



Comprendre et optimiser le compromis entre vie privée et utilité d'un point de vue fondamental

Sayan Biswas

► To cite this version:

Sayan Biswas. Comprendre et optimiser le compromis entre vie privée et utilité d'un point de vue fondamental. Artificial Intelligence [cs.AI]. Institut Polytechnique de Paris, 2023. English. NNT : 2023IPPAX092 . tel-04407120

HAL Id: tel-04407120

<https://hal.science/tel-04407120>

Submitted on 9 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Understanding and optimizing the trade-off between privacy and utility from a foundational perspective

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à INRIA et l'École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 18 octobre, 2023, par

SAYAN BISWAS

Composition du Jury :

Benjamin Nguyen Professeur, INSA Centre Val de Loire	Président
Aurélien Bellet Directeur de recherche, INRIA Lille et Télécom Paris	Rapporteur
Marco Gaboardi Professeur associé, Boston University	Rapporteur
Graham Cormode Professeur, The University of Warwick et Meta	Examineur
Ehab ElSalamouny Professeur assistant, Suez Canal University	Examineur
Anne-Marie Kermarrec Professeure, EPFL	Examinatrice
Simon Oya Professeur assistant, University of British Columbia	Examineur
Catuscia Palamidessi Directrice de recherche, INRIA Saclay et École Polytechnique	Directrice de thèse

Understanding and optimizing the trade-off between privacy and utility from a foundational perspective

By

SAYAN BISWAS

INRIA and Institut Polytechnique de Paris

Thèse de Doctorat · Doctoral Thesis

Spécialité: Informatique · Specialization: Computer Science

Rapporteurs: Aurélien BELLET
Marco GABOARDI

Directeur de thèse: Catuscia PALAMIDESSI

Examineurs: Graham CORMODE
Ehab ELSALAMOUNY
Anne-Marie KERMARREC
Benjamin NGUYEN
Simon OYA

October 2023



“To the cherished memories of past adventures and the exciting prospects of those yet to come. To Kolkata, the place where my roots are firmly planted. To Bath, the city that nurtured my growth. To Paris, where I learned the art of truly living life. To the countless inspiring individuals and enriching experiences that have graced my journey so far, and to the countless more awaiting me in the future.”

Abstract

With recent advancements in technology, the threats of privacy violations of individuals' personal data are surging like never before. While protecting the privacy of sensitive information is becoming more important than ever before, it is also crucial to uphold the utility of the data especially as data is becoming one of the most essential resources in the contemporary information-based society. Differential privacy (DP) is considered to be the gold standard of formal privacy guarantees. Its widespread applicability and uncomplicated implementation technique have led to a rapid growth in its popularity and interest in studying and applying DP to a variety of domains in academia and industry alike. Over time, the community has explored various variants of DP addressing privacy concerns in different contexts and under a variety of threat models.

Despite the prolific acceptance of DP, it is still nebulous to interpret how it interacts with and affects the utility of data, escalating the need for answers to rudimentary questions like how adding DP noise affects the utility of the shared data (e.g., the quality of service of the data owners, the statistical utility of the service providers, the accuracy of the analysis and model training performed, etc.) and does there exist some optimal DP mechanism with respect to the usefulness of data in different realms and contexts. The objective of this thesis is to address these questions and, in particular, establish a theoretical foundation to comprehensively analyze the trade-off between privacy and the utility of sensitive data from a variety of perspectives and in the context of different use cases. Aside from dissecting the age-old battle between privacy and utility, this thesis also develops privacy-preserving mechanisms to proceed in the direction of optimizing utility loss with formal privacy guarantees in diverse domains of applicability.

The first part focuses on location privacy and explores its trade-offs with utilities applicable in the context of location-based services. In particular, Chapter 3 probes the triadic trade-off between the utility for the users of location-based services and the corresponding service providers. A standard method to mitigate the privacy risks for location data is by achieving geo-indistinguishability (geo-ind). However, isolated locations are not sufficiently protected by the state-of-the-art Laplace mechanism (LAP) for geo-ind. We, therefore, focus on a mechanism based on the Blahut-Arimoto algorithm (BA) and show that it provides comprehensive location-privacy guarantees and an optimal trade-off between information leakage and quality of service. Moreover, by establishing an intriguing duality between BA and an iterative expectation-maximization method, and by showing that BA's statistical utility is better than LAP's, we propose a method for a privacy-friendly incremental collection of location data from users by service providers.

Following this, in Chapter 4, fabricates an efficient privacy-preserving querying mechanism for the navigation of electric vehicles (EVs) with a high utility. Since there are disproportionately fewer charging stations than EVs, range anxiety plays a major role in the rise in the number of queries made along the journeys to find available charging stations. In this work, we introduce the notion of approximate geo-indistinguishability (AGeoI) which allows the EVs to achieve geo-ind in a strictly bounded space (e.g., remaining within their preferred area on the map) and also ensure that the overall trajectories of the EVs also remain private. The proposed method illustrates that a very high percentage of EVs get “privacy for free”. Our method also proves effective for an accurate online prediction of charging station occupancies under privacy guarantees which is vital for efficient route planning.

The second part of this thesis studies privacy in the context of federated learning (FL). Chapter 5, in particular, aims to address this triadic interaction between the personalization, privacy guarantees, and fairness achieved by the trained models under the FL framework. Clients in FL often hold very diverse datasets representing heterogeneous communities, making it important to protect their sensitive and personal information while ensuring that the trained model is fair for all the users. To achieve this, we propose a method that provides group privacy guarantees using d -privacy (a generalization of geo-ind). Our method, besides enabling personalized model training in a federated approach and providing formal privacy guarantees, possesses significantly better group fairness measured under a variety of standard metrics than a global model trained within a classical FL template.

Continuing to study the privacy aspects of FL, Chapter 6 investigates information leakage via gradients FL from a foundational perspective. Despite being believed to be one of the first significant steps towards privacy-preserving machine learning, numerous attacks against the training data in typical FL frameworks have been discovered of late that make use of the shared gradients as well as the model. This part of the thesis provides a formal understanding to explain the working of such empirically driven gradient-inversion attacks in FL elucidating how shared gradients leak information that can be exploited to recover sensitive training data.

The last part of the thesis is dedicated to efficient and optimal techniques to price and trade private data through the emerging concept of data markets. In Chapter 7, we propose a truthful incentive mechanism that furnishes a differentially private data trading mechanism optimal with respect to the utility of the data owners and the data collectors involved in data markets. Then we extend this work in Chapter 8 to propose a model of federated data markets in which data providers form coalitions for trading their private data. We illustrate a technique to price private data and distribute the revenue fairly in such a federated environment while motivating the data providers to cooperate with their respective federations, facilitating a fair and swift private data trading process.

Résumé

Avec les récentes avancées technologiques, les menaces de violation de la vie privée concernant les données personnelles des individus se multiplient comme jamais auparavant. Si la protection de la confidentialité des informations sensibles devient plus importante que jamais, il est également crucial de préserver l'utilité des données, d'autant plus que les données deviennent l'une des ressources les plus essentielles dans la société contemporaine basée sur l'information. La protection différentielle de la vie privée est considérée comme l'étalon-or des garanties formelles de protection de la vie privée. Son applicabilité étendue et sa technique de mise en œuvre simple ont conduit à une croissance rapide de sa popularité et de l'intérêt pour l'étude et l'application de la confidentialité différentielle à une variété de domaines, tant dans les universités que dans l'industrie. Au fil du temps, la communauté a exploré diverses variantes de la protection de la vie privée pour répondre aux problèmes de confidentialité dans différents contextes et dans le cadre de divers modèles de menace.

Malgré l'acceptation prolifique du DP, il est encore difficile d'interpréter la manière dont il interagit avec les données et affecte leur utilité, ce qui accroît le besoin de réponses à des questions rudimentaires telles que la manière dont l'ajout de bruit DP affecte l'utilité des données partagées (par exemple, la qualité de service des propriétaires de données, l'utilité statistique des fournisseurs de services, la précision de l'analyse et de l'entraînement des modèles effectués, etc. L'objectif de cette thèse est de répondre à ces questions et, en particulier, d'établir une base théorique pour analyser de manière exhaustive le compromis entre la protection de la vie privée et l'utilité des données sensibles selon diverses perspectives et dans le contexte de différents cas d'utilisation. Outre la dissection de la bataille séculaire entre la vie privée et l'utilité, cette thèse développe également des mécanismes de préservation de la vie privée afin d'optimiser la perte d'utilité avec des garanties formelles de respect de la vie privée dans divers domaines d'applicabilité.

La première partie se concentre sur la confidentialité de la localisation et explore ses compromis avec les utilités applicables dans le contexte des services basés sur la localisation. En particulier, le chapitre 3 étudie le compromis triadique entre l'utilité pour les utilisateurs de services basés sur la localisation et les fournisseurs de services correspondants. Une méthode standard pour atténuer les risques d'atteinte à la vie privée liés aux données de localisation consiste à assurer l'indiscernabilité géographique (geo-ind). Toutefois, les emplacements isolés

ne sont pas suffisamment protégés par le mécanisme de Laplace (LAP) de pointe pour la géo-indiscernabilité. Nous nous concentrons donc sur un mécanisme basé sur l'algorithme de Blahut-Arimoto (BA) et montrons qu'il fournit des garanties complètes de confidentialité de la localisation et un compromis optimal entre la fuite d'informations et la qualité de service. En outre, en établissant une dualité intrigante entre l'algorithme BA et une méthode itérative de maximisation des attentes, et en montrant que l'utilité statistique de l'algorithme BA est meilleure que celle de l'algorithme LAP, nous proposons une méthode de collecte progressive, respectueuse de la vie privée, des données de localisation des utilisateurs par les fournisseurs de services.

Ensuite, dans le chapitre 4, nous fabriquons un mécanisme d'interrogation efficace et respectueux de la vie privée pour la navigation des véhicules électriques (EVs) avec une utilité élevée. Étant donné que les stations de recharge sont disproportionnellement moins nombreuses que les VE, l'anxiété liée à l'autonomie joue un rôle majeur dans l'augmentation du nombre de requêtes effectuées tout au long des trajets pour trouver des stations de recharge disponibles. Dans ce travail, nous introduisons la notion de géo-indiscernabilité approximative (AGeoI) qui permet aux VE d'atteindre la géo-ind dans un espace strictement délimité (par exemple, en restant dans leur zone préférée sur la carte) et de garantir que les trajectoires globales des VE restent également privées. La méthode proposée montre qu'un pourcentage très élevé de VE obtient la "confidentialité gratuitement". Notre méthode s'avère également efficace pour une prédiction en ligne précise de l'occupation des stations de recharge avec des garanties de confidentialité, ce qui est vital pour une planification efficace des itinéraires.

La deuxième partie de cette thèse étudie la protection de la vie privée dans le contexte de l'apprentissage fédéré (AF). Le chapitre 5, en particulier, vise à aborder cette interaction triadique entre la personnalisation, les garanties de confidentialité et l'équité obtenue par les modèles formés dans le cadre de l'apprentissage fédéré. Les clients de FL détiennent souvent des ensembles de données très divers représentant des communautés hétérogènes, d'où l'importance de protéger leurs informations sensibles et personnelles tout en veillant à ce que le modèle formé soit équitable pour tous les utilisateurs. Pour y parvenir, nous proposons une méthode qui fournit des garanties de confidentialité de groupe en utilisant d -privacy (une généralisation de geo-ind). Notre méthode, en plus de permettre la formation de modèles personnalisés dans une approche fédérée et de fournir des garanties formelles de confidentialité, possède une équité de groupe significativement meilleure, mesurée selon une variété de métriques standard, qu'un modèle global formé à l'intérieur d'un modèle FL classique. Nous fournissons des justifications théoriques pour l'applicabilité et la validation expérimentale sur des ensembles de données du monde réel pour illustrer le fonctionnement de la méthode proposée.

Poursuivant l'étude des aspects de confidentialité du FL, le chapitre 6 étudie les fuites d'informations via les gradients FL d'un point de vue fondamental. Bien qu'elle soit considérée comme l'une des premières étapes significatives vers l'apprentissage automatique préservant la vie privée, de nombreuses attaques contre les données d'apprentissage dans les cadres FL typiques ont été découvertes récemment, qui utilisent les gradients partagés ainsi que le modèle. Cette partie de la thèse fournit une compréhension formelle pour expliquer le fonctionnement de telles attaques empiriques par inversion de gradient dans l'apprentissage automatique, en élucidant la manière

dont les gradients partagés laissent échapper des informations qui peuvent être exploitées pour récupérer des données d'apprentissage sensibles.

La dernière partie de la thèse est consacrée aux techniques efficaces et optimales permettant de fixer le prix et d'échanger des données privées grâce au concept émergent des marchés de données. Dans le chapitre 7, nous proposons un mécanisme d'incitation véridique qui fournit un mécanisme d'échange de données différentiellement privées optimal en ce qui concerne l'utilité des propriétaires de données et des collecteurs de données impliqués dans les marchés de données. Nous étendons ensuite ce travail au chapitre 8 pour proposer un modèle de marchés de données fédérés dans lequel les fournisseurs de données forment des coalitions pour échanger leurs données privées. Nous illustrons une technique de tarification des données privées et de distribution équitable des revenus dans un tel environnement fédéré, tout en motivant les fournisseurs de données à coopérer avec leurs fédérations respectives, facilitant ainsi un processus d'échange de données privées équitable et rapide.

Acknowledgements

I am massively thankful and would like to express my deepest gratitude and appreciation to everyone who has made my PhD journey so incredibly enjoyable and rewarding.

To my supervisor Catuscia Palamidessi: you have been one of my biggest inspirations through my entire PhD journey. To date, I continue getting amazed by your diverse range of qualities including and not limited to your immaculate work ethic, fierce intelligence, an alluring sense of aesthetics, and extraordinary care of and encouragement for your students. Thank you for replying to my hundreds of emails over the last four years – at every possible time of the day (and night) – answering my countless questions and for the unending appreciation, encouragement, and motivation. You are one of the best teachers and guides I could ever get in my life making my PhD journey so thoroughly enjoyable with your perpetual support in every realm possible, professional and personal alike. Last but not least, thank you for financially supporting my PhD through the HYPATIA project of the European Research Council (ERC) under the Horizon 2020 research and innovation program of the European Union (Grant agreement N. 835294).

To my Mum (Ma) and late Dad (Baba): Ma, no amount of “thank you” will suffice to express my gratitude to you for raising me the way you have and for your endless selfless sacrifice and help in every stage of my life including and not limited to my PhD. Baba, you have always been there for me in every capacity whenever I needed you — thank you for all your support and encouragement. I wish you were there today to hug me and see your “Gogol” to finally complete his PhD. I am certain that you are proud of me and blessing me for my upcoming endeavours from wherever you are.

To all my extended family: Thank you for your ineffable support in every situation and for never letting me feel alone despite being thousands of miles away from you all. In particular, Bapis, without you teaching me to solve those algebraic equations and making me appreciate the essence of mathematics as a kid, I probably would have had a much harder time proving any of the theorems during my PhD. Dia, thank you for always wishing the best for me and for your continued encouragement. Mani, Tatali, Putai Didi, and Pavel Dada, thank you all for always being there for me since my childhood and making every celebration worth it.

To all my collaborators and mentors: Thank you for the plethora of email exchanges, video calls, and brainstorming sessions resulting in our successful collaborations and publications. You all have contributed immensely to my professional growth and enriched my overall PhD experience. In particular, I extend my gratitude to Kangsoo Jung for mentoring me on a variety of topics ranging from writing codes to Korean culture and to Graham Cormode, Carsten Maple,

Natasha Fernandes, and Annabelle McIver for all your valuable support, advice, and comments enabling our long-distance collaboration to go through without any hindrance.

To the entire COMÈTE team at Inria Saclay: I am truly grateful to have been a part of such a stimulating, encouraging, and fun academic environment during my PhD. I would cherish the friendships I have made within the team for a lifetime. Thank you all for being so supportive and fun. In particular, I am truly thankful to have Federica Granese as one of my best friends in my PhD life.

To my best friends: Aabesh, Anuraag, and Souvik, life (PhD and beyond) would be a quarter as enjoyable without having you three (horrible yet awesome) blokes around. On a serious note, you three occupy one of the most priceless parts of my life — thank you for always being there.

To Nikita: My PhD journey would have been incomplete without you. Thanks for your endless support, motivation, and encouragement over the last three years, being there by me through thick and thin. Moreover, thank you for always taking the time to listen to my rants, celebrating my publications, and always cheering me up – you have been the one of the most beautiful and integral parts of my PhD life.

Contents

Contents	x
List of Figures	xiv
List of Tables	xviii
I Overview	1
1 Introduction	3
1.1 Privacy in Practice	4
1.2 Privacy-Utility Trade-off	5
1.3 Contributions of this thesis	6
1.3.1 Synopsis	7
1.3.2 Publications	9
1.3.3 Select Talks	10
2 Foundations	11
2.1 Standards of Privacy	11
2.2 Federated Learning	13
2.3 Notions of Utility	14
II Location privacy	16
3 PRIVIC: A privacy-preserving method for incremental collection of location data	18
3.1 Introduction	18
3.2 Technical preliminaries	21
3.3 Related Work	24
3.4 Location-privacy with the Blahut-Arimoto algorithm	25
3.4.1 Elastic location-privacy with BA	25
3.4.2 Statistical utility: BA vs LAP	29
3.5 Duality between IBU and BA	30
3.6 PRIVIC: a privacy-preserving method for incremental data collection	31
3.7 Experimental analysis of PRIVIC	34

3.8	Vulnerability of PRIVIC	39
4	A privacy-preserving querying mechanism with high utility for electric vehicles	41
4.1	Introduction	41
4.2	Related Work	43
4.3	Approximate geo-indistinguishability (AGeol)	45
4.4	System Model	47
4.4.1	Problem Statement	47
4.4.2	Road Network Model	48
4.4.3	System Architecture	48
4.4.4	Privacy Threat Landscape	50
4.4.5	Proposed Query Model	51
4.5	Cost of privacy analysis	53
4.6	Experimental Study	55
4.6.1	Dataset Preparation	55
4.6.2	Experimental Setup	55
4.6.3	Results and Discussion	56
III	Federated learning	61
5	Advancing Personalized Federated Learning: Group Privacy, Fairness, and Beyond	63
5.1	Introduction	63
5.2	Technical Preliminaries	65
5.2.1	Personalized Federated Learning	65
5.2.2	Fairness	66
5.3	Related Works	67
5.4	An Algorithm for Private and Personalized Federated Learning	68
5.4.1	The Laplace mechanism under Euclidean distance in \mathbb{R}^n	69
5.4.2	A Heuristic for defining the Neighbourhood of a Client	70
5.5	Experiments	70
5.5.1	Characterizing privacy	70
5.5.2	Synthetic Data	71
5.5.3	Hospital Charge Data	71
5.5.4	FEMNIST Image Classification	73
5.5.5	Characterizing fairness	75
6	Characterizing the information leakage from gradient updates in federated learning	79
6.1	Introduction	79
6.2	Preliminaries	80
6.3	Federated learning setup	81
6.3.1	Overview of the federated model training	82
6.3.2	The neural network model	82

6.3.3	The NN training process	82
6.3.4	Adversarial assumptions	82
6.3.5	Measuring information leakage	83
6.4	Information leakage in simple FL models	83
6.4.1	Technical setup	84
6.4.2	Analysing gradient leaks	85
6.4.3	Leakage from bigger batches	87
6.4.4	Leakage from batches size = 2:	89
6.5	Analyzing multi-layer neural networks	90
6.5.1	Extending the TNN to a larger NN	92
6.5.2	Leakage from the penultimate layer	93
6.6	Discussion and open questions	93
IV	Private data trading	94
7	An incentive mechanism for trading personal data in data markets	96
7.1	Introduction	96
7.2	Related Work	97
7.3	Incentive mechanism for data markets	99
7.3.1	Overview of the proposed technique	99
7.3.2	Truthful price report mechanism	101
7.3.3	Optimizing the incentive mechanism	103
7.3.4	Discussion	105
7.3.5	Optimized privacy budget splitting mechanism for data providers	106
7.4	Experimental results	107
8	Establishing the price of privacy in federated data trading	109
8.1	Introduction	109
8.2	Related Work	110
8.3	Technical Preliminaries	111
8.3.1	Shapley value	111
8.4	Differentially Private Data Trading Mechanism	112
8.4.1	Mechanism outline	112
8.4.2	Earning Splitting	115
8.5	Experimental results	118
8.5.1	Number of rounds needed for data collection	119
8.5.2	Number of free riders by penalty scheme	120
8.5.3	Reduced Shapley value computation time	121
9	Conclusion and Future Work	122
	References	127

Appendices	147
A Proofs from Chapter 3	148
B Tables from Chapter 3	151
C Proofs from Chapter 4	152
D Proofs from Chapter 5	154
E Experimental settings in Chapter 5	158
F Proofs from Chapter 6	161
G Proofs from Chapter 7	171
H Examples from Chapter 7	173
I Proofs from Chapter 8	176

List of Figures

2.1	A schematic overview of the working of <i>local</i> differential privacy	12
2.2	Illustration of the working of federated learning for model training	13
3.2	Distribution of privatizing the vulnerable and the strong locations for different levels of privacy. Top-down, the rows illustrate the results for $\varepsilon = 0.4, 1.2, 1.6, 2$, respectively.	27
3.1	Gowalla check-in locations in Paris with an artificially planted vulnerable point, A , in isolation, and a strong point, B , in a crowded area.	28
3.3	Statistical utility in terms of <i>earth mover's distance</i> (EMD) between the true and the estimated distributions for locations in Paris and San Francisco, under BA and LAP.	29
3.4	Illustration of the duality between BA and IBU.	30
3.5	Illustration of the iterative process of PRIVIC	32
3.6	(a) visualizes the original locations from Gowalla dataset from Paris and San Francisco. (b) illustrates a heatmap representation of the locations in the two cities to capture the distribution of the data.	35
3.7	Visualization of the estimated true distribution of the locations in Paris ((a) and (b)) and San Francisco ((c) and (d)) by PRIVIC after its convergence; the first column is for $\beta = 0.5$ and the second column is for $\beta = 1$	37
3.8	(a) and (b) show the EMD between the true PMF of the Paris locations and its estimation by PRIVIC in each of its cycle for $\beta = 0.5$ and $\beta = 1$, respectively. . . .	37
3.9	(a) and (b) show the EMD between the true PMF of the San Francisco locations and its estimation by PRIVIC in each of its cycles for $\beta = 0.5$ and $\beta = 1$, respectively. . . .	38
3.10	(a) and (b) illustrate that EMD between the true and the estimated distributions of the locations in Paris and San Francisco, respectively, after the empirical convergence of PRIVIC for the different values of the loss parameters β	38
3.11	Effect on the privacy provided by BA to obfuscate the geo-spatially isolated location (Location A as in Figure 3.1) for different fractions of adversarial users who intentionally report their locations falsely under two different levels of formal geo-ind guarantees by BA.	40
4.1	System Architecture (EV: Electric Vehicle, RSU: Roadside Unit, MEC: Mobile-Edge Computing Unit)	49

4.2	Reported dummy and privatised locations for two respective time windows (White Pins: Privatised locations, Orange Pins: Dummy locations in 1 st Time window, Blue Pins: Dummy locations in 2 nd Time window)	52
4.3	A toy example for a static location query on discrete road network	52
4.4	A toy example for linked 3 location queries on discrete road network	52
4.5	CoP (i.e., by Definition 4.5.1, the difference in the distance an EV needs to cover to reach the nearest CS with and without local obfuscation to achieve AgeoI) for varying ε or r of AgeoI (1 st row is for sparse CS, 2 nd row is for dense CS).	57
4.6	Fraction of EVs incurring no CoP for varying ε or r of AgeoI (1 st row is for sparse CS, 2 nd row is for dense CS).	57
4.7	Impact of introducing dummy locations along with AgeoI on the CoP.	58
4.8	The Kantorovich-Wasserstein distance between the original and estimated distributions using IBU for $\varepsilon = 0.6$ and $\varepsilon = 2$ noisy distributions.	58
4.9	Estimations of the original distribution using IBU for the $\varepsilon = 0.6$ and $\varepsilon = 2$ noisy distributions.	59
5.1	A schematic overview of the personalized approach of federated model training .	65
5.2	An illustration of the implementation of the Laplace mechanism to achieve ε - d -privacy for model parameters in \mathbb{R}^2	70
5.3	Learning federated linear models with (a, b, c) one initial hypothesis and non-sanitized communication, (d, e, f) two initial hypotheses and non-sanitized communication, (g, h, i) two initial hypotheses and sanitized communication. The first two figures of each row show the parameter vectors released by the clients to the server.	72
5.4	Synthetic data: max privacy leakage among clients. Privacy leakage is constant when clients with the largest privacy leakage are not sampled (by chance) to participate in those rounds.	73
5.5	RMSE for models trained with Algorithm 6 on the Hospital Charge Dataset. Error bars show $\pm\sigma$, with σ the empirical standard deviation. Lower RMSE values are better for accuracy.	74
5.6	Hospital charge data: the empirical distribution of the privacy budget over the clients for $\nu = 3, 5$ initial hypotheses, seed = 3, r is the radius of the neighbourhood, and the total number of clients is 2062.	74
5.7	Effects of the Laplace mechanism in Proposition 5.1 with different noise multipliers as a defense strategy against the DLG attack.	75
5.8	The first two plots from the left illustrate the spatial distribution of the samples in g_1 and g_2 , respectively, and the third plot shows g_1 and g_2 superimposed together in the same space.	76

5.9	The figure shows the comparison between the personalized and non-personalized models for (from left) equal opportunity, equalized odds, and demographic parity, respectively. Experiments were performed for noise multipliers ν of 0.1, 1, 2, and 4. For all the metrics of fairness and the values of the noise multiplier, the personalized model is seen to possess improved fairness over the non-personalized model.	77
5.10	The figure shows the comparison between the personalized and non-personalized models for equal opportunity equalized odds, and demographic parity. Experiments were performed for noise multipliers ν of 0.1, 1, 2, and 4. For all metrics of fairness and values of the noise multiplier, the personalized model improved fairness over the non-personalized model.	78
6.1	Illustration of a toy neural network consisting of a single input x and a final layer with 2 nodes and a softmax activation function, interpreted as a probability distribution over labels. We include bias terms in each node in the final layer. . .	84
6.2	Larger neural network example. The final layer is fully connected although this is not depicted in the diagram for simplicity of labelling. We make no assumptions about the number of hidden layers.	91
7.1	An example of data trading process. In this figure, u_i means the i^{th} data provider and D_j means the j^{th} data consumer.	99
7.2	An example of a monotonically decreasing function $f(z)$. Let c be a parameter representing the “reported value-to-admitted ε value” ratio. For $z \geq 0$, we set $f(z)$ as $f(z) = \ln(e - cz)$ if $(e - cz) \leq 1$, and $f(x) = 0$ otherwise.	101
7.3	Graphical illustration of Theorem 7.3. We prove that $\rho(\pi_i)$ (blue hatching area) is always larger than $\rho(p_i)$ (blue rectangle area+red hatching area–green rectangle area).	102
7.4	Illustrating the payoff for c and the incentive plot for the data consumer involving one data provider reporting p_1 . The Y -intercept of μ is $\int_{p_1}^{\infty} f(z) dz$	104
7.5	Experimental result of profit under a fixed budget. The \log function represents the family $\ln(e - cp)$ and the linear function represents the family $1 - cp$. We let the parameter c range from 0 to 1. The red bin represents the optimal value of c , namely the c that gives maximum information.	108
8.1	Some examples of the privacy valuation function f illustrated with different values of K_1 and K_2 . The data consumer decides the values of the scaling factors K_1 and K_2 according to her requirement and broadcasts the determined function to the federations.	113
8.2	The number of rounds required by a federation (of varying sizes) to achieve the privacy level (i.e., information accuracy) to be reported to the data consumer is less for the catalyzing method introduced in the data collection than its non-catalyzing counterpart.	120

8.3 The number of free riders incurred by the penalty scheme (for varying sizes of the federation and the tolerance threshold) is significantly less adapting the catalyzing method for data collection than the non-catalyzing method.	121
---	-----

List of Tables

3.1	Run-time and complexity of BA and IBU in each cycle of PRIVIC	35
5.1	Hospital charge data: median and maximum local privacy budgets over the whole set of clients, averaged over 10 runs with different seeds. $\nu = 0$ means no privacy guarantee.	73
5.2	Effects of increasing the noise multiplier on the validation accuracy and standard deviation.	74
6.1	An example of a batch used in a given round of training consisting of datapoints which are <i>similar</i> as per Def. 6.4.3 for an FL model that has <i>learnt well</i> as per Def. 6.4.2. In other words, the images of the dog, the wolf, and the fox used in the training batch have very similar true-positive and true-negative probabilities.	89
8.1	Computation time of brute force and proposed pruning method	121
B.1	EMD between the true and the estimated PMFs by PRIVIC on the Paris locations.	151
B.2	EMD between the true and the estimated PMFs by PRIVIC on the San Francisco locations.	151
E.1	NN architecture adopted in the experiments of Section 5.5.4	160

Part I

Overview

“Remember, Red, hope is a good thing, maybe the best of things, and no good thing ever dies.”

– Andy Dufresne, The Shawshank Redemption

1

Introduction

Ronald H. Coase, a renowned and celebrated economist, in a talk at the University of Virginia in the early 1960s, famously quoted “Torture the data and it will confess to anything.” About half a century later, Forbes’ estimation that about 90% of the data in the world was generated in the last two years [1] suggests that we are, well and truly, headed towards a data-driven society. This huge surge of data harvesting mainly caters to a diverse range of analytics performed and technologies developed using them which, in turn, contributes to the rise of a variety of services that we indispensable cherish in our daily lives.

A flip side to this flourishing advancement of data science and technology is the progressive growth of dangerous attacks invading the privacy of the individuals whose sensitive information is often contained in the datasets produced, used, and analysed. One of the scarier facets of such blooming threats to violate individuals’ privacy is that they often involve, feature, or compromise institutions, organisations, and services (e.g., Facebook [2–4], Google [5, 6], Yahoo [7], Netflix [8], Public Transport Victoria, Government of Australia [9], etc.) that are typically inseparable from and intrinsically facilitate our contemporary lifestyle.

The right to privacy is a universally recognised human right. Criticizing the *nothing to hide argument*, former computer intelligence consultant and whistleblower Edward Snowden famously stated “Arguing that you don’t care about privacy because you have nothing to hide is no different than saying you don’t care about free speech because you have nothing to say.” An individual’s right to digital privacy involves their right to moderate the use of their personal data – the manner in which they are stored, handled, and protected from potential threats. Although data privacy often intertwines with cybersecurity, it is crucial to note that with modern standards, while leaked passwords can be modified or stolen credit cards blocked or re-issued,

privacy, once compromised, cannot be so easily reinstated. Hence, it is of utmost importance to furnish sound and applicable privacy-preserving technologies, methods, and tools. Alongside protecting the privacy of individuals' personal information, it is unequivocally acknowledged that, under such privacy guarantees, the shared data still needs to be utilisable in order to foster the services and analytics that are essential to sustain the modern technology-driven lifestyle around the globe.

1.1 Privacy in Practice

In the past, privacy breaches, although deemed to be threatening¹ indeed they were characterised and documented by the statistician Tore Dalenius [10], but not enough attention was given to this as they were thought to be practically infeasible due to the amount of information needed by an attacker in order to correlate any released statistics with particular individuals and their sensitive attributes. In fact, until the early 2000s, data anonymisation was regarded as a sufficient step for protecting individuals' privacy, and this understanding is reflected in the privacy laws that still stand today in many countries including Australia and the United States.²

Early in the 2000's some infamous privacy-compromising attacks elucidated the weaknesses of anonymisation-based techniques in the face of large-scale access to publicly available datasets. Some major examples include: In 2002, Latanya Sweeney identified the medical data of the governor of Massachusetts from anonymized hospital records by linking them with the publicly available information in a voter database [11]; In 2006, an anonymised dataset of movie ratings released by Netflix for a competition designed to improve their recommendation model was famously attacked by researchers Shmatikov and Narayanan [8]³; Recently, the US Census Bureau re-examined census data from 1940 and was able to reconstruct the details of individuals using now-available information [12]. These attacks were made possible by the scale of publicly available data on the internet and showed a gaping loophole in the anonymization-based data protection techniques which did not consider an attacker's background knowledge.

In 2006, *differential privacy (DP)* was proposed by Dwork et al. [13, 14] that emerged to solve the problem of privacy for individuals in structured datasets. The rudimentary idea behind DP relies on 'plausible deniability' for individuals in that the response of a query performed on the dataset is almost probabilistically agnostic to the presence or the absence of the record of a certain individual in the dataset and, thus, it is argued that an individual's presence in a dataset is protected. DP is now flourished to be arguably the state-of-the-art standard for data privacy due to a variety of advantages it endorses including but not limited to its well-behaved compositionability, its robustness to background knowledge, the possibility to tune the noise to harbour a desirable level of accuracy, and its straightforward applicability. These properties set it apart from anonymisation-based privacy standards that were promoted in the past and have

¹The US Census Bureau has had formal privacy requirements for statistical data releases in place since the 1920's: https://www.census.gov/history/www/reference/privacy_confidentiality/privacy_and_confidentiality_2.html.

²See Australian Privacy Act: <https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-and-the-privacy-act/> or the US HIPAA: <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html>.

³<https://www.wired.com/2009/12/netflix-privacy-lawsuit/>

been shown to be susceptible to adversarial attacks when combined with other anonymised datasets [15].

In 2012, the notion of “metric differential privacy” was introduced in order to enable DP-style reasoning in the domain of arbitrary spaces endowed with metrics. Such metric-based DP provides a generalisation and re-interpretation of DP accounting for a notion of “distance” between the secrets. Exploiting the fact that some of the main characteristics of DP rely on the metric properties of the Hamming distance between datasets, the notion of geo-indistinguishability and d -privacy were proposed [16, 17] that proved to be extremely suitable to privatize location data while harbouring a high utility without digressing from the strong and formal qualities that classical DP offers.

The aforementioned standards of privacy guarantees, although formal, are often hard to interpret for non-experts. For example, in practical scenarios, the privacy parameter of DP – typically denoted by an ϵ – is usually seen as a parameter for tweaking *utility* without associating any specific and interpretable meaning to the resulting privacy guarantee [18]. Since the value of ϵ in DP is not consequentially meaningful to an end user, the community, of late, has focussed on more “pragmatic” interpretation of data privacy like measuring the information leakage in a given mechanism [19] and defending against the rising data-reconstruction attacks [20, 21] in privacy-preserving machine learning.

In this thesis, we study the interaction privacy guarantees and the utility of data from a variety of perspectives considering a diverse range of contexts and use cases. The work presented in the thesis primarily focuses on a foundational approach to analyse the trade-off between formal cutting-edge standards of data privacy and their context-specific utility across different applications like location privacy, privacy-preserving machine learning, and economics of data privacy involving private data trading in data markets.

1.2 Privacy-Utility Trade-off

Our goal in this thesis is to explore the trade-off between privacy and utility endorsed by the application of DP and its variants and optimize it from a foundational perspective. Because of its widespread acceptance, the community (industry and academia alike) have been tempted to study and adopt DP and its variants in a wide range of domains. However, such extensive application of DP comes at the cost of degrading the utility of the data due to the addition of noise. To adhere to the rudimentary purpose of data analytics, it is imperative that the privacy guarantees on the collected and shared data do not make the utility for the users and the service providers redundant. Therefore, in order to succumb to the role of data in the development of technologies integral to our contemporary life, it is very important to understand and optimise the age-old battle between privacy and accuracy.

“Utility” is a highly metaphysical concept. In particular, the utility of data typically alters with the context and the priorities. Therefore, to fabricate a foundational understanding of how this rather philosophical concept entangles with the standards of privacy used in practice, it is crucial to quantify it. Broadly speaking, the utility of the shared data has two primary facets: a)

for the data providers (e.g., users of a service) and b) for the data collectors (e.g., the service providers). The literature has explored a bundle of formalizations of the notion of utility from both ends. For example, for location-based services (e.g., GPS, location-based games, dating apps, etc.), the most intuitive and commonly accepted notion for the utility of the users is their quality of service (QoS) with respect to some chosen distortion metric (e.g., on an average, how much extra distance one needs to cover as a result of privatizing their reported locations). In machine learning, depending on the task, utility for the users usually revolve around the trained model’s accuracy and fairness. In data markets, the utility of the participating clients is mostly related to financial incentives. For data collectors, utility fundamentally depends on the statistical precision (e.g., the accuracy of the estimation of the true distribution of the original data having observed the privatized data) of the collected data under their privacy guarantees.

Garfinkel et al. famously quoted in [22], “Differential privacy lacks a well-developed theory for measuring the relative impact of added noise on the utility of different data products, tuning equity trade-offs, and presenting the impact of such decisions.” Indeed, in complex domains, the question of *how to appropriate the noise so that the data still remains desirably utilisable?* becomes relevant. This, in turn, leads us to many open questions along this line. For example, a) in location privacy, how non-trivial is it to optimize the three-way trade-off between the privacy of the users, their QoS, and the statistical utility of the service providers? b) in machine learning, how much *really* does federated learning help in protecting the information leakage from the shared model updates and how do the model accuracy and the formal privacy guarantees interact with the more pragmatic comprehension of a user’s data privacy (e.g., defending against attacks that are able to reconstruct the training data from the available information)? c) Does there exist a sound way to price private data and enable an efficient working of data markets? These questions underlie the need for a more principled and foundational approach to understanding the nature of the privacy-utility balance which forms the kernel of this thesis.

1.3 Contributions of this thesis

In this thesis, we examine some of the riveting questions around privacy and utility posed in the previous section with the overall goal of advancing the understanding and optimization of the privacy-utility trade-off in various domains. In particular, this work makes the following contributions:

1. We have explored the extensive privacy-preserving properties of the Blahut-Arimoto algorithm and highlighted its substantial advantages over the state-of-the-art Laplace mechanism for geo-indistinguishability and, thus, established it as a prime candidate for a location privacy-preserving mechanism. Hence, bridging some key ideas from information theory and statistics, we proposed a method allowing an incremental collection of location data as a step towards optimizing the location privacy of the data owners, their quality of service, and the statistical utility for the data consumers.
2. Due to the disproportionately fewer charging stations (CSs) as compared to the surge in

the use of electric vehicles (EVs), we addressed a fundamental problem of the risk of privacy violation for EVs dynamically querying for available CSs along their journeys. To this, we theorised the notion of *approximate geo-indistinguishability* that allows us to attain geo-indistinguishability in a strictly bounded space. Hence, we proposed an efficient privacy-preserving navigation method for EVs that protect the privacy of both their individual query locations and the overall trajectories of their journeys alongside upholding a high utility for the EVs.

3. We extended the notion of geo-indistinguishability to federated learning and proposed a method leveraging advanced techniques for model personalization and addressing user privacy concerns by formalizing privacy guarantees in terms of d -privacy. Analyzing the role of d -privacy in personalized federated learning, we demonstrated a significant improvement in group fairness under formal privacy guarantees compared to the non-personalized federated learning framework and, hence, establish that our method enhances the trade-off between privacy and fairness.
4. Continuing investigating the privacy-preserving aspects of machine learning, we provided a formal characterization of the information leakage from the shared gradient updates in federated learning and analyzed the gradient-inversion type reconstruction attacks in federated learning to understand their working from a foundational perspective.
5. Shifting towards the economics of data privacy, we proposed a framework in data markets that maximizes the data providers' financial utility while optimizing a data consumer's profit and information gain w.r.t. their financial constraints. Thereafter, we considered a federated data trading environment and proposed a method to achieve an efficient working of federated data markets for trading data under differential privacy.

1.3.1 Synopsis

The technical chapters of this thesis are divided into 3 parts:

Part II: Location Privacy focuses on the privacy-utility trade-offs in domains specifically pertaining to the sharing of location data containing sensitive information.

- **Chapter 3** illustrates that geo-indistinguishability (geo-ind), a cutting-edge standard to mitigate the privacy risks for location data, alone is insufficient to cover all privacy concerns. In particular, isolated locations are not protected by the canonical Laplace mechanism (LAP), the state-of-the-art for geo-ind. We show that the Blahut-Arimoto algorithm (BA), in addition to providing geo-ind, protects the geo-spatially isolated points. Furthermore, BA provides an optimal trade-off between information leakage and quality of service and has a better statistical utility than LAP for high privacy levels. Exploiting these properties of BA and establishing its duality with the iterative Bayesian update, an instance of the expectation-maximization method, we propose *PRIVIC*, an iterative method for a privacy-friendly incremental collection of location data.

- **Chapter 4** considers the surge in popularity of electric vehicles (EVs) and, due to the fact that the number of charging stations (CSs) is disproportionately fewer than that of the EVs in use, a rise in range anxiety and an increase in queries made during journeys to find an available CS. We introduce the notion of approximate geo-ind (AGeoI) which allows the EVs to obfuscate the individual query locations while ensuring that they remain within their preferred area of interest. It is vital because journeys are often sensitive to a sharp drop in quality of service (QoS) which incurs a high cost. The proposed method combines the application of AGeoI and the generation of dummy data to provide two-fold privacy protection (individual query locations and the trajectory of the entire journeys) for EVs while preserving a high level of utility. Moreover, our method allows for a private and precise prediction of occupancies of CSs which is crucial in unprecedented traffic congestion scenarios and efficient route planning.

Part III: Federated learning analyzes the privacy-preserving aspects of federated learning and aims to deepen the understanding of the privacy-utility trade-off in federated learning.

- **Chapter 5** explores some of the vulnerabilities of federated learning (FL) such as leakage of private information, lack of personalization of the model, and the possibility of the trained model being fairer to some groups than to others. We aim to address a triadic interaction between the personalization, privacy guarantees, and fairness achieved by the trained models under FL. Clients in FL often hold very diverse datasets representing heterogeneous communities, making it important to protect their sensitive and personal information while still ensuring that the trained model is utilisable and fair. We propose a method that incorporates a generalization of geo-ind in FL that enables personalized model training in a federated approach, provides formal group privacy guarantees locally (i.e., does not have the need for clipping or a trusted curator to add noise), and possesses significantly better group fairness measured under a variety of standard metrics than a global model trained in a classical FL template.
- **Chapter 6** analyzes the potential privacy concerns in the federated framework of model training. We show that FL, although typically perceived as a stepping stone towards privacy-preserving machine learning, is vulnerable to significant information leakage via the gradients shared by the clients. Therefore, some recent work has highlighted potential gradient inversion based attacks that can reconstruct the training data despite not having direct access to them which, in turn, defeats the whole philosophy and motivation behind the development of FL. This work aims to provide a formal characterization and an elaborate comprehension of the information leakage from the shared gradient updates in FL from a foundational perspective to analyse such data reconstruction attacks relying on gradient inversion violating the clients' privacy in FL.

Part IV: Private data trading presents the application of differential privacy in digital economics and proposes efficient and practical ways to trade private data in data markets.

- **Chapter 7** introduces a truthful price report mechanism that facilitates an accurate reporting of privacy requirements by data providers in data markets and, eventually, optimizes the data consumer's profit trading differentially private data w.r.t. their budget constraints.

- **Chapter 8** considers a model of federated data markets, i.e., data markets in which data providers, who are, generally, less influential on the market than data consumers, form federations for trading their data under differential privacy guarantees. We propose an efficient technique to price private data and a revenue-distribution mechanism to distribute the utility fairly within federations while motivating the data providers to cooperate with their respective federations, facilitating a fair and swift private data trading process.

1.3.2 Publications

The following is a list of the published papers that I have co-authored during the course of my Doctoral studies which have been used in this thesis.

1. F. Galli, K. Jung, S. Biswas, C. Palamidessi, and T. Cucinotta, “Advancing personalized federated learning: Group privacy, fairness, and beyond,” *Special Issue of Springer Nature Computer Science: Recent Trends on Information Systems Security and Privacy, 2023 (to appear)*. The work carried out in this paper has been used in Chapter 5 of this thesis.
2. S. Biswas and C. Palamidessi, “PRIVIC: A privacy-preserving method for incremental collection of location data,” in *Proceedings on Privacy Enhancing Technologies (PoPETs) – PETS 2024 (to appear)*. The work carried out in this paper has been used in Chapter 3 of this thesis.
3. F. Galli, S. Biswas, K. Jung, T. Cucinotta, and C. Palamidessi, “Group privacy for personalized federated learning,” in *Proceedings of the 9th International Conference on Information Systems Security and Privacy - ICISPP*, pp. 252–263, INSTICC, SciTePress, 2023. The work carried out in this paper has been used in Chapter 5 of this thesis.
4. S. Biswas, K. Jung, and C. Palamidessi, “An incentive mechanism for trading personal data in data markets,” in *Theoretical Aspects of Computing – ICTAC 2021* (A. Cerone and P. C. Ölveczky, eds.), (Cham), pp. 197–213, Springer International Publishing, 2021. The work carried out in this paper has been used in Chapter 7 of this thesis.
5. K. Jung, S. Biswas, and C. Palamidessi, *Establishing the Price of Privacy in Federated Data Trading*, pp. 232–250. Cham: Springer International Publishing, 2021. The work carried out in this paper has been used in Chapter 8 of this thesis.

The following is a paper that I have co-authored during the course of my Doctoral studies which have been used in this thesis and is currently under review for publication.

1. U. I. Atmaca, S. Biswas, C. Maple, and C. Palamidessi, “A privacy preserving querying mechanism with high utility for electric vehicles,” in *IEEE Open Journal of Vehicular Technology*. The work carried out in this paper has been used in Chapter 4 of this thesis.
2. S. Biswas, K. Jung, and C. Palamidessi, “Tight differential privacy guarantees for the shuffle model with k -randomized response,” in *Proceedings of the 16th International Symposium on Foundations & Practice of Security (FPS – 2023)*.
3. K. Jung, S. Biswas, and C. Palamidessi, “Optimal mechanism for private data trading in federated data market,” in *Elsevier Advanced Engineering Informatics*.

The following is a list of papers that I have co-authored during the course of my Doctoral studies which got selected for presentation at some esteemed non-archival workshops.

1. S. Biswas and C. Palamidessi, “PRIVIC: A privacy-preserving method for incremental collection of location data,” in *Theory and Practice of Differential Privacy (TPDP) 2023*. Selected for poster presentation. The work carried out in this paper has been used in Chapter 3 of this thesis.
2. F. Galli, S. Biswas, K. Jung, C. Palamidessi, and T. Cucinotta, “On the adaptive sensitivity of differentially private machine learning,” in *The Fourth AAAI Workshop on Privacy-Preserving Artificial Intelligence in Conjunction with AAAI – PPAI 2023*. Selected for poster presentation.
3. F. Galli, S. Biswas, K. Jung, T. Cucinotta, and C. Palamidessi, “Group privacy for personalized federated learning,” in *Workshop on Federated Learning: Recent Advances and New Challenges in Conjunction with NeurIPS – FL-NeurIPS 2022*. Selected as **one of the 12 amongst the 103 submissions for oral presentation**. The work carried out in this paper has been used in Chapter 5 of this thesis.

The following is a list of the published papers that I have co-authored which, although have not directly been used in this thesis, have supported and contributed immensely to the overall enhancement of my Doctoral studies.

1. S. Biswas, G. Cormode, and C. Maple, “Impact of sampling on locally differentially private data collection,” in *Competitive Advantage in the Digital Economy (CADE 2022)*, vol. 2022, pp. 64–70, 2022. Winner of the **Best Paper Award**.
2. S. Biswas, K. Jung, and C. Palamidessi, “Tight differential privacy blanket for shuff model,” in *Competitive Advantage in the Digital Economy (CADE 2022)*, vol. 2022, pp. 61–63, 2022.

1.3.3 Select Talks

1. “Group privacy for personalized federated learning” at the *Workshop on Federated Learning: Recent Advances and New Challenges in Conjunction with NeurIPS 2022 – FL-NeurIPS 2022* co-located with NeurIPS 2022 in New Orleans, USA on the 02 of December, 22.
2. “Group privacy for personalized federated learning” at the *Privacy Preserving Machine Learning (PPML) Workshop* organised by Meta in Paris, France on the 09 of November, 22.
3. “A privacy-preserving method for incremental collection of location data: Differential Privacy and beyond...” at the *7th Franco-Japanese Cybersecurity Workshop* in Tokyo, Japan on the 24th of October, 2022.
4. “Three-way optimization of privacy and utility of location data” at *Atelier sur la Protection de la Vie Privée (APVP) 2022* in Chatenay sur Seine, France on 15th of June, 2022.

2

Foundations

2.1 Standards of Privacy

One of the most successful approaches to address the privacy risks of personal data while preserving their utility is *differential privacy* (DP) which mathematically guarantees that a query output does not change significantly regardless of whether a specific personal record is in a dataset or not. DP is considered to be the gold standard of formal privacy guarantees. Its widespread applicability, uncomplicated implementation techniques, and formal properties have led to a rapid growth in the popularity and interest to study and apply DP to a variety of domains in academia and industry alike. Over time, the community has explored various variants of DP addressing privacy concerns in different contexts and under a variety of threat models.

