \mid x_1) = \int_{\Theta_1} \varphi_\sigma(\theta - x_1)\, \mathrm{d}\theta = \Phi\Big(\tfrac{(x_1+x_2)/2 - x_1}{\sigma}\Big) = \Phi\Big(\tfrac{\Delta_1}{2\sigma}\Big).$$

For $i = m$

$$\Pr_{\mathcal{M}}(\Theta_m \mid x_m) = \int_{\Theta_m} \varphi_\sigma(\theta - x_m)\, \mathrm{d}\theta = 1 - \Phi\Big(\tfrac{(x_{m-1}+x_m)/2 - x_m}{\sigma}\Big) = \Phi\Big(\tfrac{\Delta_{n-1}}{2\sigma}\Big).$$

Finally, for $2 \le i \le m-1$,

$$\begin{aligned}\Pr_{\mathcal{M}}(\Theta_i \mid x_i) = \int_{\Theta_i} \varphi_\sigma(\theta - x_i)\, \mathrm{d}\theta &= \Phi\Big(\tfrac{(x_i+x_{i+1})/2 - x_i}{\sigma}\Big) - \Phi\Big(\tfrac{(x_{i-1}+x_i)/2 - x_i}{\sigma}\Big) \\ &= \Phi\Big(\tfrac{\Delta_i}{2\sigma}\Big) - \Phi\Big(-\tfrac{\Delta_{i-1}}{2\sigma}\Big) \\ &= \Phi\Big(\tfrac{\Delta_i}{2\sigma}\Big) + \Phi\Big(\tfrac{\Delta_{i-1}}{2\sigma}\Big) - 1,\end{aligned}$$

using $\Phi(-x) = 1 - \Phi(x)$. Therefore

$$\begin{aligned}\sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid x_i) &= \Phi\Big(\tfrac{\Delta_1}{2\sigma}\Big) + \sum_{i=2}^{m-1} \Big(\Phi\Big(\tfrac{\Delta_i}{2\sigma}\Big) + \Phi\Big(\tfrac{\Delta_{i-1}}{2\sigma}\Big) - 1\Big) + \Phi\Big(\tfrac{\Delta_{m-1}}{2\sigma}\Big) \\ &= 2\sum_{j=1}^{m-1} \Phi\Big(\frac{\Delta_j}{2\sigma}\Big) \;-\; (m-2),\end{aligned}$$

since each Δ_j appears exactly twice in the sum (once from its left neighbor, once from its right). Hence,

$$\text{0-RAD} \le \frac{2}{m}\sum_{j=1}^{m-1} \Phi\Big(\frac{\Delta_j}{2\sigma}\Big) \;-\; \frac{m-1}{m} \tag{A.10}$$

$$\le \frac{2(m-1)}{m}\Phi\Big(\frac{1}{(m-1)}\sum_{j=1}^{m-1} \frac{\Delta_j}{2\sigma}\Big) \;-\; \frac{m-1}{m} \tag{A.11}$$

$$\leq \frac{m-1}{m}\left(2\Phi\Big(\frac{1}{2\sigma(m-1)}\Big) - 1\right). \tag{A.12}$$

Where, Equation (A.11) follows since $\Delta_j \geq 0$, hence Φ concave, and we can apply Jensen's inequality, and Equation (A.12) since $\Delta q = 1$ therefore, $\sum_{j=1}^{m-1} \Delta_j = \Delta q = 1$.

Example A.20 (Laplace Mechanism and $aux = \{\varnothing\}$)**.** The Laplace mechanism adds Laplace noise with scale $b = \Delta q/\varepsilon$ to the query value $q(D) \in \mathbb{R}$ [16]. If $\mathcal{X} = \{x_1, \ldots, x_m\}$ if uniformly distributed and $\Delta q = 1$, analogously to Example A.19,

$$x \in \arg\max_j w(\theta, x_j)\pi_j \Leftrightarrow x \in \arg\max_j w(\theta, x_j).$$

Hence, applying Equation (5.14) we obtain

$$0\text{-RAD} \leq \frac{1}{m}\sum_{i=1}^{m}\left(\Pr_{\mathcal{M}}(\Theta_i \mid x_i) - p_{\mathcal{M}}(\Theta_i)\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\Pr_{\mathcal{M}}(\Theta_i \mid x_i) - 1\right) \tag{A.13}$$

Analogously to the Gaussian case, we use the abuse of notation $x \equiv q(D_x)$. We want to compute $\Pr_{\mathcal{M}}(\Theta_i \mid x_i)$ for $i \in [m]$. Without loss of generality we re-order $x_1 < x_2 < \cdots < x_n$, and define the gaps $\Delta_i := x_{i+1} - x_i$. For fixed θ, the maximizing density corresponds to the x_i closest to θ. Thus $\mathbb{R}$ is again partitioned into Voronoi intervals from Example A.19. Given the Laplace distribution CDF

$$F_i(x) = \begin{cases} \frac{1}{2}\exp\left(\frac{x-x_i}{b}\right) & \text{if } x < x_i \\ 1 - \frac{1}{2}\exp\left(-\frac{x-x_i}{b}\right) & \text{if } x \geq x_i \end{cases}, \tag{A.14}$$

for $i = 1$,

$$\Pr_{\mathcal{M}}(\Theta_1 \mid x_1) = F(\frac{x_1 + x_2}{2}) - F(-\infty) = 1 - \frac{1}{2}\exp\left(-\varepsilon\frac{\Delta_1}{2}\right),$$