DEFINITION 2.1.1 (Differential privacy [13, 14]). Let \mathcal{X} be a domain of secrets (e.g., personal data of users). For $\varepsilon, \delta \geq 0$, a randomizing mechanism \mathcal{R} is (ε, δ) -*differentially private* or (ε, δ) -DP if, for any pair of *adjacent*¹ datasets D_1 and D_2 on \mathcal{X} and measurable $S \subseteq \text{Range}(\mathcal{R})$, we have:

$$\mathbb{P}[\mathcal{R}(D_1) \in S] \leq e^\varepsilon \mathbb{P}[\mathcal{R}(D_2) \in S] + \delta \quad (2.1)$$

REMARK 2.1.1. When $\delta = 0$ in Equation (2.1), \mathcal{R} is said to satisfy ε -DP which is often termed as *pure* DP.

¹We recall that two datasets *adjacent* if they differ in at most one entry.

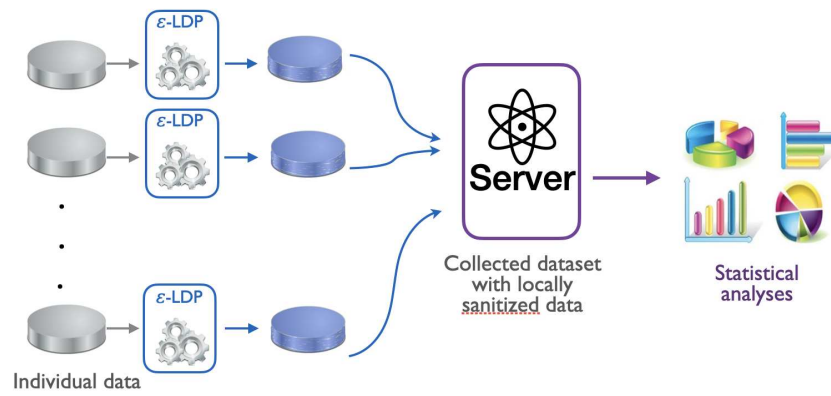


Figure 2.1: A schematic overview of the working of *local* differential privacy

The above formulation of DP is typically referred to as the *central model* of DP as the noise to the true query response is generally added by a trusted central curator before publishing or performing analytics on it. Therefore, a major drawback of the central model is that it is vulnerable to security breaches because the entire original data is stored in a central server. Moreover, there is the risk that the curator may be corrupted. Hence, a local variant of the central model has been widely popularized recently [23], where the users apply a randomizing mechanism locally on their data and send the perturbed data to the collector such that a particular value of a user's data does not have a major probabilistic impact on the outcome of the query.

DEFINITION 2.1.2 (Local differential privacy [23]). For the domain of secrets \mathcal{X} and an output set \mathcal{Y} , a randomizing mechanism \mathcal{R} satisfies ϵ -local differential privacy or ϵ -LDP if, for every $x_1, x_2 \in \mathcal{X}$ and all measurable $S \subseteq \mathcal{Y}$, we have:

$$\mathbb{P}[\mathcal{R}(x_1) \in S] \leq e^\epsilon \mathbb{P}[\mathcal{R}(x_2) \in S] \quad (2.2)$$

LDP is particularly suitable for situations where users need to communicate their personal data in exchange for some service. One such scenario is the use of location-based services where a user typically reports her location in exchange for information like the shortest path to a destination, points of interest in the surroundings, traffic information, friends nearby, etc. One of the recently popularised standards in location privacy is *geo-indistinguishability* (geo-ind) [16], which optimises the quality of service for users while preserving a generalised notion of LDP on their location data. The obfuscation mechanism of geo-ind depends on the distance between the original location of a user and a potential noisy location that they report [24, 25]. In general, *d*-privacy or *metric privacy* extends the concept of LDP to obfuscate points by capturing the essence of the distance between them. This notion of privacy is particularly useful in the context of location privacy as we will see in detail in Chapter II.

DEFINITION 2.1.3 (*d*-privacy, a.k.a. metric privacy [17]). For any space \mathcal{X} equipped with a metric $d : \mathcal{X}^2 \mapsto \mathbb{R}_{\geq 0}$ and an output space \mathcal{Y} , a randomizing mechanism $\mathcal{R} : \mathcal{X} \mapsto \mathcal{Y}$ is ϵ -*d*-private if, for every $x_1, x_2 \in \mathcal{X}$ and all measurable $S \subseteq \mathcal{Y}$, we

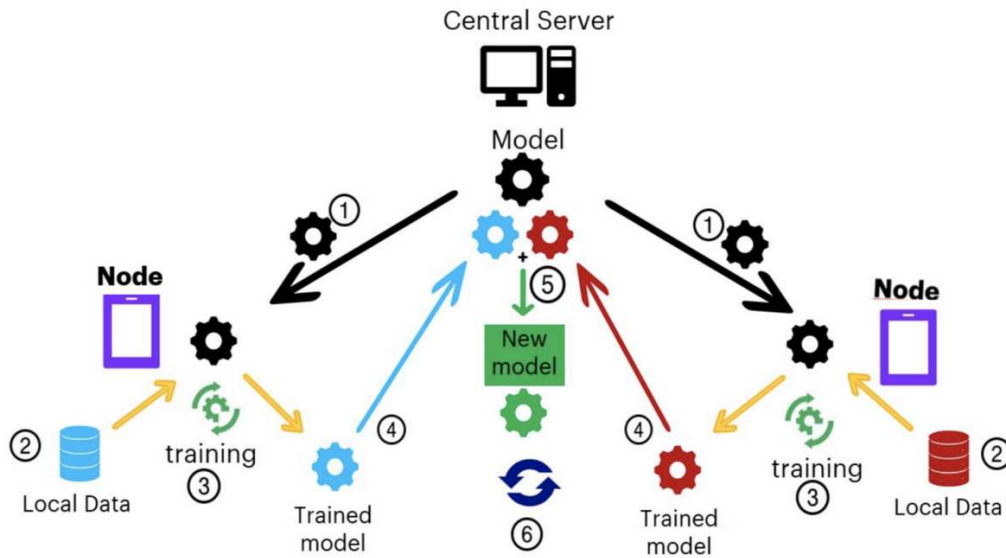


Figure 2.2: Illustration of the working of federated learning for model training

have:

$$\mathbb{P}[\mathcal{R}(x_1) \in S] \leq e^{\epsilon d(x_1, x_2)} \mathbb{P}[\mathcal{R}(x_2) \in S] \quad (2.3)$$

DEFINITION 2.1.4 (Geo-indistinguishability [16]). Setting $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$ and d as the *Euclidean metric* in Definition 2.1.3 gives rise to a cutting-edge standard for location privacy known as *geo-indistinguishability (geo-ind)*.

As a generalization of LDP, geo-ind is the cutting-edge standard for location privacy circumventing the need for a central trusted curator. It can preserve the privacy of location data amongst a set of locations with similar probability distributions without requiring a trusted third party. It provides rigorous privacy for location-based query processing and location data collection by modelling the location domain based on the Euclidean plane and capturing the essence of the ground distance between points in the domain.

Note that:

- i) If \mathcal{X} is the set of all datasets and d is the *Hamming distance* on the datasets in \mathcal{X} , we reduce down to the definition of DP.
- ii) Setting d as the *discrete metric* on any \mathcal{X} , we obtain the definition of LDP.

2.2 Federated Learning

Federated learning (FL) [26] has been in the spotlight recently as a rudimentary stepping stone towards privacy-preserving machine learning. FL is a distributed, collaborative approach to machine learning that aims to train a model without the need for the clients to share their personal data. In particular, in the classical framework of FL, a central server selects a random subset

of users and sends them a model for local optimization, following which the users locally optimize their model parameters to minimize a loss function over their own data (without sharing them with the server or any other participating client). Then these local updates are communicated back to the server which aggregates them to furnish the global update to be used by the next round's participating clients. This process is repeated until convergence. Figure 2.2 demonstrates a schematic outline of the working of FL. Despite being initially perceived as a foolproof solution to defend against major privacy concerns in machine learning, the community has exposed multiple vulnerabilities in FL (discussed and elaborated in Part III of the thesis). Therefore, there has been a surge of recent research that focuses on combining DP with FL in the quest to achieve a secure and private way of model training [27–30].

2.3 Notions of Utility

It is not preposterous to intuitively believe that, in most cases, privacy and utility of data stand against each other. As the primary goal of this thesis is to analyze this trade-off between privacy and utility in various contexts, it is important to quantify the corresponding application-specific concept of *utility*. Therefore, there has been a wide range of notions of utility considered in this thesis to incorporate the variety of use cases that we take into account. However, one of the fundamental and widely analyzed ideas to measure utility under differentially private (and, in general, noisy) data takes into account the level of statistical accuracy preserved by the data after being privatized. This is regarded as the statistical utility of the sanitized datasets and is often assessed by considering the difference between the distribution of the raw (non-privatized) data and that of the sanitized one.

While for the central model of DP, the “best guess” for the true distribution by observing the noisy dataset is the (normalized) noisy histogram itself, the problem of approximating the distribution of the original data from observing the noisy sample under local DP is slightly more non-trivial. To this purpose, two of the standard techniques to “de-noise” the locally obfuscated data, with the knowledge of the underlying privacy mechanism, to estimate true distribution are the methods of *matrix inversion* [31, 32] and *iterative Bayesian update (IBU)* [33]. IBU, an iterative method of expectation maximization, has been shown to be more robust and effective than the method of matrix inversion [34]. Amongst the diverse notions of utility under local privatizing methods (e.g., LDP, geo-indistinguishability, etc.) that have been considered in this thesis, one of the most fundamental ones, emphasized especially in the context of location privacy in Chapter II, have been comparing the distributions of the original data and that estimated by de-noising the privatized data using IBU and this notion is regarded as the statistical utility of the service providers of the privacy-preserving location-based services. The other diverse quantifications of utility (e.g., quality of service of users, accuracy in machine learning-related tasks, financial utility of parties involved in private data trading, etc.) have been introduced in the relevant chapter-specific preliminaries.

DEFINITION 2.3.1 (Full-support probability distribution). Let θ be a probability

distribution defined on the space \mathcal{X} . θ is a *full-support* distribution on \mathcal{X} if $\theta(x) > 0$ for every $x \in \mathcal{X}$.

DEFINITION 2.3.2 (Iterative Bayesian update [33]). Let C be a privacy mechanism that locally obfuscates points from the discrete space \mathcal{X} to \mathcal{Y} such that $C_{xy} = \mathbb{P}[y|x]$ for all $x, y \in \mathcal{X}, \mathcal{Y}$. Let X_1, \dots, X_n be i.i.d. random variables on \mathcal{X} following some PMF $\pi_{\mathcal{X}}$. Let Y_i denote the random variable of the output when X_i is obfuscated with C .

Let $\bar{y} = \{y_1, \dots, y_n\}$ be a realisation of $\{Y_1, \dots, Y_n\}$ and \mathbf{q} be the empirical distribution obtained by counting the frequencies of each y in \bar{y} . The *iterative Bayesian update (IBU)* estimates $\pi_{\mathcal{X}}$ by converging to its maximum likelihood estimate (MLE) with the knowledge of \mathbf{q} and C . IBU works as follows:

1. Start with any full-support PMF θ_0 on \mathcal{X} .
2. Iterate $\theta_{r+1}(x) = \sum_{y \in \mathcal{Y}} \mathbf{q}(y) \frac{\theta_r(x) C_{xy}}{\sum_{z \in \mathcal{X}} \theta_r(z) C_{zy}}$ for all $x \in \mathcal{X}$.

The convergence of IBU has been studied in [33, 34]. For a given set of observed locations, the limiting estimate $\hat{\pi}_{\mathcal{X}} = \lim_{r \rightarrow \infty} \theta_r$ is well-defined by the privacy mechanism in use, C , and the empirical distribution of the noisy locations, \mathbf{q} . We will functionally denote $\hat{\pi}_{\mathcal{X}}$ as $\text{IBU}(\mathbf{q}, C)$.

DEFINITION 2.3.3 (Earth mover's distance [35]). Let π_1 and π_2 be PMFs defined over a discrete space of locations \mathcal{X} . For a metric $d: \mathcal{X}^2 \mapsto \mathbb{R}_{\geq 0}$, the *earth mover's distance (EMD)* (aka the *Kantorovich–Wasserstein metric*) is defined as

$$\text{EMD}(\pi_1, \pi_2) = \min_{\mu \in \Pi(\pi_1, \pi_2)} \sum_{x, y} \mu(x, y) d(x, y)$$

where $\Pi(\pi_1, \pi_2)$ is the set of all joint distributions over \mathcal{X}^2 such that for any $\eta \in \Pi(\pi_1, \pi_2)$, $\sum_{x \in \mathcal{X}} \eta(x_0, x) = \pi_1(x_0)$ and $\sum_{x \in \mathcal{X}} \eta(x, x_0) = \pi_2(x_0)$ for every $x_0 \in \mathcal{X}$.

EMD is considered a canonical way to lift a distance on a certain domain to a distance between distributions on the same domain.

DEFINITION 2.3.4 (Statistical utility). Let C be a privacy mechanism that obfuscates data on the discrete space \mathcal{X} . Let $\pi_{\mathcal{X}}$ be the PMF of the original locations and let $\hat{\pi}_{\mathcal{X}}$ be its estimate by IBU. Then we define the *statistical utility* of the mechanism C as $\text{EMD}(\hat{\pi}_{\mathcal{X}}, \pi_{\mathcal{X}})$.

Part II

Location privacy

"If self is a location, so is love."

– Seamus Heaney, *The Aerodrome*

3

PRIVIC: A privacy-preserving method for incremental collection of location data

3.1 Introduction

As the need and development of various kinds of research and analysis using personal data are becoming more and more significant, the risk of privacy violations of sensitive information of the data owners is also increasing manifold. One of the most successful proposals to address the issue of privacy protection is *differential privacy (DP)* [13, 14], a mathematical property that makes it difficult for an attacker to detect the presence of a record in a dataset. This is typically achieved by answering queries performed on the dataset in a (controlled) noisy fashion. Lately, the *local variant of differential privacy (LDP)* [23] has gained popularity due to the fact that the noise is applied at the data owner's end without needing a trusted curator. LDP is particularly suitable for situations where a data owner is a user who communicates her personal data in exchange for some service. One such scenario is the use of location-based services (LBS), where a user typically sends her location in order to obtain information like the shortest path to a destination, nearby points of interest, traffic information, etc. The security and the convenience of implementing the local model directly on a user's device (tablets, smartphones, etc.) make LDP very appealing.

Typically, in exchange for their service, providers incrementally collect their users' data and then make them available to other parties which process them to provide useful statistics to companies and institutions. Obviously, the statistical precision of the collected data is essential for the quality of the analytics performed (*statistical utility*). However, injecting noise locally

into the data to protect the privacy of the users usually has a negative effect on the statistical utility. Additionally, the noise degrades the *quality of service* (QoS) as well, since, obviously, the service results from the elaboration of the information received.

Substantial research has been done to address the privacy-utility trade-off in the context of DP. In LDP, the primary focus has been to optimize the utility from the data collector’s perspective, i.e., devising mechanisms and post-processing methods that would allow deriving the most accurate statistics from the collection of the noisy data [23, 36]. In contrast, in domains such as location privacy, the focus usually has been on optimizing the QoS, i.e., the utility from the point of view of the users. In particular, this is the case for the framework proposed by Shokri et al. [37, 38].

We argue that it is important to meet the interest of all parties involved, and hence to consider both kinds of utility at the same time. Hence, the first goal of this paper is to develop a *location-privacy preserving mechanism* (LPPM) that, in addition to providing formal location-privacy guarantees, preserves as much as possible *both* the statistical utility and the QoS.

One may think that statistical utility and QoS are aligned since they both benefit from preserving as much original information as possible under the privacy constraint. However, this is not true in general: the optimization of statistical utility does not necessarily imply a significant improvement in the QoS, nor vice-versa. A counterexample is provided by Example 3.6.1 in Section 3.6. Hence, the preservation of both statistical utility and QoS is trickier than it may appear at first sight.

Geo-indistinguishability (geo-ind) [16], one of the most popular and widely used approaches to protect location-privacy, essentially obfuscates locations based on the distance between them. This idea works particularly well for protecting the precision of the location as it ensures that an attacker would not be able to differentiate between points that are close on the map by observing the reported noisy location. At the same time, it does not inject an enormous amount of noise that would be necessary to make far-away locations indistinguishable. Moreover, geo-ind has been shown to formally satisfy the basic sequential compositionality theorem [39], just like DP and its local variant. Although this approach of distance-based obfuscation seems enticing at a first glance, one of the issues it poses is that it may leave the geo-spatially isolated locations vulnerable, i.e., identifiable despite being formally geo-indistinguishable [40]. To improve the situation, [40] introduced the notion of *elastic distinguishability metrics*, which essentially leads to injecting more noise when the location to protect is isolated.

The *Blahut-Arimoto algorithm* (BA) [41, 42] from rate-distortion theory (a branch of information theory) Pareto-optimizes the trade-off between *mutual information* (MI) and average distortion. This property is appealing in the context of privacy because MI is often considered a measure of information leakage and average distortion is a commonly used metric for quantifying QoS. Moreover, BA was proven to satisfy geo-ind in [43] opening a door to study it as a potential LPPM. In this paper, we start off by exploring the privacy-preserving properties of BA and comparing them with those of the *Laplace mechanism* (LAP) [16] which is considered as the state-of-the-art mechanism for geo-ind. We show that, besides geo-ind, BA provides an elastic distinguishability metric and, hence, protects even the most isolated points in the map, unlike

LAP. We then examine the statistical utility, focusing on the estimation of the most general statistical information, namely the distribution of the original location data (true distribution). The “best” estimation is known in statistics as the *maximum likelihood estimation* (MLE), and can be computed using the *iterative Bayesian update* (IBU) [33], an instance of the *expectation maximization* (EM) method. We discover a duality between BA and IBU, which in our opinion is quite intriguing, because BA and IBU were developed in different contexts, using different concepts and metrics, and for completely different purposes. We prove experimentally that the statistical utility of BA is very good, i.e., the MLE is very close to the true distribution. We conjecture that this is probably due to the duality between the mechanism that injects the noise (BA) and the one that de-noises the noisy data (IBU). In any case, the experiments show that the statistical utility of BA outperforms that of LAP for high levels of privacy, eventually becoming comparable as the level of privacy decreases.

One important point to note is that BA requires the knowledge of the original distribution to provide the optimal mechanism. When it is fed with only an approximation of the distribution, it only provides an approximated result. We acknowledge that the distribution of the original data is usually off-limits and, even when available, it typically gets outdated over time. In any case, we can soundly assume that it is not available because it is essentially the reason for collecting the data. Hence we have a vicious circle: we want to collect data in a privacy-friendly fashion to estimate the original distribution while wanting to use a privacy mechanism that requires knowing a good approximation of the original distribution. Motivated by this dilemma, we propose PRIVIC, an incremental data collection method providing extensive privacy protection for the users of LBS’s, while retaining a high utility for both them and the service providers, and ensuring that both parties, acting in their best interest, would benefit from the end mechanism.

Finally, we prove formally the convergence of PRIVIC to the true distribution, and illustrate empirically the privacy-utility trade-off of our method. The experiments also demonstrate the efficacy of combining BA and IBU, in that the estimation of the original distribution is very accurate, especially when measured using a notion of distance between distributions compatible with the ground distance used to measure the QoS (e.g., the Earth Mover’s distance). All the experiments were performed using real location data from the Gowalla dataset for Paris and San Francisco.

Contributions

The key contributions in this chapter are:

1. We show, analytically and with experiments on real datasets, that the BA mechanism, in addition to geo-ind, provides an elastic distinguishability metric. As such, it protects the privacy of isolated locations, which the standard LAP for geo-ind fails at.
2. We prove that BA produces an invertible mechanism, which means that the MLE is unique. This is crucial to prove that the IBU always converges to the true distribution and that, therefore, we can get a good statistical utility.

3. We establish a duality between BA and IBU, thus demonstrating a connection between rate-distortion theory and the expectation-maximization method from statistics.
4. We show experimentally that BA provides a better statistical utility than LAP for high levels of privacy, eventually becoming comparable as the level of privacy decreases.
5. Since the construction of the optimal BA requires precise knowledge of the true distribution, we propose an iterative method (PRIVIC) that alternates between BA and IBU, thus getting a better and better estimation of the true distribution as more (noisy) data get collected. We show, both formally and with experiments on real location datasets, that PRIVIC converges to the true distribution. In summary, PRIVIC produces a geo-indistinguishable LPPM with an elastic distinguishability metric, which optimizes the trade-off with the QoS and provides high statistical utility.
6. We investigate the effect on the privacy guarantees of our method by considering adversarial users who report their locations falsely to compromise the privacy of the isolated locations in the map.

3.2 Technical preliminaries

Aside from the formal guarantees of DP and its variants as introduced in Section 2.1, in this chapter, we consider an additional notion of quantifying and measuring data privacy from an information theoretical perspective using *mutual information (MI)*. As such, information theoretical MI has often been explored and used in the context of privacy (and security) in the literature and we have discussed some of the key works in this area in Section 3.3.

DEFINITION 3.2.1 (Mutual information[44]). Let (X, Y) be a pair of random variables defined over the discrete space $\mathcal{X} \times \mathcal{Y}$ such that μ is the joint *probability mass function* (PMF) of X and Y , and p_X and p_Y are the marginal PMFs of X and Y , respectively, and $p_{X|Y}$ is the conditional probability of X given Y . Then the (Shannon) *entropy* of X , $H(X)$, is defined as $H(X) = - \sum_{x \in \mathcal{X}} p_X(x) \log p_X(x)$. The *residual entropy* of X given Y is defined as $H(X|Y) = \sum_{y \in \mathcal{Y}} p_Y(y) H(X|Y = y) = - \sum_{y \in \mathcal{Y}} p_Y(y) \sum_{x \in \mathcal{X}} p_{X|Y}(x|y) \log p_{X|Y}(x|y)$, and, finally, the *mutual information (MI)* is given by:

$$I(X|Y) = H(X) - H(X|Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mu(x, y) \log \frac{\mu(x, y)}{p_X(x)p_Y(y)} \quad (3.1)$$

In addition to the concept of statistical utility for the service providers as introduced in Chapter 2.3, in this chapter, we also consider the *Quality of Service (QoS)* as a measure of utility for the users of location-based services.

DEFINITION 3.2.2 (Quality of Service). For discrete spaces, \mathcal{X} and \mathcal{Y} , let $d: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ be any distortion metric (a generalization of the notion of distance). Let X be a random variable on \mathcal{X} with PMF p_X and C be any randomizing mechanism where C_{xy} is the probability of x being mapped to y by C . We define the *quality of service (QoS)* of X for C as the *average distortion w.r.t. d* , given as:

$$\text{Avg}D(X, C, d) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_X(x) C_{xy} d(x, y)$$

Now we introduce the *Blahut-Arimoto algorithm (BA)* [41, 42] that was initially developed in 1972 to present an iterative method for optimizing the information-theoretical channel capacity w.r.t. a given constraint on the average distortion. In this work, we shall resort to BA to optimize the information theoretical notion of privacy (Definition 3.2.1) and the QoS (Definition 3.2.2) of the users sharing their locations and explore its extensive location privacy-preserving properties.

DEFINITION 3.2.3 (Blahut-Arimoto algorithm [41, 42]). Let X be a random variable on the discrete space \mathcal{X} with PMF π_X and $\mathcal{C}(\mathcal{X}, \mathcal{Y})$ be the space of all mechanisms encoding \mathcal{X} to \mathcal{Y} . For a distortion $d: \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ and fixed $d^* \in \mathbb{R}^+$, we wish to find the mechanism $\hat{C} \in \mathcal{C}(\mathcal{X}, \mathcal{Y})$ that minimizes MI given the bound d^* on distortion:

$$\hat{C} = \underset{\substack{C \in \mathcal{C}(\mathcal{X}, \mathcal{Y}) \\ \text{Avg}D(X, C, d) \leq d^*}}{\text{argmin}} I(X|Y_{X,C})$$

where, for any $C \in \mathcal{C}(\mathcal{X}, \mathcal{Y})$, $Y_{X,C}$ is the random variable on \mathcal{Y} denoting the output of the encoding of X . The *Blahut-Arimoto algorithm (BA)* provides an iterative method to construct \hat{C} as follows:

1. Start with any full-support PMF c_0 on \mathcal{X} and any $C^{(0)}$.
2. Iterate:

$$C_{xy}^{(t+1)} = \frac{c_t(y) \exp\{-\beta d(x, y)\}}{\sum_{z \in \mathcal{Y}} c_t(z) \exp\{-\beta d(x, z)\}} \quad (3.2)$$

$$c_{t+1}(y) = \sum_{x \in \mathcal{X}} \pi_X(x) C_{xy}^{(t+1)} \quad (3.3)$$

where $\beta > 0$ is the negative of the slope of the *rate-distortion* function $RD(X, d^*) = \min_{C \in \mathcal{C}(\mathcal{X}, \mathcal{Y})} I(X|Y_{X,C})$ under $\text{Avg}D(X, C, d) \leq d^*$. We call β the *loss parameter*, capturing the role of d^* in BA.

REMARK 3.2.1. The equations (3.2) and (3.3) above define two transformations $\mathcal{F}: \mathcal{D}(\mathcal{X}) \rightarrow \mathcal{C}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{G}: \mathcal{C}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathcal{D}(\mathcal{X})$, where $\mathcal{D}(\mathcal{X})$ is the space of distributions on \mathcal{X} , so that $C^{(t+1)} = \mathcal{F}(c_t)$ and $c_{t+1} = \mathcal{G}(C^{(t+1)})$.

REMARK 3.2.2. In [45], Csiszár proved the convergence of BA when \mathcal{X} is finite. The limit $\lim_{n \rightarrow \infty} (\mathcal{F} \circ \mathcal{G})^n(C^{(0)})$ is the optimal mechanism \hat{C} (parametrized by β), and it is uniquely determined by the prior π_X and by the initial PMF c_0 . Note

that \hat{C} is a fixpoint of $\mathcal{F} \circ \mathcal{G}$, i.e. $\hat{C} = (\mathcal{F} \circ \mathcal{G})(\hat{C})$, and that $\hat{c} = \mathcal{G}(\hat{C})$ is a fixpoint of $\mathcal{G} \circ \mathcal{F}$.

REMARK 3.2.3. In [43], Oya et al. proved that, when d is the Euclidean metric, the mechanism \hat{C} obtained from BA with loss parameter β satisfies 2β -geo-ind.

Finally, to conclude the technical preliminaries relevant to this chapter, we recall a generalization of IBU from the literature that we use in this work. Generalized IBU (GIBU) [46] applies IBU in parallel to several empirical distributions derived from the application of (possibly different) obfuscation mechanisms to various sets of samples from the same distribution.

DEFINITION 3.2.4 (Generalized iterative Bayesian update [46]).

Let $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$, with $\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_n^{(t)})$ for every $t \in \{1, \dots, N\}$, be N datasets s.t. the entries $x_i^{(t)}$ for each $i \in \{1, \dots, n\}$ are i.i.d. samples from the discrete space \mathcal{X} following the probability distribution $\pi_{\mathcal{X}}$. Let $C^{(1)}, \dots, C^{(N)}$ be N privacy mechanisms that locally obfuscate points from \mathcal{X} to \mathcal{Y} such that the mechanism $C^{(t)}$ is applied to the dataset $\mathbf{x}^{(t)}$ and $C_{xy}^{(t)} = \mathbb{P}[y|x]$ for all $x, y \in \mathcal{X}, \mathcal{Y}$ and $i \in \{1, \dots, N\}$. Denoting the random variable of the output when $x_i^{(t)}$ is obfuscated with $C^{(t)}$ as $Y_i^{(t)}$, let $(y_1^{(t)}, \dots, y_n^{(t)})$ be a realisation of $(Y_1^{(t)}, \dots, Y_n^{(t)})$ for every $t \in \{1, \dots, N\}$.

Let $\mathcal{G} = \begin{pmatrix} C^{(1)} & \dots & C^{(N)} \end{pmatrix}$ be referred to as the *combined mechanism* a.k.a. the *output probability matrix* satisfying:

$$\mathcal{G}(x, y_i^{(t)}) = \mathbb{P}[y_i^{(t)} | x] = C_{x, y_i^{(t)}}^{(t)}$$

$$\forall x \in \mathcal{X}, i \in \{1, \dots, n\}.$$

GIBU estimates $\pi_{\mathcal{X}}$ by converging to the maximum likelihood estimate (MLE) of $\pi_{\mathcal{X}}$ with the knowledge of the noisy data and the obfuscating channels. GIBU works as follows:

1. Start with any full-support PMF θ_0 on \mathcal{X} .
2. Iterate $\theta_{r+1}(x) = \frac{1}{Nn} \sum_{t=1}^N \sum_{i=1}^n \frac{\theta_r(x) \mathcal{G}(x, y_i^{(t)})}{\sum_{z \in \mathcal{X}} \theta_r(z) \mathcal{G}(z, y_i^{(t)})}$ for all $x \in \mathcal{X}$.

Setting $\hat{\pi}_{\mathcal{X}} = \lim_{r \rightarrow \infty} \theta_r$ and $\mathbf{y}^t = (y_1^{(t)}, \dots, y_n^{(t)})$, let $\hat{\pi}_{\mathcal{X}}$ (the MLE of the prior obtained with GIBU) be functionally denoted by:

$$\text{GIBU} \left(\left(C^{(1)}, \mathbf{y}^{(1)} \right), \dots, \left(C^{(N)}, \mathbf{y}^{(N)} \right) \right).$$

In the context of the location-privacy, as addressed in this work, we obfuscate the original locations to points in the same space and, hence, in the rest of the chapter we consider the spaces of the secrets and the noisy locations to be the same, i.e., $\mathcal{X} = \mathcal{Y}$.

3.3 Related Work

The trade-off between privacy and utility has been widely studied in the literature [47, 48]. Optimization techniques for DP and utility for statistical databases have been analyzed by the community from various perspectives [49–51]. There have been works focusing on devising privacy mechanisms that are optimal to limit the privacy risk against Bayesian inference attacks while maximizing the utility [37, 38]. In [43], Oya et al. examine an optimal LPPM w.r.t. various privacy and utility metrics for the user.

In [52], Oya et al. consider the optimal LPPM proposed by Shokri et al. in [37] which maximizes a notion of privacy (the *adversarial error*) under some bound on the QoS. The construction of the optimal LPPM requires the knowledge of the original distribution, and [52] uses the EM method to estimate it and design *blank-slate models* empirically shown to outperform the traditional hardwired models. However, a problem with their approach is that there may exist LPPMs that are optimal in the sense of [37], but with no statistical utility, see Example 3.6.1 in Section 3.6. Furthermore, for the mechanisms considered in [52] the EM method may fail to converge to the true distribution. Indeed, [34] points out various mistakes in the results of [33], on which [52] intrinsically relies to prove the convergence of their method.

[53] proposed a method for generating privacy mechanisms that tend to minimize mutual information using a machine learning based approach. However, this work assumes the knowledge of the exact prior from the beginning, unlike ours. Moreover, [53] does not provide formal guarantees for location-privacy (e.g., geo-ind) which is one of the main aspects captured by our work. In [54], Zhang et al. consider the Blahut-Arimoto algorithm in the context of location-privacy. However, their proposed method also requires the knowledge of the prior distribution to construct the LPPM. Additionally, [54] focuses on measuring privacy for the trace of a single user. On the contrary, our notion of privacy assumes the collection of single check-ins (or check-ins separated in time) by a set of users.

The Laplace mechanism has been rigorously studied in the literature in various scenarios as the cutting-edge standard to achieve geo-ind [16, 39, 55] and has been proven to be optimal for one-dimensional data w.r.t. Bayesian utility [25]. Despite its wide popularity, it has been recently criticized due to its limitation to protect geo-spatially isolated points from being identified by adversaries [40]. The authors of [40] addressed this concern by proposing the idea of *elastic distinguishability metrics*.

Our work also considers mutual information (MI) as an additional privacy guarantee. MI and its closely related variants (e.g. conditional entropy) have been shown to nurture a compatible relationship with DP [56]. [57] has provided an operational interpretation of MI in terms of an attacker model. MI essentially measures the correlation between observations and secrets, and its use as a privacy metric is widespread in the literature. Some key examples are: gauging anonymity [58, 59], estimating privacy in training machine learning models with a typical cross-entropy loss function [53, 60–62], and assessing location-privacy [43].

Alongside the widespread interest in studying MI as an information theoretical notion of privacy by the community, some researchers have strongly criticized the use of Shannon entropy

and MI as measures of privacy (e.g., [63]). We do not take sides in this controversy: for us (and specifically in this work), MI is only a means to construct a mechanism that provides geo-ind under an elastic metric, which is our reference privacy notion.

A popular choice of utility metric for the users is the *average distortion*, which quantifies the expected quality loss of the service due to the noise induced by the mechanism. Such a metric has gained the spotlight in the community [16, 37, 64–66] due to its intuitive and simple nature. On the other hand, a standard notion of statistical utility for the data consumer is the precision of the estimation of the distribution on the original data from that of the noisy data. Iterative Bayesian update [31, 33] provides one of the most flexible and powerful estimation techniques and has recently become in the focus of the community [34, 46].

Incremental and privacy-friendly data collection has been explored both in the context of k -anonymity [67–69] and DP [70, 71]. However, to the best of our knowledge, the problem of providing a rather robust privacy guarantee while preserving utility for both data owners and data consumers has not been addressed by the community so far.

3.4 Location-privacy with the Blahut-Arimoto algorithm

Definition 3.2.3 shows that the BA mechanism optimizes between MI and average distortion, which is a standard choice for measuring QoS. Furthermore, Remark 3.2.3 formally links the mechanism produced by BA with geo-ind, which is our reference privacy notion.

In this section, we investigate the privacy protection offered by BA beyond geo-ind, study the statistical utility it renders, and compare it with LAP, the canonical mechanism for geo-ind.

3.4.1 Elastic location-privacy with BA

One of the concerns harboured by geo-ind is that it treats the space in a uniform way, thus making isolated locations vulnerable to an attacker that knows the prior distribution. This issue has been raised and addressed by Chatzikokolakis et al. in [40] where the authors introduce a variant of LAP based on an *elastic distinguishability metrics*, which they refer to as *elastic mechanisms*. Such mechanisms obfuscate locations not only by considering the Euclidean distance between them but also by taking into account an abstract attribute of the reported location, called *mass*, which is a parameter of the definition.

Formally, if $\mathcal{R}_{\text{elas}}$ is an elastic mechanism with privacy parameter ε defined on \mathcal{X} , then, for all $x, y \in \mathcal{X}$, $\mathcal{R}_{\text{elas}}$ must satisfy:

$$\mathbb{P}[\mathcal{R}_{\text{elas}}(x) = y] \propto \exp\{-\varepsilon d_E(x, y)\} \quad (3.4)$$

$$\mathbb{P}[\mathcal{R}_{\text{elas}}(x) = y] \propto q(y) \quad (3.5)$$

where q is the probability distribution of the reported locations.

Note that Equations 3.4 and 3.5 characterize the properties of an elastic mechanism $\mathcal{R}_{\text{elas}}$, but they *do not define what $\mathcal{R}_{\text{elas}}$ exactly is, as a function*. In fact, as a definition, Equation 3.5 would be circular, since it uses the probability mass q generated by $\mathcal{R}_{\text{elas}}$ without knowing what $\mathcal{R}_{\text{elas}}$ is. As we will see, BA solves this problem by constructing the mechanism $\mathcal{R}_{\text{elas}}$ as a fix-point of a recursive process starting from a uniform output distribution q . (To be precise the process is mutually recursive, alternating the generation of a new mechanism and a new output distribution, that, in turn, is fed into BA to generate the mechanism at the next step.)

$\mathcal{R}_{\text{elas}}$, unlike LAP, protects a point in a densely populated area (e.g. city) and a geo-spatially isolated point (e.g. island) differently by considering not only the ground distance between the true and the reported locations but also the mass of the reported location. The exact mechanism depends of course on how we define the notion of mass. A natural way, and the most meaningful from the privacy point of view, is to set the mass of y to be the probability to be reported (from any true location x). Under this definition, the interpretation of (3.5) is in the spirit of obtaining privacy by ensuring that the set of possible true locations (given the reported one) is large. In other words, given a true location x , we tend to report with higher probability those locations y that are reported with high probability from other locations as well so that it becomes harder to re-identify x as the original one. Note that this property is not incompatible with the geo-ind guarantee. However, LAP does not provide it.

Obviously, the definition of mass as the probability to be reported would be circular, because it would depend on the mechanism, which in turn is defined in terms of the mass. The authors of [40] do not explain how this mechanism could be constructed. Fortunately, the following theorem shows that an elastic mechanism of this kind can be constructed using BA. The proof is provided in Appendix A.

THEOREM 3.1. The privacy mechanism generated by BA produces an elastic location-privacy mechanism.

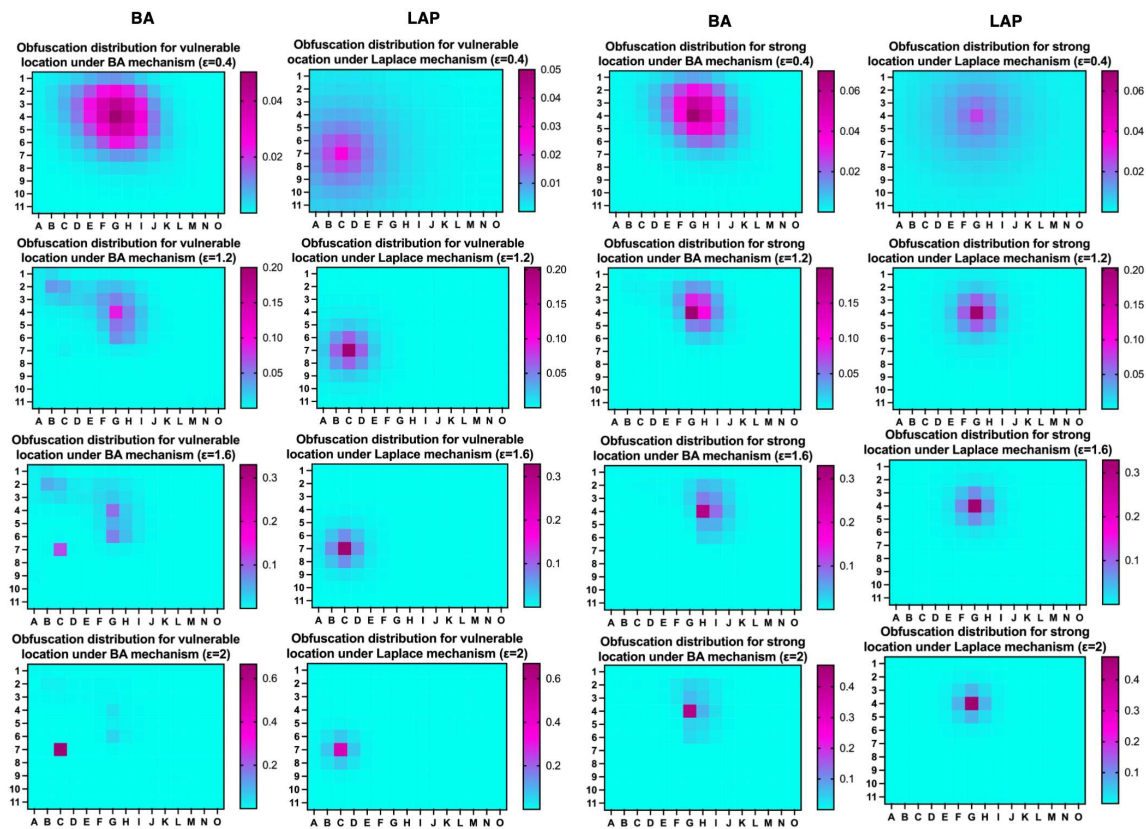
Note also that there can be many mechanisms satisfying (3.4) and (3.5) (also with the mass interpreted as probability). The one produced by the BA is the mechanism that offers the best QoS among these. Finally, a consequence of the connection with BA is that it provides an understanding of the elastic mechanism in terms of information theory and of the attacker illustrated in the previous section.

Experimental validation

Having furnished the theoretical foundation, we now enable ourselves to empirically validate that BA, indeed, satisfies the properties of the elastic mechanism unlike LAP, its state-of-the-art geo-indistinguishable counterpart. We perform experiments using real location data from the Gowalla dataset [72, 73]. We consider 10,078 Gowalla check-ins from a central part of Paris bounded by latitudes (48.8286, 48.8798) and longitudes (2.2855, 2.3909) covering an area of 8Km×6Km discretized with a 16×12 grid.

In order to demonstrate the property of an elastic mechanism, we artificially introduced an

“island” amidst the locations in Paris by choosing a grid A in a low-density area of the dataset (in the south-west region), assigning the probability mass of the grids around A to 0, and dumping this cumulative mass from the surrounding region to A , ensuring that the sum of the probability masses of all the grids remains to be 1. We call A as a *vulnerable location* in the map as it is isolated from the crowded area. To visualize the elastic behaviour of the mechanisms for locations in crowded regions, we consider another grid B in the central part of the map which has a high probability mass and has a highly populated surrounding – we refer to such a grid B as a *strong location* in the map. Figure 3.1 illustrates the selection of vulnerable and strong locations in the Paris dataset.



(a) Reporting distribution of the vulnerable location A (b) Reporting distribution of the strong location B

Figure 3.2: Distribution of privatizing the vulnerable and the strong locations for different levels of privacy. Top-down, the rows illustrate the results for $\varepsilon = 0.4, 1.2, 1.6, 2$, respectively.

For the mechanism derived from BA with a loss parameter β , we know, by Remark 3.2.3, that the privacy parameter ε is 2β , which we use to tune the privacy level of LAP in order to compare the two mechanisms under the same level of geo-ind. Figure 3.2 illustrates the probability distribution of reporting a privatized point on the map by obfuscating the vulnerable and the strong locations with different levels of geo-ind – we vary the value of ε to be 0.4, 1.2, 1.6, 2.

By comparing with the distribution of the true locations in Paris given by Figure 3.1, we observe that when the value of ε is low (privacy is high), the reported location with BA is likely to be mapped to a nearby densely populated place. For example, with $\varepsilon = 0.2$, the highest level

of privacy considered in the experiments, the location reported by BA will most probably be around the most crowded region of Paris. As ε increases, the location most likely to be reported by BA systematically moves to a densely populated region closer and closer to the true vulnerable location. LAP, on the other hand, always obfuscates every location around its true position in the map – varying the value of ε changes the spread of the distribution around the true location. As explained in the introduction, this might be problematic as the vulnerable location is known to be isolated and, hence, even being reported somewhere nearby would potentially result in its re-identification.

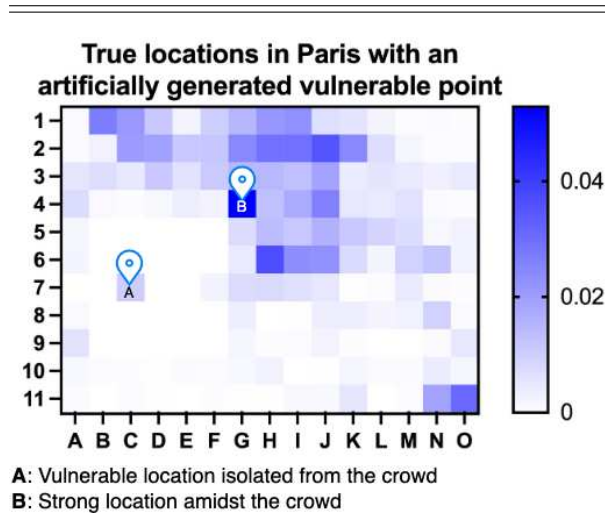


Figure 3.1: Gowalla check-in locations in Paris with an artificially planted vulnerable point, A, in isolation, and a strong point, B, in a crowded area.

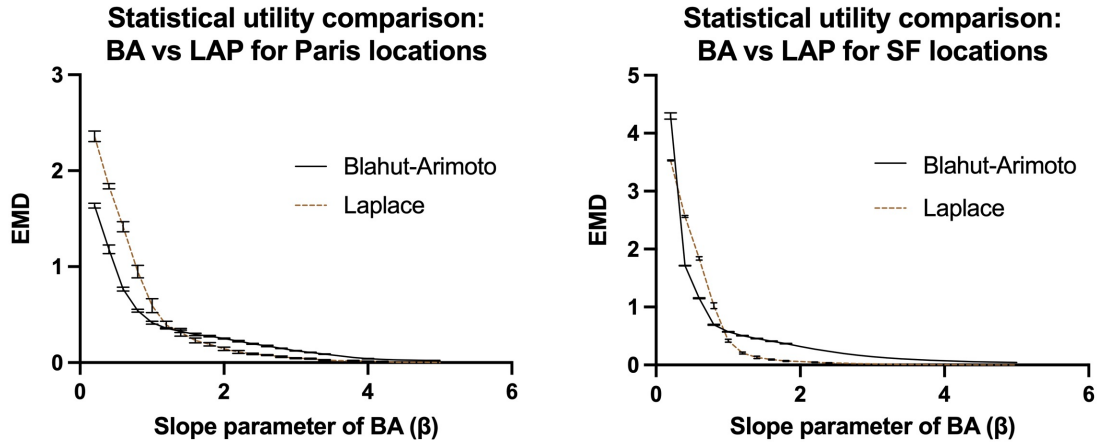
For example, we would like to highlight the setting of $\varepsilon = 1.6$ for the vulnerable location to show that the distribution of the location reported by LAP is almost completely around the true vulnerable point covering an area that is deserted, i.e., there is no realistic chance of someone being located in that region. Thus, despite providing formal 1.6-geo-ind, LAP fails to protect such a vulnerable location from being potentially identified. BA, on the other hand, does the job quite well, adhering to the principles of the elastic mechanism – it distributes the reported location in the crowded areas nearby providing a sense of camouflage amidst the many possibilities, in addition to 1.6-geo-ind.

In the case of privatizing the strong location, Figure 3.2b shows that both BA and LAP behave similarly by concealing the point around its true position. This would not give rise to a similar issue as for the vulnerable location because, by definition, the strong location B is already positioned in a highly dense region of the map and, hence, being privatized, it will still remain among the crowd with a high probability.

Focusing on the utility of individual users, we note that due to theories from Nash equilibrium [74] and Hotelling’s spatial competition [75], a huge fraction of the typical points of interest (POIs) like cinemas, theatres, restaurants, retails, etc. lie in crowded areas syncing with the distribution of population. Therefore, for an isolated point in the map that is located in some extremely unpopulated area (e.g. some forest or island far from the city), the closest POI is usually going to be in the nearest urban region, i.e., a region on the map with a high density of population. Suppose A is one such isolated location and let A_{BA} and A_{LAP} be the reported locations for A obfuscated with BA and LAP, respectively. Due to the elastic property of BA, A_{BA} is likely to be at a nearby crowded location to A, while A_{LAP} is likely to be around the true location A. Let P_{BA} and P_{LAP} be the nearest POIs from the reported locations A_{BA} and A_{LAP} , respectively. The most likely scenario is that P_{BA} and P_{LAP} are almost at a similar place under the assumption that typical POIs follow the distribution of the crowd and, therefore, a vulnerable user has to

travel a similar distance from their true position in both the cases, except that under LAP, the privacy of A will be compromised much more than that under BA.

3.4.2 Statistical utility: BA vs LAP



(a) Statistical utility for BA and Laplace on locations in Paris (b) Statistical utility for BA and Laplace on locations in SF

Figure 3.3: Statistical utility in terms of *earth mover's distance* (EMD) between the true and the estimated distributions for locations in Paris and San Francisco, under BA and LAP.

Now we proceed to empirically compare the statistical utility of BA and LAP by performing experiments on the locations obtained from the Gowalla dataset for two different cities: Paris and San Francisco. In addition to the same setting for the Gowalla check-ins in Paris as considered in the experiments of Section 3.4.1, here we also test for 123,025 check-in locations from the Gowalla dataset in a northern part of San Francisco bounded by latitudes (37.7228, 37.7946) and longitudes (-122.5153, -122.3789) covering an area of 12Km \times 8Km discretized with a 24 \times 17 grid. The locations were privatized with BA and LAP under varying levels of privacy – the loss parameter, β , for BA ranged from 0.2 to 5.0, which implies that the value of the geo-ind parameter, ε , ranged from 0.4 (very high level of privacy) to 10.0 (almost no privacy). To account for the randomness in the process of generating the sanitized locations, 5 simulations were run for each value of the privacy parameter for obfuscating every location in both datasets.

Figure 3.3 reveals that BA possesses a significantly better statistical utility than LAP for a high level of privacy (for $\beta \in (0.4, 1.4]$ and $\beta \in (0, 1)$, i.e., ε up to 2.8 and 2, in Paris and San Francisco datasets, respectively). As the level of privacy decreases, the EMD of BA becomes worse than that of LAP. We conjecture that this is the price to pay for the added privacy provided by the elasticity of the mechanism. Eventually, the EMD between the true and the estimated PMFs converge to 0 in both mechanisms, as we would expect, fostering the maximum possible statistical utility with, practically, no privacy guarantee.

Summarizing the results from Sections 3.4.1 and 3.4.2, we can establish that:

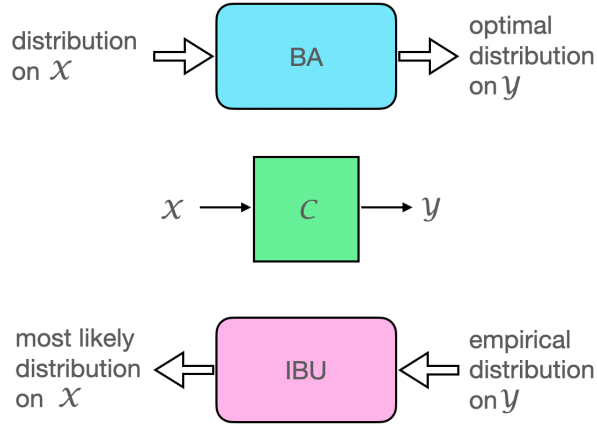


Figure 3.4: Illustration of the duality between BA and IBU.

- in addition to providing a formal geo-ind guarantee, BA also gives an LPPM with an elastic distinguishability metric to enhance the privacy of vulnerable locations.
- BA optimizes the trade-off between QoS and MI.
- the statistical utility for high levels of privacy is significantly better for BA than LAP.

Therefore, we conclude that BA is a key contender for providing a comprehensive notion of location-privacy while preserving the utility of the data for both the users and the service providers.

3.5 Duality between IBU and BA

We now explore a relationship between BA and IBU which we found rather intriguing. For a metric space (X, d) , let X be a random variable on X with PMF π_X . Recalling the iteration of BA from (3.3) and (3.2):

$$c_t(y) = \sum_{x \in X} \pi_X(x) C_{xy}^{(t)} \text{ and } C_{xy}^{t+1} = \frac{c_t(y) \exp\{-\beta d(x, y)\}}{\sum_{z \in X} c_t(z) \exp\{-\beta d(x, z)\}}$$

Hence, we obtain:

$$c_{t+1}(y) = \sum_{x \in X} \pi_X(x) C_{xy}^{(t+1)} = \sum_{x \in X} \pi_X(x) \frac{c_t(y) \exp\{-\beta d(x, y)\}}{\sum_{z \in X} c_t(z) \exp\{-\beta d(x, z)\}} \quad (3.6)$$

Comparing it with the iteration of IBU as in Definition 2.3.2, we observe that (3.6) BA is dual to IBU. Indeed, consider an exponential mechanism of the form $C = c \exp\{-\beta d(x, y)\}$. Flipping the roles of x and y in (3.6), and replacing the input distribution π_X with the empirical distribution in output to C , we obtain the iterative step of IBU.

Due to this duality between BA and IBU (illustrated in Figure 3.4) and taking advantage of the fact that BA converges [45], i.e., $\lim_{t \rightarrow \infty} c_t$ exists, we obtain that also IBU converges.

3.6 PRIVIC: a privacy-preserving method for incremental data collection

To ensure that the produced mechanism is truly optimal, BA needs a good approximation of the prior distribution. In the beginning, we cannot assume to have such knowledge, but as the service providers incrementally collect data from their users, we can use these data to refine the estimation of the prior and get a better mechanism. These data, however, are obfuscated by the privacy mechanism and, hence, it is not obvious that the estimation of the prior really improves in the process. We show that this is the case, and, summarizing all results obtained for BA so far, we propose a method that facilitates the service providers to incrementally collect data and gradually achieve a high statistical utility with respect to the QoS. We shall refer to our proposed method for **PRIV**acy-preserving Incremental Collection of location data as *PRIVIC*.

The goal of PRIVIC is to construct an obfuscation mechanism that guarantees formal geo-ind, acts as an elastic mechanism, and eventually optimizes between MI and QoS, while producing, at the same time, a good estimation of the distribution of the data.

We shall consider locations sampled from a finite space $\mathcal{X} = \{x_1, \dots, x_m\}$. Let the *true distribution* or *true PMF* on \mathcal{X} (from which the users' locations are sampled) be $\pi_{\mathcal{X}}$. Note that *we do not assume the knowledge of $\pi_{\mathcal{X}}$ in our method*. We assume that the new locations are sampled independently from the previous ones. This hypothesis is reasonable if the collection of the new data is enough separated in time from the previous one, otherwise, we would have a potential correlation between samplings due to the possibility that a user sends repeated check-ins from spatially closed locations. In any case, geo-ind, like DP, satisfies the property of sequential compositionality [39], which means that privacy degradation is under control.

In this work, to achieve geo-ind, we shall adhere to the Euclidean metric d_E to measure the ground distance between locations.

PRIVIC proceeds as follows (cf. also Figure 3.5):

1. Set θ_0, c_0 to be the uniform distributions on \mathcal{X} , i.e., $\theta_0(x) = c_0(x) = 1/|\mathcal{X}|$ for all $x \in \mathcal{X}$.
2. In step $t \geq 1$:
 - i) For a fixed the maximum average distortion, set $\hat{\mathcal{C}}^{(t)} = \text{BA}(\theta_{t-1}, c_0)$.
 - ii) Sample a new set of locations $\mathbf{x}^{(t)}$ from the (unknown) true distribution and obfuscate them locally by the mechanism $\hat{\mathcal{C}}^{(t)}$ to get $\mathbf{y}^{(t)}$, thus obtaining the empirical distribution of the reported locations $\mathbf{q}_t = \{\mathbf{q}_t(x) : x \in \mathcal{X}\}$.
 - iii) $\mu_t = \text{IBU}(\hat{\mathcal{C}}^{(t)}, \theta_{t-1}, \mathbf{q}_t)$.
 - iv) if $t = 1$ then $\theta_t = \mu_t$ else $\theta_t = \mu_t \oplus \theta_{t-1}$ (combination of previous and new estimation proportional to the respective number of samples).
3. $\hat{\pi}_{\mathcal{X}} = \text{GIBU}(\left(\hat{\mathcal{C}}^{(1)}, \mathbf{y}^{(1)}\right), \dots, \left(\hat{\mathcal{C}}^{(N)}, \mathbf{y}^{(N)}\right))$.

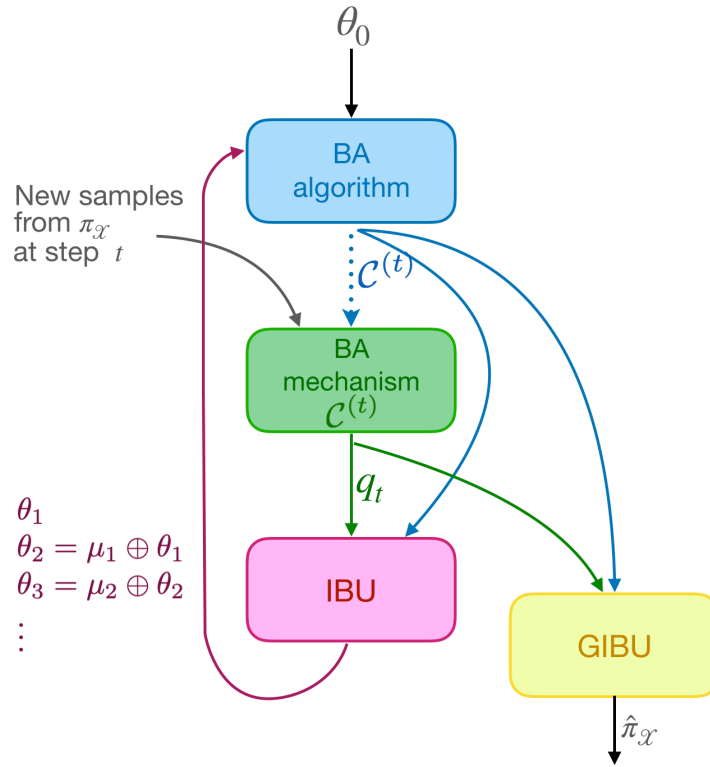


Figure 3.5: Illustration of the iterative process of PRIVIC

REMARK 3.6.1. The initial distribution θ_0 does not need to be a uniform distribution, any fully-supported distribution would suffice for the process to eventually converge to an optimal mechanism. However, starting with a uniform distribution allows us to avoid any bias in the mechanisms produced in the intermediate steps.

REMARK 3.6.2. We believe that the last step (3) is not really necessary: The combination of all estimations should already be the MLE of the true distribution, and this is also what we have witnessed in the experiments. However, applying this last step allows us to *formally prove* the convergence to the MLE, using the results for GIBU in [46].

In the practical implementation of PRIVIC, we use the precision parameters δ_{BA} , δ_{IBU} , and δ_{GIBU} to set the threshold of empirical convergence of BA, IBU, and GIBU, respectively. Let the privacy mechanism generated this way after N iterations, for fixed parameters c_0 , β , δ_{BA} , δ_{IBU} , and δ_{GIBU} , be functionally represented as $\hat{\mathcal{C}}_{\text{BA}}(\theta_0, N)$.

Concerning statistical utility, it is important to ensure that IBU converges to the true distribution. As a matter of fact, IBU always converges to an MLE but the MLE may not be unique [46]. More precisely, there can be more than one distribution that is the most likely input to the obfuscation mechanism, for a given empirical distribution on the noisy data. Thus, even though IBU converges, it may converge to a distribution different from the true one. This is a problem in

Algorithm 1: PRIVIC

Input: Loss parameter: β , No. of iterations: N , precision of BA: δ_{BA} , precision of IBU: δ_{IBU} , precision of GIBU: δ_{GIBU} ;
Output: Optimal channel: \hat{C} , Estimation of true PMF: $\hat{\pi}_X$;
 $\theta_0(x) \leftarrow 1/|\mathcal{X}|$;
 $c_0 \leftarrow 1/|\mathcal{X}|$;
 $t \leftarrow 0$;
while $t \leq N$ **do**
 $\hat{C}^{(t+1)} = \text{BA}(\theta_t, c_0, \beta, \delta_{BA})$;
 $\mathbf{y}^{(t)} \leftarrow (y_1^{(t)}, \dots, y_n^{(t)})$: New noisy locations reported by users after obfuscating their newly sampled true locations with $\hat{C}^{(t)}$;
 $\mathbf{q} \leftarrow \{q(x) : x \in \mathcal{X}\}$: Empirical PMF obtained from \mathcal{L} by the service provider;
 $\mu \leftarrow \text{IBU}(\hat{C}^{(t+1)}, \theta_t, \mathbf{q}, \delta_{IBU})$;
 if $t = 0$ then $\theta_{t+1} \leftarrow \mu$ else $\theta_{t+1} \leftarrow \mu \oplus \theta_t$;
 $t \leftarrow t + 1$;
 $\hat{\pi}_X \leftarrow \text{GIBU}((\hat{C}^{(1)}, \mathbf{y}^{(1)}), \dots, (\hat{C}^{(N)}, \mathbf{y}^{(N)}), \delta_{GIBU})$;
 $\hat{C} \leftarrow \text{BA}(\hat{\pi}_X, c_0, \beta, \delta_{BA})$;
Return: $\hat{C}, \hat{\pi}_X$

Algorithm 2: Blahut-Arimoto algorithm (BA)

Input: PMF: π , initial mechanism: $C^{(0)}$, loss parameter: β , precision: δ_{BA} ;
Output: mechanism giving minimum mutual information for maximum avg. distortion encapsulated by β : \hat{C} ;
Function $\text{BA}(\pi, c_0, \beta, \delta_{BA})$:
 $t \leftarrow 0$;
 while $\delta_{BA} \leq |C^{(t)} - C^{(t-1)}|$ **do**
 $C_{xy}^{(t+1)} \leftarrow \frac{c_t(y) \exp\{-\beta d_E(x, y)\}}{\sum_{z \in \mathcal{X}} c_t(z) \exp\{-\beta d_E(z, y)\}}$;
 $c_{t+1}(y) \leftarrow \sum_{x \in \mathcal{X}} \pi(x) C_{xy}^{(t+1)}$;
 $t \leftarrow t + 1$
 $\hat{C} \leftarrow C^{(t)}$;
 Return: \hat{C}

the method by Oya et al. in [52] which computes the obfuscation mechanism via the algorithm of Shokri et al. [76]. The resulting mechanism optimizes the trade-off between distortion and a Bayesian notion of privacy, but may not have a unique MLE, as illustrated in the example below. They probably did not realize the problem, because they relied on the flawed results by [33] according to which every mechanism would have a unique MLE.

The following example is a simplified version of the example given in [46] (Sections 3.1 and 3.2.) which was aimed at showing the non-uniqueness of the MLE, and consequent convergence to the wrong distribution, in a more general setting. However, for the scope of our work, a simpler variant suffices.

EXAMPLE 3.6.1. Consider three collinear locations, a , b and c , where b lies in between a and c at a unit distance from each of them. Assume that the prior

Algorithm 3: iterative Bayesian update (IBU)

Input: Privacy mechanism: C , Full-support PMF: ϑ_0 , empirical PMF from observed data: \mathbf{q} , precision: δ_{IBU} ;
Output: MLE of true PMF: θ ;
Function IBU($C, \vartheta_0, \mathbf{q}, \delta_{IBU}$):
 Set $t \leftarrow 0$;
 while $\delta_{IBU} < |\vartheta_t - \vartheta_{t-1}|$ **do**
 $\vartheta_{t+1}(x) \leftarrow \sum_{y \in \mathcal{X}} \mathbf{q}(y) \frac{C_{xy} \vartheta_t(x)}{\sum_{z \in \mathcal{X}} C_{zy} \vartheta_t(z)}$;
 $t \leftarrow t + 1$
 $\theta \leftarrow \vartheta_t$;
 Return: θ ;

distribution on these three locations is uniform and that the constraint on the utility is that it should not exceed $2/3$. Then a mechanism that optimizes the QoS in the sense of [76] is the one that maps all locations to b . However, this mechanism has no statistical utility, as the b 's do not provide any information about the original distribution. Indeed, given n obfuscated locations (i.e., n b 's) all distributions on a, b and c of the form $k_a/n, k_b/n, k_c/n$ with $k_a + k_b + k_c = n$, have the same likelihood to be the original one.

Fortunately, our method does not have this problem, because the BA produces an invertible mechanism, and invertibility implies the uniqueness of the MLE [46]. In particular, we are now able to show the convergence of PRIVIC as a whole using the results of [46].

THEOREM 3.2. For any $t \geq 1$, the mechanism generated by BA over \mathcal{X} at the t 'th iteration, seen as a stochastic matrix, is invertible.

Proof. In Appendix A. □

THEOREM 3.3. PRIVIC converges to the unique MLE of the true distribution.

Proof. In Appendix A. □

To evaluate the statistical utility of $\hat{C}_{BA}(\theta_0, N)$ (cf. Section 3.7), we will measure the EMD between the true and the estimated PMFs at the end of N iterations of PRIVIC. Thus, the quantity $EMD(\hat{\pi}_{\mathcal{X}}, \pi_{\mathcal{X}})$ parameterizes the utility of $\hat{C}_{BA}(\theta_0, N)$ for the service providers. We use the same Euclidean distance as the underlying metric for computing, both, the EMD and the average distortion – this consistency threads together and complements the notion of *utility* of the service providers and that from the sense of the QoS of the users.

3.7 Experimental analysis of PRIVIC

In this section, we describe the empirical results obtained by carrying out experiments to illustrate and validate the working of our proposed method. Standard Python packages were used to run the experiments in a MacOS Ventura 13.2.1 environment with an Intel core i9 processor and 32 GB of RAM. Like in the previous experiments to compare the statistical utilities of BA and LAP, as elaborated in Section 3.4.2, we use real locations from the same regions in Paris and San Francisco from the Gowalla dataset [72, 73]. In particular, we consider Gowalla check-ins from (i) a northern part of San Francisco bounded by latitudes (37.7228, 37.7946) and longitudes (-122.5153, -122.3789) covering an area of 12Km×8Km discretized with a 24×17 grid; (ii) a central part of Paris bounded by latitudes (48.8286, 48.8798) and longitudes (2.2855, 2.3909) covering an area of 8Km×6Km discretized with a 16×12 grid. In this setting, we work with 123,108 check-in locations in San Francisco and 10,260 check-in locations in Paris. Figure 3.6a shows the particular points of check-in from Paris and San Francisco and Figure 3.6b highlights their distribution.

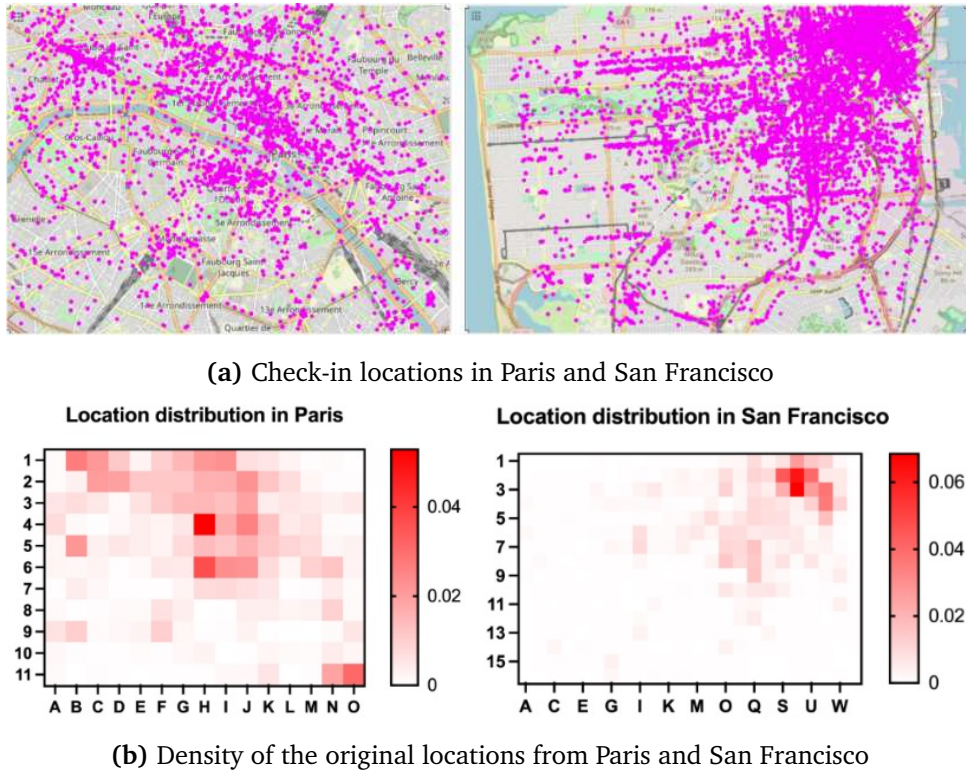


Figure 3.6: (a) visualizes the original locations from Gowalla dataset from Paris and San Francisco. (b) illustrates a heatmap representation of the locations in the two cities to capture the distribution of the data.

Table 3.1: Run-time and complexity of BA and IBU in each cycle of PRIVIC

Dataset	BA		IBU	
	Mean run-time (sec.)	Complexity	Mean run-time (sec.)	Complexity
Paris	3.256	$O(n^2)$	1.30	$O(n^2)$
San Francisco	16.805	$O(n^2)$	128.192	$O(n^2)$

Framework: MacOS Ventura 13.2.1 with Intel core i9 processor and 32 GB RAM

We implemented PRIVIC on the locations from Paris and San Francisco separately to judge

its performance on real data with very different priors. In both cases, we ran our mechanism until it empirically converged. 15 cycles of PRIVIC were required for the Paris dataset where each cycle comprised 8 iterations of BA until it converged to generate the privacy mechanism and 10 iterations of IBU until it converged to the MLE of the prior. For the San Francisco dataset, PRIVIC needed 8 cycles to converge with 5 iterations of BA and IBU each to converge in every cycle. The complexities and the run-times of BA and IBU are summarised in Table 3.1. In both cases, we assigned the value of the loss parameter signifying the QoS of the users, β , to be 0.5 and 1. This was done to test the performance of PRIVIC in estimating the true PMF under two different levels of privacy. Each experiment was run for 5 rounds of simulation to calibrate the randomness of the sampling and obfuscation. In each cycle of PRIVIC, across all the settings, BA was initiated with the uniform marginal c_0 and a uniform distribution over the space of locations as the “starting guess” of the true distribution.

With $\beta = 1$, BA produces a geo-indistinguishable mechanism that injects less local noise than that obtained with $\beta = 0.5$. As a result, PRIVIC obtains a more accurate estimate of the true PMF for the $\beta = 1$ than for $\beta = 0.5$. However, in both cases, the EMD between the true and the estimated distributions is very low, indicating that the PRIVIC mechanism is able to preserve a good level of statistical utility. Moreover, for both Paris and San Francisco, PRIVIC seems to significantly improve its estimation of the true PMF with every iteration until it converges to the MLE. Comparing Figures 3.7 and 3.6b, we see that the estimations of the true distributions of the locations in Paris and San Francisco by IBU under PRIVIC for both the settings of the loss parameter are fairly accurate. However, as we would anticipate, the statistical utility for $\beta = 1$ is better than that for $\beta = 0.5$.

Now we shift our attention to analyze the performance of PRIVIC in preserving the statistical utility and its long-term behaviour of the two datasets. Figure 3.8 shows us the EMD between the true distribution of the locations in Paris and its estimate by IBU under PRIVIC in each of its 15 cycles under the two settings of privacy ($\beta = 0.5, 1$). One of the most crucial observations here is that the EMD between the true and the estimated PMFs seems to decrease with the number of iterations and it finally converges, implying that the estimation of PMFs given by PRIVIC seems to improve at the end of each cycle and, eventually, it converges to the MLE of the prior of the noisy locations, giving the estimate of the true PMF. This, empirically, suggests the convergence of the entire method. This is a major difference from the work of [52] which, as we pointed out before, has the potential of encountering an LPPM which is optimal according to the standards set by Shokri et al. in [76] but the EM method used to estimate the true distribution would fail to converge for that mechanism as illustrated in Example 3.6.1. We observe a very similar trend for the San Francisco dataset. Figure 3.9 shows the statistical utility of the mechanism generated by PRIVIC under each of its 8 cycles for $\beta = 0.5$ and $\beta = 1$. The explicit values of the EMD between the true and the estimated PMFs on the location data from Paris and San Francisco for both the settings of the loss parameter can be found in Tables B.1 and B.2 in Appendix B.

In the next part of the experiments, we set ourselves to dissect the trend of the statistical utility harboured by PRIVIC w.r.t. the level of geo-ind it guarantees. We recall that the higher the value of β , the lesser the local noise that is injected into the data, and, hence, the worse will be

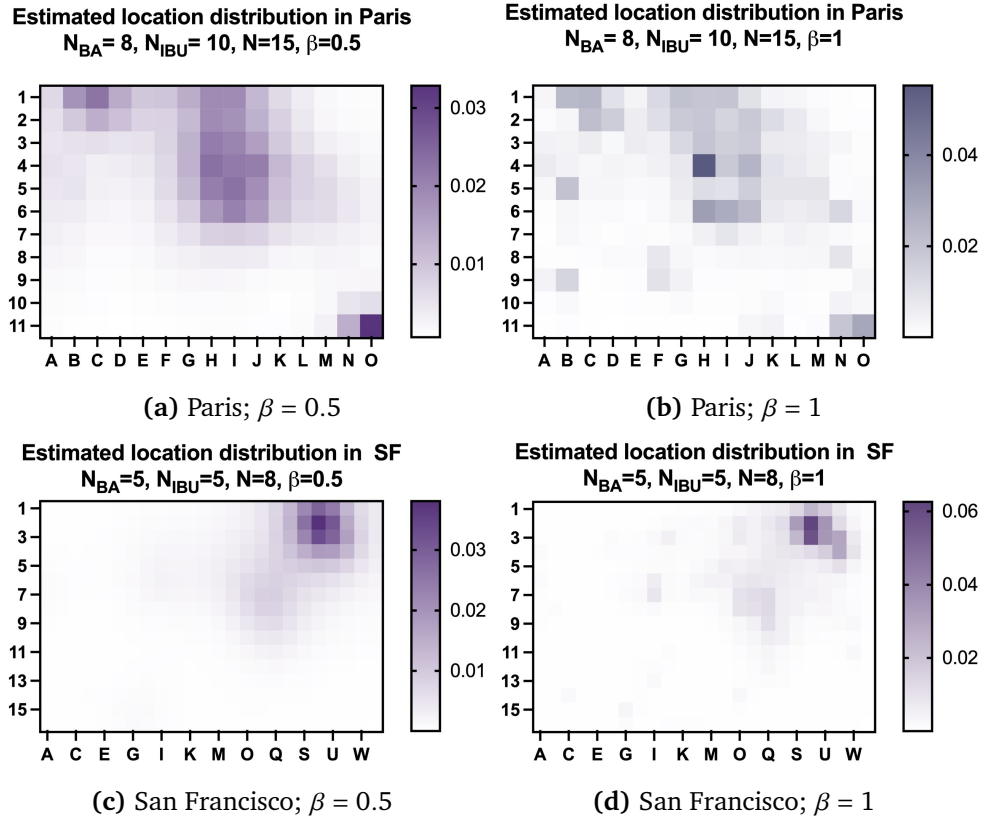


Figure 3.7: Visualization of the estimated true distribution of the locations in Paris ((a) and (b)) and San Francisco ((c) and (d)) by PRIVIC after its convergence; the first column is for $\beta = 0.5$ and the second column is for $\beta = 1$.

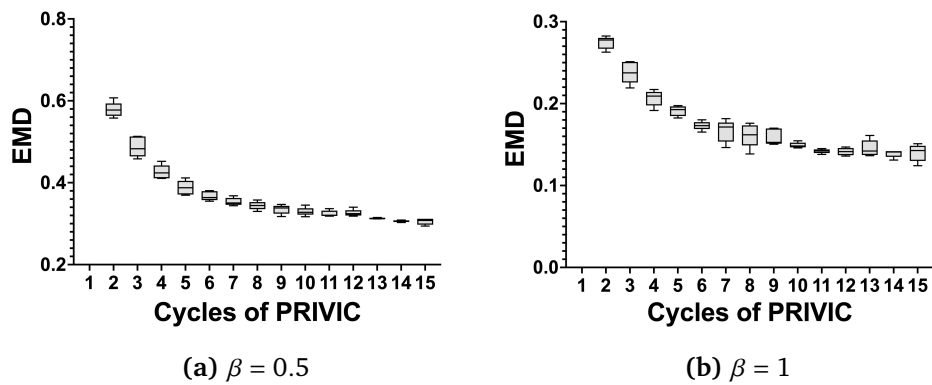


Figure 3.8: (a) and (b) show the EMD between the true PMF of the Paris locations and its estimation by PRIVIC in each of its cycle for $\beta = 0.5$ and $\beta = 1$, respectively.

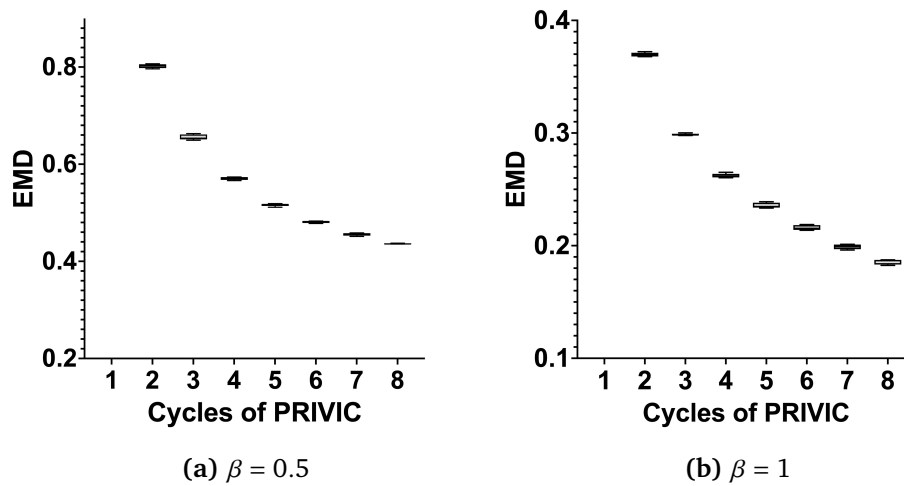


Figure 3.9: (a) and (b) show the EMD between the true PMF of the San Francisco locations and its estimation by PRIVIC in each of its cycles for $\beta = 0.5$ and $\beta = 1$, respectively.

the statistical utility, staying consistent with our observations in Figure 3.7. We continue working with the location data from Paris and San Francisco obtained from the Gowalla dataset in the same framework as described before. We consider β taking the values 0.1, 0.3, 0.5, 0.7, 0.9, 1, and for each value of the loss parameter, we run PRIVIC on both datasets using the same number of iterations as in the previous experiments. We adhere to 5 rounds of simulation for each β to account for the randomness generated in the obfuscation process.

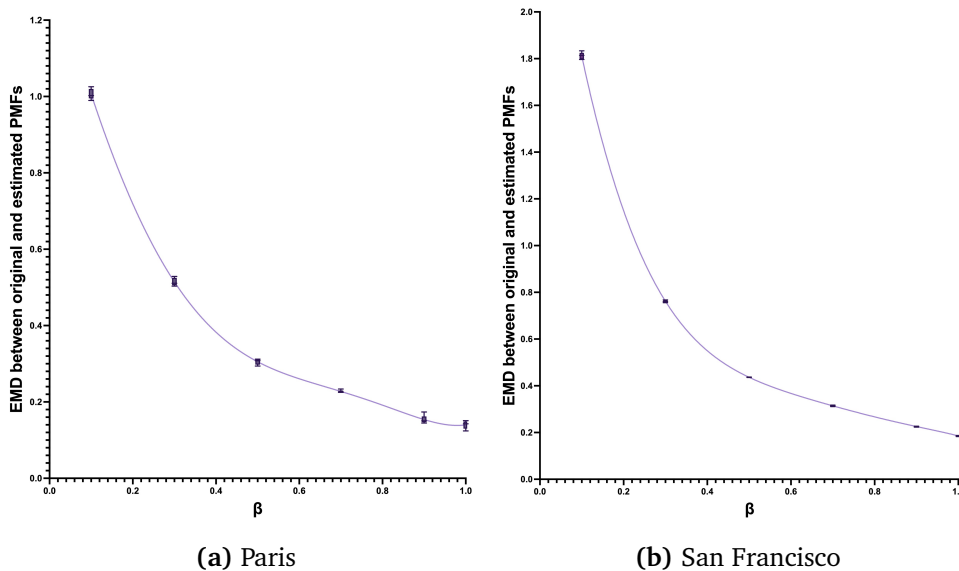


Figure 3.10: (a) and (b) illustrate that EMD between the true and the estimated distributions of the locations in Paris and San Francisco, respectively, after the empirical convergence of PRIVIC for the different values of the loss parameters β .

Figure 3.10 shows us that the difference between the true and the estimated PMFs under PRIVIC starts by sharply decreasing and then eventually stabilizes with an increase in the value of the loss parameter. In other words, as the intensity of the local noise decreases, we will end

up estimating the unique MLE of the original distribution while optimizing MI and the users' QoS. Both the location datasets result in a Pareto curve showing a similar trend. This depicts an improvement of the estimated PMF until it converges to the true distribution. This observation complements the Pareto-optimality of MI with the maximum average distortion as studied in *rate-distortion theory* [44], and thus, we empirically weave together the two ends of utility with the information theoretical notion of privacy under PRIVIC.

Discussion

As a justification for the applicability and the working of our method, in a setting where the service providers periodically collect location data from clients, it is reasonable to assume that, over time, they would like to maximize their utility by accurately approximating the true distribution of the population for improving their service in various aspects (crowd management, security enhancement, WLAN hotspot positioning, etc.). BA, in addition to guaranteeing geo-ind, acts as an elastic location-privacy mechanism and optimizes between MI and the data owners' QoS when it initiates with the true prior. Therefore, as every iteration of PRIVIC improves the estimation of the original distribution, as seen in Figures 3.3a and 3.3b, which is used as the starting distribution in its next cycle, the overall privacy protection and its trade-off with QoS of the users will also improve, motivating the users and the service providers comply with PRIVIC to act in their best interests and, in turn, engaging them in a positive feedback loop to maximize the corresponding privacy and utility goals.

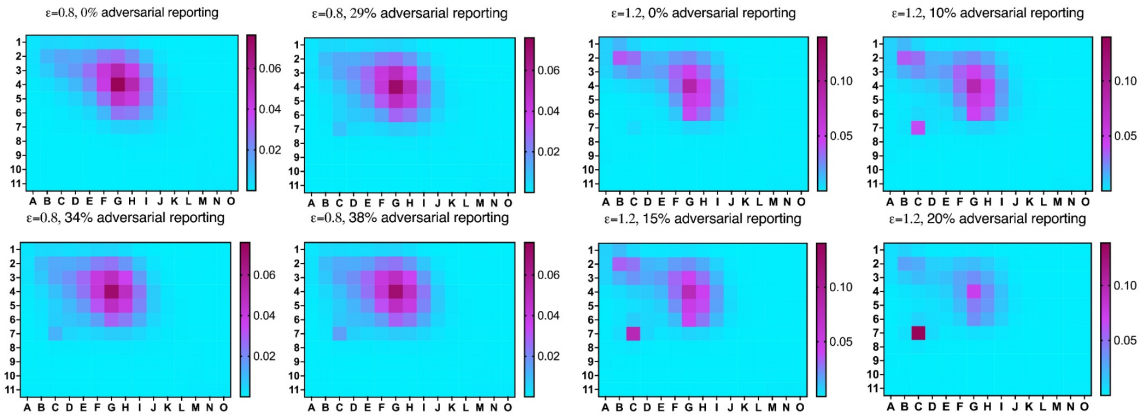
3.8 Vulnerability of PRIVIC

In this section, we illustrate a potential vulnerability of PRIVIC when a subset of colluded users (*adversarial users*) intentionally deviate from the correct use of the protocol.

The attack consists in falsely reporting their location in order to alter the estimation of the true distribution and, consequently, the obfuscation mechanism produced by BA. Specifically, we study two cases: (i) adversaries reporting a crowded location (*strong location*) and adversaries reporting an isolated location (*vulnerable location*).

We used the real locations from the Paris dataset with the geo-spatially isolated “island” (as illustrated by Figure 3.1 presented in Section 3.4) representing a strong and a vulnerable location in the map denoted by points A and B in the figure, respectively. We performed the experiments with two different levels of formal geo-indistinguishability ($\epsilon = 0.8$ and $\epsilon = 1.2$) considering different fractions of “adversarial data submissions” (i.e., adversarial users reporting their locations falsely to compromise the privacy of other users) in each case.

Although, in both cases (i) and (ii), we observed that with an increase in the fraction of adversaries, the probability mass assigned by BA (used to obfuscate the corresponding points locally) becomes higher in and around the corresponding reported points, the impact of privacy differs across both settings. For (i), the obfuscation distribution happens to be weighed heavily around the true crowded location (point B) by both BA and LAP (as illustrated by Figures 3.2b



(a) Obfuscation distribution of a vulnerable location in the map using BA satisfying 0.8-geo-ind with (left-to-right, top-to-bottom) 0%, 29%, and 34%, and 39% adversarial users, respectively. (b) Obfuscation distribution of the vulnerable location in the map using BA satisfying 1.2-geo-ind with (left-to-right, top-to-bottom) 0%, 10%, 15%, and 20% adversarial users, respectively.

Figure 3.11: Effect on the privacy provided by BA to obfuscate the geo-spatially isolated location (Location A as in Figure 3.1) for different fractions of adversarial users who intentionally report their locations falsely under two different levels of formal geo-ind guarantees by BA.

and 3.2a) even without any adversarial users. This trend was seen to continue even when we assumed different levels of adversaries.

However, case (ii) represents a much more serious attack. As the number of adversarial users increases, BA and, in turn, PRIVIC become less potent to be able to protect the privacy of honest users who are genuinely located in an isolated location on the map. We also observe that BA and PRIVIC start behaving more like LAP. In particular, Figures 3.11a and 3.11b illustrate that the obfuscation distribution generated by BA satisfying geo-ind with $\epsilon = 0.8$ and $\epsilon = 1.2$, respectively, of the (non-adversarial) users located in point A assigns more and more weight to and around point A which, as a result, makes them more and more identifiable. This evaluation of the vulnerability of PRIVIC under adversarial data submission essentially exposes a weakness of the elastic distinguishability metric. We plan to address this aspect and aim to make PRIVIC more robust against adversarial users in our future works.

4

A privacy-preserving querying mechanism with high utility for electric vehicles

4.1 Introduction

Air pollution is one of the immediate issues that the world is experiencing [77–79]. In the United Kingdom in 2019, 27% of all greenhouse gas emissions come from transportation, as the largest emitting sector [80–82]. Hence, the transportation industry and academic communities are increasingly interested in developing alternative energy vehicles to reduce emissions. Automobile manufacturers are introducing a new generation of electric vehicles (EVs) that often employ connected and automated driving functions [83].

EVs are regarded as one of the most promising means of reducing emissions and reliance on fossil fuels. Along with environmental benefits, EVs provide superior energy efficiency to conventional vehicles [84]. As the cost of batteries continues to decrease, the large-scale adoption of EVs is becoming more viable [85]. Despite the advantages and competitive cost, many customers remain concerned about running out of battery power before reaching their destination or waiting for their EVs to charge. The primary obstacles to EV adoption are the availability of chargers and the range that can be travelled on a single charge, often referred to as *range anxiety* in the literature [86].

There has been some recent focus on forecasting how busy the charging stations (CS) are in certain areas to ensure that the EVs can plan their journeys conveniently [87, 88]. However, the existing research in this direction, primarily founded upon machine learning based methods,

does not address the privacy concerns involved in such predictive techniques and does not consider situations where there may arise unprecedented traffic congestion (e.g. due to a one-off concert or an event). DP has been one of the cutting-edge approaches for protecting the privacy of personal data while allowing for analysing and exploring its utility. However, as discussed in Section 2.1, such central DP models possess a risk of a single point of failure and are vulnerable to having an adversarial curator. To circumvent the need for such a central and trusted dependency, LDP was proposed which has been getting a lot of attention lately.

On the other hand, future vehicles are getting more sophisticated in their sensory, onboard computation, and communication capacities. Furthermore, the emergence of Mobile Edge Computing (MEC) also changes the Intelligent Transportation Systems (ITS) by providing a platform to assist computationally heavy tasks by offloading the computation to the Edge cloud [89]. This architecture often employs three tiers, with the vehicle on the first, MEC on the second, and standard cloud services on the third [90]. Figure 4.1 shows the system architecture for the location privacy framework proposed in this chapter.

ITS provides a platform containing distributed and resource-constrained systems to support real-time vehicular functions where these functions' efficacy relies on the data shared across entities. However, the risk of privacy disclosure and tracking increases due to data sharing [91]. Privacy-preserving schemes are developed using established techniques such as group signature, anonymity, and pseudonymity [92, 93]. However, it is possible to identify privatised data with adequate background information. Hence, DP approaches have emerged as the gold standard of data privacy because they provide a formal privacy guarantee independent of a threat actor's background knowledge and computing capability [94].

As introduced in Section 2.1, geo-ind has been emerging as a gold standard for privacy-preserving location-based services (LBS) and can be implemented directly on the user's device (tablet, smartphone, etc.). The fact that the users can control their explicit privacy-protection level for various LBS makes it very appealing. However, a drawback of injecting noise locally to the datum is that it deteriorates the QoS due to the lack of accuracy of the data. However, vehicles are located on the road network under normal circumstances. For vehicular location queries, the classical geo-ind mechanism may result in publishing unrealistic privatised locations such as houses, parks, or lakes or may even perturb the real location to a very faraway point with some probability. Thus, there is a need for an adapted model of geo-ind for vehicular applications. This chapter proposes a novel privacy model called *approximate geo-indistinguishability (AGeoI)*, which is based on the notion of geo-ind, by using a discrete road network graph.

Our proposed method, in addition to allowing the EVs to have formal privacy guarantees on their queries to locate the nearest CS, enables the users to estimate the live occupancy of the CS efficiently allowing convenient journey planning (e.g., avoid going through a route aiming for a CS which has a dense crowd of EVs looking for a CS around it).

Contributions

Our key contributions in this chapter are:

1. We present the notion of *approximate* geo-indistinguishability (AGeoI), a formal standard of location-privacy in a bounded co-domain, by generalising the classical paradigm of geo-ind and adapting it to a graphical environment. We illustrate its applicability by proving that it satisfies the compositionality theorem. Moreover, we show that the truncated Laplace mechanism satisfies AGeoI by deriving the appropriate privacy parameters.
2. We propose a two-fold privacy-preserving navigation method for EVs dynamically querying for CS on road networks – the method protects against threats to individual locations of their queries with formal AGeoI guarantees and we provide protection against adversaries tracing the trajectories of the EVs by interpolating their query locations in an online setting.
3. We perform experiments on real vehicular journey data from San Francisco with real locations of charging stations from the area under two settings of their sparsity in the road network, and show that our method ensures that a very high fraction of EVs enjoys “privacy for free” and that the cost of utility-loss for the EVs is very low compared to the formal gain in privacy.
4. We demonstrate that our method, aside from providing formal location-privacy guarantees, allows the EVs to predict the live occupancy of the charging stations based on the sanitised queries received by the server, enabling the users to conveniently and efficiently plan their journeys.

4.2 Related Work

Both corporate and academic communities have recently piqued interest in advancing EVs and charging infrastructure to improve the transportation system’s sustainability. Despite the advancements, the EV sector confronts challenges that delay the adoption process, such as range anxiety, an absence of convenient and available charging infrastructure, and waiting time to charge [95, 96]. An offline static map of CS is insufficient to resolve these obstacles since EVs may need to reserve a charging station when a trip is planned or query the available stations based on their battery state. Thus, live vehicular and charging station data is utilised in querying and reservation/scheduling mechanisms [97–99]. Encryption techniques can be used in such mechanisms to prevent external intrusions, but they cannot preserve users’ privacy from malicious servers and third-party providers.

Several data types are considered in these mechanisms, including real-time location, intended route, battery level, and station availability, to ensure the drivers are not detoured from their intended route [97, 100]. Although disclosing such information poses privacy concerns for the driver’s location and vehicle tracking, the privacy requirements of such mechanisms are not sufficiently studied in the literature. Existing methods for planning charging points for EV journeys are considered mechanisms for confidentiality and integrity, but the drivers’ location privacy is regarded as an issue of trust in the third-party service provider [101, 102].

This problem can be addressed by several approaches based on the threat model of the system. Location anonymity is achieved through cloaking an area [103, 104]. This approach can

only be applied to the Edge of our system model to provide anonymity to a group of EVs, but we consider the Edge as an honest-but-curious threat actor and aim to preserve individual vehicles' privacy locally. Thus, such techniques are not trivially applicable to our considered threat model. Furthermore, anonymity techniques do not provide a formal privacy guarantee [105]. Similarly, mix-network approaches cannot be applied because there is no guarantee that multiple vehicles will be present in an Edge's coverage in any timestamp due to vehicles' movement [106].

An applicable approach to download the charging station's live map on EVs to search for the nearest or on-the-route available charging station has been considered and studied by the community [107]; however, the communication overhead of this technique is predicted to be much higher than the vehicles' location-based inquiry since it will require downloading a recent snapshot of the map for each query and, therefore, has been criticised in the literature [108]. Moreover, due to the absence of data sharing, such methods hinder the statistical utility of the location data for the servers that may be useful for a variety of purposes (e.g. providing vital statistics to industries and institutions for optimally placing the CS on the map based on the query densities) and prevent the EVs from receiving any information about the traffic around and occupancy of certain CS restricting them to plan their journeys accordingly.

DP methods are increasingly being deployed to preserve location privacy in a variety of domains, including automotive systems. The studies in [109, 110] proposed models by deploying a geo-ind based mechanism on the Edge for LBS. However, their approach did not consider preserving vehicles' location privacy against the Edge. An approach that compliments the problem we aim to address in this work was proposed by Qiu et al. in [111] where the authors proposed a technique to crowd-source a task in a vehicular network while preserving geo-ind of the location of the vehicles offering Mobility as a Service in the spatial network to solve a task at a publicly known location in the map (e.g. taxi services). The problem formulation in this work is the inverse of what we aim to achieve in this chapter. As a result, the work in [111] cannot be extended to address the privacy concerns induced by multiple queries dynamically made along the journey.

In [112], Cunningham et al. studied the problem of trajectory sharing under DP and proposed a mechanism to tackle it. However, this work assumes the setting of an offline trajectory sharing which breaks down in the practical environment where the trajectories are being shared online as there is no prior information or limitation on the number of queries made by an EV during a journey and their respective locations. Therefore, the method proposed by the authors in [112] cannot be directly adapted to our dynamic environment closely simulating the real-world scenario for such a use case.

Of late, a major direction of research is along the lines of studying the statistical utility of differentially private data. A standard notion of statistical utility, which is extended to a variety of contexts, is the precision of the estimation of the distribution on the original data from that of the noisy data. Iterative Bayesian update (IBU) [31, 33], an instance of the famous *expectation maximization (EM)* method from statistics, provides one of the most flexible and powerful estimation techniques and has recently become in the focus of the community [34, 46]. In this work, we use IBU to approximate the distribution of the true locations of the queries made

to the server and based on that, the users of the EVs can predict the availability of the CS around them in real-time and plan their route accordingly.

4.3 Approximate geo-indistinguishability (AGeoI)

In the classical framework of GeoI [16], the space of the noisy data is, in theory, unbounded under the planar Laplace mechanism. Under a certain level of GeoI that is achieved, the planar Laplace mechanism ensures a nonzero probability of obfuscating an original location to a privatized one which may be quite far, thus inducing a possibility of a substantial deterioration in the QoS of the users. This loss of QoS can be more sensitive in the context of the navigation of EVs, where it is extremely important to prioritize a bounded domain where a user is willing to drive; this may be a result of time constraints, the rising cost of fuel, geographical boundaries (e.g. international borders), etc. – giving rise to an idea of *area of interest* for each EV. This motivated us to extend the classical GeoI to a more generalized approximate paradigm, inspired by the approach of the development of approximate DP from its pure counterpart.

Let \mathcal{X} and \mathcal{Y} be the spaces of the real and noisy locations equipped with distance metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, respectively. In general, $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ may be different and unrelated. However, for simplicity, here we assume $\mathcal{X} \subseteq \mathcal{Y}$ and, therefore, $d_{\mathcal{X}} = d_{\mathcal{Y}} = d$, and we proceed to define the notion of *approximate geo-indistinguishability*. It is worth noting here that, to an extent, we abuse the formal notion of “metric” as d is not required to be symmetric, i.e., there may exist $x_1, x_2 \in \mathcal{Y}$ such that $d(x_1, x_2) \neq d(x_2, x_1)$.

DEFINITION 4.3.1 (Approximate geo-indistinguishability). A mechanism \mathcal{K} is *approximately geo-indistinguishable (AGeoI)* or (ε, δ) -*geo-indistinguishable* if for every measurable $S \subseteq \mathcal{Y}$, any pair of secrets $x, x' \in \mathcal{X}$, and for $\varepsilon, \delta \in \mathbb{R}_{\geq 0}$ satisfying $\delta e^{\varepsilon d(x, x')} \in [0, 1]$:

$$\mathbb{P}_{\mathcal{K}}[y \in S|x] \leq e^{\varepsilon d(x, x')} \mathbb{P}_{\mathcal{K}}[y \in S|x'] + \delta e^{\varepsilon d(x, x')} \quad (4.1)$$

One of the biggest advantages of DP and all of its variants that are accepted by the community is the property of compositionality, where the level of privacy can be formally derived with a repeated number of queries. Thus, we now enable ourselves to investigate the working of the compositionality theorem with the AGeoI which we defined, to stay consistent with the literature [105].

THEOREM 4.1. [Compositionality Theorem for AGeoI] Let mechanisms \mathcal{K}_1 and \mathcal{K}_2 be $(\varepsilon_1, \delta_1)$ and $(\varepsilon_2, \delta_2)$ geo-indistinguishable, respectively. Then their composition is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$ -geo-indistinguishable. In other words, for every $S_1, S_2 \subseteq \mathcal{Y}$ and all $x_1, x'_1, x_2, x'_2 \in \mathcal{X}$:

$$\mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2}[(y_1, y_2) \in S_1 \times S_2 | (x_1, x_2)] \leq e^{\varepsilon_1 d(x_1, x'_1) + \varepsilon_2 d(x_2, x'_2)} \mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2}[(y_1, y_2) \in S_1 \times S_2 | (x'_1, x'_2)] + \left(\delta_1 e^{\varepsilon_1 d(x_1, x'_1)} + \delta_2 e^{\varepsilon_2 d(x_2, x'_2)} \right)$$

Proof. In Appendix C. □

We now proceed to generalize the conventional planar Laplace mechanism [113] to define the *truncated Laplace mechanism* extended to a generic metric space.