for $i = m$,

$$\Pr_{\mathcal{M}}(\Theta_m \mid x_m) = 1 - F(\frac{x_m + x_{m-1}}{2}) = 1 - \frac{1}{2}\exp\left(-\varepsilon\frac{\Delta_{m-1}}{2}\right),$$

and for the reminder $2 \leq i < m$:

$$\begin{aligned}\Pr_{\mathcal{M}}(\Theta_m \mid x_m) =& F\left(\frac{x_i + x_{i+1}}{2}\right) - F\left(\frac{x_{i-1} + x_i}{2}\right) \\ =& 1 - \frac{1}{2}\exp\left(-\varepsilon\frac{\Delta_i}{2}\right) + \frac{1}{2}\exp\left(-\varepsilon\frac{\Delta_{i-1}}{2}\right).\end{aligned}$$

Hence,

$$\sum_{i=1}^{m}\Pr_{\mathcal{M}}(\Theta_i \mid x_i) = m - \frac{1}{2}\left(\mathrm{e}^{-\frac{\varepsilon\Delta_1}{2}} + \mathrm{e}^{-\frac{\varepsilon\Delta_{m-1}}{2}} + \sum_{i=2}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_i}{2}} + \mathrm{e}^{-\frac{\varepsilon\Delta_{i-1}}{2}}\right) = m - \sum_{j=1}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}}$$

since each Δ_j appears exactly twice in the sum (once from its left neighbor, once from its right). Hence,

$$\begin{aligned}
\text{0-RAD} \leq & \frac{1}{m}\left(m-1-\sum_{j=1}^{m-1} \mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}}\right) && \text{(A.15)}\\
\leq & \frac{m-1}{m}-\frac{1}{m}\sum_{j=1}^{m-1} \mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}} && \text{(A.16)}\\
\leq & \frac{m-1}{m}-\frac{m-1}{m}\mathrm{e}^{-\frac{1}{m-1}\sum_j \frac{\varepsilon\Delta_j}{2}} && \text{(A.17)}\\
\leq & \frac{m-1}{m}\left(1-\mathrm{e}^{-\frac{\varepsilon}{2(m-1)}}\right). && \text{(A.18)}
\end{aligned}$$

Where, Equation (A.17) follows since $\Delta_j \geq 0$, hence we can apply Jensen's inequality, and Equation (A.18) since $\Delta q = 1$ therefore, $\sum_{j=1}^{m-1} \Delta_j = \Delta q = 1$.

A.4. Additional Details for Chapter 6

In Example 6.4, we argue that the distribution in Table 6.1 does not present perfect correlation, i.e., variables that perfectly determine each other. Here we attach the corresponding proof:

Step 1: Check if X_2 is determined by X_1. For X_2 to be fully determined by X_1, each value of X_1 must correspond to a unique value of X_2. Consider the conditional probabilities:

$$P(X_2 = 0 \mid X_1 = 0) = \frac{P(X_1 = 0, X_2 = 0)}{P(X_1 = 0)} = \frac{\frac{1}{r^2}}{\frac{1+r}{r^3}} = \frac{r}{1+r} \neq 0, 1,$$

$$P(X_2 = 1 \mid X_1 = 0) = \frac{P(X_1 = 0, X_2 = 1)}{P(X_1 = 0)} = \frac{\frac{1}{r^3}}{\frac{1+r}{r^3}} = \frac{1}{1+r} \neq 0, 1.$$

Both values of X_2 are possible given $X_1 = 0$. Similarly, for $X_1 = 1$:

$$P(X_2 = 0 \mid X_1 = 1) = \frac{\frac{r-1}{r^2}}{\frac{r^3-r-1}{r^3}} = \frac{r(r-1)}{r^3-r-1} \neq 0, 1,$$

$$P(X_2 = 1 \mid X_1 = 1) = \frac{\frac{r^3-r^2-1}{r^3}}{\frac{r^3-r-1}{r^3}} = \frac{r^3-r^2-1}{r^3-r-1} \neq 0, 1.$$

Hence, multiple values of X_2 are possible given X_1, so X_2 is *not* fully determined by X_1.

Step 2: Check if X_1 is determined by X_2. For X_1 to be fully determined by X_2, each value of X_2 must correspond to a unique value of X_1. For $X_2 = 0$:

$$P(X_1 = 0 \mid X_2 = 0) = \frac{P(X_1 = 0, X_2 = 0)}{P(X_2 = 0)} = \frac{1/r^2}{1/r} = \frac{1}{r} \neq 0, 1,$$

$$P(X_1 = 1 \mid X_2 = 0) = \frac{P(X_1 = 1, X_2 = 0)}{P(X_2 = 0)} = \frac{(r-1)/r^2}{1/r} = \frac{r-1}{r} \neq 0, 1.$$

For $X_2 = 1$:

$$P(X_1 = 0 \mid X_2 = 1) = \frac{P(X_1 = 0, X_2 = 1)}{P(X_2 = 1)} = \frac{1/r^3}{(r-1)/r} = \frac{1}{r^2(r-1)} \neq 0, 1,$$

$$P(X_1 = 1 \mid X_2 = 1) = \frac{P(X_1 = 1, X_2 = 1)}{P(X_2 = 1)} = \frac{(r^3 - r^2 - 1)/r^3}{(r-1)/r} = \frac{r^3 - r^2 - 1}{r^2(r-1)} \neq 0, 1.$$

Thus, X_1 is *not* fully determined by X_2.