DEFINITION 4.3.2 (Truncated Laplace mechanism). The *truncated Laplace mechanism* \mathcal{L} on a space \mathcal{X} equipped with, not necessarily symmetric, distance metric d truncated to a radius r , is defined as:

$$\mathbb{P}_{\mathcal{L}}[y|x] = \begin{cases} c e^{-\varepsilon d(y,x)} & \text{if } d(x,y) \leq r \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

where c is the truncated normalization constant defined such that $\int_{y \in \mathcal{Y}} \mathbb{P}_{\mathcal{L}}[y|x] dy = 1$, and ε is the desired privacy parameter. Let us call r the *radius of truncation* for \mathcal{L} .

Note that for a discrete domain \mathcal{Y} , c is defined by normalizing $\sum_{y \in \mathcal{Y}} \mathbb{P}_{\mathcal{L}}[y|x] = 1$, and, in this case, \mathcal{L} is a truncated geometric mechanism [114] extended to a generic metric space.

LEMMA 4.2. For every $x_1, x_2 \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have $e^{-\varepsilon d(x_1, x_2)} \mathbb{P}_{\mathcal{L}}[y|x_1] - \mathbb{P}_{\mathcal{L}}[y|x_2] \leq 1$.

Proof. In Appendix C. □

THEOREM 4.3. \mathcal{L} satisfies (ε, δ) -geo-indistinguishability where

$$\delta = \max \left\{ \max_{\substack{y \in \mathcal{Y} \\ x_1, x_2 \in \mathcal{X}}} e^{-\varepsilon d(x_1, x_2)} \mathbb{P}_{\mathcal{L}}[y|x_1] - \mathbb{P}_{\mathcal{L}}[y|x_2], 0 \right\}.$$

Proof. Trivially $\delta e^{\varepsilon d(x_1, x_2)} > 0$ for any $x_1, x_2 \in \mathcal{X}$ as $\delta > 0$. Moreover, Lemma 4.2 ensures that $\delta e^{\varepsilon d(x_1, x_2)} \leq 1$. Hence, for every $y \in \mathcal{Y}$ and for all $x_1, x_2 \in \mathcal{X}$, we have:

$$\begin{aligned} e^{-\varepsilon d(x_1, x_2)} \mathbb{P}_{\mathcal{L}}[y|x_1] - \mathbb{P}_{\mathcal{L}}[y|x_2] &\leq \delta \\ \Rightarrow \mathbb{P}_{\mathcal{L}}[y|x_1] - e^{\varepsilon d(x_1, x_2)} \mathbb{P}_{\mathcal{L}}[y|x_2] &\leq \delta e^{\varepsilon d(x_1, x_2)} \end{aligned}$$

□

The explicit process of sampling private locations satisfying AGeol from a given set of original locations through a truncated Laplace mechanism on a discrete location space has been described in Algorithms 4 and 5.

Algorithm 4: Discrete and truncated Laplace mechanism (DTLap)

Input: Discrete domain of original locations: \mathcal{X} , Discrete domain of private locations: \mathcal{Y} , Desired privacy parameter: ε , Desired truncation radius: r ; **Output:** Channel C satisfying (4.2);

Function DTLap($\mathcal{X}, \mathcal{Y}, \varepsilon, r$):

```

    Set  $C \leftarrow$  empty channel;
    Set  $Y \leftarrow$  empty list;
    for  $x \in \mathcal{X}$  do
         $c_x = \frac{1}{\sum_{\substack{y \in \mathcal{Y} \\ d(x,y) \leq r}} e^{-\varepsilon d(x,y)}};$ 
        for  $y \in \mathcal{Y}$  do
            if  $d(x, y) \leq r$  then
                 $C[x, y] = 0$ 
            else
                 $C[x, y] = c_x e^{-\varepsilon d(x,y)}$ 
    Return:  $C$ ;

```

Algorithm 5: Sampling private locations with DTLap (DTLapSamp)

Input: Discrete domain of original locations: \mathcal{X} , Discrete domain of private locations: \mathcal{Y} , Desired privacy parameter: ε , Desired truncation radius: r ; Vector of original locations: X ;

Output: Corresponding vector of private locations: Y ;

Function DTLapSamp($\mathcal{X}, \mathcal{Y}, \varepsilon, r, X$):

```

     $C = \text{DTLAP}(\mathcal{X}, \mathcal{Y}, \varepsilon, r)$ ;
    Set  $Y \leftarrow$  empty list;
    for  $x \in X$  do
        Randomly sample  $y \in \mathcal{Y} \sim C[x, :]$ ;
        Append  $y$  to  $Y$ 
    Return:  $Y$ ;

```

4.4 System Model

This section details our privacy-preserving model for finding an optimal charging station in the Internet of Vehicles (IoV) as a use case of the proposed AGEoI technique. We begin with a discussion of the location privacy problems inherent in finding optimal CS in the IoV. This is followed by road networking modelling, a description of the system architecture for differentially private location sharing, the trust relationship between system tiers, and the privacy threat model.

4.4.1 Problem Statement

EVs have emerged as crucial components of future sustainable transportation systems, aimed at reducing CO2 emissions. Consequently, they have received considerable attention from both academia and industry [96]. However, due to their limited battery capacity, EVs often need to visit CS during journeys. This requirement leads to range anxiety among some drivers, where they fear that their vehicles lack sufficient battery power to reach their intended destinations. Range anxiety has been identified as a significant barrier to the widespread adoption

of EVs [115]. While CS are not always readily available, as it takes time to sufficiently charge EVs, the implementation of a CS booking service can help alleviate range anxiety.

To minimize charging wait times, EVs can access CS booking services through third-party providers, enabling them to discover the nearest and readily available CS. This can be achieved through static or live location queries. However, location sharing raises privacy challenges, such as vehicle tracking. GeoI technique provides a formal privacy guarantee for location queries. However, it is not highly applicable to this use case for two reasons. It does not consider the feasible locations where a vehicle can be present, and it does not stop vehicle tracking in the case of linked queries during the vehicle trajectory. Thus, a tailored privacy-preserving mechanism is facilitated by combining the proposed AGeoI technique with dummy location generation.

4.4.2 Road Network Model

Similar to [111], the road network G is represented as a weighted directed graph $G = (N, E, W)$, where N is the set of nodes, $E \subseteq N^2$ is the set of edges, and $W : N^2 \rightarrow \mathbb{R}^+$ is the set of weights representing the minimum travelling distance between any two nodes. The nodes and edges correspond to junctions and road segments of the network, respectively. Each edge $e \in E$ is addressed by the pair of respective starting node, ending node, and a weight representing the travelling distance through that edge, i.e., $e = (N_e^s, N_e^e, w_e) \in N$, where the direction of the traffic is from N_e^s to N_e^e on e . For any $i \in N$ and $j \in N$, let the sequence of edges (e_1, \dots, e_r) denote a *path* from node i to node j if $N_{e_1}^s = i$ and $N_{e_r}^e = j$. Hence, let $C(i, j)$ represent the set of paths that connect node i to node j . Then W is a $N \times N$ matrix, where

$$W_{ij} = \begin{cases} \min_{p \in C(i, j)} \sum_{e \in p} w_e & \text{if } C(i, j) \neq \phi \\ \infty & \text{otherwise} \end{cases}$$

Essentially W_{ij} is the shortest travelling distance from node i to node j in the network. We shall address the quantity W_{ij} as the *traversal distance* between nodes i and j in the graph G and denote it as $d_G(i, j)$ for every $(i, j) \in N^2$. Note that, as G is a directed graph, d_G may not be symmetric.

4.4.3 System Architecture

IoV applications are revolutionising transportation systems by mitigating human errors, enhancing travel convenience, and reducing energy, operational, and environmental costs [116, 117]. EVs have emerged as a viable technology for lowering carbon emissions and travel costs [118]. However, range anxiety is one of the major challenges of their wide adoption. Vehicular location data can be utilised to optimise the vehicle charging plan and mitigate range anxiety. Third-party services can assist users by recommending available CS in close proximity. Nevertheless, relying on these third-party providers raises significant privacy concerns within the honest-but-curious service provider threat model, necessitating users to trust them.

The system architecture, illustrated in Figure 4.1, incorporates vehicles within an ITS that

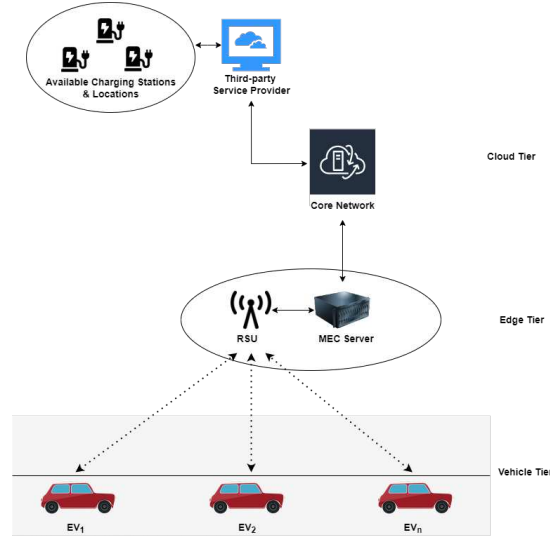


Figure 4.1: System Architecture (EV: Electric Vehicle, RSU: Roadside Unit, MEC: Mobile-Edge Computing Unit)

operates on a three-tier architecture. This architecture comprises Roadside Units (RSUs) connected to a Mobile Edge Computing (MEC) Server, which is connected to the Core Cloud through a secure communication channel. The Core Cloud facilitates the connection between vehicles and third-party services, including the charging station recommender system, which is the main focus of our work. However, guaranteeing the complete trustworthiness of the cloud architecture and third-party service providers in handling vehicular location data is not feasible, aligning with the honest-but-curious threat model. Consequently, our proposed architecture only shares privatised vehicular location data. The subsequent sections delve into a comprehensive description of the roles and functions of each system component.

Vehicle Tier

We fix a road network G with nodes $G(N)$ and edges $G(E)$. We choose an arbitrary edge $I \in G$, and focus on the queries made by the EVs in I 's range of coverage, $R(I)$, provided by its RSU. An EV u employs a local obfuscation technique to protect its true location $x^u \in R(I)$ to $x_1^u \in R(I)$ within the coverage area $R(I)$ of a specific edge. When an EV moves from the area of coverage of one Edge cloud to another, we can assume the queries and the privacy threats against the Edge to reset as each Edge communicates with the Cloud-based services and the third-party service providers.

Edge Tier

Given the substantial volume of data generated and exchanged between vehicles and infrastructure, the installation of edge clouds in close proximity to vehicles becomes essential to host off-board vehicular services, which require low access latency from onboard vehicular services [119]. In addition to performing essential data processing and forwarding functions, the

Edge also serves as a layer for data aggregation. Moreover, it enables the deployment of supplementary privacy-preserving measures before sharing the data with third-party entities.

Cloud Tier

It is expected to provide computation and storage capabilities for top-level processes, including data-sharing interfaces for third-party services.

Third-party Service Provider

It is the external party to ITS and is expected to enhance the quality of the function for finding the available CS for the vehicles by receiving search queries compromised of privatised and dummy location vectors for the respective vehicles.

Communication Channel

ITS comprises a network of RSU, vehicle on-board electronic control units (ECU), and distributed cloud computing and storage services. Wireless communications are enabled for V2V (Vehicle-to-Vehicle), V2I (Vehicle-to-Infrastructure) and V2X (Vehicle-to-Everything) by the technologies such as IEEE 802.11p DSRC/WAVE (Dedicated Short Range Communication/Wireless Access in Vehicular Environments), cellular advances such as C-V2X, and the long-term evolution for vehicles (LTE-V) [120]. Confidentiality of the wireless communication channel is secured by public key infrastructure (PKI) encryption methods which are beyond the scope of this work.

4.4.4 Privacy Threat Landscape

In real-time IoV location-based applications, it is often necessary for users to share their location information with the service provider in order to access location-specific services. However, this raises privacy concerns as it can potentially expose sensitive information about individuals' movements and activities. To address these concerns, data perturbation techniques can be employed to protect the privacy of users while still allowing them to access the services they need. These techniques introduce uncertainty or noise into the data, preventing an attacker from identifying the precise location of an individual. However, real-world solutions often rely on user consent, access control, and non-disclosure agreement-based mechanisms instead of providing formal privacy guarantees. Thus, there are existing privacy challenges related to shared location data, including journey tracing and location identification.

Furthermore, apart from these major privacy challenges, vehicular location data may also be susceptible to unauthorised use, data inference, retention, or insider privacy breaches within the service provider when formal privacy guarantees are lacking. The third-party provider is typically considered an honest-but-curious adversary model, assuming it is honest in accurately executing the protocol required to provide location data. However, there is a possibility that the provider may be curious about inferring users' private information based on the acquired location data [121].

This study aims to offer a formal privacy guarantee for location-based querying that can be utilised by vehicles throughout their trajectories to effectively address the associated privacy challenges with this process. To achieve this, the system is considered in three categories: (i) the vehicle users (data subject), (ii) ITS encompassing the Edge and Core Cloud Tiers (data controller and data processor), and (iii) the third party that receives the privatised data through the deployed privacy-preserving mechanisms. The third party is assumed to be an EV charging management system, which may operate under a registration-based approach for a specific area. Our focus is on mitigating the following two major sources of threats that have the potential to compromise the privacy of EVs.

Location identification: It is vital to ensure that the privatized version of the true location of the EV is within a certain radius of interest w.p. 1, making sure that the reported location is within a feasible and drivable distance away, and most importantly, within the area of coverage of the Edge where its true location lies. Therefore, we defined AGeoI as an extension of GeoI. Thus, to ensure the privacy of any given query in the road network, the EVs locally obfuscate their true locations using the truncated Laplace mechanism with their desired parameter ε and the radius of truncation r , which, in turn, decide the value of δ .

Journey tracing:

EVs may inquire about the nearest available charging station without proceeding with the query and raise further queries along the journey. Subsequently, additional queries may be raised at different points during the journey. In our model, we aim to capture this realistic setting by allowing multiple queries to be made by the EV within a single journey. However, this introduces a potential threat of approximately tracing the trajectory of the EV's journey by interpolating the locations of the queries, despite each individual location being AGeoI-protected. This is due to the fact that the obfuscated location of each query is not distinguishable from the real location, but they are not too far off from each other with a very high probability. Consequently, if a large number of queries are made within a single journey, it becomes relatively straightforward to approximate the trajectory of the EV's journey.

Cunningham et al. [112] proposed a mechanism to securely share trajectories under LDP. However, the authors in [112] assumed a model of offline sharing of the entire trajectory and, hence, sanitising it with the proposed mechanism to engender LDP guarantees. In our setting, this method cannot be directly implemented as we consider a dynamic environment where the queries made by the EVs are in real-time, with the server not having any prior knowledge of the number or the location of the queries made by a certain EV. Therefore, the mechanism of [112] cannot trivially be extended in the online location-sharing environment, and hence, the threat of adversaries able to reconstruct the journey of a particular EV with a high number of queries remains as a concern.

4.4.5 Proposed Query Model

During the journey, an EV u located within the coverage of an Edge I locally obfuscates its true location $x^u \in R(I)$ to $x_1^u \in R(I)$ using the truncated Laplace mechanism, guaranteeing AGeoI, and generate $m - 1$ feasible dummy locations $\{x_2^u, \dots, x_m^u\} \in R(I)^{m-1}$, i.e., locations that cannot

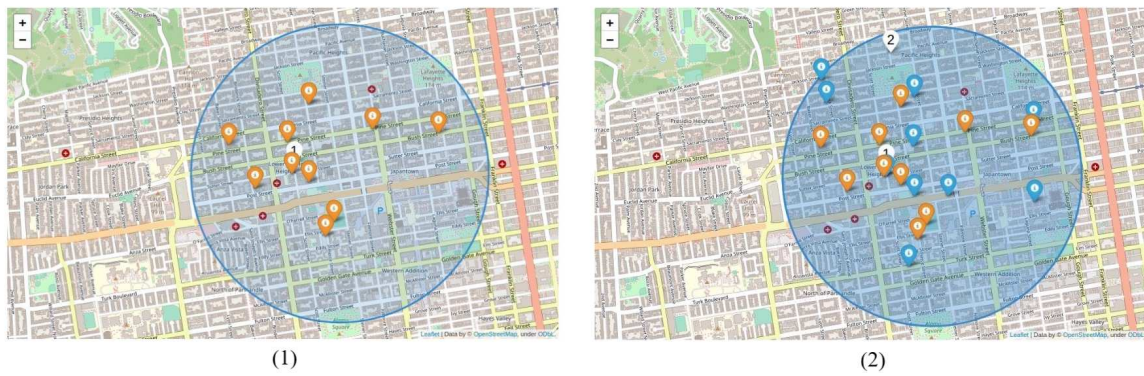


Figure 4.2: Reported dummy and privatised locations for two respective time windows (White Pins: Privatised locations, Orange Pins: Dummy locations in 1st Time window, Blue Pins: Dummy locations in 2nd Time window)

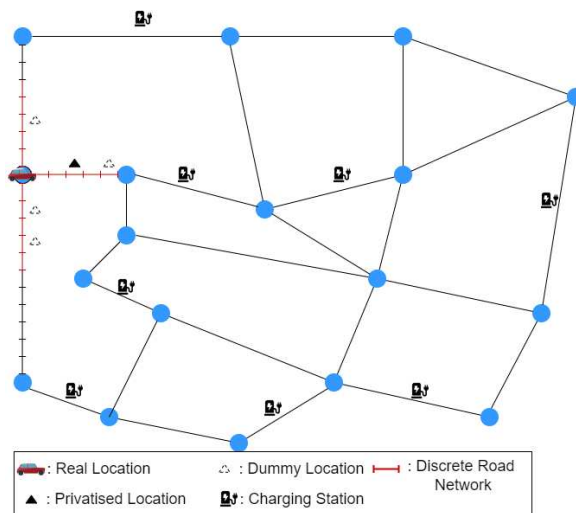


Figure 4.3: A toy example for a static location query on discrete road network

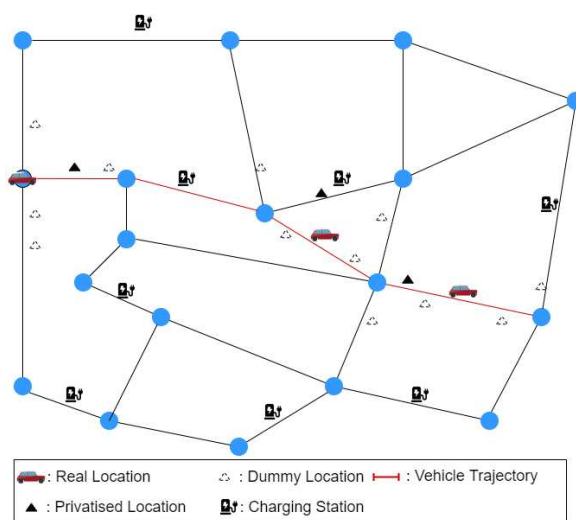


Figure 4.4: A toy example for linked 3 location queries on discrete road network

be trivially identified as being artificially generated given the query of the previous time stamp w.r.t. realistic speed limits, travelling conditions, etc. For the first query that u makes along its journey, it can generate any random $m - 1$ dummy locations in $R(I)$. Thus, each CS query of u consists of reporting the vector of m locations, $l_u = (x_1^u, \dots, x_m^u) \in R(I)^m$, to I for the Edge to process and communicate the query to the Cloud services and the third-parties to find the nearest available CS in $R(I)$ for u . This approach ensures that the adversary will have at least m possible trajectories that the EV could have realistically followed at every time stamp, making it highly improbable for the Edge and the third party to be able to conclude which of them was the actual journey as, after k queries made along a single journey, each interpolated trajectory will have a probability of at least $1/m^k$ of being the real one.

Figure 4.2 illustrates 10 reported dummy locations along with the privatized location for two consecutive time windows. Notably, the dummy locations in the subsequent time window can be feasibly linked to at least one of the preceding dummy locations. For a clearer understanding of the proposed privacy-preserving mechanism, Figures 4.3 and 4.4 present simplified examples. The former demonstrates a static query on a discrete road network, while the latter showcases linked dynamic queries on the same network.

At any given time, the Edge collects all the reported locations from the querying EVs, shuffles them by effacing the links between the location vectors and the corresponding EVs, and sends this jumbled collection of all the reported locations in the network to know their respective nearest available CS to the third-party service provider. After receiving the response, the Edge, which internally keeps the record of the IDs of the EVs against their queried locations, assigns the corresponding vector of locations of the nearest available CS to each EV and communicates them back to the respective vehicles.

In other words, at time t , if the Edge receives the location vectors from k_t querying EVs as $\mathcal{L}(t) = \{l_{u_1}, \dots, l_{u_{k_t}}\}$, the Edge is responsible for shuffling all the individual location points in these reported vectors and forward the scrambled collection of locations $\mathcal{L}'(t) = \{x_i^u : u \in \{u_1, \dots, u_{k_t}, i \in [m]\}\}$ to the Cloud and the third-party, while internally keeping a track of the IDs of the EVs to reconnect the query-response back to the corresponding users. Setting \hat{x} as the location of the nearest available charging station from location x in $R(I)$, the Edge receives $\mathcal{R}(t) = \{\hat{x}_i^u : u \in \{u_1, \dots, u_{k_t}, i \in [m]\}\}$ as the response from the third-party service provider handling the CS data real-time. After this, matching the stored IDs of the EVs with the locations of the CS, the Edge communicates the response vector $\hat{l}_u = (\hat{x}_i^u : i \in [m])$ back to the corresponding EV u . Then the EV can choose to navigate to $\operatorname{argmin}_{x \in \hat{l}_u} \{d_G(x, x_u)\}$, where x_u is the real location of EV u . The overview of this mechanism is explained in Figure 4.1.

4.5 Cost of privacy analysis

DEFINITION 4.5.1 (Cost of privacy). Suppose an EV u at location x^u chooses to locally obfuscate its real location of a query as x_1^u using the truncated Laplace mechanism $\mathcal{L}_{\varepsilon, r}$ satisfying (ε, δ) -geo-indistinguishability with a corresponding radius of truncation r . Then we define the *cost of privacy (CoP)* of EV u as

$\text{CoP}(u, \mathcal{L}_{\varepsilon, r}) = \mathbf{c}(x^u, \hat{x}_1^u) - \mathbf{c}(x^u, \hat{x}^u)$, where \hat{x}^u and \hat{x}_1^u are the nearest available CS in the network to x^u and x_1^u , respectively, and $\mathbf{c} : G(N)^2 \mapsto \mathbb{R}^+$ is any cost function that reflects the “cost” of the commute from locations x to y in the network.

In other words, CoP, as in Definition 4.5.1, essentially captures the *extra* cost that an EV needs to cover as a result of the privatized location it reports to the Edge satisfying AGeoI, as opposed to its true location. For the purpose of simplicity of the analysis, we considered the cost function as the travelling distance in the network, i.e., $\mathbf{c} = d_G$. However, in practice, any suitable cost function could be used (e.g. fuel efficiency, time, etc.) could be used as \mathbf{c} , depending on the context and requirement of the architecture.

To formally characterize and analyze the CoP of the EVs in the network, inspired from the classical version of *Voronoi decomposition*, we extend the concept in the setting of our road network in the network coverage for a fixed Edge w.r.t. graph-traversal distance, d_g .

DEFINITION 4.5.2 (Voronoi decomposition). Let G be the graph representing the road network equipped with travelling distance d_G . Let the set of CS in G be $C_G = \{c_1, \dots, c_{n_G}\}$. Then the *Voronoi decomposition* on G w.r.t. C_G is defined as $\mathbf{V}_G = \{V_i : i \in [n_G]\}$ such that $V_i \cap V_j = \emptyset$ for any $i \neq j$ and $\bigcup_{i \in [n_G]} V_i = G$, where

$$V_i = \{x \in G : d_G(x, c_i) \leq d_G(x, c_j) \forall j \in [n_G], j \neq i\}$$

DEFINITION 4.5.3 (Closed ball around a location). For any $x \in G$ and $r \in \mathbb{R}_{\geq 0}$, the *closed ball* of x of radius r is defined as $\beta_r(x) = \{y \in G : d_G(x, y) \leq r\}$

DEFINITION 4.5.4 (Fenced Voronoi decomposition). For any $r \in \mathbb{R}_{\geq 0}$ and charging station i , let the *r-fenced Voronoi decomposition* on road network G be defined as $V_G^{-r} = \{V_i^{-r} : i \in [n_G]\}$ such that $V_i^{-r} \cap V_j^{-r} = \emptyset$ for $i \neq j$ and $V_i^{-r} = \{x \in V_i : B_r(x) \subseteq V_i\}$. In other words, V_i^{-r} essentially constructs an area contained within V_i restricted by a *fence* at a distance r from the edge of V_i .

THEOREM 4.4. Suppose an EV u positioned at x^u on G obfuscates its location using AGeoI with any radius of truncation $r \in \mathbb{R}_{\geq 0}$. Let \hat{x}^u be the location of the nearest available charging station to the true location x^u . Then $\mathbb{P}[\text{CoP}(u, \mathcal{L}_{\varepsilon, r}) = 0] = 1$ for every $x^u \in V_{\hat{x}^u}^{-r}$. In other words, if an EV lies in the r -fenced Voronoi decomposition for its nearest available CS, it has a *zero cost for privacy* w.p. 1.

Proof. Immediate from Definition 4.5.4. □

THEOREM 4.5. Suppose an EV u lies in $V_{\hat{x}^u} \setminus V_{\hat{x}^u}^{-r}$ and it uses AGeoI to obfuscate its true location x^u to x_1^u with a radius of truncation r and privacy parameter ε for making a private query to the Edge. Then $\mathbb{P}[\text{CoP}(u, \mathcal{L}_{\varepsilon, r}) = 0] =$

$\sum_{x_1^u \in V_{\hat{x}^u}} c e^{-\varepsilon d_G(x^u, x_1^u)}$, where c is the normalizing constant of the truncated Laplace mechanism as in Definition 4.3.2.

Proof. To compute $\mathbb{P}[\text{CoP}(u, \mathcal{L}_{\varepsilon, r}) = 0]$, we only need to exclude the possibilities where the reported location of the EV lies outside the Voronoi decomposition of the station \hat{x}^u , which, essentially, is $\sum_{x_1^u \in V_{\hat{x}^u}} c e^{-\varepsilon d_G(x^u, x_1^u)}$. \square

4.6 Experimental Study

This section presents the experimental study with the objectives as follows: (i) to validate proposed theoretical claims and solutions empirically; (ii) to use the method to find the nearest available charging station for EVs as a case study; (iii) to investigate the cost of privacy in real-time settings; and (iv) to conduct a real-time CS occupancy prediction technique from the noisy vehicle distribution.

4.6.1 Dataset Preparation

The road network data extracted from OpenStreetMap [122]. The cost of privacy is calculated as the additional routing distance caused by noise in vehicular locations during queries to identify the optimal charging station. The cost of privacy depends on the sparsity of CS. We prepared two datasets: one with 404 existing charging station locations in San Francisco obtained from the United States Department of Energy [123], and another by merging existing and planned charging station locations with on-street and off-street parking locations from DataSF [124], resulting in 716 independently distributed locations.

The EPFL mobility dataset includes GPS records of 536 taxi trajectories in San Francisco over four weeks [125]. The dataset provides information such as the taxi identifier, latitude, longitude, occupancy state (vacant or occupied), and a UNIX epoch timestamp. Leveraging the occupancy information, we were able to split the complete taxi trajectories into individual customer trajectories, resulting in over 450,000 exported trajectories. For our study, we randomly selected 536 trajectories from each taxi.

4.6.2 Experimental Setup

A group of EVs sends out location queries to find the closest available CS during their journeys on the road network G . The edges of the road network G are truncated into discrete segments with an equal k travel distance, similar to the work in [111]. DTLap is utilised to generate the privacy channel by using the Laplace mechanism for the user's desired values for privacy budget ε and truncation radius r . Following this, DTLapSamp is used to generate privatised locations with respect to the users' real locations.

A location query contains a privatised location and $m - 1$ dummy locations as a vector and is collected by the Edge for sending them to the third party through the core cloud as a single vector of all locations. The third party responds to the locations in the vector with the closest

available CS for each, and the Edge sends vehicle location vectors to the related vehicles without being able to differentiate privatised and dummy locations.

For IBU to approximate the original distribution of the query locations of the EVs in the road network in order to predict the availability of the CS and, thus, assist the users in planning their journeys appropriately, we note that each original query location goes through two independent steps of sanitization: a) locally using the truncated Laplace mechanism to achieve AGeoI and b) generating the realistic dummy locations in the area of coverage of the Edge to ensure protection against attacks reconstructing their journeys. Setting the domain \mathcal{X} as the area of coverage of the RSU of the fixed Edge that we focus on, while the former is a straightforward use of the channel \mathcal{L} , the latter can be thought of as $m - 1$ independent applications of the uniform channel \mathcal{U} , where $U : \mathcal{X}^2 \mapsto \mathbb{R}$ with $\mathcal{U}_{x,y}$ denoting $\mathbb{P}_{\mathcal{U}}[y|x] = 1/|\mathcal{X}|$, by each EV. Therefore, after accounting for the normalization, the channel incorporating the local obfuscation and the generation of the dummy locations used by each EV reduces down to $\frac{1}{m}\mathcal{L} + \frac{m-1}{m}\mathcal{U}$ which we use as the privacy channel to implement IBU.

The first set of experiments examines the CoP for randomly selected 536 vehicle traces, where each trace contains a series of GPS coordinates and 3 randomly selected points along each for the real locations of the queries. The discrete road network is generated by setting the distance $k = 100$ meters. The parameters of ε and r are varied in the range of 0.2 to 2, and 1 to 20, respectively.

The privatised location, together with the dummy locations, is sent to the third-party for a query to prevent the third-party from tracking the vehicle. The area of Edge coverage, rather than the vehicle's area of interest, is considered for dummy location generation rather than the vehicle's area of interest, as the centre of mass may give away the true location. The second set of experiments examined the impact of dummy locations on the CoP.

The location queries could be used for real-time predictive analysis on the optimisation of the smart power grid, managing staff, and determining where new CS should be deployed. Thus, service providers can have the utility of the datasets (e.g., train ML models, etc.) with DP-based methods while the privacy of individuals is preserved. The third set of experiments utilises the IBU method to retrieve the true distribution of locations of the queries from the noisy distribution, which includes privatised and dummy locations.

4.6.3 Results and Discussion

Cost of Privacy: empirical results

DP approaches introduce a trade-off between privacy and data utility, with a higher level of privacy requiring a greater level of noise. The efficacy of the respective service may correspondingly decrease due to the fall in data utility, and this difference in the *quality of service* is referred to as the 'cost of privacy' (CoP) in this study. In particular, in the context of the use case considered in this work, the CoP is formalised in Definition 4.5.1.

The following results are achieved by carrying out the experiments for 3 linked queries of 536 randomly selected vehicle trajectories for varying values of ε or r ranging from 0.2 to 2, and

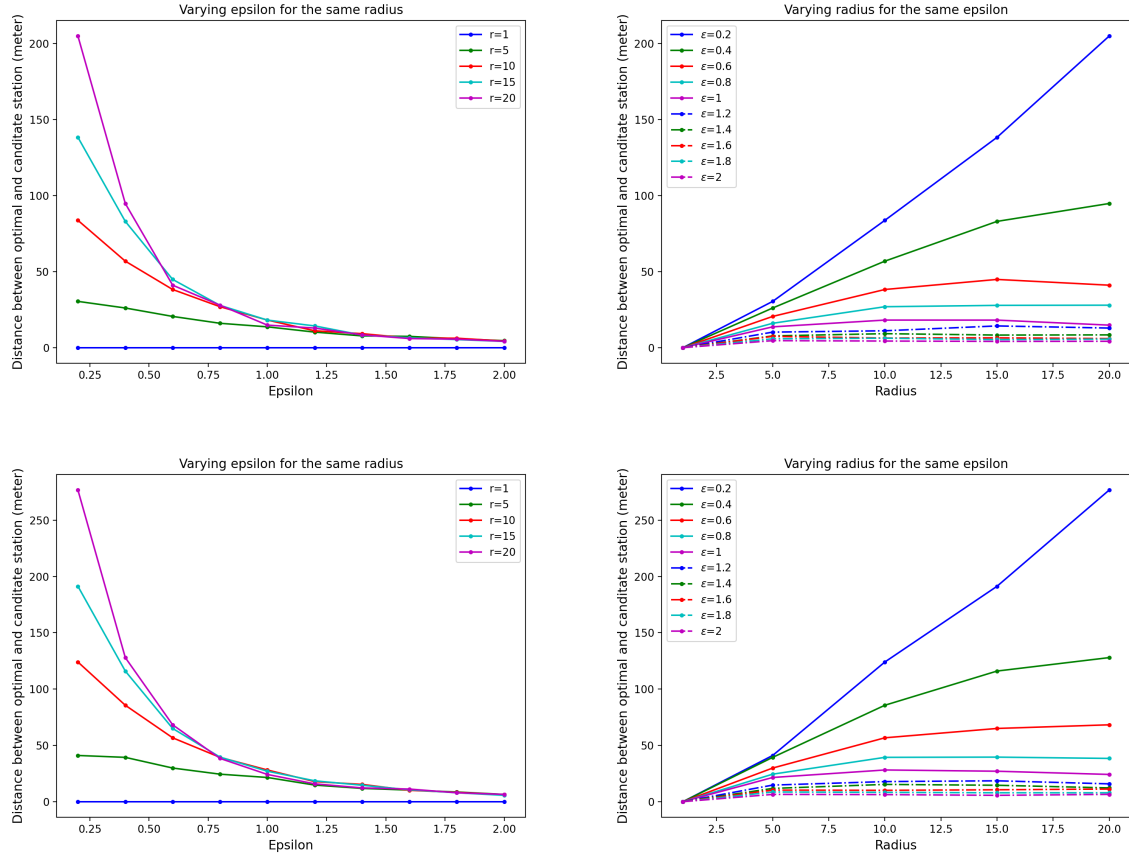


Figure 4.5: CoP (i.e., by Definition 4.5.1, the difference in the distance an EV needs to cover to reach the nearest CS with and without local obfuscation to achieve AgeoI) for varying ϵ or r of AGEoI (1st row is for sparse CS, 2nd row is for dense CS).

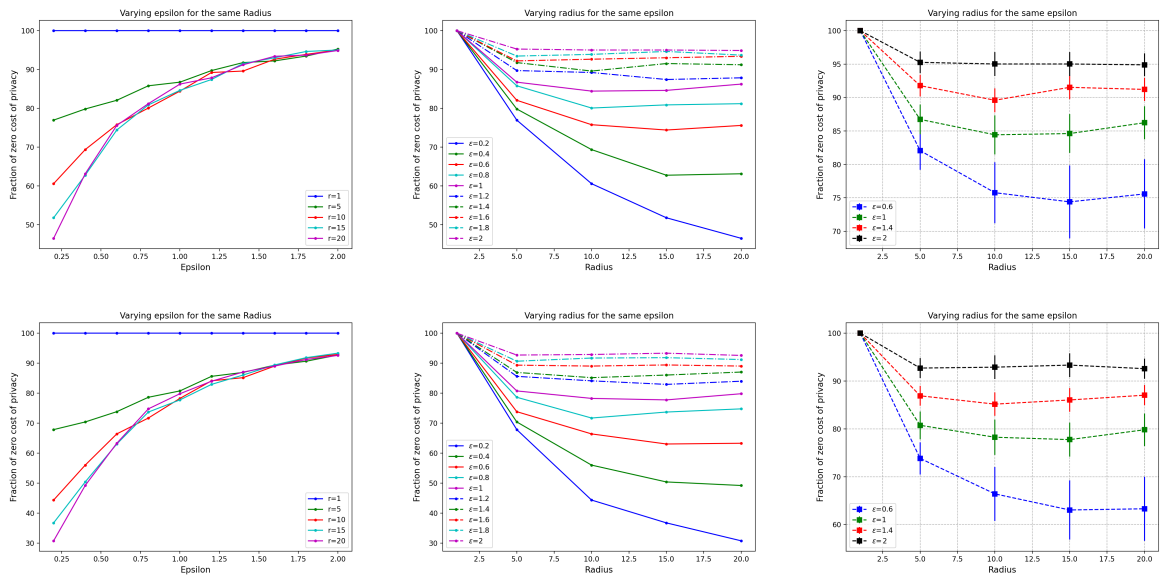


Figure 4.6: Fraction of EVs incurring no CoP for varying ϵ or r of AGEoI (1st row is for sparse CS, 2nd row is for dense CS).

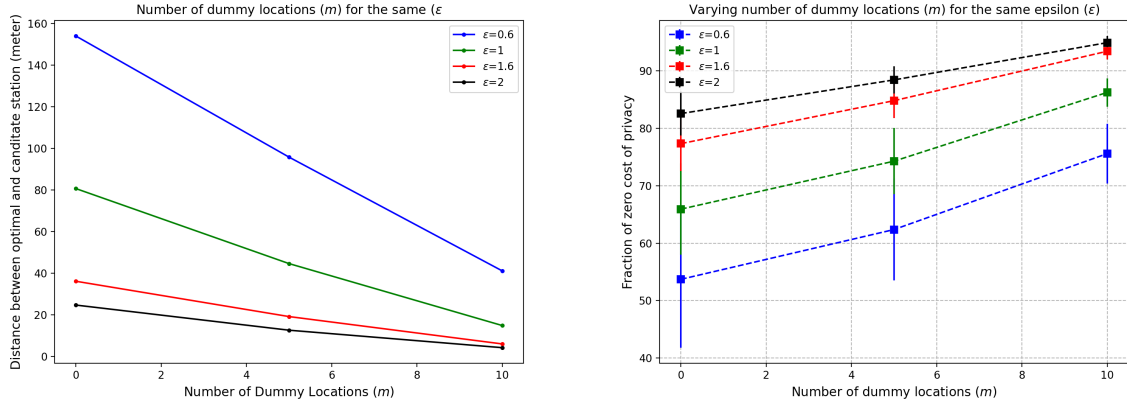


Figure 4.7: Impact of introducing dummy locations along with AGeoI on the CoP.

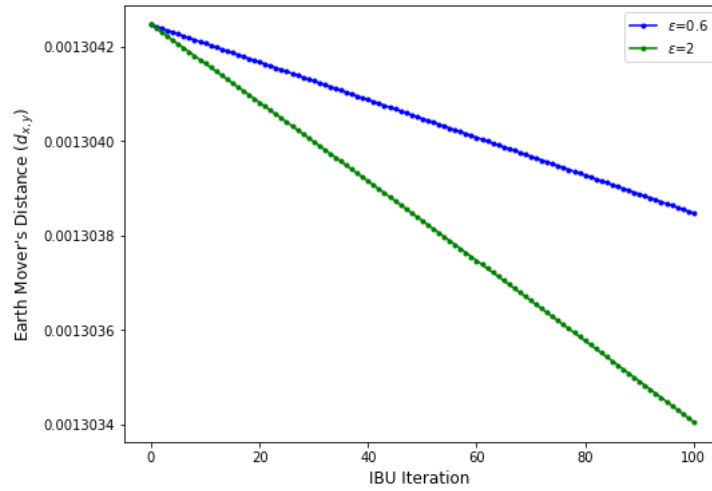


Figure 4.8: The Kantorovich-Wasserstein distance between the original and estimated distributions using IBU for $\epsilon = 0.6$ and $\epsilon = 2$ noisy distributions.

1 to 20, respectively. Figure 4.5 demonstrates the CoP in terms of the extra travelling distance due to the privacy-preserving mechanism, where a similar pattern is observed for both of the datasets. Another observation is that a high frequency of queries resulted in no cost for privacy preservation. Figure 4.6 shows the fraction of the queries with “privacy for free” where both datasets followed similar patterns. Vehicle queries contain dummy locations and their privatised true locations. It is possible that the dummy locations can sometimes provide a better utility, but our experiments consider the utility of a privatised location as the worst-case for analysis.

Figure 4.5 shows that our method provides a negligible cost of utility loss for the formal privacy gain enjoyed by the EVs. By increasing the truncation radius, an abrupt drop in the distance between the location of the nearest available charging station for the true location of the query and that of the privatised one implies that the cost of the extra travel distance needed to be taken due to the AGeoI guarantee is almost negligible. A similar trend is seen for the varying ϵ with a fixed radius. As the level of privacy decreases, the fraction of EVs in the network enjoying *privacy for free* grows to be more than 60% for a radius of truncation of merely

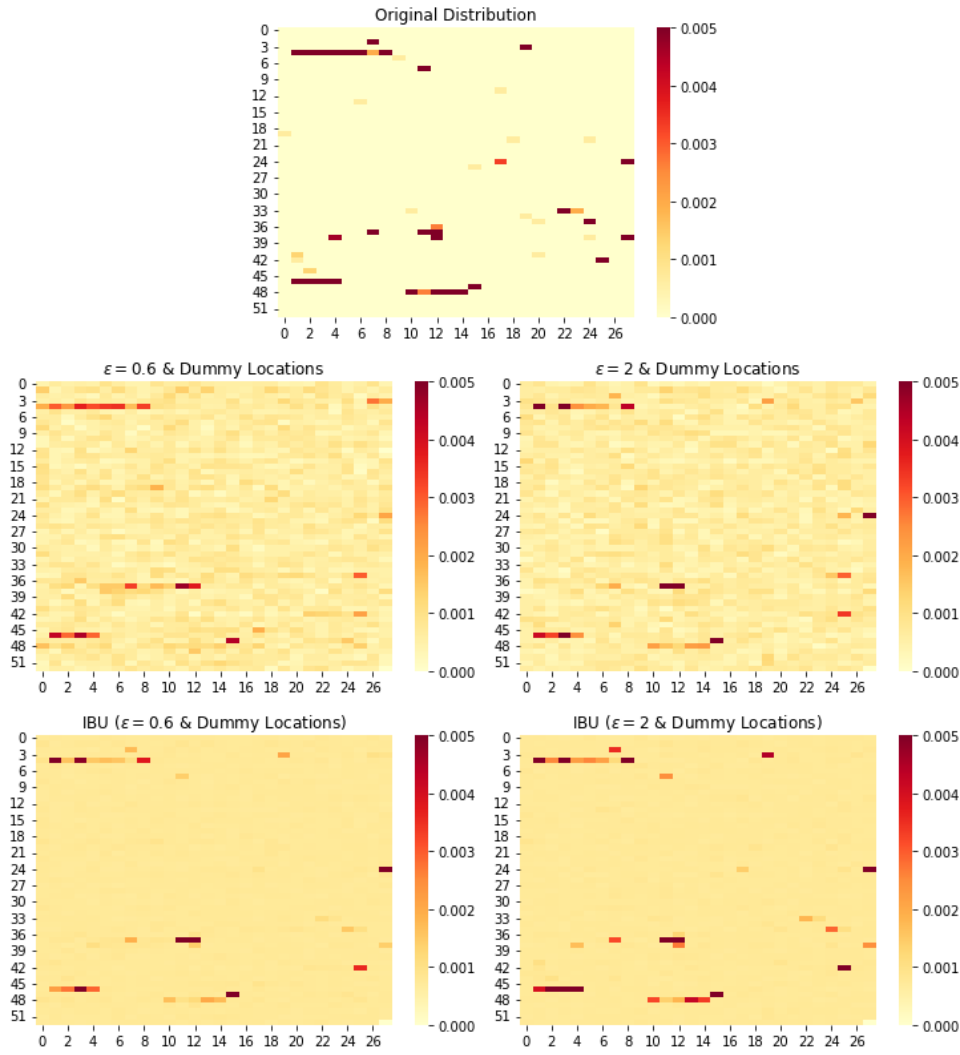


Figure 4.9: Estimations of the original distribution using IBU for the $\varepsilon = 0.6$ and $\varepsilon = 2$ noisy distributions.

10 road segments, where each segment is 100 meters long, for $\varepsilon \geq 0.5$. However, more than 90% of the EVs achieve a zero cost of privacy for $\varepsilon \geq 1.5$, irrespective of the truncation radius as illustrated in Figure 4.6. Due to increasing perturbation in disclosed location, the width of the confidence interval for zero cost of privacy increases, as seen in Figure 4.6. The likelihood of achieving zero cost of privacy fluctuates over a wider range and it does not monotonically decrease with the growing radius due to rising randomness.

Impact of Dummy Data Generation

Considering an adversary interested in finding the true locations of the EVs, $(\alpha, \beta]$ -*identifiability* is defined for any location x as $\mathbb{P}[d(y, x) < \alpha] > \beta$, where y is any guessed location by the adversary. With the proposed method, with a sufficiently small radius of truncation to obfuscate the true location using the truncated Laplace and generating $m - 1$ dummy locations in the area of coverage of the Edge, the probability of hitting the true x within an error of α is $\mathbb{P}[d(x, y) < \alpha] = \frac{1}{m} c e^{-\varepsilon \alpha} = \beta$, where c is the normalising constant.

There has been some work in this area from the perspective of just GeoI [109–111, 126, 127] or just from the standpoint of generating dummy locations exploiting anonymisation techniques [128, 129]. One of the first major concerns in using only GeoI is when we allow dynamic and multiple queries along the journey of the EVs, as individual locations, despite being privatised, can still be interpolated to approximate the entire trace. If only dummy locations are used, however, any estimated (or observed) y could be the real location w.p. $\frac{1}{m-1}$, as there is no formal privacy guaranteed, i.e., every location x has, is $(0, 1/m - 1)$ -identifiable among $(m - 1)$ dummy locations. With potential parallel processing, brute-force attacks are just one way that it has been shown that anonymisation techniques are not sufficient to protect privacy [130].

Figure 4.7 illustrates how the CoP increases with an increase in the noise due to the lack of dummy locations under the same level of identifiability. To achieve the same (α, β) -identifiability with just AGeoI without dummy locations, the parameter ε needs to be scaled by $\frac{1}{\ln m}$, i.e., more noise needs to be added, which results in having a worse trade-off between privacy and CoP for the same level of privacy.

Real-time Predictive Study

Predicting the availability of CS is a crucial component of EV trip planning and can help ease range anxiety. Some existing methods adhered to machine learning-based approaches to develop such prediction models [131–135]. The main consideration of these models is that drivers can book timeslots for CS and the prediction is made based on factors such as past usage of CSs, traffic density, and some other features such as weather conditions. However, such consideration may be limited to facilitating effective EV journey planning, given that traffic is highly dynamic and subject to unexpected changes. Due to the traffic, EVs may be late for their scheduled charging time, and another EV cannot be navigated to charge from the same station, despite the fact that it is empty. Hence, a real-time predictive analysis would be critical to determine the likelihood of a CS being available when an EV arrives. Our proposed method provides privacy-preserved live traffic distribution of the querying vehicles. IBU is applied to estimate the real-time distribution of the traffic and, hence, the statistical distance between the estimated and the original distributions are shown in Figure 4.9. We considered two different levels of AGeoI with $\varepsilon = 0.6$ and $\varepsilon = 2$ and IBU was run for 100 iterations. The results demonstrate that the distance between the original and the estimated distributions of the traffic is decreasing. The accuracy of the estimation of the original distribution from the noisy locations is illustrated by the heatmaps of Figure 4.9 depicting the original, noisy, and estimated traffic distributions. This essentially highlights the high statistical utility of our proposed method and, specifically, helps in the prediction of how likely a CS will be available when the vehicle arrives and the traffic, in general.

Part III

Federated learning

“Learn continually. There’s always one more thing to learn.”

– Steve Jobs

Advancing Personalized Federated Learning: Group Privacy, Fairness, and Beyond

5.1 Introduction

The widespread collection of user data in modern machine learning has raised concerns regarding privacy violations and the potential disclosure of sensitive personal information [136, 137]. To address these concerns, Federated Learning [26] was introduced as a collaborative machine learning paradigm, where users' devices train a global predictive model without transmitting raw data to a central server. While FL offers promises of preserving user privacy and maintaining model performance, the heterogeneity of data distributions among clients can lead to challenges such as reduced model utility and convergence issues during training. In response, personalized federated learning approaches have emerged, aiming to tailor models to clusters of users with similar data distributions [138–140].

Furthermore, it has been demonstrated that avoiding the release of users' raw data alone does not provide sufficient protection against potential privacy violations [141–143]. To address this issue, researchers have explored the application of Differential Privacy (DP) [13, 14] to federated learning, providing privacy guarantees for users participating in the optimization process. DP mechanisms introduce randomness in the model updates released by clients, making each user's contribution to the final model probabilistically indistinguishable up to a certain likelihood factor. To bound this factor, the domain of secrets (i.e., the parameter space in FL) is artificially constrained, either to offer central [144, 145] or local DP guarantees [146, 147]. However, constraining the optimization process to a subset of \mathbb{R}^n can have negative effects, such as when the optimal model parameters for a particular cluster of users lie outside such a

bounded domain.

To address the challenges of personalization and local privacy protection, we propose the adoption of a more general notion of DP called d -privacy or metric-based privacy [17] which has been in the spotlight of late mainly in the context of location-privacy [55, 148, 149]. This concept of privacy does not require a bounded domain and provides guarantees based on the distance between any two points in the parameter space. Therefore, assuming that clients with similar data distributions have similar optimal fitting parameters, d -privacy offers strong indistinguishability guarantees. Conversely, privacy guarantees degrade gracefully for clients with significantly different data distributions.

In addition to addressing privacy concerns, this work also investigates the impact of our method on fairness aspects in federated model training, extending the analysis conducted in [150]. As machine learning-based decision systems become more prevalent, it has become apparent that many of these systems exhibit gender and racial biases that disproportionately affect minority populations [151, 152]. Therefore, beyond protecting user privacy, it is crucial to explore cutting-edge machine learning algorithms that can potentially mitigate this pervasive lack of fairness among participating clients. However, we have observed that systems aiming to protect privacy while ensuring fairness often involve a trade-off between the two [153]. This trade-off arises because privacy protection techniques based on DP tend to minimize the impact of outliers or minorities within the overall dataset. In other words, the application of d -privacy, a metric-based generalization of DP, to personalized FL could potentially compromise the fairness of the machine learning model. This work presents extensive experimental results demonstrating that the use of personalized FL under group privacy guarantees not only significantly improves fairness compared to the classical (non-personalized) FL framework, but also maintains a relatively small trade-off between privacy and fairness.

Contributions

The key contributions of this chapter are:

1. We propose an algorithm for collaborative training of machine learning models, leveraging advanced techniques for model personalization and addressing user privacy concerns by formalizing privacy guarantees in terms of d -privacy.
2. Our research focuses on studying the Laplace mechanism under Euclidean distance and providing a closed-form expression for its generalization in \mathbb{R}^n , as well as an efficient sampling procedure.
3. We show that personalized federated learning under formal privacy guarantees improves group fairness significantly compared to the non-personalized federated learning framework and, hence, establish that our method enhances the trade-off between privacy and fairness.

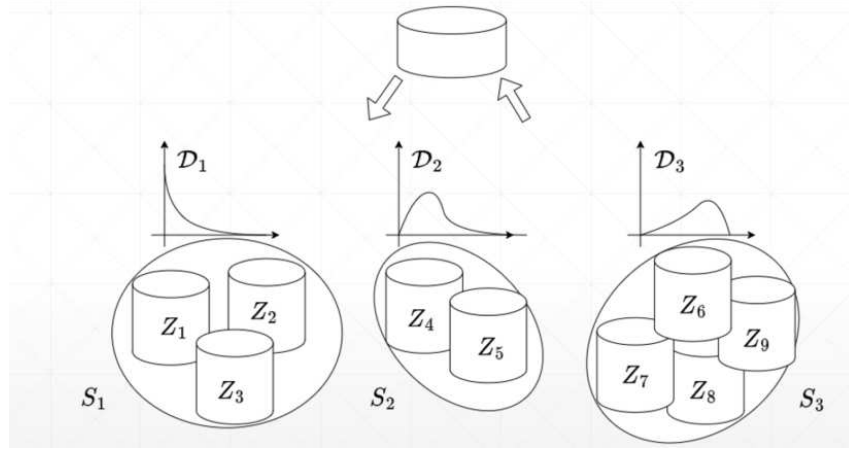


Figure 5.1: A schematic overview of the personalized approach of federated model training

5.2 Technical Preliminaries

5.2.1 Personalized Federated Learning

The problem of personalized federated learning falls within the framework of stochastic optimization, and we adopt the notation from [138] to determine the set of minimizers $\theta_j^* \in \mathbb{R}^n$ with $j \in \{1, \dots, k\}$ of the cost functions

$$F(\theta_j) = \mathbb{E}_{z \sim \mathcal{D}_j} [f(\theta_j; z)], \quad (5.1)$$

where $\{\mathcal{D}_1, \dots, \mathcal{D}_k\}$ are the data distributions which cannot be accessed directly but only through a collection of client datasets $Z_c = \{z | z \sim \mathcal{D}_j, z \in \mathbb{D}\}$ for some $j \in \{1, \dots, k\}$ with $c \in C = \{1, \dots, N\}$ the set of clients, and \mathbb{D} a generic domain of data points. C is partitioned in k disjoint sets

$$S_j^* = \{c \in C \mid \forall z \in Z_c, z \sim \mathcal{D}_j\} \quad \forall j \in \{1, \dots, k\} \quad (5.2)$$

The mapping $c \rightarrow j$ is unknown and we rely on estimates S_j of the membership of Z_c to compute the empirical cost functions

$$\begin{aligned} \tilde{F}(\theta_j) &= \frac{1}{|S_j|} \sum_{c \in S_j} \tilde{F}_c(\theta_j; Z_c); \\ \tilde{F}_c(\theta_j; Z_c) &= \frac{1}{|Z_c|} \sum_{z_i \in Z_c} f(\theta_j; z_i) \end{aligned} \quad (5.3)$$

The cost function $f: \mathbb{R}^n \times \mathbb{D} \mapsto \mathbb{R}_{\geq 0}$ is applied on $z \in \mathbb{D}$, parametrized by the vector $\theta_j \in \mathbb{R}^n$. Thus, the optimization aims to find, $\forall j \in \{1, \dots, k\}$,

$$\tilde{\theta}_j^* = \underset{\theta_j}{\operatorname{argmin}} \tilde{F}(\theta_j) \quad (5.4)$$

5.2.2 Fairness

With the recent surge of interest in building ethical ways to train machine learning models, the topic of fairness in machine learning has been in the spotlight and, correspondingly, various metrics and algorithms to quantify and establish fairness in model training have been studied from a variety of perspectives and in different contexts [154–156]. Most fairness metrics consider the simple case of having a *privileged* group and an *unprivileged* group in the population. Under this assumption, typically one attribute of the dataset is selected as a sensitive attribute (e.g., gender, race, etc.) that defines the privileged and the unprivileged groups. The goal of fairness in machine learning is to ensure fair and non-discriminated results regardless of the membership in a sensitive attribute. The two main notions of fairness considered by the community are individual fairness and group fairness: *Individual fairness* [157] claims that similar individuals should be treated similarly, and *group fairness* requires that different demographic subgroups should receive equal treatment with respect to their sensitive attributes. While both notions of fairness are important, in this work, we focus on group fairness because our goal is to analyze and mitigate the potential bias against certain groups (e.g. demographic groups) through personalization techniques. In particular, we considered the following metrics for evaluating group fairness as a part of this work.

Without loss of generality, we shall use $S = 1$ to represent the privileged group and $S = 0$ to represent the unprivileged group in the rest of the chapter.

DEFINITION 5.2.1. *Equal opportunity* [158] is satisfied by a system if its prediction \hat{Y} is conditionally independent of the sensitive attribute S given the target label Y

$$\mathbb{P}[\hat{Y} = 1|Y = 1, S = 1] = \mathbb{P}[\hat{Y} = 1|Y = 1, S = 0] \quad (5.5)$$

In other words, equal opportunity is satisfied if the system produces equal true positive rates across the privileged and unprivileged groups.

DEFINITION 5.2.2. A system satisfies *equalized odds* [158] if its prediction \hat{Y} is conditionally independent of the sensitive attribute S given the target label Y ,

$$\mathbb{P}[\hat{Y} = 1|Y = y, S = 1] = \mathbb{P}[\hat{Y} = 1|Y = y, S = 0], \quad y \in \{0, 1\} \quad (5.6)$$

In other words, in order to ensure having equalized odds, a system requires the privileged and unprivileged groups to have equal true positive rates and equal false positive rates.

DEFINITION 5.2.3. *Demographic parity* [157] is achieved by a system when the prediction \hat{Y} of the target label Y is statistically independent of the sensitive attributes S , i.e.,

$$\mathbb{P}[\hat{Y}|S = 1] = \mathbb{P}[\hat{Y}|S = 0] \quad (5.7)$$

In practice, under the aforementioned metrics, we might not require the difference between

the privileged and unprivileged groups to be exactly zero but we aim to minimize this gap.

5.3 Related Works

Federated optimization has demonstrated suboptimal performance when the local datasets consist of samples from non-congruent distributions, resulting in the inability to simultaneously minimize both client-level and global objectives. In previous studies [138–140], researchers examined various meta-algorithms for personalization, but the assertion of preserving user privacy relies solely on clients releasing updated models or model updates, rather than transferring raw data to the server, which can have significant consequences. To address this issue, several works have focused on the privatization of the (federated) optimization algorithm within the framework of DP [27, 144, 145, 159], which adopt DP to provide defences against an *honest-but-curious* adversary. However, even in this setting, there is no guarantee of protection against sample reconstruction from the local datasets using client updates, as highlighted in [143]. Various strategies have been explored to offer local privacy guarantees, either through cryptographic approaches [160] or within the framework of local DP [146, 161, 162]. Specifically, in [162], the authors tackle the problem of personalized and locally differentially private federated learning, but only for the case of simple convex, 1-Lipschitz cost functions of the inputs. It is worth noting that this assumption is unrealistic in the majority of machine learning models and excludes many statistical modelling techniques, particularly neural networks. Finally, some research focused on designing architectures capable of providing private computing environments for remote users [163], often making use of trusted platform modules, secure processors [164], or similar mechanisms [165] improving efficiency by enforcing encryption on network transmissions, rather than memory accesses. For example, the latter work conceptualizes an architecture that could be leveraged to deploy a server that can only reveal the data being processed to clients that instantiated the server. It shall be noted, however, that cryptographic guarantees of security are orthogonal to the privacy notions of differential privacy and its generalizations.

Of late, a great deal of attention has been devoted to studying and understanding the aspects of fairness in machine learning [153, 166–171]. Most of the research on fairness focuses on developing techniques to mitigate bias in machine learning algorithms. These techniques can be categorized into three main approaches: pre-processing, in-processing, and post-processing. Pre-processing techniques [172, 173] aim to generate a less biased dataset by modifying the values or adjusting the sampling process. In the case of in-processing techniques [174, 175], the objective function is optimized while taking into account discrimination-aware regularizers. Post-processing techniques [176, 177] involve adjusting the trained model to produce fairer outcomes. However, it is worth noting that the majority of these studies primarily target centralized machine learning models as opposed to FL. Furthermore, there is a lack of research exploring the interplay between accuracy and fairness [169, 170] or privacy and fairness [153, 178]. In particular, to the best of our knowledge, disproportionately fewer works have focused on investigating the relationship between privacy and fairness. [153] formally proved that privacy and fairness can be at odds with each other with non-trivial accuracy. A few recent works on group

fairness in FL have emerged [167, 168] but they do not consider the facet of privacy-fairness trade-off.

5.4 An Algorithm for Private and Personalized Federated Learning

We propose an algorithm for personalized federated learning with local guarantees to provide group privacy (Algorithm 6). Locality refers to the sanitization of the information released by the client to the server, whereas group privacy refers to indistinguishability with respect to a neighbourhood of clients, defined with respect to a certain distance metric. Thus we proceed to define *neighbourhood* and *group*.

DEFINITION 5.4.1. For any model parametrized by $\theta_0 \in \mathbb{R}^n$, we define its r -*neighbourhood* as the set of points in the parameter space which are at a L_2 distance of at most r from θ_0 , i.e., $\{\theta \in \mathbb{R}^n : \|\theta_0 - \theta\|_2 \leq r\}$. Clients whose models are parametrized by $\theta \in \mathbb{R}^n$ in the same r -neighbourhood are said to be in the same *group*, or *cluster*.

Algorithm 6 is motivated by the Iterative Federated Clustering Algorithm (IFCA) [138] and builds on top of it to provide formal privacy guarantees. The main differences lie in the introduction of the SanitizeUpdate function described in Algorithm 7 and k -means for server-side clustering of the updated models.

Algorithm 6: An algorithm for personalized federated learning with formal privacy guarantees in local neighbourhoods.

Input: number of clusters: k , initial hypotheses: $\theta_j^{(0)}, j \in \{1, \dots, k\}$, number of rounds: T , number of users per round: U , number of local epochs: E , local step size: s , user batch size: B_s , noise multiplier: ν , local dataset held by user c : Z_c ;

Server-side loop;

for $0 \leq t \leq T - 1$ **do**

$C^{(t)} \leftarrow \text{SAMPLEUSERSUBSET}(U)$;

$\text{BROADCASTPARAMETERVECTORS}(C^{(t)}; \theta_j^{(t)}, j \in \{1, \dots, k\})$;

Client-side loop;

for $c \in C^{(t)}$ **do**

in parallel;

$\bar{j} = \text{argmin}_{j \in \{1, \dots, k\}} F_c(\theta_j^{(t)}; Z_c)$;

$\theta_{\bar{j}, c}^{(t)} \leftarrow \text{LOCALUPDATE}(\theta_{\bar{j}}^{(t)}; s; E; Z_c)$;

$\hat{\theta}_{\bar{j}, c}^{(t)} \leftarrow \text{SANITIZEUPDATE}(\theta_{\bar{j}, c}^{(t)}; \nu)$;

$\{S_1, \dots, S_k\} = \text{k-means}(\hat{\theta}_{j, c}^{(t)}, c \in C^{(t)}; \theta_j^{(t)}, j \in \{1, \dots, k\})$;

$\theta_j^{(t+1)} \leftarrow \frac{1}{|S_j|} \sum_{c \in S_j} \hat{\theta}_{j, c}^{(t)}, \quad \forall j \in \{1, \dots, k\}$;

Algorithm 7: SanitizeUpdate obfuscates a vector $\theta \in \mathbb{R}^n$ with a Laplacian noise tuned on the radius of a certain neighbourhood and centred in 0.

```

function SANITIZEUPDATE( $\theta_{\bar{j},c}^{(t)}; \nu$ )
     $\delta_c^{(t)} = \theta_{\bar{j},c}^{(t)} - \theta_{\bar{j}}^{(t)}$ 
     $\varepsilon = \frac{n}{\nu \|\delta_c^{(t)}\|}$ 
    Sample  $\rho \sim \mathcal{L}_\varepsilon(0)$ 
     $\hat{\theta}_{\bar{j},c}^{(t)} = \theta_{\bar{j},c}^{(t)} + \rho$ 
    return  $\hat{\theta}_{\bar{j},c}^{(t)}$ 
end function

```

5.4.1 The Laplace mechanism under Euclidean distance in \mathbb{R}^n

Algorithm 7's SanitizeUpdate is based on a generalization of the Laplace mechanism under Euclidean distance to \mathbb{R}^n , introduced in [179] for geo-indistinguishability in \mathbb{R}^2 . The motivation to adopt the L_2 norm as a measure of distance is twofold. First, clustering is performed on θ with the k -means algorithm under Euclidean distance. Since we define clusters or groups of users based on how close their model parameters are under L_2 norm, we are looking for a d -privacy mechanism that obfuscates the reported values within a certain group and allows the server to discern among users belonging to different clusters. Second, parameters that are sanitized by equidistant noise vectors in L_2 norm are also equiprobable by construction and lead to the same bound in the increase of the cost function in first-order approximation, as shown in Proposition 5.2. The Laplace mechanism under Euclidean distance in a generic space \mathbb{R}^n is defined in Proposition 5.1.

PROPOSITION 5.1. Abusing the notation $\mathbb{P}[\cdot]$ to denote the PDF, let $\mathcal{L}_\varepsilon: \mathbb{R}^n \mapsto \mathbb{R}^n$ be the Laplace mechanism with distribution $\mathcal{L}_{x_0, \varepsilon}(x) = \mathbb{P}[\mathcal{L}_\varepsilon(x_0) = x] = K e^{-\varepsilon d(x, x_0)}$ with d being the Euclidean distance. If $\rho \sim \mathcal{L}_{x_0, \varepsilon}(x)$, then:

1. $\mathcal{L}_{x_0, \varepsilon}$ is ε - d -private and $K = \frac{\varepsilon^n \Gamma(\frac{n}{2})}{2\pi^{\frac{n}{2}} \Gamma(n)}$
2. $\|\rho\|_2 \sim \gamma_{\varepsilon, n}$ s.t. $\mathbb{P}[\|\rho\|_2 = r] = \frac{\varepsilon^n e^{-\varepsilon r} r^{n-1}}{\Gamma(n)}$
3. The i^{th} component of ρ has variance $\sigma_{\rho_i}^2 = \frac{n+1}{\varepsilon^2}$

where $\Gamma(n)$ is the Gamma function defined for positive reals as $\int_0^\infty t^{n-1} e^{-t} dt$ which reduces to the factorial function whenever $n \in \mathbb{N}$.

PROPOSITION 5.2. Let $y = f(x, \theta)$ be the fitting function of a machine learning model parameterized by θ , and $(X, Y) = Z$ the dataset over which the RMSE loss function $F(Z, \theta)$ is to be minimized, with $x \in X$ and $y \in Y$. If $\rho \sim \mathcal{L}_{0, \varepsilon}$, the bound on the increase of the cost function does not depend on the direction of ρ , in first-order approximation, and:

$$\begin{aligned} \|F(Z, \theta + \rho)\|_2 - \|F(Z, \theta)\|_2 &\leq \\ \|J_f(X, \theta)\|_2 \|\rho\|_2 + o(\|J_f(X, \theta)\rho\|_2) \end{aligned} \tag{5.8}$$

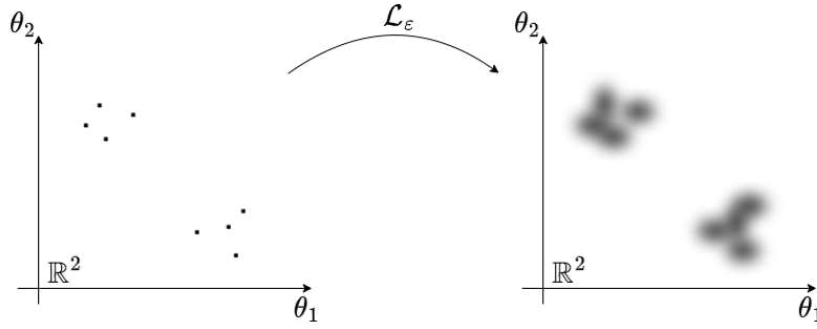


Figure 5.2: An illustration of the implementation of the Laplace mechanism to achieve ε - d -privacy for model parameters in \mathbb{R}^2

The results in Proposition 5.1 allow to reduce the problem of sampling a point from Laplace to i) sampling the norm of such point according to the result in Item 2 of Proposition 5.1 and then ii) sample uniformly a unit (directional) vector from the hypersphere in \mathbb{R}^n . Much like DP, d -privacy provides a means to compute the total privacy parameters in case of repeated queries, a result known as the Compositionality Theorem for d -privacy 5.3. Although it was known as a folk result, we provide formal proof in Appendix D.

THEOREM 5.3. Let \mathcal{K}_i be (ε_i) - d -private mechanism for $i \in \{1, 2\}$. Then their independent composition is $(\varepsilon_1 + \varepsilon_2)$ - d -private.

5.4.2 A Heuristic for defining the Neighbourhood of a Client

At the t^{th} iteration, when a user c calls the `SanitizeUpdate` routine in Algorithm 7, it has already received a set of hypotheses, optimized $\theta_{\bar{j}}^{(t)}$ (the one that fits best its data distribution), and got $\theta_{\bar{j},c}^{(t)}$. It is reasonable to assume that clients whose datasets are sampled from the same underlying data distribution $\mathcal{D}_{\bar{j}}$ will perform an update similar to $\delta_c^{(t)}$. Therefore, we enforce points which are within the $\delta_c^{(t)}$ -neighbourhood of $\hat{\theta}_{\bar{j},c}^{(t)}$ to be indistinguishable. To provide this guarantee, we tune the Laplace mechanism such that the points within the neighbourhood are $\varepsilon \|\delta_c^{(t)}\|_2$ differentially private. With the choice of $\varepsilon = n/(\nu \delta_c^{(t)})$, one finds that $\varepsilon \|\delta_c^{(t)}\|_2 = n/\nu$, and we call ν the *noise multiplier*. It is straightforward to observe that the larger the value of ν gets, the stronger is the privacy guarantee. This results from the norm of the noise vector sampled from the Laplace distribution being distributed according to Equation (D.6) whose expected value is $\mathbb{E}[\gamma_{\varepsilon,n}(r)] = n/\varepsilon$.

5.5 Experiments

5.5.1 Characterizing privacy

In the following Section, we provide a number of experimental validations of our algorithm on different tasks and datasets. In particular, we aim to evaluate and assess the trade-off in

training personalized federated learning models under formal local privacy guarantees. Detailed experimental settings are discussed in Appendix E.

5.5.2 Synthetic Data

We generate data according to $k = 2$ different distributions: $y = x^T \theta_i^* + u$ and $u \sim \text{Uniform}[0, 1)$, $\forall i \in \{1, 2\}$ and $\theta_1^* = [+5, +6]^T$, $\theta_2^* = [+4, -4.5]^T$. We then assess how training progresses as we move from the Federated Averaging [180] (Figure 5.3a, 5.3b, 5.3c), to IFCA (Figure 5.3d, 5.3e, 5.3f), and finally Algorithm 6 (Figure 5.3g, 5.3h, 5.3i). When using Federated Averaging, there seems to be an obvious problem, that is using one single hypothesis is not enough to capture the diversity in the data distributions, resulting in the final parameters settling somewhere in between the optimal parameters (Figure 5.3b). On the contrary, using IFCA, shows that having multiple initial hypotheses helps in improving the performance when the clients have heterogeneous data, as the optimized clients-parameters almost overlap the optimal parameters (Figure 5.3e). Adopting our algorithm shows that on top of providing formal guarantees, we can still achieve great results in terms of proximity to the optimal parameters (Figure 5.3h) and reduction of the loss function (Figure 5.3i). Figure 5.4 provides the maximum value of privacy leakage clients incur per cluster. Further details about the experimental settings are provided in Appendix E.

5.5.3 Hospital Charge Data

This experiment is performed on the Hospital Charge Dataset by the Centers for Medicare and Medicaid Services of the US Government [181]. The healthcare providers are considered the set of clients willing to train a machine learning model with federated learning. The goal is to predict the cost of a service given where it is performed in the country, and what kind of procedure it is. More details on the preprocessing and training settings are included in Appendix E. To assess the trade-off between privacy, personalization and accuracy, a different number of initial hypotheses has been checked, as it is not known apriori how many distributions generated the data. Accuracy has been evaluated at different levels of the noise multiplier ν . Note that, using Algorithm 6 with 1 hypothesis results in the Federated Averaging algorithm. Figure 5.5 shows that adopting multiple hypotheses drastically reduces the RMSE loss function. This is especially true when moving from 1 to 3 hypotheses. Additionally, we highlight how increasing the number of hypotheses also helps in curbing the effects of the noise multiplier even when it reaches high levels, on the right-hand side of the picture, making a compelling case for adopting formal privacy guarantees when a slight increase in the cost function is admissible. Figure 5.6 provides the empirical privacy leakage distribution of the clients involved in a particular training configuration. Table 5.1 shows privacy leakage statistics over multiple rounds and for all configurations.

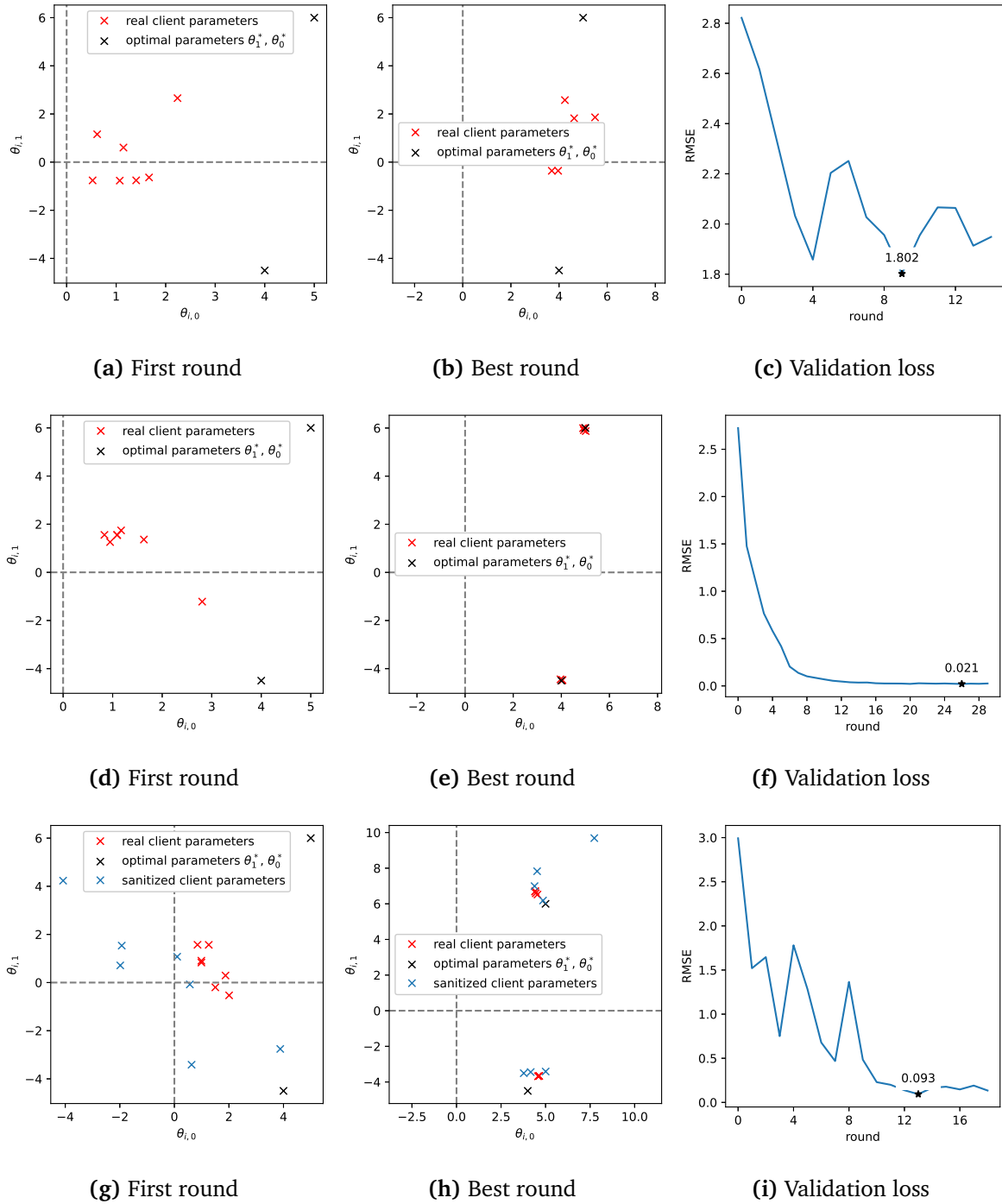


Figure 5.3: Learning federated linear models with (a, b, c) one initial hypothesis and non-sanitized communication, (d, e, f) two initial hypotheses and non-sanitized communication, (g, h, i) two initial hypotheses and sanitized communication. The first two figures of each row show the parameter vectors released by the clients to the server.

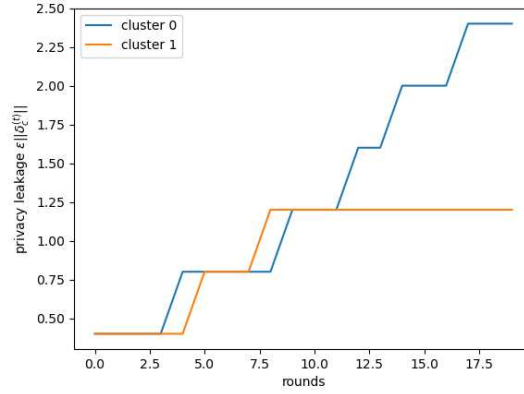


Figure 5.4: Synthetic data: max privacy leakage among clients. Privacy leakage is constant when clients with the largest privacy leakage are not sampled (by chance) to participate in those rounds.

ν	Hypotheses			
	7	5	3	1
0	–, –	–, –	–, –	–, –
0.1	517.0, 1551.0	418.0, 1342.0	473.0, 1386.0	528.0, 1540.0
1	36.3, 126.5	40.7, 127.6	44.0, 138.6	49.5, 147.4
2	15.4, 57.8	14.3, 54.5	22.0, 69.3	21.5, 66.6
3	7.7, 32.3	8.4, 36.7	12.5, 40.0	12.1, 40.0
5	5.7, 21.3	5.9, 22.0	5.5, 21.6	5.3, 20.9

Table 5.1: Hospital charge data: median and maximum local privacy budgets over the whole set of clients, averaged over 10 runs with different seeds. $\nu = 0$ means no privacy guarantee.

5.5.4 FEMNIST Image Classification

This task consists of image-based character recognition on the FEMNIST dataset [182]. Details on the experimental settings are in Appendix E. With the choice of the range of noise multipliers ν the corresponding value for the privacy leakage $\varepsilon \|\delta_c^{(t)}\|_2 = n/\nu$ would be enormous, considering a CNN with $n = 206590$ parameters, providing no meaningful theoretical privacy guarantees. This is a common issue for local privacy mechanisms [183], and it comes from the linear dependence of the expected value of the norm of the noise vector on n : $\mathbb{E} [\gamma_{\varepsilon, n}(r)] = n/\varepsilon$. Still, it is possible to validate, in practice, whether this particular generalization of the Laplace mechanism can protect against a *specific* attack: DLG [143]. Figure 5.7 and Table 5.2 report the results of varying the noise multiplier values. When $\nu = 10^{-3}$ the ground truth image is fully reconstructed. Up to $\nu = 10^{-1}$ we see that at least partial reconstruction is possible. For $\nu \geq 1$ we see that, experimentally, the DLG attack fails to reconstruct input samples when we protect the client-server communication with the mechanism in Proposition 5.1.

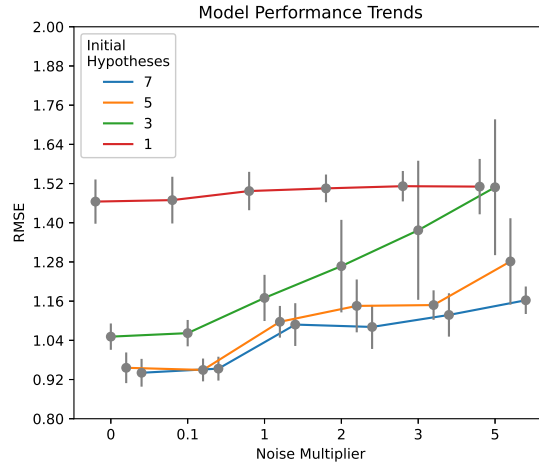


Figure 5.5: RMSE for models trained with Algorithm 6 on the Hospital Charge Dataset. Error bars show $\pm\sigma$, with σ the empirical standard deviation. Lower RMSE values are better for accuracy.

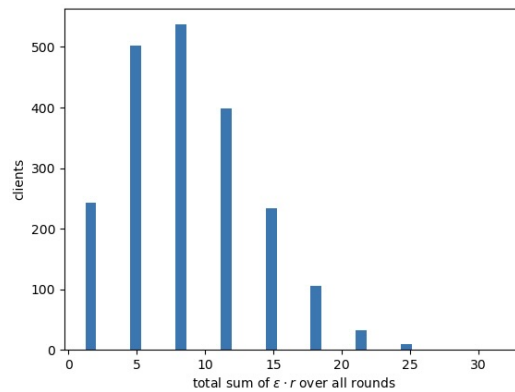


Figure 5.6: Hospital charge data: the empirical distribution of the privacy budget over the clients for $\nu = 3, 5$ initial hypotheses, seed = 3, r is the radius of the neighbourhood, and the total number of clients is 2062.

ν	Cross Entropy loss		RMSE loss	
	Average Accuracy	Standard Deviation	Average Accuracy	Standard Deviation
0	0.832	± 0.012	0.801	± 0.001
0.001	0.843	± 0.006	0.813	± 0.014
0.01	0.832	± 0.017	0.805	± 0.008
0.1	0.834	± 0.026	0.808	± 0.019
1	0.834	± 0.014	0.814	± 0.012
3	0.835	± 0.017	0.825	± 0.010
5	0.812	± 0.016	0.787	± 0.003
10	0.692	± 0.002	0.687	± 0.014
15	0.561	± 0.005	0.622	± 0.003

Table 5.2: Effects of increasing the noise multiplier on the validation accuracy and standard deviation.

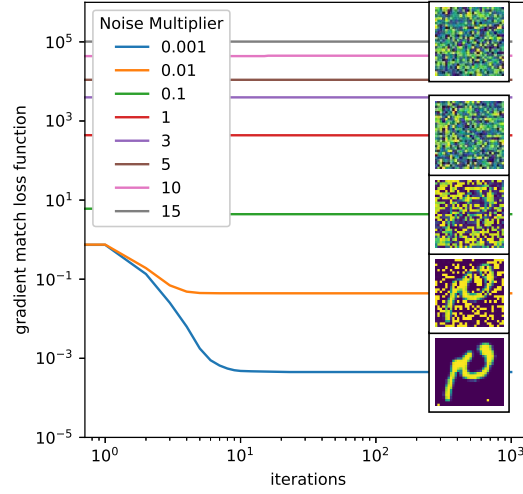


Figure 5.7: Effects of the Laplace mechanism in Proposition 5.1 with different noise multipliers as a defense strategy against the DLG attack.

5.5.5 Characterizing fairness

In this section, we analyze how group fairness improves with the personalization of the trained models under d -privacy guarantees when there are two groups with different data distributions. Experiments were performed on synthetic data and the FEMNIST image classification dataset that was used in Section 5.5.1. To ensure a thorough evaluation, we considered a variety of group fairness metrics in the experiments. In particular, we measured the fairness with respect to equal opportunity [158], equalized odds [158], and demographic parity [157] as explained in Section 5.2.2.

In particular, in Figures 5.9 and 5.10, the X-axis denotes the noise multiplier ν representing the amount of d -private noise added to the local updates as explained in Section 5.4.2 and the Y-axis denotes the difference in fairness between the privileged and unprivileged groups with respect to the different metrics of group fairness that we considered.

Synthetic data

Synthetic data was generated in a method similar to that in Section 5.5.2 with the following modifications for enabling ourselves to investigate the aspect of group fairness fostered by our method: i) Total number of users is 1000 and each user holds 10 samples. 800 users have data that is generated according to distributions $y = x^T \theta_1 + u$ and $u \sim \text{Uniform}[0, 1]$, $\forall i \in \{1, 2\}$, and set as a privileged majority group g_1 . The remaining 200 users have data that is generated according to distribution $y = x^T \theta_2 + 15 + u$ and $u \sim \text{Uniform}[0, 1]$, $\forall i \in \{1, 2\}$, and set as an unprivileged minority group g_2 . In this case, the sensitive attribute considered to evaluate fairness is the group id G where $G \in \{g_1, g_2\}$. ii) For binary classification, we set labels by using the $z = \text{Sigmoid}(Y)$, $\forall y, \hat{y} \in Y$. In the case of g_1 , we assign the label 1 if the value of z is greater than or equal to 0.5 and assign the label 0 otherwise. On the other

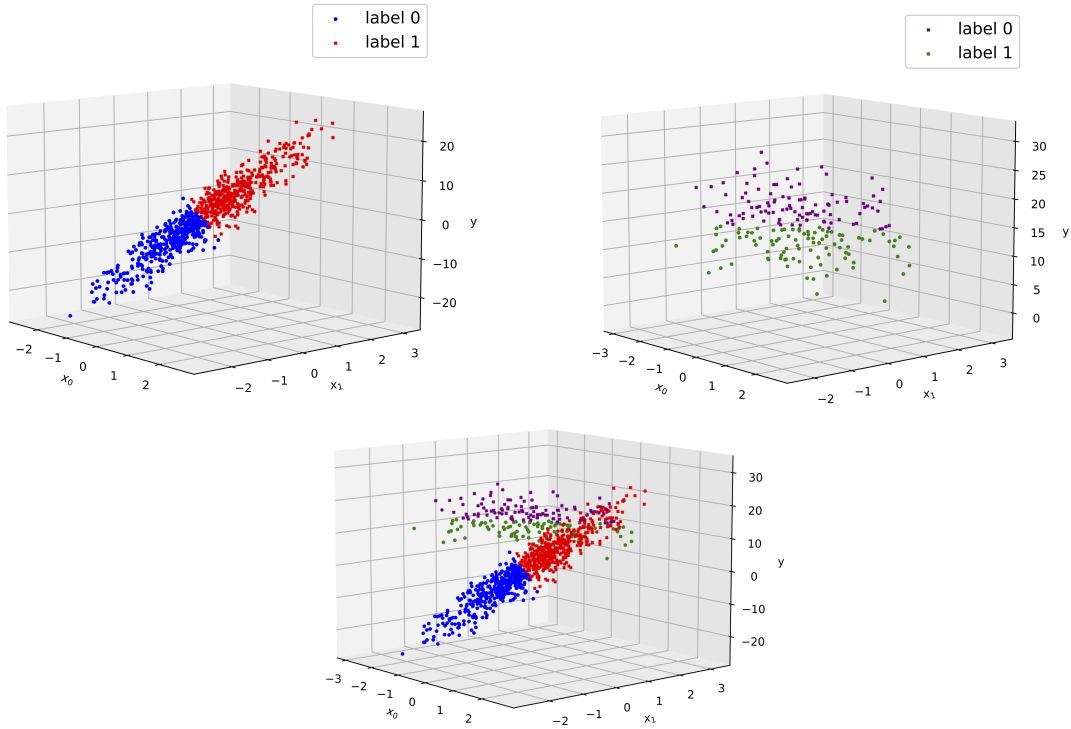


Figure 5.8: The first two plots from the left illustrate the spatial distribution of the samples in g_1 and g_2 , respectively, and the third plot shows g_1 and g_2 superimposed together in the same space.

hand, in the case of g_2 , the label 1 is assigned when the $z = \text{Sigmoid}(Y - 15)$, $\forall y, \hat{y} \in Y$ is less than or equal to 0.5, and the label 0 is assigned otherwise. This setting is to simulate a situation in which discrimination occurs depending on sensitive attributes in the real world such as minorities would have experienced a higher loan rejection rate than white applicants with the same property [184]. Thus, in our experiment, label 1 could be interpreted as 'loan approved' and label 0 as 'loan denied'. The data generated in this way are shown in Figure 5.8.

We compared the fairness for two cases: one with a single hypothesis (no personalization) and the other with the number of hypotheses as 2 (with personalization) in the framework of Algorithm 6. The experimental results are demonstrated in Figure 5.9.

The results illustrated by Figure 5.9 assert that the personalization of models (i.e., Algorithm 6) enhances the group fairness under all the metrics and the levels of formal privacy guarantees compared to that of the non-personalized model. A major reason behind this significant improvement of fairness by the personalized model is that unlike the non-personalized model, which trains using data from both groups that are biased towards the majority group g_1 , the personalized model training optimizes for each group's data distribution without disregarding the effect of the minority group g_2 . We also observe that fairness deteriorates as the value of the noise multiplier increases, as we would expect. This is presumably due to the decreasing influence of the minority group g_2 as the amount of noise insertion increases. This is consistent with the philosophy behind and the definition of DP and its variants. Furthermore, interestingly we

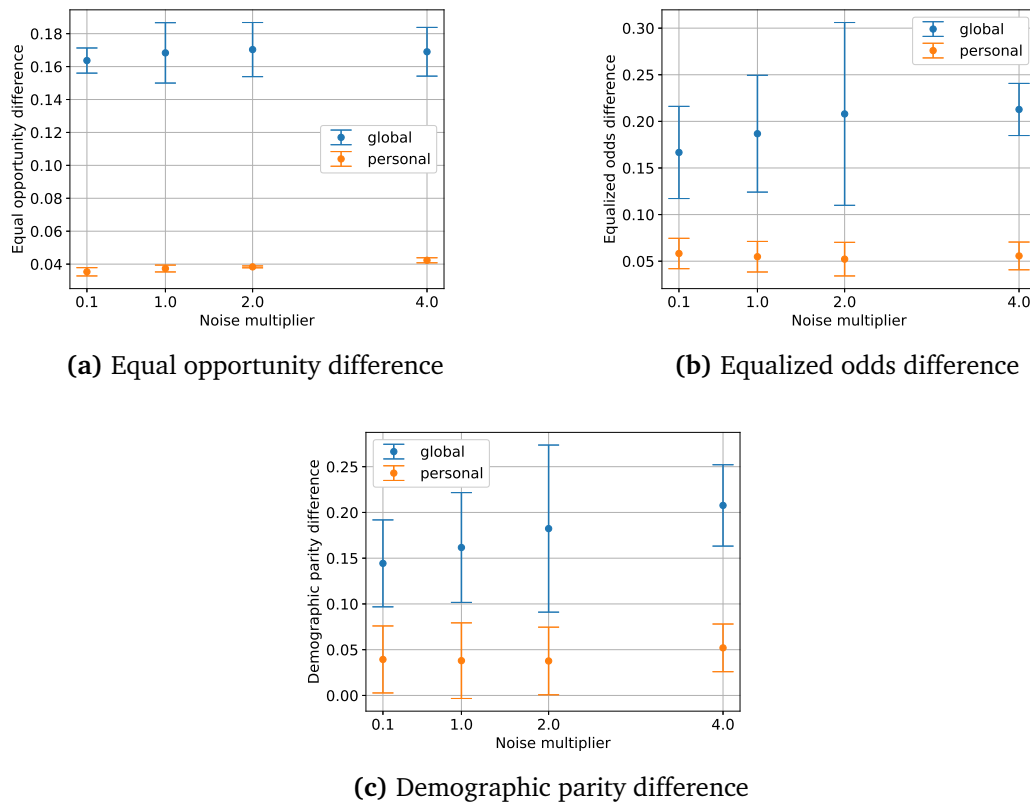


Figure 5.9: The figure shows the comparison between the personalized and non-personalized models for (from left) equal opportunity, equalized odds, and demographic parity, respectively. Experiments were performed for noise multipliers ν of 0.1, 1, 2, and 4. For all the metrics of fairness and the values of the noise multiplier, the personalized model is seen to possess improved fairness over the non-personalized model.

observe that the personalized model ensures better fairness than the non-personalized model even with the highest level of privacy protection. This shows that personalization in FL under d -privacy can be a comprehensive solution towards privacy-preserving and ethical machine learning as it provides both privacy guarantees and enhanced fairness.

FEMNIST Image Classification

To evaluate the fairness of our method on real datasets, we considered FEMNIST image classification dataset in the same form as in Section 5.5.4. As in experiments performed with the synthetic data in Section 5.5.5, the size of the groups considered privileged and unprivileged were different denoting the existence of a majority and a minority in the population. In this part, the rotated images are set as the unprivileged group g_2 with a total number of sampled users of 382 forming only 20% of all users. and the un-rotated images are used to represent the privileged group g_1 with a total number of users of 1736. Like in the case of synthetic data considered before, the group membership was used to denote the sensitive attribute. In the case

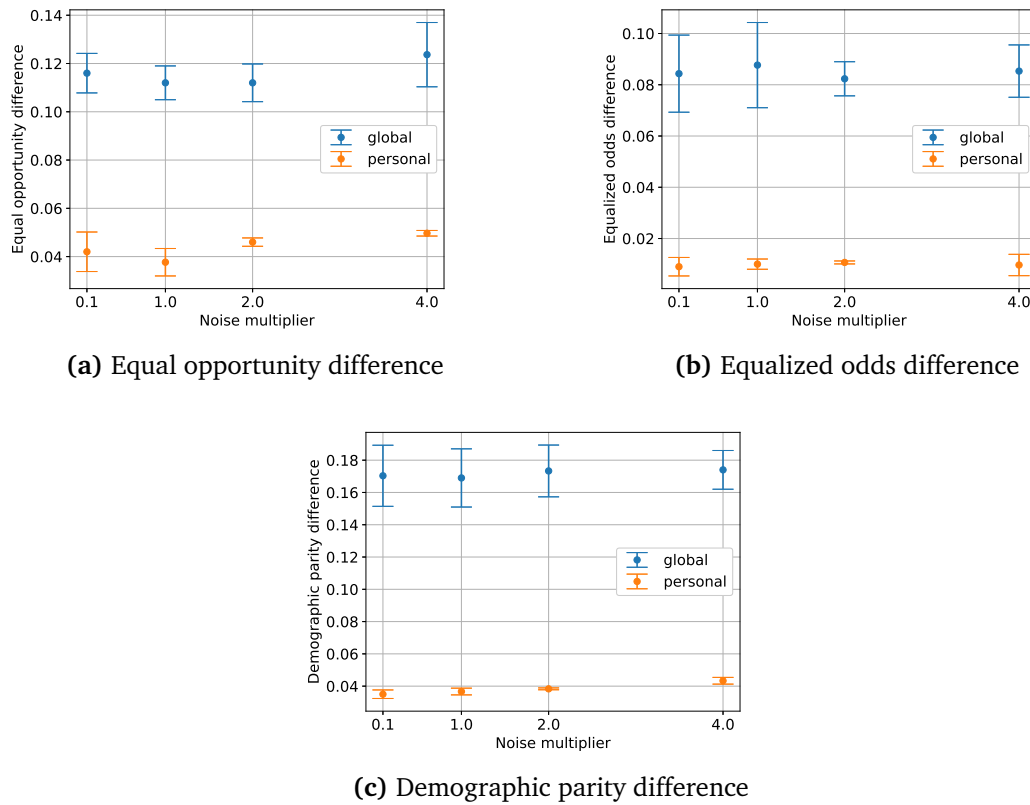


Figure 5.10: The figure shows the comparison between the personalized and non-personalized models for equal opportunity equalized odds, and demographic parity. Experiments were performed for noise multipliers ν of 0.1, 1, 2, and 4. For all metrics of fairness and values of the noise multiplier, the personalized model improved fairness over the non-personalized model.

of g_1 , we assign label 1 if the FEMNIST image label is even and 0 if it is odd. And for the g_2 , we assign label 0 if the FEMNIST image label is even and assign 1 if it is odd. The experimental results are given by Figure 5.10.

We observe that the personalized model training harbours significantly better group fairness across all metrics compared to its non-personalized counterpart. The change in fairness due to the amount of noise added was not as notable as in the case of the synthetic dataset but it was still observed to deteriorate with an increase in the value of the noise multiplier. Personalized model training in FL under the highest level of privacy is still observed to have better fairness across all the metrics than (non-personalized) models trained in a classical FL framework even with no privacy, similar to what we observed in the experiments with the synthetic data.

6

Characterizing the information leakage from gradient updates in federated learning

6.1 Introduction

Substantial work in the area of privacy-preserving ML has explored the applicability of differential privacy (DP) [13, 14] in central ML model training [27]. However, such models require a trusted central server which has access to and processes the (sensitive) training data held by the clients. To circumvent the need for a centrally trusted curator, local differential privacy (LDP) [23] was proposed, in which data owners locally obfuscate their data before communicating them to the server. Unfortunately, the use of LDP has been shown to have a detrimental effect on the accuracy of ML models [185–189]. Hence, in order to ameliorate the trust model in ML-related tasks (e.g., by consolidating the possibility of having an honest-but-curious or an adversarial server) while maintaining an acceptable level of model accuracy, the notion of *federated learning (FL)* [26] was proposed.

FL is a distributed, collaborative approach to ML where participating clients process their data and train models locally, sharing only updates of the training process. Following this, a server aggregates the local updates across all the participating clients and communicates the aggregated update back to all the clients and the process is repeated. This, in turn, furnishes a *global* model by aggregating the local updates at each round of training. Generally, the goal of FL is privacy-aware optimization of a statistical model's parameters by minimizing a cost function over a collection of datasets which are distributed among a set of clients.

Although the ideas behind FL appear to address privacy concerns in the central model for

ML (caused by data sharing), many recent works have identified attacks which allow the reconstruction of training data using only the shared gradients used in FL [20, 21, 190]. In particular, Zhu et al. [20] were one of the pioneers in the area, illustrating a simple gradient-matching-based attack in order to reconstruct the training data, which are supposedly held locally by the participating clients. The attack is effective in an environment where the clients use simple datasets (e.g., low-resolution images) with small batch sizes to train their local models. This work was improved by Yin et al. [21] where the authors explored more sophisticated techniques and used advanced computational resources to construct a gradient-inversion attack in vision-based FL. Their attack is effective in reconstructing more complex training data (images with high-resolution) of the participating clients even for high batch sizes. Our work is motivated by these attacks, and aims to formally study how privacy breaches occur in FL by understanding information leaks caused by gradient-sharing from a foundational perspective.

We commence by studying the information leakage fostered by the shared local updates for a simple example of a model trained by each client in every round. Later we extend our investigation to more realistic models under more pragmatic settings that are in use.

Contributions

The key contributions of this chapter are:

1. We provide a formal characterization of the information leakage from the shared gradient updates in federated learning.
2. We analyze the gradient-inversion type attacks in federated learning to understand their working from a foundational perspective.
3. To the best of our knowledge, we are the first to formally dissect and study such attacks in federated learning for a pragmatic multi-batch environment.

6.2 Preliminaries

We can describe a neural network in our setting as a function $F_{NN} : \mathcal{X} \rightarrow \mathbf{d}_{\mathcal{Y}}$ taking inputs \mathcal{X} to distributions over labels \mathcal{Y} . We will assume $\mathcal{X} = \mathbb{R}^n$ for some fixed n .

As we aim to investigate some of the state-of-the-art gradient inversion attacks in FL given by [20, 21], we adhere to the learning environment and model heuristics considered by them. In particular, the nodes in the hidden layers each contain a bias term and use a ReLU activation function, and the final layer uses a softmax activation to compute the output probabilities.

We denote by x_i the i 'th component of the input vector \mathbf{x} , and by $z_{k,\ell}$ the output from the k 'th hidden node in layer ℓ . The bias at hidden node $z_{k,\ell}$ is $b_{k,\ell}$. We write $w_{j,k,\ell}$ for the weight from node j in layer $\ell - 1$ to node k in layer ℓ .¹ Denote by \mathbf{z}_{ℓ} the vector of outputs from hidden layer ℓ , and by $\mathbf{w}_{k,\ell}$ the weight vector into node k at layer ℓ .

¹For nodes which are not connected we designate a weight of 0.

DEFINITION 6.2.1 (ReLU). The ReLU activation function $f_R : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ is defined:

$$f_R(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} . \quad (6.1)$$

DEFINITION 6.2.2 (Softmax probability). The softmax function $\sigma : \mathbb{R}^k \rightarrow (0, 1]^k$ is defined

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}} . \quad (6.2)$$

Observe that $\sum_{i=1}^k \sigma(\mathbf{z})_i = 1$.

DEFINITION 6.2.3 (Cross-entropy loss). Given two probability distributions $\pi : \mathbf{d}_Y$ and $\rho : \mathbf{d}_Y$, the cross-entropy loss $\mathcal{L} : \mathbf{d}_Y \times \mathbf{d}_Y \rightarrow \mathbb{R}$ is given by

$$\mathcal{L}(\pi, \rho) = - \sum_{y \in \mathcal{Y}} \pi_y \log(\rho_y) . \quad (6.3)$$

DEFINITION 6.2.4 (Gradient updates). The gradient updates for a neural network consist of the vector of partial derivatives of each weight and bias w.r.t. the loss function \mathcal{L} . For m weights $\mathbf{w}_{k,\ell}$ into node k at layer ℓ we have:

$$\nabla \mathbf{w}_{k,\ell}(\mathbf{x}) = \left(\frac{\partial \mathcal{L}}{\partial w_{k,1}}, \frac{\partial \mathcal{L}}{\partial w_{k,2}}, \dots, \frac{\partial \mathcal{L}}{\partial w_{k,m}} \right) \quad (6.4)$$

and for m biases \mathbf{b}_ℓ at layer ℓ we have

$$\nabla \mathbf{b}_\ell(\mathbf{x}) = \left(\frac{\partial \mathcal{L}}{\partial b_1}, \frac{\partial \mathcal{L}}{\partial b_2}, \dots, \frac{\partial \mathcal{L}}{\partial b_m} \right) . \quad (6.5)$$

From the above, we can compute the value of the output from node k at layer ℓ as:

$$z_{k,\ell} = f_R(\mathbf{w}_{k,\ell} \cdot \mathbf{z}^{(\ell-1)} + b_{k,\ell}) \quad (6.6)$$

6.3 Federated learning setup

A classical cross-device FL setup involves clients training local models using their personal data and sharing the local gradient updates with a server. Each client's model is typically assumed to have an identical architecture. We assume an honest-but-curious adversary (e.g., the server) who can be trusted with the computations they are responsible for (e.g., aggregation of the shared gradients) but are efficient enough to exploit all the information they have access to (i.e., the architecture of the global model, the gradients shared with the server at each round of the federated model training, etc.) in order to reconstruct the personal data of the clients used for the local training which, in essence, defeats the whole point and motivation behind FL.

It is crucial to note that the architecture of the neural network for each of the participating

clients will determine the extent of the information that is leaked to the adversary. Furthermore, it is noteworthy to observe that with the knowledge of the model architecture used for local training and the gradients shared at each round, any scope of randomness is eradicated for the adversary and, therefore, the network is deterministic from the point of view of the attacker. Our simple leakage model is then to describe conditions under which the NN is an invertible function.

6.3.1 Overview of the federated model training

We consider the task of training a global model for classification tasks in a federated environment involving a set of clients C and a server S . Each client in C has a common architecture and begins with the same model parameters, which are broadcast in the first round. Thereafter, each client who is sampled to participate in a round of training trains their own model using their local (private) data before communicating the gradient updates to S . The gradients of all the participating clients in a given round are aggregated by the server and communicated back to all of the clients, who update their local models using the aggregated gradient updates. The process continues iteratively until the convergence of the collaboratively trained global model.

6.3.2 The neural network model

We assume that each client is equipped with a deep neural network, in which both the architecture and the initial parameters of the network (the weights and the biases) are known and common to all clients (and also known to the server). The weights and biases are initialized randomly. We assume a classical deep NN architecture, with n -dimensional inputs, r hidden layers of various dimensions (not necessarily fully connected) and an m -dimensional fully-connected output layer. Each node in the hidden layer contains a ReLU activation and a bias term. The output layer is passed through a softmax function (ie. as its activation) so that the output of the network is a probability distribution over m labels. Finally, we assume the network is trained using a cross-entropy loss function.

6.3.3 The NN training process

At each round of training, each client takes input from their local (private) data and passes it through the network. They then use gradient descent with backpropagation to compute updates to the weights and biases of their NN based on their own input data. They pass their proposed updates to the server, which collates them and averages them and returns these new updates to each client. Each client then updates their network based on the weights and biases sent by the server.

6.3.4 Adversarial assumptions

We assume that the server behaves as an honest-but-curious adversary trying to infer information about the sensitive training data. In particular, the adversary is equipped with:

- a) knowledge of the architecture of the network including the activation functions at each layer and the loss function used to train the network; and
- b) knowledge of the initial weights and biases of the network.
- c) *no* trivial access to the training data (including the labels) or the output of any of the client's networks.

Note that some of the modern FL paradigms use decentralized techniques (e.g., secure aggregation) to compute the average gradient update in each round circumventing the need for a server. However, recent works like [191] have highlighted effective side-channel attacks that can *disaggregate* the averaged gradient updates reducing the threat model to that we considered in this work.

6.3.5 Measuring information leakage

To measure the information leakage from the system, we adopt the *information channel* model of Shannon [192] to model the flow of information from secrets to observations made by the adversary. We represent the channel as a matrix $C : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ over secrets \mathcal{X} and observations \mathcal{Y} , where $C_{x,y} = \mathbb{P}[Y = y | X = x]$. We usually think of C as a function from \mathcal{X} to distributions over \mathcal{Y} , written $C : \mathcal{X} \rightarrow \mathbf{d}_{\mathcal{Y}}$. In our model the observations are the gradient updates revealed to server, and the secrets are the inputs x .

Observing that all sources of randomness in the NN model (the weights and biases) are known to the adversary, we deduce that the information channel C is in fact *deterministic* (i.e. consisting of 0's and 1's only). Of particular interest is the case in which the information channel is the identity channel \mathbb{I} , which means that the channel C leaks *everything*, or equivalently when C defines an invertible function. We will focus on this scenario in this chapter.

More generally, when C is not the identity, then the adversary must make a guess at the secret; her probability of success can be defined using Bayes' rule as laid out in other works [19, 193]. We leave this approach to future work.

6.4 Information leakage in simple FL models

At first, we consider the objective of the collaboratively trained model to be binary classification with labels A or B and, hence, we aim to understand and analyze a simple setting where each client trains a neural network with one layer (i.e., with no hidden layer) and with a batch size of 1. We begin with an analysis of a toy neural network (TNN) depicted in Figure 6.1. The inputs are n -dimensional non-negative real vectors $x \in \mathbb{R}_{\geq 0}^n$ and the output layer consists of 2 nodes, each containing a bias term and a softmax activation function which produces a distribution over labels $\{A, B\}$. Each client uses a cross-entropy loss function to train their network.

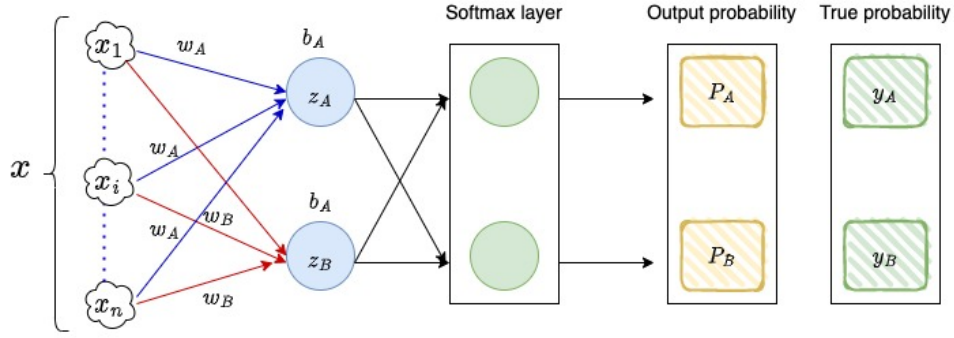


Figure 6.1: Illustration of a toy neural network consisting of a single input x and a final layer with 2 nodes and a softmax activation function, interpreted as a probability distribution over labels. We include bias terms in each node in the final layer.

6.4.1 Technical setup

To investigate the information leakage by the gradients and develop a foundational understanding and a formal characterization of gradient-inversion attacks in FL, w.l.o.g. we fix a client $c \in C$ and focus our analysis on the local model trained by them. Let $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}_{\geq 0}^n$, $|\mathbf{x}| \neq 0$ be the training data (input) used by c and let $y(\mathbf{x}) \in \{A, B\}$ be the true label of \mathbf{x} . We assume the probability of the labels is determined exactly, so we write $y_A(\mathbf{x}) \in \{0, 1\}$ for the probability that \mathbf{x} has label A , and correspondingly for label B . Denote by z_A and z_B the nodes of the final layer. Abusing notation, we also write $z_A(\mathbf{x})$ for the output from node z_A , and likewise for node z_B . Let $\mathbf{w}_A \in \mathbb{R}^n$ and $\mathbf{w}_B \in \mathbb{R}^n$ be the vectors of weights that feed into nodes z_A , z_B respectively. Let $b_A, b_B \in \mathbb{R}$ be the biases introduced in z_A and z_B , respectively. For input \mathbf{x} , let $p_A(\mathbf{x}) \in [0, 1]$ and $p_B(\mathbf{x}) \in [0, 1]$ represent the output probabilities given by the model.

From Def. 6.2.2 we derive:

$$p_A(\mathbf{x}) = \frac{e^{z_A(\mathbf{x})}}{e^{z_A(\mathbf{x})} + e^{z_B(\mathbf{x})}} \quad (6.7)$$

$$p_B(\mathbf{x}) = \frac{e^{z_B(\mathbf{x})}}{e^{z_A(\mathbf{x})} + e^{z_B(\mathbf{x})}} \quad (6.8)$$

Using Def. 6.2.3 we derive:

$$\mathcal{L}(p, y)(\mathbf{x}) = -y_A(\mathbf{x}) \log(p_A(\mathbf{x})) - y_B(\mathbf{x}) \log(p_B(\mathbf{x})) . \quad (6.9)$$

From Def. 6.2.4, we derive the following closed form for the gradient update vectors:

$$\nabla \mathbf{w}_A(\mathbf{x}) = (p_A(\mathbf{x}) - y_A(\mathbf{x}))\mathbf{x} \quad (6.10)$$

$$\nabla \mathbf{w}_B(\mathbf{x}) = (p_B(\mathbf{x}) - y_B(\mathbf{x}))\mathbf{x} \quad (6.11)$$

$$\nabla \mathbf{b}(\mathbf{x}) = (p_A(\mathbf{x}) - y_A(\mathbf{x}), p_B(\mathbf{x}) - y_B(\mathbf{x})) \quad (6.12)$$

REMARK 6.4.1. From (6.10) and (6.11), and knowing that $(y_A(\mathbf{x}), y_B(\mathbf{x})) =$

$(0, 1)$ or $(1, 0)$, we observe that corresponding components of $\nabla \mathbf{w}_A(\mathbf{x})$ and $\nabla \mathbf{w}_B(\mathbf{x})$ have different signs.

6.4.2 Analysing gradient leaks

Gradients on biases leak everything

To lay down a rudimentary insight, we start by analysing the simplest setting: in the most primitive form of FL, recall that $\nabla \mathbf{w}_A(\mathbf{x})$, $\nabla \mathbf{w}_B(\mathbf{x})$ and $\nabla \mathbf{b}(\mathbf{x})$ are observed by the adversary. We observe immediately that the gradients on the bias terms can be used to deduce the values of \mathbf{x} exactly.

LEMMA 6.1. Given a binary classification NN with no hidden layers, batch size 1 and a final layer containing bias terms, the gradient update using a cross-entropy loss function reveals the values of the input \mathbf{x} exactly.

Proof. In Appendix F. □

To analyze the information leakage caused by these gradient updates, we first define the notion of *gradient clones*.

DEFINITION 6.4.1 (Gradient clones). We call $\mathbf{x}, \mathbf{x}' \in \mathbb{R}$ *gradient clones* if $\mathbf{x} \neq \mathbf{x}'$, $|\mathbf{x}|^2 + |\mathbf{x}'|^2 \neq 0$, and $\nabla \mathbf{w}_k(\mathbf{x}) = \nabla \mathbf{w}_k(\mathbf{x}')$ for $k = A, B$.

Essentially, gradient clones are inputs \mathbf{x} which produce the same observations. Gradient clones exist whenever the information channel C (representing the leakage to the adversary) is *not* an invertible function. The results of this section pinpoint the precise conditions under which gradient clones exist in our TNN, under the assumption that only the weight updates are shared with the adversary. In the rest of the chapter, for any vectors \mathbf{v} and \mathbf{v}' , let $\mathbf{v} \cdot \mathbf{v}'$ denote their dot product.

LEMMA 6.2. If \mathbf{x} and \mathbf{x}' are gradient clones, there is some $K \in \mathbb{R}$ such that $\mathbf{x}' = K\mathbf{x}$.

Proof. In Appendix F. □

REMARK 6.4.2. If the space of inputs is $\mathbb{R}_{\geq 0}^n$ or $\mathbb{R}_{\leq 0}^n$ and if \mathbf{x}' is a gradient clone of \mathbf{x} , then $y(\mathbf{x}') = B$ can be trivially rejected by the comparison of the *signs* of the components of $\nabla \mathbf{w}_B(\mathbf{x})$ and $\nabla \mathbf{w}_B(\mathbf{x}')$ and, hence, we will have $\mathbf{x}' = K\mathbf{x}$ where $K > 0$. In the subsequent analysis, we assume that the domain of inputs is known to us as $\mathbb{R}_{\geq 0}^n$ (e.g., image data).

A noteworthy implication of Lemma 6.2 emerges when we consider images as inputs: two input images generate identical gradient updates iff they differ solely in their lighter or darker

shading. In practical terms, this implies an alarming lack of privacy protection for potentially sensitive images.

LEMMA 6.3. Setting $\gamma = \mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B)$ and $\Delta = b_B - b_A$, \mathbf{x} has a gradient clone iff there is some $K \in \mathbb{R}$ satisfying the equation

$$e^{K\gamma} - Ke^\gamma = (K - 1)e^\Delta. \quad (6.13)$$

The corresponding gradient clone of \mathbf{x} would be \mathbf{x} scaled by that K .

Proof. In Appendix F. □

THEOREM 6.4. For any $K > 1$, if $\mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B) \leq \frac{\ln K}{K-1}$, then \mathbf{x} has no gradient clone scaled by that K .

Proof. In Appendix F. □

THEOREM 6.5. \mathbf{x} cannot have a gradient clone \mathbf{x}' with $|\mathbf{x}'| > |\mathbf{x}|$ if there exists some $\lambda \in \mathbb{R}^+$ for which

$$\mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B) = 1 + \lambda + W_0 \left(\frac{e^\Delta}{e^{(1+\lambda)e^\Delta}} \right)$$

where $\Delta = b_B - b_A$ and W_0 is the principal branch of the Lambert W -function.

Proof. In Appendix F. □

THEOREM 6.6. \mathbf{x} has a gradient clone \mathbf{x}' iff all of the following hold:

- a) $\frac{e^\Delta}{e^\gamma + e^\Delta} - \frac{1}{\gamma} W \left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta} \right) \neq 1$
- b) $e^{-1}(e^\gamma + e^\Delta) < \gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}$
- c) $\mathbf{x}' = K\mathbf{x}$ with $K = \frac{e^\Delta}{e^\gamma + e^\Delta} - \frac{1}{\gamma} W \left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta} \right)$

where $\gamma = \mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B)$, $\Delta = b_B - b_A$, and $W(\cdot)$ is the Lambert W -function.

Proof. In Appendix F. □

COROLLARY 6.7. \mathbf{x} has at most two gradient clones in \mathbb{R}^n .

Proof. In Appendix F. □

REMARK 6.4.3. The results derived in this section extend naturally to arbitrary numbers of labels (i.e. multinary classification tasks).

To sum up, the implications of the above results can be summarised as:

1. For any observations $\nabla \mathbf{w}_A, \nabla \mathbf{w}_B$ there are at most 2 inputs (gradient clones) which could have produced those observations (Corollary 6.7).
2. Inputs \mathbf{x}, \mathbf{x}' which produce identical observations are always scalar multiples of each other. (Lemma 6.2).
3. Given a neural network and a set of observations produced in a single round, the adversary can test whether there is exactly 1 (or exactly 2) inputs which could have produced those observations (Theorem 6.6).

REMARK 6.4.4. It is important to note that the focus of the aforementioned test (Thm. 6.6) is to check whether the channel is invertible or not, i.e., it checks whether the shared gradient update is unique w.r.t. the inputs resulting in them. We do not use this result to find the explicit value of the gradient clone.

6.4.3 Leakage from bigger batches

We now expand our analysis to models trained on batches of size greater than one. Following [21], we consider batches consisting of labels which are all distinct and use the signs of the gradient updates to recover the labels of the training batch.

For a batch B of size k , the average gradient update obtained by training the local model by the client for every element of the batch (i.e., $\frac{1}{k} \sum_{\mathbf{w} \in \nabla \mathbf{w}_B} \mathbf{w}$) is communicated to the server. Hence, we naturally extend the definition of a *gradient clone* to batches: a batch $B' \in \mathbb{R}_{\geq 0}^{n \times k}$ be called a *gradient clone* of $B \in \mathbb{R}_{\geq 0}^{n \times k}$ if $|B| = |B'|$, they comprise of the same set of labels, and $\frac{1}{k} \sum_{\mathbf{w} \in \nabla \mathbf{w}_B} \mathbf{w} = \frac{1}{k} \sum_{\mathbf{w} \in \nabla \mathbf{w}_{B'}} \mathbf{w}$.

Therefore, setting $\mathcal{L} = \{L_1, \dots, L_m\}$ as the set of all possible classification labels, let $B = \{\mathbf{x}_1, \dots, \mathbf{x}_k\} \subseteq \mathbb{R}_{\geq 0}^{n \times k}$, satisfying $m > k$, be a particular training batch used as an input to the local network trained by a client. Note that it is reasonable to assume that in real-life settings for classification tasks, the total number of possible classifying labels is higher than the size of the batches used for training. W.l.o.g., let the true labels associated with the batch B be $L(B) = \{L_1, \dots, L_k\} \subset \mathcal{L}$ where $L_i = y(\mathbf{x}_i)$ such that $y(\mathbf{x}_i) = y(\mathbf{x}_j)$ iff $i = j$. Using the same line of argument as presented in [21], let us assume that the adversary is able to recover the set of labels L from having observed the shared gradient updates.

LEMMA 6.8. If batches $B = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ and $B' = \{\mathbf{x}'_1, \dots, \mathbf{x}'_k\}$ are gradient clones, we must have $X\pi = X'\pi'$ where:

$$X = \begin{pmatrix} x_{11} & \dots & x_{k1} \\ \vdots & \vdots & \vdots \\ x_{1n} & \dots & x_{kn} \end{pmatrix}, X' = \begin{pmatrix} x'_{11} & \dots & x'_{k1} \\ \vdots & \vdots & \vdots \\ x'_{1n} & \dots & x'_{kn} \end{pmatrix},$$

$$\pi = \begin{pmatrix} p_{11} - 1 & p_{21} & \cdots & p_{k1} & \cdots p_{m1} \\ p_{12} & p_{22} - 1 & \cdots & p_{k2} & \cdots p_{m2} \\ \vdots & \vdots & \ddots & \vdots & \\ p_{1k} & p_{2k} & \cdots & p_{kk} - 1 & \cdots p_{mk} \end{pmatrix},$$

$$\pi' = \begin{pmatrix} p'_{11} - 1 & p'_{21} & \cdots & p'_{k1} & \cdots p'_{m1} \\ p'_{12} & p'_{22} - 1 & \cdots & p'_{k2} & \cdots p'_{m2} \\ \vdots & \vdots & \ddots & \vdots & \\ p'_{1k} & p'_{2k} & \cdots & p'_{kk} - 1 & \cdots p'_{mk} \end{pmatrix},$$

and $p_{js} = \mathbb{P}[\mathbf{x}_s = L_j]$ (i.e., the output probability of the s^{th} input of batch X to be classified as label L_j) and, similarly, $p'_{js} = \mathbb{P}[\mathbf{x}'_s = L_j]$ for $s = 1, \dots, k$ and $j = 1, \dots, m$.

Proof. In Appendix F. □

As B and B' have the same labels, ideally we would have $p_{js} \approx p'_{js}$ for every $j \in \{1, \dots, m\}, s \in \{1, \dots, k\}$ (e.g., towards the beginning of training) and, in turn, implies $\pi \approx \pi'$. Therefore, Lemma 6.8 reduces to:

$$(X - X')\pi = \mathbf{0}_{n \times k} \quad (6.14)$$

An immediate consequence of this is that if π has a right-inverse, then the input batch becomes unique w.r.t. the shared gradient updates averaged over the batch, implying the lack of privacy protection for the training data in FL.

DEFINITION 6.4.2. Setting the output probability of classifying \mathbf{x}_s as its true label L_s (i.e., p_{ss}) as σ_s for all $s \in \{1, \dots, k\}$, we say that the FL model *learns well* if the output probabilities of obtaining the *wrong labels* be small and almost the same, i.e., $p_{js} \approx \frac{1-\sigma_s}{m-1}$ for each $j \in \{1, \dots, m\}$ s.t. $j \neq s$. In such a case, for each $s \in \{1, \dots, k\}$, let $p_{js} \approx \frac{1-\sigma_s}{m-1}$ be denoted by r_s .




Hence, under the assumption that the trained model learns well, we have:

$$\pi = \underbrace{\begin{pmatrix} \sigma_1 - 1 & r_1 & \cdots & \cdots & r_1 \\ r_2 & \sigma_2 - 1 & \cdots & \cdots & r_2 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ r_k & r_k & \cdots & \sigma_k - 1 & \cdots r_k \end{pmatrix}}_m \quad (6.15)$$

DEFINITION 6.4.3. We say that the batch $B = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ consists of *similar data* if $\sigma_1 \approx \dots \approx \sigma_k$ which, in turn, would result in $r_1 \approx \dots \approx r_k$.

REMARK 6.4.5. Definition 6.4.3 essentially says that we call a batch of data *similar* if each input gives approximately the same softmax probability distribution pivoted around its corresponding true label.

Table 6.1: An example of a batch used in a given round of training consisting of datapoints which are *similar* as per Def. 6.4.3 for an FL model that has *learnt well* as per Def. 6.4.2. In other words, the images of the dog, the wolf, and the fox used in the training batch have very similar true-positive and true-negative probabilities.

Batch	Output probability (softmax)		
	“Dog”	“Wolf”	“Fox”
	0.70	0.13	0.17
	0.12	0.71	0.17
	0.13	0.15	0.72

THEOREM 6.9. If the data in the batch that is used to train a certain round are *similar* and when the FL model *learns well*, the spatial arrangement of the members of the batch used for training is unique with respect to the shared gradients.

Proof. In Appendix F. □

REMARK 6.4.6. An immediate consequence of Theorem 6.9 is that if one of the elements of the training batch is identified, every other element in it can be reconstructed from the shared gradient updates (irrespective of the size of the batch or the number of dimensions of the data).

6.4.4 Leakage from batches size = 2:

Let us consider the case where clients use batches of size 2 to train their local models. Hence, by Lemma 6.8, $B = \{\mathbf{x}_1, \mathbf{x}_2\}$ and $B' = \{\mathbf{x}'_1, \mathbf{x}'_2\}$ are gradient clones iff:

$$\begin{pmatrix} x_{11} & x_{21} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{pmatrix} \begin{pmatrix} p_{11} - 1 & p_{21} & \cdots & p_{m1} \\ p_{12} & p_{22} - 1 & \cdots & p_{m2} \end{pmatrix} = \\
 \begin{pmatrix} x'_{11} & x'_{21} \\ \vdots & \vdots \\ x'_{1n} & x'_{2n} \end{pmatrix} \begin{pmatrix} p'_{11} - 1 & p'_{21} & \cdots & p'_{m1} \\ p'_{12} & p'_{22} - 1 & \cdots & p'_{m2} \end{pmatrix} \quad (6.16)$$

LEMMA 6.10. For batch size = 2, π (and, similarly, π') is right-invertible.

Proof. In Appendix F. □

REMARK 6.4.7. A very similar line of argument used in the proof of Lemma 6.10 can be extended to show that π (and, similarly, π') is right invertible for batch size = 3 and, hence, all the subsequent results of this section can be extended to batches of size 3. We conjecture that this right-invertibility of π holds for batches of size any arbitrary k .

COROLLARY 6.11. For batch size = 2, $X' = XD$ where $D = \pi\pi'^{-1}$.

Proof. Immediate from Lemma 6.8. □

REMARK 6.4.8. Corollary 6.11 essentially suggests that for a given input batch, a potential gradient clone would be a linear transformation of it. It is worth noting that D , as defined in Corollary 6.11, is a $k \times k$ matrix for a general batch of size k (specifically, 2×2 in Corollary 6.11). Hence, for batch size 1 (i.e., $k = 1$), this reduces down to Lemma 6.2 and is consistent with the results presented in the earlier part of this chapter considering batches of size 1.

THEOREM 6.12. $d_E(\mathbf{x}_i, \mathbf{x}'_i) = d_E(\delta_{1i}\mathbf{x}_1, -\delta_{2i}\mathbf{x}_2)$ for $i = 1, 2$, where $D = \mathbf{I}_{2 \times 2} + \delta$.

Proof. In Appendix F. □

REMARK 6.4.9. One important implication of Theorem 6.12 is that we can derive the distance between the corresponding elements of the real training batch and a potential gradient clone of it and, hence, can deduce a bound on the space that needs to be pruned to eventually identify the real training batch.

6.5 Analyzing multi-layer neural networks

We now extend our analysis to a larger neural network depicted in Figure 6.2. We make the following assumptions: a) the final layer is fully connected (but not necessarily the hidden layers); b) ReLU activations are used on hidden layers with softmax on the final layer; c) n -dimensional real-valued inputs, k hidden layers and an ℓ -dimensional output layer, cross-entropy loss and batch size 1. We make no assumptions about the non-negativity of the inputs \mathbf{x} , assuming that they can take on any values in \mathbb{R} . We will assume ReLU activation functions throughout the network with a softmax on the final layer. We assume n -dimensional real-valued inputs, r hidden layers and an ℓ -dimensional output layer. We again assume that the loss function used in training is the cross entropy loss and that the batch size is 1.

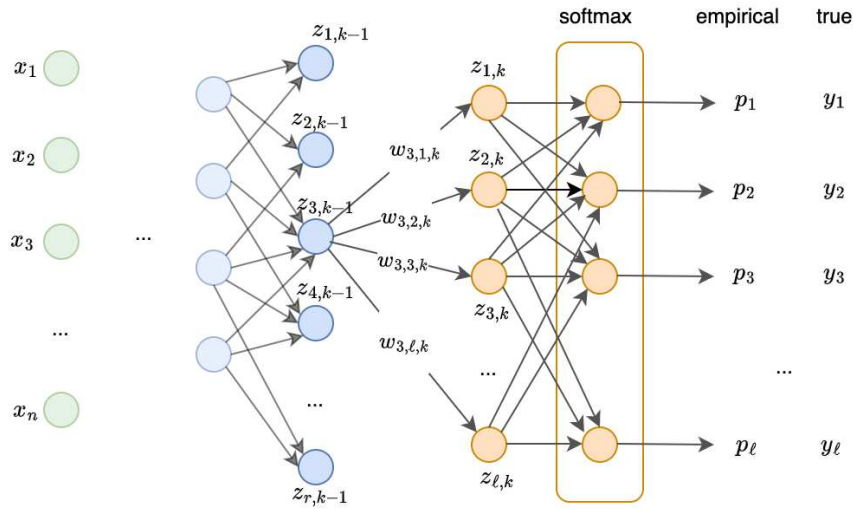


Figure 6.2: Larger neural network example. The final layer is fully connected although this is not depicted in the diagram for simplicity of labelling. We make no assumptions about the number of hidden layers.

From Def. 6.2.2 we derive:

$$p_i(\mathbf{x}) = \frac{e^{z_{i,k}(\mathbf{x})}}{\sum_{j=1}^{\ell} e^{z_{j,k}(\mathbf{x})}} \quad (6.17)$$

Using Def. 6.2.3 we derive:

$$\mathcal{L}(p, y)(\mathbf{x}) = - \sum_{i=1}^{\ell} y_i(\mathbf{x}) \log(p_i(\mathbf{x})) . \quad (6.18)$$

From Def. 6.2.4 we derive the following closed form for the gradient update vectors from the penultimate layer to the final layer:

$$\nabla \mathbf{w}_{i,k}(\mathbf{x}) = (p_i(\mathbf{x}) - y_i(\mathbf{x})) \mathbf{z}_{k-1} \quad (6.19)$$

where $\mathbf{w}_{i,k}$ is the weight vector into node i of layer r .

Finally, we derive a recursive form for the gradient update vectors between hidden layers up to the penultimate layer as follows. First, using (6.6) we derive:²

$$\frac{\partial z_{k,\ell}}{\partial \mathbf{w}_{k,\ell}} = \mathbf{z}_{\ell-1} \quad (6.20)$$

Then, for any hidden layer h we observe that:³

²This holds when the ReLU activation takes non-zero values. When it is zero, the derivative is also zero.

³The second line holds when the ReLU activation is non-zero. When it is zero, it contributes a zero weight.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial z_{i,h}} &= \sum_j \frac{\partial z_{j,h+1}}{\partial z_{i,h}} \frac{\partial \mathcal{L}}{\partial z_{j,h+1}} \\
&= \sum_j w_{i,j,h+1} \frac{\partial \mathcal{L}}{\partial z_{j,h+1}}
\end{aligned} \tag{6.21}$$

where the summation is taken over all of the nodes connected to $z_{i,h}$ in layer $h + 1$.

When h is the final layer we derive:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial z_{i,h}} &= \sum_j \frac{\partial \mathcal{L}}{\partial p_j(\mathbf{x})} \frac{\partial p_j(\mathbf{x})}{\partial z_{i,h}} \\
&= p_i(\mathbf{x}) - y_i(\mathbf{x})
\end{aligned} \tag{6.22}$$

Finally, we can compute the gradients of the weights in terms of the values in the network as follows:

$$\begin{aligned}
\nabla w_{i,h}(\mathbf{x}) &= \frac{\partial z_{i,h}}{\partial w_{i,h}} \frac{\partial \mathcal{L}}{\partial z_{i,h}} \\
&= z_{h-1} \frac{\partial \mathcal{L}}{\partial z_{i,h}} \quad [\text{using (6.20)}]
\end{aligned} \tag{6.23}$$

where the $\frac{\partial \mathcal{L}}{\partial z_{i,h}}$ term can be “unrolled” recursively (using the above equations) into a closed form in terms of z , p , y and w values from subsequent layers in the network.

In particular when h is the first hidden layer we have that:

$$\begin{aligned}
\nabla w_{i,h}(\mathbf{x}) &= \frac{\partial z_{i,h}}{\partial w_{i,h}} \frac{\partial \mathcal{L}}{\partial z_{i,h}} \\
&= \mathbf{x} \frac{\partial \mathcal{L}}{\partial z_{i,h}}
\end{aligned} \tag{6.24}$$

6.5.1 Extending the TNN to a larger NN

The TNN analyzed in Section 6.4 can be viewed as a subnetwork of the larger NN that we consider in this section. In particular, we can think of the inputs \mathbf{x} of the TNN as the outputs z_{k-1} on the penultimate layer in the larger NN. From this perspective, the TNN results give conditions upon which we can deduce the output values in the penultimate layer. In this section we will show how knowledge of these penultimate layer values leaks the values of the secrets \mathbf{x} in the larger NN.

Importantly, we note that the nodes in the penultimate layer contain ReLU activation functions, and thus all outputs z_{k-1} are non-negative. In this section we will make no assumptions about the non-negativity of the inputs \mathbf{x} , assuming that they can take on any values in \mathbb{R} .

6.5.2 Leakage from the penultimate layer

The analysis of Section 6.4 showed that the inputs x_i are leaked by the release of the gradients of the final layer (the derivatives of the loss function \mathcal{L} w.r.t. the weights and biases of the final layer). In our larger NN example (cf. Figure 6.2), this corresponds to a leaking of the vector \mathbf{z}_{k-1} , where here we are treating the $z_{i,k-1}$ as if they are inputs to the final layer (in orange). In other words, Lemma 6.1 tells us that the release of the gradients w.r.t. the biases on \mathbf{z}_k and the weights \mathbf{w}_k leaks the values \mathbf{z}_{k-1} exactly. We now show how this leaks the input values the \mathbf{x} .

LEMMA 6.13. Let \mathbf{z}_{k-1} be the vector of outputs from the penultimate layer $k - 1$ of a neural network. Assume a single non-zero value $z_{j,k-1}$ is revealed. Then every $z_{i,k-1}$ for $i \neq j$ can be deduced by an adversary with knowledge of the network and gradients.

Proof. In Appendix F. □

Lemma 6.13 tells us that knowledge of just one node of the penultimate layer is sufficient to leak the whole layer, even if the bias gradients are not revealed. Next, we show that knowledge of any hidden layer permits an adversary to learn the values in the previous layer.

THEOREM 6.14. Let \mathbf{z}_{k-1} denote the vectors of outputs from penultimate layer of the NN, and let \mathbf{x} be the vector of inputs. Then an adversary with knowledge of \mathbf{z}_{k-1} learns the values in \mathbf{x} exactly.

Proof. In Appendix F. □

Theorem 6.14 and Lemma 6.13 show that an adversary only requires knowledge of a single output value in the penultimate layer to learn the values of the inputs exactly. In Section 6.4 we derived conditions upon which the penultimate layer could be learned (by treating it as the input layer). Therefore we have characterized the conditions under which an adversary can learn the values of the inputs in a deep neural network using a softmax final activation function restricted to batches of size 1.

6.6 Discussion and open questions

The analyses carried out in this paper have opened up interesting avenues for future work. One intriguing future direction to pursue would be trying to formalize the trade-off between the *diversity* of a training batch and the corresponding information leakage that it incurs which may help the clients to optimally sample the batches to be used for training their local models. Another direction would be to widen the scope of our analysis by considering arbitrary loss and activation functions and introduce an information-channel modelling of the FL architecture to understand how an optimal defence model would look (e.g., using differential privacy and quantitative information flow) to optimize the privacy-utility trade-off in FL.

Part IV

Private data trading

“Privacy is something you can sell, but you can’t buy it back.”

– Bob Dylan

An incentive mechanism for trading personal data in data markets

7.1 Introduction

Nowadays, as we are heading towards an information-based society, data is becoming one of the most essential resources contributing to the advancements in technology. In the past, data broker companies such as Acxiom collected personal data and sold them to companies that needed data. However, as users are becoming more and more aware of the value of their personal data and with a rise in concern about their privacy, people are less and less willing to let their data be collected for free. In this scenario, the model of *data markets* is starting to emerge in order to obtain high-quality personal information in exchange for compensation. Liveen [194] and Datacoup [195] are examples of prototypes of data market services where the data providers can obtain additional revenue from selling their data while the data consumers can collect the desired personal data.

The problem of privacy violation by personal data analysis is one of the major issues in such data markets. As the population is becoming increasingly aware of the negative consequences of privacy breaches (e.g., the Cambridge Analytica scandal [4]), people are reluctant to release their data unless they are properly sanitised. To facilitate this, techniques like noise insertion [13, 14], synthetic data generation [196], secure multi-party computation [197], and homomorphic encryption [198] are being actively studied.

Differential privacy provides a privacy protection framework based on solid mathematical foundations and enables quantified privacy protection according to the amount of noise insertion. However, like all privacy-protection methods, it deteriorates the data utility. If the data

provider inserts too much noise because of privacy concerns, the data consumer cannot proceed with the analytics with the required performance. Thus, the privacy protection and data utility depend on the amount of noise insertion while applying DP and this amount of noise is determined by the DP parameter ϵ . Thus, determining the appropriate value of the parameter ϵ is a fundamental problem in differential privacy. It is difficult to establish the appropriate value of ϵ because it depends on many factors that are difficult to quantify, like the attitude towards the privacy of the data provider, which may be different from person to person.

In this work, we propose an incentive mechanism to encourage the data providers to join the data market and motivate them to share more accurate data. The amount of noise insertion depends on the data providers' privacy preference and the incentives provided to them by data consumers. In the scope of this work, we assume a model where the data consumers decide on incentives to pay to the data provider by considering the profit to be made from the collected data. By sharing some of the consumers' profit as an incentive, the data providers can get fair prices for providing their valuable information to be used in the subsequent analytics. The proposed mechanism consists of the truthful price report mechanism and an optimization method within budget constraints. The truthful price report mechanism guarantees that the data provider takes the optimal profit when she reports her privacy price to the data consumer honestly. Based on a data provider's reported privacy price, a data consumer can maximize her profit within a potential budget constraint.

Contributions

The key contributions of this chapter are:

1. We propose an incentive mechanism that guarantees that the data provider maximizes her benefit when she reports her privacy price honestly.
2. We propose an optimization method to maximize the data consumer's profit and information gain with a fixed financial budget for data collection.
3. We propose a method of splitting the privacy budget for the data providers allowing them to maximize their utility gain within a fixed privacy budget dealing with multiple data consumers.

7.2 Related Work

Methods for choosing ϵ

In DP, the parameter ϵ is the knob to control the privacy-utility trade-off. The smaller the ϵ , the higher the privacy protection level and the more it deteriorates the data utility. Conversely, a larger ϵ decreases privacy protection and enhances data utility. However, there is no gold standard to determine the appropriate value of ϵ . Apple has been promoting the use of differential

privacy to protect user data since iOS 10 was released, but the analysis of [199] showed the ϵ value was set at approximately 10 without any particular reason. The work of [200] showed that the privacy protection level set by an arbitrary ϵ can be infringed by inference using previously disclosed information and proposed a method to set the value of ϵ considering posterior probability. Much research has been conducted to study and solve this problem [201–204]. Although a lot of work is being done in this area, the problem of determining a reasonable way of choosing an optimal value for ϵ still remains open as there are many factors to consider in deciding the value of ϵ . In this chapter, we propose a technique to determine an appropriate value of ϵ by setting a price for the privacy of the data provider.

Pricing mechanism

One of the solutions to find an appropriate value of ϵ , as has been explored in the literature, is to price it according to the data accuracy [205–210]. In [205], the strength of the privacy guarantee and the accuracy of the published results are considered to set the appropriate value of ϵ and a simple model to assign the value of ϵ that can satisfy data providers and consumers was suggested. In [206], the author proposed a compensation mechanism via auction in which data providers are rewarded based on the accuracy of the reported differentially private data and the data consumers' budget. It is the most similar work to our study. The main differences between our work and Ghosh and Roth's paper are as follows:

- a) We define a truthful price report mechanism and formally show that a data provider gets the best profit when she reports her privacy price honestly.
- b) We propose an optimized incentive mechanism to maximize the data consumer's profit with a fixed expense budget, and a privacy budget splitting method to maximize the data provider's utility gain in a multi-data consumer environment.

In [209] the authors design a mechanism that can estimate statistics accurately without compromising the user's privacy. They propose a Bayesian incentive and privacy-preserving mechanism that guarantees privacy and data accuracy. The study of [210] proposes a Stackelberg game to maximize mobile users who provide their trajectory data.

Several techniques for pricing data assuming a data market environment have been studied in [211–218]. In [211], the authors suggested a data pricing mechanism to establish a balance between the privacy guarantees and the price of data in data markets. [212] proposed the data market model for IoTs and showed that the proposed pricing model has a global optimal point. In [213], the authors proposed a theoretical framework for determining prices to noisy responses to queries in differentially private data markets. However, this research cannot flexibly reflect the requirements of the data market. [215] highlighted an ϵ -choosing method based on Rubinstein bargaining and assumed a market manager that mediates a data provider and consumer in the data trading.

It is realistic to consider personal data as a digital asset and is reasonable to attempt to find a bridge between privacy protection level and price according to the value of ϵ in DP. Existing

studies are attempting to find an equilibrium between data providers and consumers under the assumption that both are reasonable individuals. In this work, we follow a direction similar to existing studies and focus on the incentive mechanism that motivates a data provider to report her privacy price honestly. In particular, we consider that the value of differentially private data increases non-linearly with respect to the increase of the value of ε .

7.3 Incentive mechanism for data markets

7.3.1 Overview of the proposed technique

Data markets aim at collecting personal data legally with the consent of the users holding the data. Typically in such scenarios, a data provider can sell her own data and get paid for it, and a data consumer can collect personal data for analysis by paying a price, resulting in a win-win situation. Naturally, the data consumer would want to collect personal data as accurately as possible at the lowest possible expenditure and the data provider would want to sell her data at a price as high as possible while protecting sensitive information. In general, every effective privacy-preserving technique affects the utility of the data negatively. In the particular case of DP, the levels of utility and privacy are determined by the parameter ε and, hence, the price of the differentially private data is affected directly by the value of ε . Determining the appropriate value of ε and the actual price of the data is critical to the success of the data market. However, this is not trivial especially when each data provider has different privacy needs [219] as is, typically, the case. We propose an incentive mechanism to find the price of the data and the value of ε that can satisfy both the data provider and the data consumer.

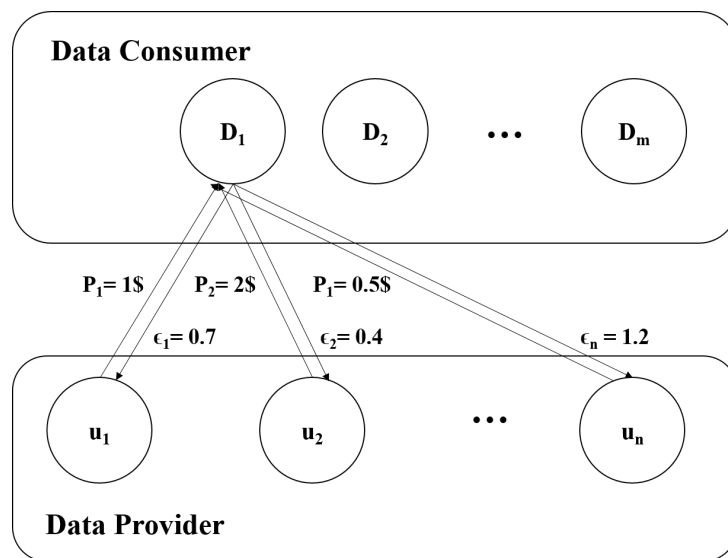


Figure 7.1: An example of data trading process. In this figure, u_i means the i^{th} data provider and D_j means the j^{th} data consumer.

We consider a scenario with n data providers, u_1, \dots, u_n , and m data consumers, D_1, \dots, D_m

where each provider and consumer proceeds with the deal independently (we use the term “data provider” and “data producer” interchangeably, in the same sense). The term “ ε unit price” (e.g., 1\$ per ε value 0.1) will be used to express the price per unit ε (the DP parameter used to obfuscate the reported data) as a measure of the accuracy of information. We recall that as ε increases, the data becomes less private and more information can be obtained from it, and vice versa. Thus, the price per unit ε represents the “value” of the provider’s information¹. The price of ε is expected to differ from one data provider to another because each individual has a different privacy requirement. We denote the ε unit price reported by u_i as p_i and her true ε unit price as π_i . Note that p_i may not be the same as π_i as the users reporting their ε unit price might render a piece of false information to the data consumers trying to insinuate that their data needs a higher (or lower) level of privacy protection than they really do.

Figure 7.1 illustrates how the process works. At first, every data consumer broadcasts a function f to the data providers, which represents the amount of data (expressed in ε units) the consumer is willing to buy for a given ε unit price. Each consumer has her own function, and it can differ from one consumer to another. We will call it *ε -allocating function*. We assume f to be monotonically decreasing as the consumers naturally prefer to buy more data from those data producers who are willing to offer them for less price (i.e., low value of ε) or, correspondingly, a low level of privacy requirement (i.e., more accurate data). Note that the product $p_i f(p_i)$ represents the total amount that will be paid by the data producer to the consumer if they agree on the trade. The function f , however, has also a second purpose: as we show in Section 7.3.2, it is designed to encourage providers to demand the price that they really consider the true price of their privacy requirement rather than asking for more to manipulate the system.

Thanks to the truthful price report mechanism (cf. Section 7.3.2), the data providers report the prices of their data honestly to the data consumers in accordance with the published f . In the example illustrated by Figure 7.1, u_1 reports her ε price per 0.1 as 1\$ and u_2 reports her ε price per 0.1 as 2\$. Finally, the data consumer checks the price reported by the data provider and determines the total price and value of ε to be obtained from each provider using f . In this example, the data consumer D_1 determines ε_1 to be 0.7 and ε_2 to be 0.4. Then the data providers select the consumers to whom to sell their data in order to maximize their profits and confirm the deal with the value of their ε they would use to obfuscate their data and the total price they would receive in compensation. In the example in Figure 7.1, D_1 pays 7\$ to u_1 and 8\$ to u_2 . Finally, the data providers add noise to their data based on the determined ε and share the sanitized data with the respective consumers, and the consumers pay the corresponding prices to the providers. We assume that data providers and consumers keep the promise of the value of ε and compensation decided in the deal, once confirmed.

This process can be repeated until the data consumers exhaust all their budget or achieve the targeted amount of information. The task of allocating a suitable budget in each round and how determining the amount of needed information are also important topics, but they are out of the scope of the work pursued in this chapter.

¹The ε unit price can be of any form including but not limited to being financial. The method we propose is independent of the nature of the price, so we do not need to specify it.

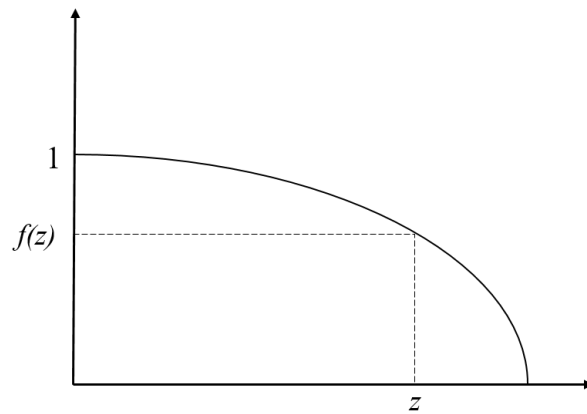


Figure 7.2: An example of a monotonically decreasing function $f(z)$. Let c be a parameter representing the “reported value-to-admitted ε value” ratio. For $z \geq 0$, we set $f(z)$ as $f(z) = \ln(e - cz)$ if $(e - cz) \leq 1$, and $f(x) = 0$ otherwise.

7.3.2 Truthful price report mechanism

For the correct functioning of the data trading, the data provider should be honest and demand her true privacy price. However, she may be motivated to report a higher price in the hope to persuade the data consumer that the information is “more valuable” and be willing to pay more. Note that the true privacy price of each data provider is personal information that only the provider herself knows and is not obliged to disclose. To mitigate this problem, we propose a truthful price report mechanism to ensure that the data providers report their ε unit prices honestly. The purpose of the mechanism is to provide incentives so that the providers are guaranteed to get the greatest profit when they report their true price.

When the data provider reports her price p_i , the data consumer determines the amount of ε to purchase using $f(p_i)$, where f is the ε -allocating function introduced in Section 7.3.1. We recall that f is a monotonically decreasing function chosen by the data consumer. We assume that the domain of f (i.e., the ε unit price) is normalized to take values in the interval $[0, 1]$. The total price for the data estimated by the consumer is the product of the ε price unit and the amount to be purchased, namely, $p_i f(p_i)$. To this value, the consumer adds an *incentive* $\int_{p_i}^{\infty} f(z) dz$, the purpose of which is to make it convenient for the data provider to report the true price (we assume that the data provider knows f and the strategy of the consumer in advance). It is worth noting that the incentive should be finite and, hence, the contribution of $f(z)$ should vanish as z goes to ∞ . An example of such a function is illustrated in Figure 7.2.

DEFINITION 7.3.1 (Payment offer). The data consumer sets the *offer* $\mu(p_i)$ to the provider u_i as follows:

$$\mu(p_i) = p_i f(p_i) + \int_{p_i}^{\infty} f(z) dz$$

We now illustrate how this strategy achieves its purpose of convincing the consumer to report her true price. We start by defining the *utility* that the data provider obtains by selling her data

as the difference between the offer and the true price of her data, represented by the product of the true ε unit price and the amount to be sold, namely $\pi_i f(p_i)$.

DEFINITION 7.3.2 (Utility of the data provider). The utility $\rho(p_i)$, of the provider u_i , for the reported price p_i , is defined as:

$$\rho(p_i) = \mu(p_i) - \pi_i f(p_i)$$

We are now going to show that the proposed mechanism guarantees truthfulness. The basic reason is that each provider u_i achieves the best utility when reporting the true price. Namely, $\rho(\pi_i) \geq \rho(p_i)$ for any $p_i \in \mathbb{R}^+$ and equality is achieved when $p_i = \pi_i$, where we recall that π_i is the true price of the provider u_i . The only technical condition is that the function f is monotonically decreasing. Under this assumption, we formally derive the required results via Lemmas (see also Figure 7.3 to get the intuition of the proof):

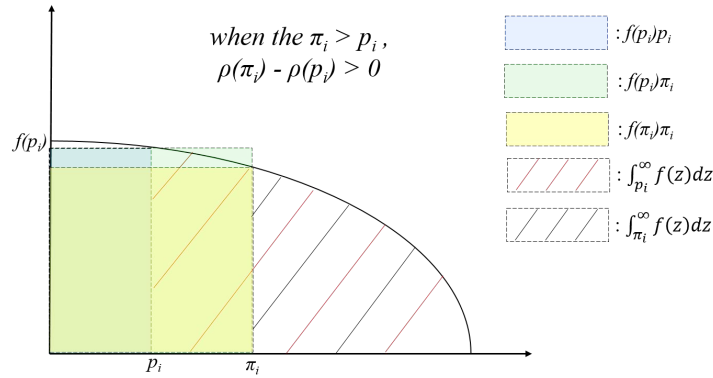


Figure 7.3: Graphical illustration of Theorem 7.3. We prove that $\rho(\pi_i)$ (blue hatching area) is always larger than $\rho(p_i)$ (blue rectangle area + red hatching area – green rectangle area).

LEMMA 7.1. If u_i reports a price greater than her true price, i.e., $p_i \geq \pi_i$, then her utility will be less than the utility for the true price, i.e., $\rho(p_i) \leq \rho(\pi_i)$.

Proof. In Appendix G □

LEMMA 7.2. If u_i reports a price smaller than her true price, i.e., $p_i \leq \pi_i$, then her utility will be less than the utility for the true price, i.e., $\rho(p_i) \leq \rho(\pi_i)$.

Proof. In Appendix G □

Combining Lemma 7.1 and Lemma 7.2 gives the announced result. We assume that each data producer is a rational individual, i.e., capable of identifying the best strategy to maximize her utility.

THEOREM 7.3. If every data producer acts rationally, then the proposed incentive mechanism guarantees the truthfulness of the system.

Proof. Immediate from Lemma 7.1 and Lemma 7.2. \square

7.3.3 Optimizing the incentive mechanism

In this section, we propose an optimization mechanism to identify an optimal function f for the data consumer concerning the following two desiderata:

- (a) *Maximum Information:* Maximize the total information gain of the data consumer with a fixed budget.
- (b) *Maximum Profit:* Maximize the total profit of the data consumer with a fixed budget².

We start by introducing the notions of total information and profit for the consumer. Note that, by the sequential compositionality of DP [105], the total information is the sum of the information obtained from each data provider.

DEFINITION 7.3.3 (Total information). The *total information* $I(\mathbf{u})$ obtained by the data consumer by concluding trades with each of the data providers of the tuple $\mathbf{u} = (u_1, \dots, u_n)$ is given as $I(\mathbf{u}) = \sum_{i=1}^n f(p_i)$.

As for the profit, we can reasonably assume it to be monotonically increasing with the amount of information obtained, and that the total profit is the sum of the profits obtained with each individual trading. The latter is naturally defined as the difference between the benefit (a.k.a. *payoff*) obtained by re-selling or processing the data and the price paid to the data providers to gather them.

DEFINITION 7.3.4 (Payoff and profit). The *payoff* function for the data consumer, denoted by $\tau(\cdot)$, is the benefit that the data consumer receives by processing or selling the information gathered from the different data providers. The argument of $\tau(\cdot)$ is ε representing the amount of information received. We assume $\tau(\varepsilon)$ to be monotonically increasing with ε .

The *total profit* for the data consumer is given by $\sum_{i=1}^n (\tau(\varepsilon_i) - \mu(\varepsilon_i))$, where $\varepsilon_i = f(p_i)$, i.e., the ε -value allocated to u_i .

We shall consider a family of functions \mathcal{F} , parameterized by c , to which the ε -allocating function f belongs. The parameter c reflects the data consumer's willingness to collect the information and, for technical reasons, we assume f to be continuous, differentiable, and concave w.r.t. c . For each data provider, different values of c will give different f , that, in turn, will give rise to a different incentive curve as per equation (7.2) which the data consumer should adhere to for compensating for the information obtained from that data provider.

As described in previous sections, the ε -allocating functions should be monotonically decreasing with the ε unit price as the consumer is motivated to buy more information from the consumers that offer it at a lower price. This property also ensures, by Theorem 7.3, that the

²Budget here refers to the budget of the data consumer to pay the data providers.

prices reported by the data producers will be their true prices. Hence, we have:

$$\mathcal{F} \subseteq \{f : f(c, p) \text{ is continuous, differentiable, and concave on } c, \text{ and decreasing with } p\} \quad (7.1)$$

Note that we have added the parameter c as an additional argument in f , so f can now be perceived to have two arguments.

EXAMPLE 7.3.1. An example of such class \mathcal{F} is that of Figure 7.2:

$$\mathcal{F} = \{\ln(e - cp) : c \in \mathbb{R}^+\}.$$

EXAMPLE 7.3.2. Another example is:

$$\mathcal{F} = \{1 - cp : c \in \mathbb{R}^+\}.$$

After the prices p_1, \dots, p_n have been reported by the data producers u_1, \dots, u_n , the data consumer will try to choose an optimal c maximizing her profit. Figure 7.4 illustrates an example showing one data provider's incentive graph and the payoff for the data consumer. We shall analyse the possibility to choose an optimal c , that, in turn, leads to an optimal $f(c, \cdot)$ addressing scenarios (a) and (b).

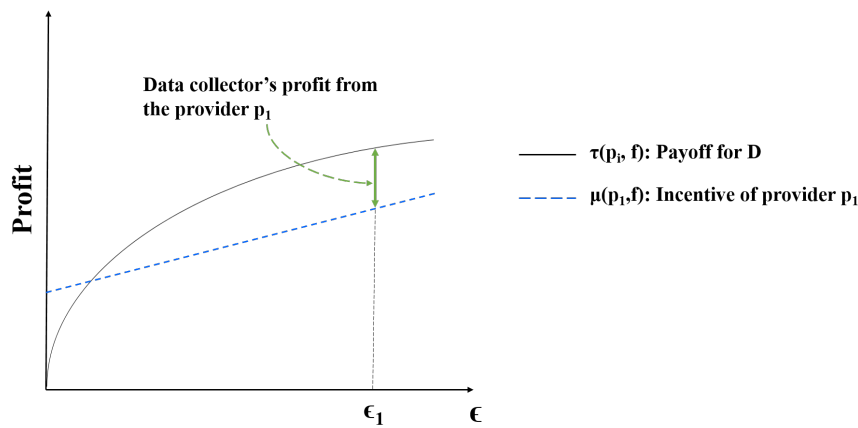


Figure 7.4: Illustrating the payoff for c and the incentive plot for the data consumer involving one data provider reporting p_1 . The Y -intercept of μ is $\int_{p_1}^{\infty} f(z)dz$.

To fit into the context of DP, we assume that τ (the data consumer's payoff function) is additive, i.e.,

$$\tau(a + b) = \tau(a) + \tau(b) \quad \text{for every } a, b \in \mathbb{R}^+. \quad (7.2)$$

This is a reasonable assumption that goes well along with the sequential compositionality property of DP, at least for small values of ε ³. We start by showing that the two desiderata (a) and (b) are equivalent:

³From a technical point of view, the additive property holds also for large values of ε . However, from a practical point of view, for large values of ε , for instance, 200 and 400, then the original information is almost entirely revealed in both cases and would not make sense to pay twice the price of 200 ε units to achieve 400 ε units.

THEOREM 7.4. If $\tau(\cdot)$ is additive, then maximizing information and maximizing profit (desiderata (a) and (b)) are equivalent, in the sense that a ε -allocating function $f(\cdot, \cdot)$ that maximizes the one, maximizes also the other.

Proof. In Appendix G.

COROLLARY 7.5. If $\tau(\cdot)$ is additive, then the optimal choice of $f(\cdot, \cdot)$ w.r.t. the selected family of functions will maximize both the information gain and the profit for the data consumer.

Proof. Immediate from Theorem 7.4. □

We now consider the complexity problem for finding the optimal $f(\cdot, \cdot)$. Due to the assumptions made in Equation 7.1, and to the additivity of τ , we can apply the method of the Lagrangians to find such $f(\cdot, \cdot)$ (cf. Appendix G).

THEOREM 7.6. If τ is additive, then there exists a c that gives an optimal **profit-maximizing** function $f(c, \cdot) \in \mathcal{F}$, for a fixed budget, and we can derive such c via the method of the Lagrangians.

Proof. In Appendix G. □

THEOREM 7.7. There exists a c that gives an optimal **information-maximizing** function $f(c, \cdot) \in \mathcal{F}$, for a fixed budget, and we can derive such c via the method of the Lagrangians.

Proof. In Appendix G.

To demonstrate how the method works, we show how to compute the specific values of c on the two classes \mathcal{F} of Examples 7.3.1 and 7.3.2. Such c gives the optimal ε -allocating function $f(c, \cdot)$, maximizing $\mathcal{I}(\mathbf{u})$ for a given budget. The derivations are described in detail in Appendix H. In each example, p_i is the reported ε unit price of u_i .

EXAMPLE 7.3.3. Let $\mathcal{F} = \{\ln(e - cp) : c \in \mathbb{R}^+\}$. The optimal parameter c is the solution of the equation $\ln\left(\prod_{i=1}^n e^{p_i} (e - cp_i)^{\frac{e}{c}}\right) = B + \frac{n(e-1)}{c}$.

EXAMPLE 7.3.4. Let $\mathcal{F} = \{1 - cp : c \in \mathbb{R}^+\}$. The optimal parameter c is the solution of the equation $c^2 \sum_{i=1}^n p_i^2 + 2Bc - n = 0$.

7.3.4 Discussion

In our model, for the scenario that we have considered so far, the parameter c is determined by the number of providers and the budget. We observe that in both Examples 7.3.3 and 7.3.4, if n increases then c increases, and vice versa. This seems natural because in the families of both these examples c the incentive that the consumer is going to propose decreases monotonically

with c . This means that the larger the offer, the smaller the incentive that the consumer needs to be paying. In other words, the examples confirm the well-known market law according to which the price decreases when the offer increases, and vice versa.

We note that we have been assuming that there is enough offer to satisfy the consumer's demand. If this hypothesis is not satisfied, i.e., if the offer is smaller than the demand, then the situation is quite different: now the data provider can choose to whom to sell her data. In particular, the data consumer who sets a lower c will have a better chance to buy data because, naturally, the provider prefers to sell her data to the data consumers who give a higher incentive. In the following section, we explore in more detail the process, from the perspective of the data provider, in the case in which the demand is higher than the offer.

7.3.5 Optimized privacy budget splitting mechanism for data providers

After optimizing an incentive mechanism for a given data consumer dealing with multiple data providers, we focus on the flip side of the setup. We assume a scenario in which a given data provider has to provide her data to multiple data consumers, and that there is enough demand so that she can sell all her data.

Let there be m data consumers, D_1, \dots, D_m seeking to obtain data from the user u . By truthful price report mechanism, as discussed in Section 7.3.2, u reports her true price to D_i for every $i \in \{1, \dots, m\}$. As discussed in Section 7.3.3, D_i computes her optimal ε -allocating function f_i and requests data from u , differentially privatized with $\varepsilon = f_i(\pi)$. After receiving f_1, \dots, f_m , u would like to provide her data in such a way that maximizes the utility received after sharing her data. Note that, under the assumption that the data consumers collude with each other, the privacy landscape of this setting is equivalent to that of the data provider dealing with one data consumer over m rounds.

DEFINITION 7.3.5. We say that the data provider has made a *deal* with the data consumer D_i if, upon reporting the true per-unit price of her information, π , she agrees to share her data privatized with privacy parameter $\varepsilon = f_i(\pi)$.

It is important to note here that u is not obliged to deal with any data consumer D_i even after receiving f_i . Realistically, u has a privacy budget of $\varepsilon_{\text{total}}$, which she would not exceed at any price. Let $S = \{i_1, \dots, i_k\}$ be an arbitrary subset $\{1, \dots, m\}$. By the sequential composition property of DP, the final privacy budget exhausted by u by sharing her data with an arbitrary set of data consumers D_{i_1}, \dots, D_{i_k} (or, equivalently, to one data consumer in rounds i_1, \dots, i_k) is $\varepsilon_S = \sum_{j \in S} f_j(\pi)$. u 's main intention is to share her data in such a way that ensures $\varepsilon_S \leq \varepsilon_{\text{total}}$ for all subset S of $\{1, \dots, n\}$, while maximizing $\sum_{j \in S} \rho_i(\pi, f_j)$, i.e., the total utility received. Reducing it down to the 0/1 knapsack problem, we propose that u should be dealing with $\{D_{i_1}, \dots, D_{i_k}\}$ where $S^* = \{i_1, \dots, i_k\} \subseteq \{1, \dots, m\}$, chosen as

$$S^* = \operatorname{argmax}_S \left\{ \sum_{j \in S} \rho(\pi, f_j) \mid S \subseteq \{1, \dots, m\}, \sum_{j \in S} f_j(\pi) \leq \varepsilon_{\text{total}} \right\} \quad (7.3)$$

We show the pseudocode for the ε allocation algorithm and the entire process in Algorithms 1 and 2.

Algorithm 8: Optimized privacy budget splitting algorithm

Input: $\{\varepsilon_1, \dots, \varepsilon_n\}$ stored in array w , $\{p_1, \dots, p_n\}$ stored in array v , $\varepsilon_{\text{total}}$;
Output: List of data consumer $\{D_1 \dots D_k\}$ that is selected to sell data;
initiate Two-dimension array m ;
while $i \leq n$ **do**
 while $j \leq \varepsilon_{\text{total}}$ **do**
 if $w[i] > \varepsilon_{\text{total}}$ **then**
 $m[i, j] := m[i-1, j]$
 else
 $m[i, j] := \max(m[i-1, j], m[i-1, j-w[i]] + v[i])$
 end
 end
end
backtrack using the final solution m and find the index of the data consumer ;
return List of selected data consumer ;

Algorithm 9: The proposed data trading process

Input: the data provider $\{u_1, \dots, u_n\}$, the data consumer $\{D_1, \dots, D_m\}$;
Output: List of the data provider and consumer pair that trade is completed ;
while $i \leq m$ **do**
 D_i calculate the parameter c to optimize the $f_i(\cdot)$;
 D_i inform the $f_i(\cdot)$ to the data provider
while $j \leq n$ **do**
 u_j report price p_j to the data consumer
 while $i \leq m$ **do**
 while $j \leq n$ **do**
 D_i calculate the ε_j based on p_j ;
 D_i inform the ε_j to the u_j
 end
 end
 u_j perform the **Optimized ε allocation algorithm** to maximize the utility

7.4 Experimental results

In this section, we highlight the results of some experiments we perform to verify that the proposed optimization method can find the best profit for the data consumer. For the experiments, we consider the families \mathcal{F} of Examples 7.3.3 and 7.3.4, namely $\mathcal{F} = \{\ln(e - cp) : c \in \mathbb{R}^+\}$ and $\mathcal{F} = \{1 - cp : c \in \mathbb{R}^+\}$. For these two families, the corresponding optimal parameter c is formally derived in Appendix H.

The experimental variables are set as follows: we assume that there are 10 data consumers and the total number of data providers n is set from 1000 to 2000 at an interval of 500. The data provider's ε unit price is distributed as per $\mathcal{N}(1, 1)$ and, to align with the scaling, we map

the values of the ε unit price less than 0 and more than 2 to 0 and 2, respectively. We set the unit value ε to 0.1 and the maximum value of ε of the data providers to 3. We consider the budget of the data consumer to be 60, 90, and 120 dealing 1,000, 1,500, and 2,000 data providers, respectively. We assumed that the data consumer earned a profit of 10 per 0.1 epsilon and assigned the parameter c from 1 to 10 for comparison.

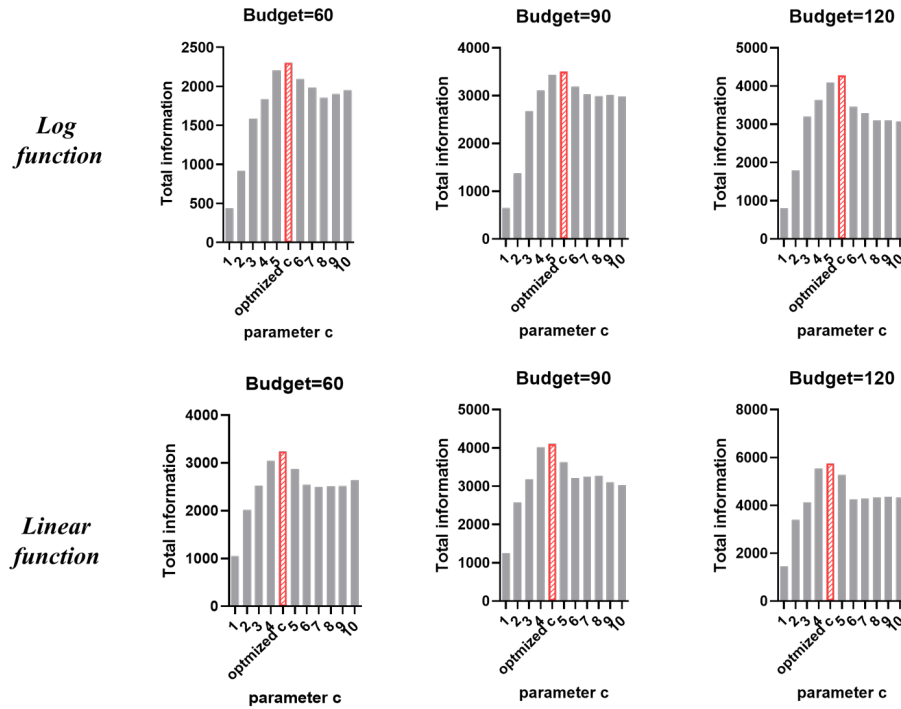


Figure 7.5: Experimental result of profit under a fixed budget. The *log* function represents the family $\ln(e - cp)$ and the *linear* function represents the family $1 - cp$. We let the parameter c range from 0 to 1. The red bin represents the optimal value of c , namely the c that gives maximum information.

The results are shown in Fig. 7.5. For instance, in the case of the log family $\ln(e - cp)$, the optimal parameter c is 5.36, and in the case of the linear family $1 - cp$, the optimal parameter c is 4.9. It is easy to verify that the optimal values of c correspond to those determined by solving Equations (H.2) and (H.3) for Example 7.3.3 and Equations (H.9) and (H.10) for Example 7.3.4 given in in Appendix H.

8

Establishing the price of privacy in federated data trading

8.1 Introduction

With the emerging trend of data markets (e.g., Datacoup [195], Liveen [194], etc.) incorporating DP [13, 14] for trading private data from clients in exchange for compensation in order to obtain high-quality analytics, as studied in Chapter 7, we envision a data trading framework in which groups of data providers ally to form federations in order to increase their bargaining power following the traditional model of trade unions. At the same time, federations guarantee that the members respect their engagement concerning the trade. Another important aspect of such federations is that the value of the collection of all data is usually different from the sum of the values of all members' data. It could be larger, for instance, because the accuracy of the statistical analyses increases with the size of the dataset, or could be smaller, for instance, because of some discount offered by the federation. Data consumers are supposed to make a collective deal with a federation rather than with individual data providers, and, from their perspective, this approach can be more reliable and efficient than dealing with individuals. Thus, data trading through federations can benefit both parties. Given such a scenario, the two crucial questions that arise which we address in this work are: (a) how to determine the price of the collective data according to the privacy preferences of each member, and (b) how to determine the individuals' contribution to the overall value obtained from the collectively reported data in order to establish a fair distribution of the earnings within the federations.

Contributions

The key contributions of this chapter are:

1. We propose a method to determine the price of collective differentially private data in a federated data trading environment.
2. We propose a distribution model based on game theory. More precisely, we use the notion of *Shapley values* from the theory of cooperative games to determine the contribution of each participant of a federation in the collective earnings and use it to ensure that each member of the federation receives fair compensation according to their contribution.

8.2 Related Work

As discussed in Chapter 7, data markets such as Datacoup[195] and Liveen[194] need to guarantee satisfactory privacy protection in order to encourage the data owners to participate. On this note, one of the key questions explored by the community is how to appropriately price data obfuscated by a privacy-preserving mechanism. In the context of DP, as the accuracy of data typically depends on the value of the noise parameter ϵ , this question is linked to the problem of how to establish the value of ϵ . Researchers have debated how to choose this value since the introduction of DP and there have been several proposals along this line [220–223]. In particular, [221] showed that the privacy protection level by an arbitrary ϵ can be infringed by inference attacks, and it proposed a method for setting ϵ based on the posterior belief. [222] considered the relation between DP and t -closeness, a notion of group privacy which prescribes that the earth movers distance between the distribution in any group E and the distribution in the whole dataset does not exceed the threshold t , and showed that both ϵ -DP and t -closeness are satisfied when the $t = \max_E \frac{|E|}{N} \left(1 + \frac{N-|E|-1}{|E|} e^\epsilon\right)$ where N is the number of records of the database.

Several other works have studied how to price the data according to the value of ϵ [224–231]. The primary focus of such studies is to determine the price and value of the ϵ according to the data consumer's budget, the accuracy requirement of the subsequent analytics, the privacy preference of the data providers, and the relevance of the data. In particular, the study in [230] assumed a dynamic data market and proposed an incentive mechanism for data owners to truthfully report their privacy preferences. In [228], the authors proposed a framework to find the balance between financial incentive and privacy in personal data markets where data owners sell their own data and suggested the main principles to achieve reasonable data trading. Ghosh and Roth [225] proposed a pricing mechanism based on auctions that maximizes the data accuracy under a budget constraint or minimizes the budget for the fixed data accuracy requirement under DP guarantees.

Our study differs from previous work in that, unlike the existing approaches assuming a one-to-one data trading between data consumers and providers, we consider an environment involving a data consumer and a federation of data providers.

8.3 Technical Preliminaries

In this work, we assume the local model of DP (i.e., each data provider obfuscates their own data locally to foster DP guarantees). When the domain of data points is finite, one of the simplest and most widely used mechanisms for LDP is *k-Randomized Response (k-RR)* [32]. In this work, we assume a setting where all data providers use the *k-RR* mechanism to locally obfuscate their data.

DEFINITION 8.3.1 (*k-Randomized Response* [32]). Let \mathcal{X} be the domain of secrets of size $k < \infty$. For a given LDP parameter ε and given an input $x \in \mathcal{X}$, the *k-Randomized Response (k-RR)* mechanism returns $y \in \mathcal{X}$ with probability

$$\mathbb{P}[y|x] = \frac{1}{k - 1 + e^\varepsilon} \begin{cases} e^\varepsilon & \text{if } y = x \\ 1 & \text{otherwise} \end{cases}$$

8.3.1 Shapley value

When participating in data trading through a federation, *Pareto efficiency* and *symmetry* are the important properties for the intra-federation earning distribution. Pareto efficiency means that at the end of the distribution process, no change can be made without making participants worse off. Symmetry means that all players who make the same contribution must receive the same share. Obviously, the share should vary according to the member's contribution to the collective data.

Shapley value [232, 233] is a concept from game theory named in honour of Lloyd Shapley who introduced it. Shapley value, typically applied in cooperative games, introduces a method to distribute the total gain that satisfies Pareto efficiency, symmetry, and differential distribution according to a player's contribution. Thus, all participants have the advantage of being fairly incentivized. The solution based on the Shapley value is unique. Due to these properties, the Shapley value is regarded as a cutting-edge approach for designing a fair incentive distribution method based on individual contributions.

Let $N = \{1, \dots, n\}$ be a set of players involved in a cooperative game and $M \in \mathbb{R}^+$ be a financial revenue received (e.g., from the data consumer). Let $v : 2^N \mapsto \mathbb{R}^+$ be the characteristic function, mapping each subset $S \subseteq N$ to the total expected sum of payoffs the members of S can obtain by cooperation. (i.e., $v(S)$ is the total collective payoff of the players in S). According to the Shapley value, the benefit received by player i in the cooperative game is given as follows:

$$\psi_i(v, M) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

We observe that $v(A) > v(B)$ for any subsets $B \subset A \subseteq N$, and hence, $v(S \cup \{i\}) - v(S)$ is positive. We call this quantity the *marginal contribution* of player i in a given subset S . Note that $\psi_i(v, M)$ is the expected marginal contribution of player i over all subsets $S \subseteq N$.

In this work, we use the Shapley value to distribute the earnings according to the contributions of the data providers in the federations.

8.4 Differentially Private Data Trading Mechanism

8.4.1 Mechanism outline

Overview: We focus on an environment with multiple federations of data providers and one data consumer who interacts with the federations in order to obtain the data (differentially privatized with varying values of ε) in exchange for financial revenues. We assume that federations and the data consumer are aware that the data providers independently use the k -RR mechanism to obfuscate their sensitive information with their desired levels of privacy (which can differ from provider to provider). Our method provides a sensible way of splitting the earnings of a federation using the Shapley value. In addition, it also motivates an individual to cooperate with the federation they are a part of and penalises intentional and recurring non-cooperation.

Notations and set-up: Let $\mathcal{F} = \{F_1, \dots, F_k\}$ be a set of k federations of data providers, where each federation F_i has n_{F_i} members for each $i \in \{1, \dots, k\}$. For a federation $F \in \mathcal{F}$, let its members be denoted by $F = \{p_1^F, \dots, p_{n_F}^F\}$. And finally, for every federation F , let $p_*^F \in F$ be an elected representative of F interacting with the data consumer. This approach to communication benefits both the data consumer and the data providers because (a) the data consumer minimizes her communication cost by interacting with just one representative of the federation, and (b) the reduced communication induces an additional layer of privacy.

We assume that each member p of a federation, F has a maximum privacy threshold ε_p^T with which she, independently, obfuscates her data using the k -RR mechanism. We also assume that p has d_p data points to potentially report.

We know from Equation (8) of [234] that if there are m data providers reporting d_1, \dots, d_m data points, independently privatizing them using k -RR mechanism with the privacy parameters $\varepsilon_1, \dots, \varepsilon_m$, the federated data of all the m providers also follow a k -RR mechanism with the privacy parameter being:

$$\varepsilon = \ln \left(\frac{\sum_{i=1}^m d_i}{\sum_{i=1}^m \frac{d_i}{k-1+e^{\varepsilon_i}}} + 1 - k \right).$$

We call the quantity $d_p \varepsilon_p^T$ the *information limit* of data provider $p \in F$, and

$$\varepsilon_F^T = \ln \left(\frac{\sum_{p \in F} d_p}{\sum_{p \in F} \frac{d_p}{k-1+e^{\varepsilon_p^T}}} + 1 - k \right) \quad (8.1)$$

the *maximum information threshold of the federation F* .

We now introduce the concept of a *valuation function* f mapping financial revenues to information, representing the amount of information to be obtained for a given price. It is reasonable

to have f as strictly monotonically increasing and continuous. In this work, we shall focus on the effect of the privacy parameter and, hence, we shall regard the number of data points to be reported by each data provider as a constant and let only ε can vary. We shall call f the *privacy valuation function*.

As f is strictly monotonically increasing and continuous, it is also invertible. Hence, f^{-1} maps a certain privacy parameter ε to the financial revenue evaluated with selling data privatized using k -RR with ε as the parameter. Noting that f essentially determines the privacy parameter of a DP mechanism (k -RR, in this case), it is reasonable to assume that f should be not only increasing but also increasing exponentially for a linear increase of money. In fact, when ε is high, it hardly makes any difference to further increase its value. For example, when ε increases from 200 to 250, it practically makes no difference to the data as they were already non-private with $\varepsilon = 200$. On the other hand, if we increase ε from 0 to 50, it creates a huge difference, conveying much more information. Therefore, it makes sense to set f to increase exponentially with a linear increase of the financial revenue.

$f(M) = K_1(e^{K_2 M} - 1)$ is an example of a privacy valuation function that we consider which takes the financial revenue $M \in \mathbb{R}^+$ as its argument, satisfying the reasonable assumptions of evaluating the DP parameter that should be used to privatize the data in exchange of the financial revenue of M . Here the scaling factors $K_1 \in \mathbb{R}^+$ and $K_2 \in \mathbb{R}^+$ are decided by the data consumer according to her requirements.

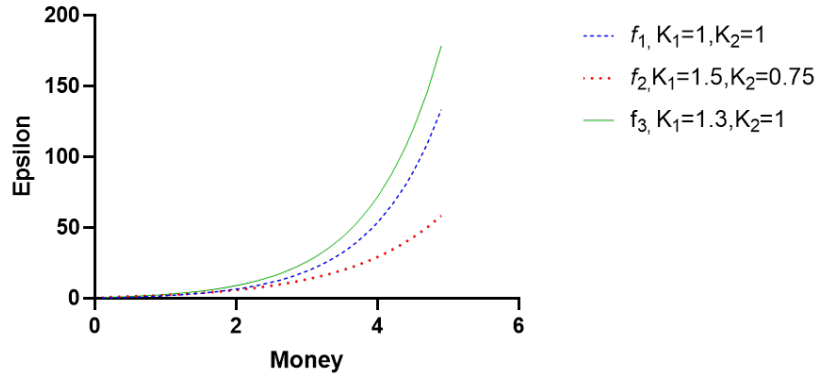


Figure 8.1: Some examples of the privacy valuation function f illustrated with different values of K_1 and K_2 . The data consumer decides the values of the scaling factors K_1 and K_2 according to her requirement and broadcasts the determined function to the federations.

Finalizing and achieving the deal: Before the data trading commences, the data consumer, D , truthfully broadcasts her financial budget, B , and a privacy-valuation function, f , to all the federations. At this stage, each federation computes its maximum privacy threshold. In particular, for a federation F with members $F = \{p_1, \dots, p_n\}$, and a representative p_* , p_i reports d_{p_i} and $\varepsilon_{p_i}^T$ to p_* for all $i \in \{1, \dots, n\}$. p_* computes the maximum information threshold of federation F , ε_F^T , as given by (8.1). Hence, p_* places a bid to D to obtain $\$M$, which maximises the earning for F under the constraint of their maximum privacy threshold and the maximum

budget available from D , i.e., p_* wishes to maximize M within the limits $M \leq B$ and $f(M) \leq \varepsilon_F^T$. Thus, p_* bids for sending data privatized using the k -RR mechanism with ε_F^T in exchange of $f^{-1}(\varepsilon_F^T)$.

At the end of this bidding process by all the federations, D ends up with $\varepsilon = \{\varepsilon_{F_1}^T, \dots, \varepsilon_{F_k}^T\}$, the maximum privacy thresholds of all the federations. D must ensure that $\sum_{i=1}^k f^{-1}(\varepsilon_{F_i}^T) \leq B$, adhering to her financial budget. In all probability, $\sum_{i=1}^k f^{-1}(\varepsilon_{F_i}^T)$ is likely to exceed B in a pragmatic setup. D needs a way to finalize the deal with the federations staying within her financial budget while maximizing her information gain, i.e., maximizing $\sum_{i=1}^k d_{F_i} \varepsilon_{F_i}$, where d_{F_i} is the total number of data points obtained from the i^{th} federation F_i , and ε_{F_i} is the overall privacy parameter of the k -RR differential privacy with the combined data of all the members of F_i .

A way D could close the deal with the federations is by proposing to receive information obfuscated with $w^* \varepsilon_{F_i}^T$ using k -RR mechanism to $F_i \forall i \in \{1, \dots, k\}$, where

$$w^* = \operatorname{argmax} \left\{ \sum_{i \in \{1, \dots, k\}} f^{-1}(w \varepsilon_{F_i}^T) \leq B, w \in [0, 1] \right\}, \quad (8.2)$$

i.e., proportional to every federation's maximum privacy threshold ensuring that the price to be paid to the federations is within D 's budget. Note that $w \in [0, 1]$ guarantees that $w \varepsilon_F^T \leq \varepsilon_F^T$ for every federation F , making the proposed privacy parameter possible to achieve by every federation as it is within their respective maximum privacy thresholds. Let the combined privacy parameter for federation F_i proposed by D to successfully complete the deal be denoted by $\varepsilon_{F_i}^P = w^* \varepsilon_{F_i}^T \forall i \in \{1, \dots, k\}$. We call this $\varepsilon_{F_i}^P$ to be the *promised* level of privacy to be achieved by each federation to participate in the data market.

The above method to scale down the maximum privacy parameters to propose a deal maximizing D 's information gain is just one of the possible approaches. In theory, any method that ensures the total price to be paid to all the federations in exchange for their data is within D 's budget and the privacy parameters proposed are within the corresponding privacy budgets of the federations could be implemented to furnish a revised set of privacy parameters and, in turn, the price associated with them. When all the federations are informed about the revised privacy parameters desired of them and they agree to proceed with the private-data trading with the data consumer by achieving the revised privacy parameter by combining the data of their members, we say *the deal has been sealed* between the federations and the data consumer.

Once the deal is sealed between the federations and the data consumer, F_i is expected to provide data gathered from its members with an overall obfuscation with the privacy parameter $\varepsilon_{F_i}^P$ using the k -RR mechanism, in exchange for a price $M^i = f^{-1}(\varepsilon_{F_i}^P)$ for every $i \in \{1, \dots, k\}$. Failing to achieve this parameter of privacy for any federation results in a failure to uphold the conditions of the *deal* and makes the deal void for that federation, with no price received.

A rational assumption made here is that if a certain federation F fails to gather data from its members such that the overall k -RR privacy parameter of F is less than ε_F^P , then F doesn't receive any partial compensation for its contribution, as it would incur an increase in communication

cost and time for the data consumer in proceeding to this stage and seal a new deal with F , instead of investing the revenue to a more responsible federation.

The rest of the process consists in collecting the data and it takes place within every federation F which has sealed the deal with the consumer. At the t^{th} round, for $t \in \{1, 2, \dots\}$, any member p of F has the freedom of contributing $d_p^t \leq d_p - \sum_{i=1}^{t-1} d_p^i$ data points privatized using k -RR mechanism with any parameter ε_p^t . The process continues until the overall information collected achieves a privacy level of at least ε_F^P . Let \mathcal{T} denote the number of rounds needed by F to achieve the required privacy level. As per the deal sealed between F and D , F needs to submit $D_F = \sum_{p \in F} \sum_{i=1}^{\mathcal{T}} d_p^i$ data points to D such that the overall k -RR privacy level of the collated data, ε_F , is at least ε_F^P , and in return F receives a financial revenue of M from D .

8.4.2 Earning Splitting

We use the Shapley value to estimate the contribution of each data provider of the federation, in order to split the whole earning M which F would receive from D at the end of the trade. Let $\psi : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}^+$ be the valuation function used for evaluating the Shapley values of the members after each contribution. If a certain member, p , of F reports d differentially private data points with privacy parameter ε , $\psi_i(v)$ should give the share of *contribution* made by p over the total budget, M , of F , to be split across all its members. It is assumed that each member, p , of F computes her Shapley value, knows what share of revenue she would receive by contributing her data privatized with a chosen privacy parameter, and uses this knowledge to decide on ε_p^t at every round t , depending on her financial desire. In our model, characteristic function $v(S)$ is as follows:

$$v(S) = \begin{cases} M, & \text{if } \varepsilon_F \geq \varepsilon_F^P \\ 0, & \text{if } \varepsilon_F < \varepsilon_F^P \end{cases}$$

where n is the number of data provider in subset S .

EXAMPLE 8.4.1. As an example, let us assume that there are p_1, p_2, p_3 , and each provider's contribution $\sum_{t=1}^{\mathcal{T}} d_p^t \frac{e^{\varepsilon_p^t}}{k-1+e^{\varepsilon_p^t}}$ are 1.0, 0.5 and 0.3. And we assume that ε_F^P is 1.4 and financial revenue of M is 60. In this case, the calculation of each provider's revenue using Shapley value is as follows:

Case 1) Only one data provider participates:

$$\begin{aligned} p_1 : v(p_1) &= 0 \\ p_2 : v(p_2) &= 0 \\ p_3 : v(p_3) &= 0 \end{aligned}$$

Case 2) Two providers participate: $v(p_1+)=0, v(p_2)=0$,

$$\begin{aligned} p_1 : v(p_1 + p_2) - v(p_2) &= M, v(p_1 + p_3) - v(p_3) = M \\ p_2 : v(p_1 + p_2) - v(p_1) &= M, v(p_2 + p_3) - v(p_3) = 0 \\ p_3 : v(p_1 + p_3) - v(p_1) &= 0, v(p_2 + p_3) - v(p_2) = 0 \end{aligned}$$

Case 3) All providers participate:

$$p_1 : v(p_1 + p_2 + p_3) - v(p_2 + p_3) = M$$

$$p_2 : v(p_1 + p_2 + p_3) - v(p_1 + p_3) = M$$

$$p_3 : v(p_1 + p_2 + p_3) - v(p_1 + p_2) = 0$$

According to the above results, the share of each user, according to their Shapley values, is as follows:

$$\psi_1(v) = \frac{0!2!}{3!}0 + \frac{1!1!}{3!}M + \frac{1!1!}{3!}M + \frac{2!0!}{3!}M = \frac{4M}{6} = 40$$

$$\psi_2(v) = \frac{0!2!}{3!}0 + \frac{1!1!}{3!}M + \frac{1!1!}{3!}0 + \frac{2!0!}{3!}M = \frac{2M}{6} = 20$$

$$\psi_3(v) = \frac{0!2!}{3!}0 + \frac{1!1!}{3!}0 + \frac{1!1!}{3!}0 + \frac{2!0!}{3!}0 = \frac{0M}{6} = 0$$

In this example, p_3 has no effect on achieving the ε_F^P . Thus, p_3 is excluded from the revenue distribution. If the revenue were distributed proportionally, without considering the Shapley values, the revenue of p_1 would be 33, p_2 is 17, and p_3 is 10. It would mean p_1 and p_2 would receive lower revenues even though their contribution are sufficient to achieve the ε_F^P , irrespective of the participation of p_3 . The Shapley value enables the distribution of revenues only for those who have contributed to achieving the goal.

One of the problems of computing the Shapley values is the high computational complexity involved. If there is a large number of players, i.e., the size of a federation is large, the total number of subsets to be considered becomes considerably large, engendering a limitation to real-world applications. To overcome this, we use a *pruning technique* to reduce the computational complexity of the mechanism. A given federation F receives revenue M only when $\varepsilon_F \geq \varepsilon_F^P$, as per the deal sealed with the data consumer. Therefore, it is not necessary to calculate for Shapley values for the cases where $\varepsilon_F < \varepsilon_F^P$, since such cases do not contribute towards the overall Shapley value evaluated for the members of F .

The data trading between the data consumer and the federations would typically continue periodically for a length of time. For example, Acxiom, a data broker company, periodically collects and manages personal data related to daily life, such as consumption patterns and occupations. Periodic data collection has a higher value than one-time data collection because it can track temporal trends. For simplicity of explanation, let us assume that the trading occurs every year. Hence, we consider a yearly period to illustrate the final two steps of our proposed mechanism – *swift data collection* and the *penalty scheme*. This would ensure that the data collection process is as quick as possible for every federation every year. Additionally, this would motivate the members to cooperate and act in the best interests of their respective federations by not, unnecessarily, withholding their privacy contributions to hinder achieving the privacy goals of their group, as per the deal finalized with D .

Let $R \in \mathbb{N}$ be the *tolerance period*. For a member $p \in F$, we denote $d(m)_p^i$ to be the number of data points reported by p in the i^{th} round of data collection of year m and we denote $\varepsilon(m)_p^i$ to be the corresponding privacy parameter used by p to obfuscate the data points. Let T_m be the number of rounds of data collection needed in year m by federation F to achieve their privacy

goal. We denote the total number of data points reported by p in the year m by $d(m)_p$, and observe that $d(m)_p = \sum_{i=1}^{T_m} d(m)_p^i$. Let $\varepsilon(m)^P$ denote the value of the privacy parameter of the combined k -RR mechanism of the collated data that F needs, in order to successfully uphold the condition of the deal sealed with D .

DEFINITION 8.4.1 (Contributed privacy level). For a given member $p \in F$, we define the *contributed privacy level* of p in year m as

$$\varepsilon(m)_p = \sum_{i=1}^{T_m} \varepsilon(m)_p^i$$

DEFINITION 8.4.2 (Privacy saving). For a given member $p \in F$, we define the *privacy saving* of p over a tolerance period R (given by a set of some previous years), decided by the federation F , as

$$\Delta_p = \sum_{m \in R} \left(d(m)_p \varepsilon_p^T - d(m)_p \varepsilon(m)_p \right)$$

Swift data collection: It is in the best interest of F , and all its members, to reduce the communication cost, time, and resources over the data collection rounds, and achieve the goal of ε^P as soon as possible, to catalyze the trade with D , and receive the financial revenue. We aim to capture this through our mechanism, and enable the members not to “hold back” their data well below their capacity.

To do this, in our model we design the Shapley valuation function, ψ , such that for $p \in F$, in year m , $\psi(N_p \varepsilon(m)_p^{t+1}, d(m)_p, M) = \psi(\varepsilon(m)_p^t, d(m)_p, M)$, where $N_p \in \mathbb{Z}^+$ is the *catalyzing parameter* of the data collection, decided by the federation, directly proportional to Δ_p . In particular, for $p \in F$, and a tolerance period R decided, in prior, by F , it is a reasonable approach to make $N_p \propto \Delta_p$, as this would mean that any member $p \in F$, reporting $d(m)_p$ data points, would need to use N_p times higher value of ε in the $(t+1)^{st}$ round of data collection in the year m , as compared to that in the t^{th} round for the same number of data points reported to get the same share of the benefit of the federation’s overall revenue, where N_p is decided by how much privacy savings p has had over a fixed period of R .

This is made to ensure that if a member of a federation has been holding back her information by using high values of privacy parameters over a period of time, she should need to compensate in the following year by helping to quicken up the process of data collection of her federation. This should motivate the members of F to report their data with a high value of the privacy parameter in earlier rounds than later, staying within their privacy budgets, so that the number of rounds needed to achieve $\varepsilon(m)^P$ is reduced.

Penalty scheme: It is desirable to have every member of any given federation cooperate with the other members of the same federation and facilitate the trading process in the best interest of the federation, to the best of their ability. That is why, in our mechanism, we incorporate an idea of a *penalty scheme* for the members of a federation who are being selfish by keeping a substantial

gap between their maximum privacy threshold and their contributed privacy level, wishing to enjoy benefits of the revenue at an unfair cost of other members providing information privatized with almost their maximum privacy threshold. To prevent such non-cooperation and attempted “free ride”, we design a penalty scheme in the mechanism.

DEFINITION 8.4.3 (Free rider). We call a certain member $p \in F$ to be a *free rider* if $\Delta_p \geq \delta_F$, for some $\delta_F \in \mathbb{R}^+$. Here, δ_F is a threshold decided by the federation F beforehand and informed to all the members of F .

Thus, in the ideal case, every member of F would have their privacy savings to be 0 if everyone contributed information to the best of their abilities, i.e., provided data obfuscated with their maximum privacy parameter. But to capture a pragmatic scenario as a federation, a threshold amount of privacy savings is tolerated for every member. Under the penalty scheme, if a certain member $p \in F$ qualifies as a free rider, she is excluded from the federation and is given a demerit point by the federation, that can be recorded by a central system keeping track of every member of every federation, preventing p from getting admission to any other federation for her tendency to free ride. This would mean p and has the responsibility of trading with the data consumer by herself. We could define the Shapley valuation function used to determine the share of p 's contribution such that $f^{-1}(\varepsilon_p^T) < \psi(v, M)$, implying that the revenue to be received by p dealing directly with D , providing one data point obfuscated with her maximum privacy threshold with respect to the privacy valuation function f , would be giving a much lower revenue than what p would receive being a member of federation F . Imposing the penalty scheme is expected to drive every member of a given federation to be cooperating with the interests of the federation and all the other fellow members to the best of their abilities, preventing potential free riders. Our proposed mechanism is elucidated by Algorithms 10, 11 and 12.

THEOREM 8.1. If the privacy valuation function used in order to impose the *penalty scheme* to any member p of a federation F is $f(m) = K_1(e^{K_2 m} - 1)$, the Shapley valuation function ψ chosen must satisfy

$$\frac{\ln\left(\frac{\varepsilon_p^T}{K_1} + 1\right)}{K_2} < \psi\left(\varepsilon_p^T, \frac{\ln\left(\frac{w^* \varepsilon_p^T}{K_1} + K\right)}{K_2}\right),$$

where $K = \frac{1}{K_1} \sum_{\substack{p' \neq p \\ p' \in F}} d(m)_{p'} \varepsilon_{p'}^T + 1$ and w^* is the scaling parameter as given by (8.2).

Proof. In Appendix I. □

8.5 Experimental results

Algorithm 10: Federation based data trading algorithm

Input: Federation F , Data consumer D ;
Output: ε_F^P and M ;
 D broadcasts total budget B and f ;
Federation F computes the $\varepsilon_F^T = \sum_{i=1}^n d_{p_i} \varepsilon_{p_i}^T$;
 p_* places a bid to D to obtain revenue M ;
 F and D “seal the deal” to determine the ε_F^P and M ;
while $\varepsilon_F \leq \varepsilon_F^P$ and $t \leq T$ **do**
 SWIFTDataCollection(F, ε_F^P);
 p_* computes the overall privacy ε_F
if $\varepsilon_F \geq \varepsilon_F^P$ **then**
 F receives the revenue M ;
 p_* computes the Shapley value $\psi_i(v, M)$;
 p_i get their share of the revenue M
else
 deal fails

Algorithm 11: Swift data collection algorithm

Input: $F = \{p_1, \dots, p_{n_F}\}, \varepsilon_F^P$;
Output: $\varepsilon(m)_p^t$;
Function SwiftDataCollection(F, ε_F^P):
 while $i \leq n_F$ **do**
 Compute Δ_{p_i} ;
 Compute the catalyzing parameter N_{p_i} ;
 Determine the $\varepsilon(m)_{p_i}^t = N_{p_i} \varepsilon(m)_{p_i}^{t-1}$

In this section, we show the experimental results that illustrate the efficient working of the proposed method in order to obtain the promised ε and reduce the computation time for Shapley value evaluation. The number of data providers constituting the federation is set to 25, 50, 75, and 100, respectively. The ε_p^T is sampled independently for each member p following $\mathcal{N}(5, 1)$ and we only consider the values in the range $[1, 10]$. The hardware environment used for the experiments was an Intel(R) i5-9400H CPU with 16 GB of memory.

8.5.1 Number of rounds needed for data collection

Achieving the ε_F^P is the key to the participation of F in the data market. If ε_F^P is not achieved as the collated information level for the federation, there is no revenue from the data consumer. Thus, it is important to encourage the data providers to report sufficient data in order to reach

Algorithm 12: Penalty scheme

Input: $F = \{p_1, \dots, p_{n_F}\}, \Delta_F = \{\Delta_{p_1}, \dots, \Delta_{p_{n_F}}\}, \delta_F$;
Output: Updated F ;
while $i \leq n_F$ **do**
 if $\Delta_{p_i} \geq \delta_F$ **then**
 $F \setminus \{i\}$

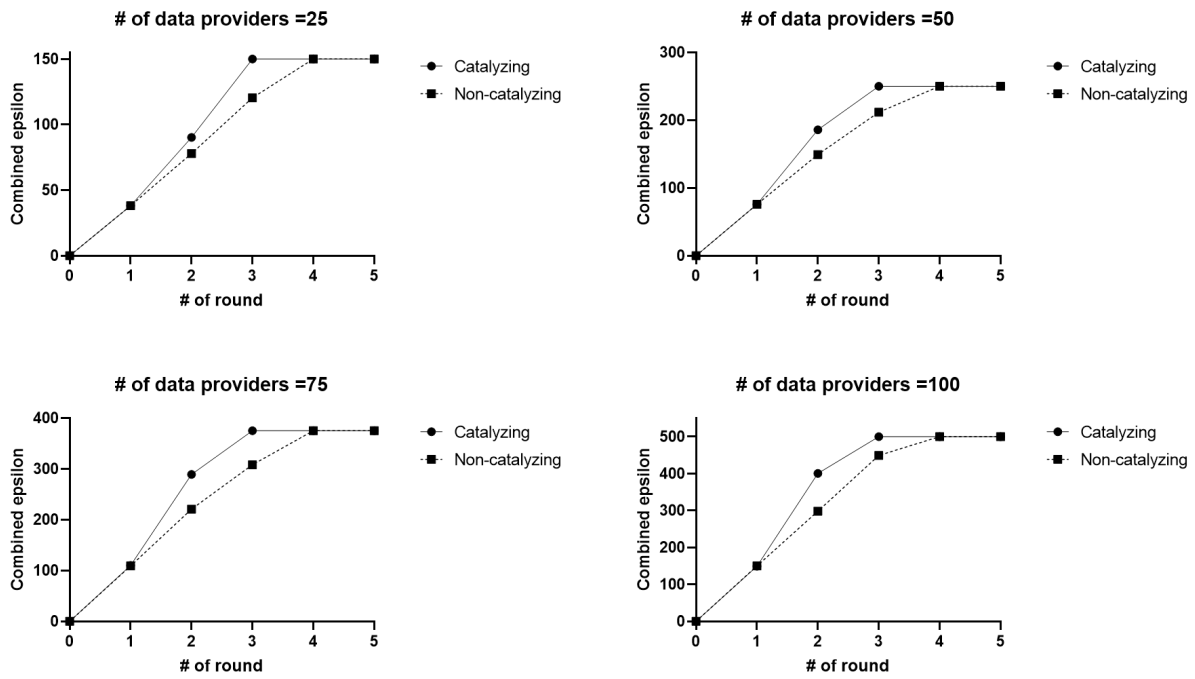


Figure 8.2: The number of rounds required by a federation (of varying sizes) to achieve the privacy level (i.e., information accuracy) to be reported to the data consumer is less for the catalyzing method introduced in the data collection than its non-catalyzing counterpart.

the goal of the deal sealed with the data consumer. The swift data collection is a way to catalyze the process of obtaining data from the members of every federation F minimising the number of rounds of data collection to achieve ε_F^P . Furthermore, we set $N_p = \frac{\Delta_p}{d(m)_p \varepsilon_p^T}$ for a certain member p in federation F to motivate the data providers who have larger privacy savings to provide more information per round.

In the experiment, ε_F^P is set to be 125, 250, 375 and 500, respectively. Data provider p determines $\varepsilon(m)_p^t$ randomly in the first round, and then computes $\varepsilon(m)_p^t$ according to N_p , for every p in the federation. We compare two cases, the catalyzing method and the non-catalyzing method.

As illustrated in Figure 8.2, we see that both catalyzing data collection and its non-catalyzing counterpart achieve the promised epsilon values within 5 rounds of data collection, but it can be seen that the catalyzing method achieves ε_F^P earlier because data providers decide the privacy level used to obfuscate their data with considering their privacy savings resulting in a *swift data collection*.

8.5.2 Number of free riders by penalty scheme

The penalty scheme that prevents free riders is based on the premise that trading data by participating in a federation is more beneficial than trading data directly with data consumers (Theorem 8.1). We evaluated the number of free riders in the catalyzing and non-catalyzing methods according to the increase of the threshold δ_F in the experiment.

Figure 8.3 shows that the number of free riders increases in both techniques as the threshold

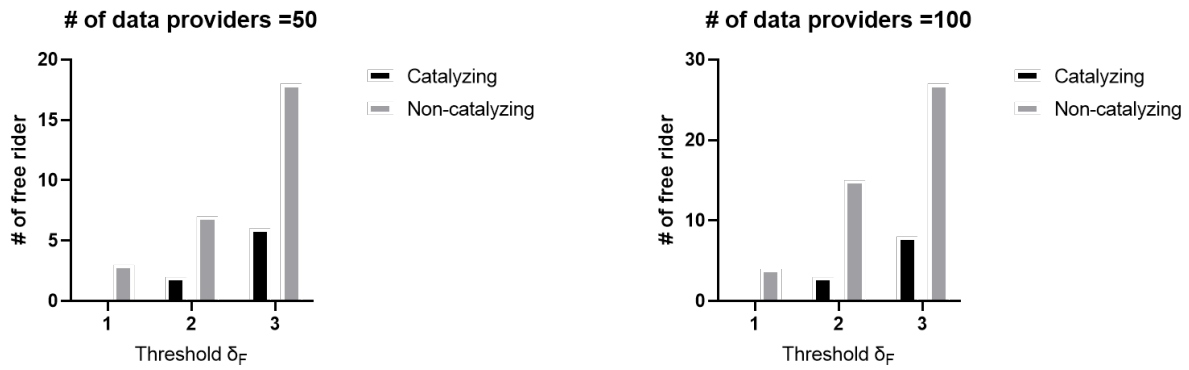


Figure 8.3: The number of free riders incurred by the penalty scheme (for varying sizes of the federation and the tolerance threshold) is significantly less adapting the catalyzing method for data collection than the non-catalyzing method.

value δ_F is set to 1, 2, and 3. However, the non-catalyzing method allows more free riders than the catalyzing method which changes the accuracy of the reported information according to privacy saving Δ_P . In other words, the catalyzing method and penalty scheme help to keep members in the federation by inducing them to reach the target epsilon quicker and, in turn, enables the method to work efficiently.

8.5.3 Reduced Shapley value computation time

Table 8.1: Computation time of brute force and proposed pruning method

# of data providers	brute force(Sec)	pruning method (Sec)
15	0.003	0.0007
18	0.02	0.001
21	0.257	0.0049
24	2.313	0.009
27	19.706	0.019

As mentioned in Section 8.4.2, one of the limitations of Shapley value evaluation is to compute it for all combinations of subsets. Through this experiment, we demonstrate that the proposed pruning technique reduces the computation time for calculating the Shapley values. We compared the computation times of the proposed method with the brute force method that calculates all the cases by increasing the number of data providers in the federation from 15 to 27. Table 8.1 shows that the computation time of Shapley value evaluation increases exponentially because the total number of subsets to be considered does the same. The proposed pruning technique can calculate the Shapley values in less time by removing unnecessary computations.

Conclusion and Future Work

In this thesis, we have explored the relationship between privacy and utility from a foundational standpoint. Primarily, we focused on the interaction of differentially private systems and their close variants with the utility of the data providers (often the users of various data-driven services) and the data consumers (often the providers of such services) in a variety of domains. We used the insights gained through the theoretical and experimental analysis to investigate the optimality of privacy guarantees with a wide range of context-specific notions for utility. This chapter is aimed to bring together and reconcile the ideas explored in this thesis to show how the core theme of understanding and optimizing the privacy-utility trade-off has been addressed by this thesis and how these open up new avenues for future work. In summary:

- i) Part II of the thesis elaborately studied the triadic interaction between the privacy of the users, their quality of service, and the statistical utility of the collected noisy data in the context of location-based services.
- ii) Part III of the thesis analysed privacy-preserving aspects of the collaborative and federated approach to model training and dissected the interplay of privacy with the utility in the context of ethical and trustworthy machine learning.
- iii) Part IV of the thesis furnished an efficient and practical model of private data trading by enhancing the functionality of data markets and optimizing the privacy requirements of the clients with the financial utility of the data collectors.

Part II: Optimizing location-privacy preserving mechanisms

Chapter 3 bridged together some key ideas from rate-distortion theory and expectation maximization in statistics by establishing a duality between the Blahut-Arimoto algorithm and the

iterative Bayesian update and unravelled the comprehensive privacy-preserving properties of the Blahut-Arimoto algorithm, especially for location data. Benefitting from all the privacy-preserving and utility-friendly characteristics of the Blahut-Arimoto algorithm and its systematic entanglement with the iterative Bayesian update, PRIVIC, a mechanism for incremental collection of location-data, was proposed which instituted a rudimentary stepping stone heading towards a multi-faceted trade-off between the formal standard of location-privacy, privacy from an information theoretical perspective, quality of service for the users, and the statistical precision of the noisy data.

Chapter 4 theorised a bounded paradigm for geo-indistinguishability called approximate geo-indistinguishability and used it to propose a location privacy-preserving navigation protocol for the use case of electric vehicles making repeated queries during their journeys looking for available charging stations. Our method protects the privacy of both the specific query locations and the trace of the entire journey while fostering a high quality of service for the users and aiding in efficient route planning and traffic prediction by considering real-time queries. Considering a practical use case that is enduring a day-to-day surge in its social relevance, this study highlighted a method to obtain “privacy for free”, substantially enhancing the privacy-utility trade-off in the domain of navigation-based queries for vehicles especially when the points of interest (e.g., charging stations for electric vehicles) are sparse in the map.

The analyses carried out in Part II presents an in-depth understanding and illuminates the potential for enhancing the privacy-utility trade-off pertaining to both the users and service providers in location-based services. However, in addition to taking a step towards optimizing the location-privacy guarantees with different parameters of utility, this work unfurls a range of open questions that would be extremely interesting to address in future works. One such is to characterize the vulnerability of PRIVIC in handling malicious users falsely reporting their locations to compromise the privacy of the honest users located in isolated points in the map as presented in Section 3.8 of Chapter 3. In particular, we have seen that when all the users who are involved in sharing the locations and adhering to PRIVIC are honest, the Blahut-Arimoto algorithm, through its intrinsic elastic distinguishability metric, protects the locations isolated from the crowd much better than the Laplace mechanism, the canonical mechanism for geo-indistinguishability, does. However, considering adversarial users aiming to break the extensive privacy guarantees offered by PRIVIC, we highlight a major vulnerability of the Blahut-Arimoto algorithm, the elastic distinguishability metric, and, in turn, PRIVIC. Therefore, one of the major future directions to pursue is to propose a generalized and more robust version of PRIVIC that can handle such malicious data sharing. One possible way to formalise this idea could be by incorporating some penalty scheme, along the lines of what has been proposed in Section 8.4.2 of Chapter 8, to identify and exclude dishonest users intending to break the system.

Another important future avenue leading on from the work carried out in Part II is to re-evaluate the extensive privacy guarantees and the optimization of privacy with the different kinds of utility that PRIVIC offers by considering a correlation between the shared data and a change in the prior distribution. In practice, location-based services like *Pokémon Go* and *Tinder* collect and, in turn, use location data from users at different timesteps that are (obviously)

correlated and the original distribution changes over time [235, 236]. Hence, an important goal moving forward is to examine the functionality of PRIVIC capturing such realistic scenarios where the original distribution of the users' locations alters and the collected data are not necessarily i.i.d. in nature. An added advantage of being potentially able to handle correlations between the shared locations in the context of location-privacy preserving mechanisms is that it would open up complimentary avenues of future work from Chapter 4 and enable us to extend geo-indistinguishability and approximate geo-indistinguishability to traces as opposed to just individual locations. There have been a few approaches explored in the literature [112, 237] to privatize correlated locations and trajectories as opposed to independently sampled individual locations. We plan to incorporate these with our work to propose more robust, pragmatic, and efficient location-privacy-enhancing technologies with an optimal privacy-utility trade-off.

Part III: Investigating privacy guarantees of federated learning

Chapter 5 investigated personalized federated learning with metric privacy guarantees enabling a privacy-preserving collaborative paradigm for federated model training catering to the diversity present amongst the clients and their datasets. Our mechanism shows particular promise for machine learning models with a relatively small number of parameters. For larger models, experimental results demonstrate the effectiveness of the Laplace mechanism to defend against the gradient-inversion attacks like DLG. Additionally, we also evaluate the fairness of the trained personalized federated learning models under d -privacy using various group fairness metrics and concluded that personalized models significantly improve group fairness across all evaluated metrics and privacy levels, effectively mitigating biases towards the majority group, unlike non-personalized models. Remarkably, even with the highest level of privacy protection, personalized models consistently maintain superior fairness compared to their non-personalized counterparts, making a strong case for personalized model training in federated learning under d -privacy as a comprehensive solution for privacy-preserving and ethical machine learning. On the other hand, Chapter 6 presented a thorough foundational analysis to study the information leakage caused by the classical paradigm of (non-personalized) federated learning and, hence, dissected the working of the gradient-inversion-based data-reconstruction attacks threatening the very philosophy behind federated learning.

The work pursued in Part III of this thesis, alongside bespeaking various absorbing aspects of privacy-preserving machine learning through the recently popularized federated approach, unfolds many interesting open questions to be studied in the future. One of the major future lines of work would be to incorporate more practical notions of user-oriented privacy (e.g., vulnerability against being identified in the training set in a black-box setting) with personalized federated learning with d -privacy guarantees. Chapter 5 provided a solid ground to believe that combining metric privacy with personalized model training in a federated environment provides a win-win situation from the perspective of the model's accuracy, privacy, and fairness standards. However, the information about the clusters could potentially unravel some gaping holes in shielding personalized models from membership inference attacks from malicious third parties. Such a study would be a gripping avenue of future work to comprehensively understand

the multimodal privacy-utility trade-off involved in personalized federated learning. A more ambitious goal would be to appropriately modify the IFCA algorithm¹ to minimize the strength of membership inference attacks under a given level of privacy budget and an acceptable level of accuracy and fairness. A third line of study which can be really appealing to pursue in the future could be to combine the approach of explaining the information leakage from shared gradient updates, as presented in Chapter 6, with the privacy-preserving personalized federated learning method as proposed in Chapter 5 and, more ambitiously, to analyse the potential risks of membership inference attacks that personalized model training entails from a foundational and information theoretical perspective and fabricate appropriate defence mechanisms accordingly.

Part IV: Optimally trading private data in data markets

Chapter 7 envisioned a scenario where data consumers trade directly with the data providers to obtain (privatized) data in exchange for financial compensation. This work started by proposing a truthful incentive mechanism to ensure that data providers do not indulge in dishonest reporting of their privacy requirements in the quest of earning more revenue which, eventually, enabled optimizing the privacy budgets of users, their financial benefits from the trade, and the profit of the data consumers. Chapter 8 considered an environment in which data providers form federations dealing with the data consumers. This chapter proposed a method for efficiently and swiftly trading locally differentially private data in such federated data markets and fairly splitting the earnings within federations by factoring in the contributions of each member. In addition, this chapter introduced the notion of a penalty scheme to penalize the free riders hindering the coherent and smooth functioning of federated data markets. The work exhibits an interdisciplinary approach by combining aspects of economics, game theory, and differential privacy to address the burning topic of appropriate pricing of differentially private data and a fair approach to motivate users to comply with the working of federated data markets.

The methods proposed in Part IV evolve some exciting prospects for potential future works. One very interesting avenue to explore in the future would be to integrate the notion of the truthful incentive mechanism of Chapter 7 in auditing differential privacy noise to ensure that genuine privacy requirements of users are satisfied. Another exciting line of future works following the theme of Chapter 8 would be to incentivize clients participating in federated model training (especially, in a cross-silo setting) to comply with the system and, in turn, with the help of the game theoretical methods used in Chapter 8 to fairly split the earnings within the participating silos. Finally, as elaborated earlier, it would be intriguing to integrate the idea of a penalty scheme, as introduced in Chapter 8, with PRIVIC (in Chapter 3) to deal with adversarial users keen to break the privacy of isolated locations in the maps.

Final remark To conclude, we remark that while the kernel of the investigations in this thesis has been to study and enhance the trade-off between privacy and utility to ensure that users of data-driven services and the corresponding service providers churn out the maximum possible

¹the algorithm used for personalized model training in Chapter 5

benefit they can whilst upholding the fundamental right to privacy in a formal way, many interesting and diverse applications of differential privacy and its variants are yet to be explored which, in their own unique ways, will open up different notions of utility and, corresponding, privacy-utility trade-offs. Some of the findings of this thesis have immense potential to be applied in a wide variety of domains in pursuit of achieving an optimal mechanism for privacy and utility. Hence, we ambitiously envision that building up on the work carried out in this thesis and with further advancements in research and technology, given the required thresholds for privacy guarantees and for a number of notions of utility, we will be able to systematically construct a system involving the optimal mechanism to satisfy these required goals in a way that both the users and the providers of the service involved are motivated to conform to the working of the system.

References

- [1] B. Marr. *How much data do we create every day? the mind-blowing stats everyone should read* (2022). URL <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/>.
- [2] Wikipedia contributors. *2021 facebook leak* — *Wikipedia, the free encyclopedia*. https://en.wikipedia.org/w/index.php?title=2021_Facebook_leak&oldid=1157893507 (2023). [Online; accessed 13-July-2023].
- [3] J. Silverstein. *Hundreds of millions of facebook user records were exposed on amazon cloud server* (2019). URL <https://www.cbsnews.com/news/millions-facebook-user-records-exposed-amazon-cloud-server/>.
- [4] Wikipedia contributors. *Facebook–cambridge analytica data scandal* — *Wikipedia, the free encyclopedia*. https://en.wikipedia.org/w/index.php?title=Facebook%E2%80%93Cambridge_Analytica_data_scandal&oldid=1162404364 (2023). [Online; accessed 13-July-2023].
- [5] Wikipedia contributors. *2018 google data breach* — *Wikipedia, the free encyclopedia*. https://en.wikipedia.org/w/index.php?title=2018_Google_data_breach&oldid=1148797801 (2023). [Online; accessed 13-July-2023].
- [6] D. MacMillan and R. MacMillan. *Google exposed user data, feared repercussions of disclosing to public* (2018). URL <https://www.wsj.com/articles/google-exposed-user-data-feared-repercussions-of-disclosing-to-public-1539017194>.
- [7] Wikipedia contributors. *Yahoo! data breaches* — *Wikipedia, the free encyclopedia*. https://en.wikipedia.org/w/index.php?title=Yahoo!_data_breaches&oldid=1086724827 (2022). [Online; accessed 10-May-2022].
- [8] A. Narayanan and V. Shmatikov. *Robust de-anonymization of large sparse datasets*. In *Proceedings of the Symposium on Security and Privacy (SP)*, pp. 111–125 (IEEE, 2008).
- [9] O. e Office of the Victorian Information Commissioner. *Disclosure of myki travel information*. URL https://ovic.vic.gov.au/wp-content/uploads/2019/08/Report-of-investigation_disclosure-of-myki-travel-information.pdf.
- [10] T. Dalenius. *Towards a methodology for statistical disclosure control*. *Statistik Tidskrift* **15**, 429 (1977).

- [11] L. Sweeney. *k-anonymity: A model for protecting privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(5), 557 (2002).
- [12] S. Garfinkel, J. M. Abowd, and C. Martindale. *Understanding database reconstruction attacks on public data*. Communications of the ACM **62**(3), 46 (2019).
- [13] C. Dwork, F. McSherry, K. Nissim, and A. Smith. *Calibrating noise to sensitivity in private data analysis*. In S. Halevi and T. Rabin, eds., *Theory of Cryptography*, pp. 265–284 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).
- [14] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. *Our data, ourselves: Privacy via distributed noise generation*. In S. Vaudenay, ed., *Advances in Cryptology - EUROCRYPT 2006*, pp. 486–503 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).
- [15] N. Fernandes, M. Dras, and A. McIver. *Processing text for privacy: An information flow perspective*. In K. Havelund, J. Peleska, B. Roscoe, and E. P. de Vink, eds., *Formal Methods - 22nd International Symposium, FM 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 15-17, 2018, Proceedings*, vol. 10951 of *Lecture Notes in Computer Science*, pp. 3–21 (Springer, 2018). URL https://doi.org/10.1007/978-3-319-95582-7_1.
- [16] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. *Geo-indistinguishability: Differential privacy for location-based systems*. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, CCS '13*, p. 901–914 (Association for Computing Machinery, New York, NY, USA, 2013). URL <https://doi.org/10.1145/2508859.2516735>.
- [17] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi. *Broadening the scope of differential privacy using metrics*. In *International Symposium on Privacy Enhancing Technologies Symposium*, pp. 82–102 (Springer, 2013).
- [18] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth. *Differential privacy: An economic method for choosing epsilon*. In *IEEE 27th Computer Security Foundations Symposium, CSF 2014, Vienna, Austria, 19-22 July, 2014*, pp. 398–410 (IEEE Computer Society, 2014). URL <https://doi.org/10.1109/CSF.2014.35>.
- [19] M. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith. *The Science of Quantitative Information Flow*. Information Security and Cryptography (Springer, Springer Nature, United States, 2020).
- [20] L. Zhu, Z. Liu, , and S. Han. *Deep leakage from gradients*. In *Annual Conference on Neural Information Processing Systems (NeurIPS)* (2019).
- [21] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov. *See through*

- gradients: Image batch recovery via gradinversion*. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16332–16341 (IEEE Computer Society, Los Alamitos, CA, USA, 2021). URL <https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01607>.
- [22] S. L. Garfinkel, J. M. Abowd, and S. Powazek. *Issues encountered deploying differential privacy*. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, pp. 133–137 (2018).
- [23] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. *Local privacy and statistical minimax rates*. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438 (2013).
- [24] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. *Optimal geo-indistinguishable mechanisms for location privacy*. In *Proceedings of the 21th ACM Conference on Computer and Communications Security (CCS 2014)* (2014).
- [25] N. Fernandes, A. McIver, and C. Morgan. *The laplace mechanism has optimal utility for differential privacy over continuous queries*. In *36th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2021.*, pp. 1–12 (IEEE, 2021). URL <https://doi.org/10.1109/LICS52264.2021.9470718>.
- [26] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. *Communication-Efficient Learning of Deep Networks from Decentralized Data*. In A. Singh and J. Zhu, eds., *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282 (PMLR, 2017). URL <https://proceedings.mlr.press/v54/mcmahan17a.html>.
- [27] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. *Deep Learning with Differential Privacy*. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16*, pp. 308–318 (ACM, New York, NY, USA, 2016). URL <http://doi.acm.org/10.1145/2976749.2978318>.
- [28] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. *Learning Differentially Private Recurrent Language Models*. In *International Conference on Learning Representations* (2018). URL <https://openreview.net/forum?id=BJ0hF1Z0b>.
- [29] H. B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, N. Papernot, and P. Kairouz. *A general approach to adding differential privacy to iterative training procedures* (2019). [1812.06210](https://arxiv.org/abs/1812.06210).
- [30] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers. *Protection against reconstruction and its applications in private federated learning* (2019). URL <https://arxiv.org/pdf/1812.00984>.
- [31] R. Agrawal, R. Srikant, and D. Thomas. *Privacy preserving olap*. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 251–262 (2005).

- [32] P. Kairouz, K. Bonawitz, and D. Ramage. *Discrete distribution estimation under local privacy*. In *International Conference on Machine Learning*, pp. 2436–2444 (PMLR, 2016).
- [33] D. Agrawal and C. C. Aggarwal. *On the design and quantification of privacy preserving data mining algorithms*. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '01, p. 247–255 (Association for Computing Machinery, New York, NY, USA, 2001). URL <https://doi.org/10.1145/375551.375602>.
- [34] E. ElSalamouny and C. Palamidessi. *Full convergence of the iterative bayesian update and applications to mechanisms for privacy protection*. CoRR **abs/1909.02961** (2019). **1909.02961**, URL <http://arxiv.org/abs/1909.02961>.
- [35] L. V. Kantorovich. *Mathematical Methods of Organizing and Planning Production*, vol. 6 (INFORMS, 1960).
- [36] Úlfar Erlingsson, V. Pihur, and A. Korolova. *Rappor: Randomized aggregatable privacy-preserving ordinal response*. In *Proceedings of the 21st ACM Conference on Computer and Communications Security* (Scottsdale, Arizona, 2014). URL <https://arxiv.org/abs/1407.6981>.
- [37] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec. *Protecting location privacy: Optimal strategy against localization attacks*. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, p. 617–627 (Association for Computing Machinery, New York, NY, USA, 2012). URL <https://doi.org/10.1145/2382196.2382261>.
- [38] R. Shokri. *Privacy games: Optimal user-centric data obfuscation*. *Proceedings on Privacy Enhancing Technologies* **2015**(2), 299 (2015). URL <https://doi.org/10.1515/popets-2015-0024>.
- [39] F. Galli, S. Biswas, K. Jung, C. Palamidessi, and T. Cucinotta. *Group privacy for personalized federated learning*. arXiv preprint arXiv:2206.03396 (2022).
- [40] K. Chatzikokolakis, C. Palamidessi, and M. Stronati. *Constructing elastic distinguishability metrics for location privacy*. *Proceedings on Privacy Enhancing Technologies* **2015**(2), 156 (2015). URL <https://doi.org/10.1515/popets-2015-0023>.
- [41] R. E. Blahut. *Computation of channel capacity and rate-distortion functions*. *IEEE Trans. Inform. Theory* **18**, 460 (1972).
- [42] S. Arimoto. *An algorithm for computing the capacity of arbitrary discrete memoryless channels*. *IEEE Trans. Inf. Theory* **18**, 14 (1972).
- [43] S. Oya, C. Troncoso, and F. Pérez-González. *Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms*. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1959–1972 (ACM, 2017). URL <http://doi.acm.org/10.1145/3133956.3134004>.

- [44] C. E. Shannon. *A mathematical theory of communication*. The Bell System Technical Journal **27**(3), 379 (1948).
- [45] I. Csiszar. *On the computation of rate-distortion functions (corresp.)*. Information Theory, IEEE Transactions on **20**, 122 (1974).
- [46] E. ElSalamouny and C. Palamidessi. *Generalized iterative bayesian update and applications to mechanisms for privacy protection*. In *2020 IEEE European Symposium on Security and Privacy (EuroS P)*, pp. 490–507 (2020).
- [47] J. Brickell and V. Shmatikov. *The cost of privacy: Destruction of data-mining utility in anonymized data publishing*. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, p. 70–78 (Association for Computing Machinery, New York, NY, USA, 2008). URL <https://doi.org/10.1145/1401890.1401904>.
- [48] S. Ioannidis, A. Montanari, U. Weinsberg, S. Bhagat, N. Fawaz, and N. Taft. *Privacy trade-offs in predictive analytics*. In *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '14, p. 57–69 (Association for Computing Machinery, New York, NY, USA, 2014). URL <https://doi.org/10.1145/2591971.2592011>.
- [49] A. Ghosh, T. Roughgarden, and M. Sundararajan. *Universally utility-maximizing privacy mechanisms*. SIAM Journal on Computing **41**(6), 1673 (2012). <https://doi.org/10.1137/09076828X>, URL <https://doi.org/10.1137/09076828X>.
- [50] M. Gupte and M. Sundararajan. *Universally optimal privacy mechanisms for minimax agents*. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, p. 135–146 (Association for Computing Machinery, New York, NY, USA, 2010). URL <https://doi.org/10.1145/1807085.1807105>.
- [51] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. *Optimizing linear counting queries under differential privacy*. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, p. 123–134 (Association for Computing Machinery, New York, NY, USA, 2010). URL <https://doi.org/10.1145/1807085.1807104>.
- [52] S. Oya, C. Troncoso, and F. Pérez-González. *Rethinking location privacy for unknown mobility behaviors*. In *2019 IEEE European Symposium on Security and Privacy (EuroS P)*, pp. 416–431 (2019).
- [53] M. Romanelli, K. Chatzikokolakis, and C. Palamidessi. *Optimal obfuscation mechanisms via machine learning*. In *2020 IEEE 33rd Computer Security Foundations Symposium (CSF)*, pp. 153–168 (2020).
- [54] W. Zhang, M. Li, R. Tandon, and H. Li. *Online location trace privacy: An information theoretic approach*. IEEE Transactions on Information Forensics and Security **14**(1), 235 (2019).

- [55] U. I. Atmaca, S. Biswas, C. Maple, and C. Palamidessi. *A privacy preserving querying mechanism with high utility for electric vehicles* (2022). [2206.02060](#).
- [56] P. Cuff and L. Yu. *Differential privacy as a mutual information constraint*. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, CCS '16, pp. 43–54 (ACM, New York, NY, USA, 2016). URL <http://doi.acm.org/10.1145/2976749.2978308>.
- [57] B. Köpf and D. A. Basin. *An information-theoretic model for adaptive side-channel attacks*. In P. Ning, S. D. C. di Vimercati, and P. F. Syverson, eds., *Proceedings of the 2007 ACM Conference on Computer and Communications Security (CCS 2007)*, pp. 286–296 (ACM, 2007).
- [58] Y. Zhu and R. Bettati. *Anonymity vs. information leakage in anonymity systems*. In *Proc. of ICDCS*, pp. 514–524 (IEEE Computer Society, 2005).
- [59] K. Chatzikokolakis, C. Palamidessi, and P. Panangaden. *Anonymity protocols as noisy channels*. *Information and Computation* **206**(2–4), 378 (2008). URL <http://hal.inria.fr/inria-00349225/en/>.
- [60] M. Abadi and D. G. Andersen. *Learning to protect communications with adversarial neural cryptography*. CoRR **abs/1610.06918** (2016). [1610.06918](#), URL <http://arxiv.org/abs/1610.06918>.
- [61] A. Tripathy, Y. Wang, and P. Ishwar. *Privacy-preserving adversarial networks*. In *57th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2019, Monticello, IL, USA, September 24-27, 2019*, pp. 495–505 (IEEE, 2019). URL <https://doi.org/10.1109/ALLERTON.2019.8919758>.
- [62] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal. *Context-aware generative adversarial privacy*. *Entropy* **19**(12) (2017).
- [63] P. Syverson. *Why i'm not an entropist*. In B. Christianson, J. A. Malcolm, V. Matyáš, and M. Roe, eds., *Security Protocols XVII*, pp. 213–230 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).
- [64] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. *Optimal geo-indistinguishable mechanisms for location privacy*. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS '14*, p. 251–262 (Association for Computing Machinery, New York, NY, USA, 2014). URL <https://doi.org/10.1145/2660267.2660345>.
- [65] K. Chatzikokolakis, C. Palamidessi, and M. Stronati. *Constructing elastic distinguishability metrics for location privacy*. *Proceedings on Privacy Enhancing Technologies* **2015**(2), 156 (2015). URL <https://doi.org/10.1515/popets-2015-0023>.
- [66] K. Chatzikokolakis, E. ElSalamouny, and C. Palamidessi. *Efficient utility improvement for location privacy*. *Proceedings on Privacy Enhancing Technologies* **2017**(4), 308 (2017). URL <https://doi.org/10.1515/popets-2017-0051>.

- [67] J.-W. Byun, T. Li, E. Bertino, N. Li, and Y. Sohn. *Privacy-preserving incremental data dissemination*. J. Comput. Secur. **17**(1), 43–68 (2009).
- [68] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li. *Secure anonymization for incremental datasets*. In W. Jonker and M. Petković, eds., *Secure Data Management*, pp. 48–63 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).
- [69] A. Anjum, G. Raschia, M. Gelgon, A. Khan, N. Ahmad, M. Ahmed, S. Suhail, M. M. Alam, et al. τ -safety: A privacy model for sequential publication with arbitrary updates. *computers & security* **66**, 20 (2017).
- [70] Y. Wang, X. Wu, and D. Hu. *Using randomized response for differential privacy preserving data collection*. In *EDBT/ICDT Workshops*, vol. 1558, pp. 0090–6778 (2016).
- [71] M. E. Gursoy, A. Tamersoy, S. Truex, W. Wei, and L. Liu. *Secure and utility-aware data collection with condensed local differential privacy*. IEEE Transactions on Dependable and Secure Computing **18**(5), 2365 (2021).
- [72] J. Leskovec and A. Krevl. *SNAP Datasets: Stanford large network dataset collection*. <http://snap.stanford.edu/data/loc-Gowalla.html> (2014).
- [73] E. Cho, S. A. Myers, and J. Leskovec. *Friendship and mobility: user movement in location-based social networks*. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090 (2011).
- [74] J. Nash. *Non-cooperative games*. Annals of Mathematics **54**(2), 286 (1951). URL <http://www.jstor.org/stable/1969529>.
- [75] E. Gal-or. *Hotelling’s spatial competition as a model of sales*. Economics Letters **9**(1), 1 (1982). URL <https://www.sciencedirect.com/science/article/pii/0165176582900891>.
- [76] R. Shokri, G. Theodorakopoulos, G. Danezis, J.-P. Hubaux, and J.-Y. Le Boudec. *Quantifying location privacy: The case of sporadic location exposure*. In S. Fischer-Hübner and N. Hopper, eds., *Privacy Enhancing Technologies*, pp. 57–76 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).
- [77] E. Ferrero, S. Alessandrini, and A. Balanzino. *Impact of the electric vehicles on the air pollution from a highway*. Applied Energy **169**, 450 (2016).
- [78] K. Hampshire, R. German, A. Pridmore, and J. Fons. *Electric vehicles from life cycle and circular economy perspectives*. Version **2**, 25 (2018).
- [79] R. Zhang and S. Fujimori. *The role of transport electrification in global climate change mitigation scenarios*. Environmental Research Letters **15**(3), 034019 (2020).
- [80] R. Hickman and D. Banister. *Looking over the horizon: Transport and reduced co2 emissions in the uk by 2030*. Transport Policy **14**(5), 377 (2007).

- [81] S. Kufeoglu and D. K. K. Hong. *Emissions performance of electric vehicles: A case study from the united kingdom*. *Applied Energy* **260**, 114241 (2020).
- [82] E. P. Seymour Millen. *Transport and environment statistics 2021 annual report* (May 2021). URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984685/transport-and-environment-statistics-2021.pdf.
- [83] T. Gersdorf, P. Hertzke, P. Schaufuss, and S. Schenk. *Mckinsey electric vehicle index: Europe cushions a global plunge in ev sales* (2020).
- [84] R. Hensley, S. Knupfer, and D. Pinner. *Electrifying cars: How three industries will evolve*. *McKinsey Quarterly* **3**(2009), 87 (2009).
- [85] P. Gao, R. Hensley, and A. Zielke. *A road map to the future for the auto industry*. *McKinsey Quarterly*, Oct pp. 1–11 (2014).
- [86] M. Lombardi, K. Panerali, S. Rousselet, and J. Scalise. *Electric vehicles for smarter cities: the future of energy and mobility*. In *World Economic Forum*. http://www3.weforum.org/docs/WEF_2018_20Electric_For_Smarter_Cities.pdf (2018).
- [87] A. Ostermann, Y. Fabel, K. Ouan, and H. Koo. *Forecasting charging point occupancy using supervised learning algorithms*. *Energies* **15**(9), 3409 (2022).
- [88] A. Sao, N. Tempelmeier, and E. Demidova. *Deep information fusion for electric vehicle charging station occupancy forecasting*. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 3328–3333 (2021).
- [89] L. Gillam, K. Katsaros, M. Dianati, and A. Mouzakitis. *Exploring edges for connected and autonomous driving*. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 148–153 (IEEE, 2018).
- [90] C. Maple, M. Bradbury, A. T. Le, and K. Ghirardello. *A connected and autonomous vehicle reference architecture for attack surface analysis*. *Applied Sciences* **9**(23), 5101 (2019).
- [91] D. Hahn, A. Munir, and V. Behzadan. *Security and privacy issues in intelligent transportation systems: Classification and challenges*. *IEEE Intelligent Transportation Systems Magazine* **13**(1), 181 (2019).
- [92] X. Lin and X. Li. *Achieving efficient cooperative message authentication in vehicular ad hoc networks*. *IEEE Transactions on Vehicular Technology* **62**(7), 3339 (2013).
- [93] N. Kumar, R. Iqbal, S. Misra, and J. J. Rodrigues. *An intelligent approach for building a secure decentralized public key infrastructure in vanet*. *Journal of Computer and System Sciences* **81**(6), 1042 (2015).
- [94] P. Zhao, G. Zhang, S. Wan, G. Liu, and T. Umer. *A survey of local differential privacy for securing internet of vehicles*. *The Journal of Supercomputing* **76**(11), 8391 (2020).

- [95] T. Franke and J. F. Krems. *Understanding charging behaviour of electric vehicle users*. Transportation Research Part F: Traffic Psychology and Behaviour **21**, 75 (2013).
- [96] R. R. Kumar and K. Alok. *Adoption of electric vehicle: A literature review and prospects for sustainability*. Journal of Cleaner Production **253**, 119911 (2020).
- [97] Z. Tian, T. Jung, Y. Wang, F. Zhang, L. Tu, C. Xu, C. Tian, and X.-Y. Li. *Real-time charging station recommendation system for electric-vehicle taxis*. IEEE Transactions on Intelligent Transportation Systems **17**(11), 3098 (2016).
- [98] W. Zhang, H. Liu, F. Wang, T. Xu, H. Xin, D. Dou, and H. Xiong. *Intelligent electric vehicle charging recommendation based on multi-agent reinforcement learning*. In *Proceedings of the Web Conference 2021*, pp. 1856–1867 (2021).
- [99] R. Flocea, A. Hîncu, A. Robu, S. Senocico, A. Traciu, B. M. Remus, M. S. Răboacă, and C. Filote. *Electric vehicle smart charging reservation algorithm*. Sensors **22**(8), 2834 (2022).
- [100] G. Wang, W. Li, J. Zhang, Y. Ge, Z. Fu, F. Zhang, Y. Wang, and D. Zhang. *sharedcharging: Data-driven shared charging for large-scale heterogeneous electric vehicle fleets*. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **3**(3), 1 (2019).
- [101] PlugShare. *Privacy policy* (2023). URL <https://company.plugshare.com/privacy.html>.
- [102] ChargePoint. *Privacy and cookie policy for europe* (2023). URL https://eu.chargepoint.com/privacy_policy.
- [103] C.-Y. Chow, M. F. Mokbel, and W. G. Aref. *Casper* query processing for location services without compromising privacy*. ACM Transactions on Database Systems (TODS) **34**(4), 1 (2009).
- [104] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li. *Achieving k-anonymity in privacy-aware location-based services*. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 754–762 (IEEE, 2014).
- [105] C. Dwork and A. Roth. *The algorithmic foundations of differential privacy*. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014). URL <https://doi.org/10.1561/04000000042>.
- [106] N. Guo, L. Ma, and T. Gao. *Independent mix zone for location privacy in vehicular networks*. IEEE Access **6**, 16842 (2018).
- [107] S. Amini, J. Lindqvist, J. I. Hong, M. Mou, R. Raheja, J. Lin, N. Sadeh, and E. Tochb. *Caché: Caching location-enhanced content to improve user privacy*. SIGMOBILE Mob. Comput. Commun. Rev. **14**(3), 19–21 (2011).
- [108] P. Asuquo, H. Cruickshank, J. Morley, C. P. A. Ogah, A. Lei, W. Hathal, S. Bao, and Z. Sun. *Security and privacy in location-based services for vehicular and mobile communications:*

- An overview, challenges, and countermeasures.* IEEE Internet of Things Journal **5**(6), 4778 (2018).
- [109] L. Zhou, L. Yu, S. Du, H. Zhu, and C. Chen. *Achieving differentially private location privacy in edge-assistant connected vehicles.* IEEE Internet of Things Journal **6**(3), 4472 (2018).
- [110] L. Luo, Z. Han, C. Xu, and G. Zhao. *A geo-indistinguishable location privacy preservation scheme for location-based services in vehicular networks.* In *International Conference on Algorithms and Architectures for Parallel Processing*, pp. 610–623 (Springer, 2019).
- [111] C. Qiu, A. C. Squicciarini, C. Pang, N. Wang, and B. Wu. *Location privacy protection in vehicle-based spatial crowdsourcing via geo-indistinguishability.* IEEE Transactions on Mobile Computing pp. 1–1 (2020).
- [112] T. Cunningham, G. Cormode, H. Ferhatosmanoglu, and D. Srivastava. *Real-world trajectory sharing with local differential privacy.* Proc. VLDB Endow. **14**(11), 2283–2295 (2021).
- [113] K. Chatzikokolakis, E. ElSalamouny, and C. Palamidessi. *Efficient utility improvement for location privacy.* Proceedings on Privacy Enhancing Technologies **2017**(4), 308 (2017).
- [114] A. Ghosh, T. Roughgarden, and M. Sundararajan. *Universally utility-maximizing privacy mechanisms* (2009). 0811.2841.
- [115] E. Bulut and M. C. Kisacikoglu. *Mitigating range anxiety via vehicle-to-vehicle social charging system.* In *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–5 (IEEE, 2017).
- [116] W. Duan, J. Gu, M. Wen, G. Zhang, Y. Ji, and S. Mumtaz. *Emerging technologies for 5g-iov networks: Applications, trends and opportunities.* IEEE Network **34**(5), 283 (2020).
- [117] H. Ji, O. Alfarraj, and A. Tolba. *Artificial intelligence-empowered edge of vehicles: Architecture, enabling technologies, and applications.* IEEE Access **8**, 61020 (2020).
- [118] H. Patil and V. N. Kalkhambkar. *Grid integration of electric vehicles for economic benefits: A review.* Journal of Modern Power Systems and Clean Energy **9**(1), 13 (2020).
- [119] L. Gillam, K. Katsaros, M. Dianati, and A. Mouzakitis. *Exploring edges for connected and autonomous driving.* In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 148–153 (2018).
- [120] H. Seo, K.-D. Lee, S. Yasukawa, Y. Peng, and P. Sartori. *Lte evolution for vehicle-to-everything services.* IEEE communications magazine **54**(6), 22 (2016).
- [121] A. Paverd, A. Martin, and I. Brown. *Modelling and automatically analysing privacy properties for honest-but-curious adversaries.* Tech. Rep (2014).
- [122] OpenStreetMap contributors. *Planet dump retrieved from <https://planet.osm.org> . <https://www.openstreetmap.org>* (2022).

- [123] A. F. D. C. U.S. Department of Energy. *Alternative fuels data center: Data downloads* (2023). URL https://afdc.energy.gov/data_download/.
- [124] M. T. Agency. *Datasf* (2023). URL https://data.sfgov.org/browse?Department-Metrics_Publishing-Department=Municipal+Transportation+Agency+.
- [125] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. *Crawdad data set epfl/mobility (v. 2009-02-24)* (2009).
- [126] X. Li, Y. Ren, L. T. Yang, N. Zhang, B. Luo, J. Weng, and X. Liu. *Perturbation-hidden: Enhancement of vehicular privacy for location-based services in internet of vehicles*. *IEEE Transactions on Network Science and Engineering* **8**(3), 2073 (2021).
- [127] L. Zhang, X. Meng, K.-K. R. Choo, Y. Zhang, and F. Dai. *Privacy-preserving cloud establishment and data dissemination scheme for vehicular cloud*. *IEEE Transactions on Dependable and Secure Computing* **17**(3), 634 (2018).
- [128] J. Freudiger, M. Raya, M. Félegyházi, P. Papadimitratos, and J.-P. Hubaux. *Mix-zones for location privacy in vehicular networks*. In *ACM Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS)*, CONF (2007).
- [129] 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 Comput. Surv.* **54**(1) (2021).
- [130] H. Zang and J. Bolot. *Anonymization of location data does not work: A large-scale measurement study*. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pp. 145–156 (2011).
- [131] C. Hecht, J. Figgenger, and D. U. Sauer. *Predicting electric vehicle charging station availability using ensemble machine learning*. *Energies* **14**(23), 7834 (2021).
- [132] A. Nait-Sidi-Moh, A. Ruzmetov, M. Bakhouya, Y. Naitmalek, and J. Gaber. *A prediction model of electric vehicle charging requests*. *Procedia Computer Science* **141**, 127 (2018).
- [133] T.-Y. Ma and S. Faye. *Multistep electric vehicle charging station occupancy prediction using hybrid lstm neural networks*. *Energy* **244**, 123217 (2022).
- [134] A. Almaghrebi, F. Aljuheshi, M. Rafaie, K. James, and M. Alahmad. *Data-driven charging demand prediction at public charging stations using supervised machine learning regression methods*. *Energies* **13**(16), 4231 (2020).
- [135] R. Luo, Y. Zhang, Y. Zhou, H. Chen, L. Yang, J. Yang, and R. Su. *Deep learning approach for long-term prediction of electric vehicle (ev) charging station availability*. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 3334–3339 (IEEE, 2021).

- [136] D. Le Métayer and S. J. De. *PRIAM: a Privacy Risk Analysis Methodology*. In G. Livraga, V. Torra, A. Aldini, F. Martinelli, and N. Suri, eds., *Data Privacy Management and Security Assurance* (Springer, Heraklion, Greece, 2016). URL <https://hal.inria.fr/hal-01420983>.
- [137] NIST. *Nist privacy framework core*. URL <https://www.nist.gov/system/files/documents/2021/05/05/NIST-Privacy-Framework-V1.0-Core-PDF.pdf>.
- [138] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran. *An efficient framework for clustered federated learning*. *Advances in Neural Information Processing Systems* **33**, 19586 (2020).
- [139] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh. *Three approaches for personalization with applications to federated learning*. *arXiv preprint arXiv:2002.10619* (2020).
- [140] F. Sattler, K.-R. Müller, and W. Samek. *Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints*. *IEEE transactions on neural networks and learning systems* **32**(8), 3710 (2020).
- [141] B. Hitaj, G. Ateniese, and F. Perez-Cruz. *Deep models under the gan: information leakage from collaborative deep learning*. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 603–618 (2017).
- [142] M. Nasr, R. Shokri, and A. Houmansadr. *Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning*. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753 (IEEE, 2019).
- [143] L. Zhu, Z. Liu, and S. Han. *Deep leakage from gradients*. *Advances in Neural Information Processing Systems* **32** (2019).
- [144] G. Andrew, O. Thakkar, B. McMahan, and S. Ramaswamy. *Differentially private learning with adaptive clipping*. *Advances in Neural Information Processing Systems* **34** (2021).
- [145] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. *Learning differentially private recurrent language models*. In *International Conference on Learning Representations* (2018). URL <https://openreview.net/forum?id=BJ0hF1Z0b>.
- [146] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei. *Ldp-fed: Federated learning with local differential privacy*. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, pp. 61–66 (2020).
- [147] Y. Zhao, J. Zhao, M. Yang, T. Wang, N. Wang, L. Lyu, D. Niyato, and K.-Y. Lam. *Local differential privacy-based federated learning for internet of things*. *IEEE Internet of Things Journal* **8**(11), 8836 (2020).
- [148] S. Biswas and C. Palamidessi. *Privic: A privacy-preserving method for incremental collection of location data* (2023). [2206.10525](#).
- [149] N. Fernandes, A. McIver, C. Palamidessi, and M. Ding. *Universal optimality and robust utility bounds for metric differential privacy*. In *2022 IEEE 35th Computer Security Foundations Symposium (CSF)*, pp. 348–363 (2022).

- [150] F. Galli, S. Biswas, K. Jung, T. Cucinotta, and C. Palamidessi. *Group privacy for personalized federated learning*. In *Proceedings of the 9th International Conference on Information Systems Security and Privacy - ICISSP*, pp. 252–263 (SciTePress - INSTICC, 2023).
- [151] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. *Fairness in criminal justice risk assessments: The state of the art*. *Sociological Methods & Research* **50**(1), 3 (2021).
- [152] A. Chouldechova. *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. *Big data* **5**(2), 153 (2017).
- [153] S. Agarwal. *Trade-offs between fairness and privacy in machine learning*. *ijcai 2021 workshop on ai for social good*. 2021 (2022).
- [154] S. Verma and J. Rubin. *Fairness definitions explained*. In *Proceedings of the international workshop on software fairness*, pp. 1–7 (2018).
- [155] R. Hanna and L. Linden. *Measuring discrimination in education*. Tech. rep., National Bureau of Economic Research (2009).
- [156] K. Makhoulf, S. Zhioua, and C. Palamidessi. *On the applicability of machine learning fairness notions*. *ACM SIGKDD Explorations Newsletter* **23**(1), 14 (2021).
- [157] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. *Fairness through awareness*. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226 (2012).
- [158] M. Hardt, E. Price, and N. Srebro. *Equality of opportunity in supervised learning*. *Advances in neural information processing systems* **29** (2016).
- [159] R. C. Geyer, T. Klein, and M. Nabi. *Differentially private federated learning: A client level perspective*. arXiv preprint arXiv:1712.07557 (2017).
- [160] K. A. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. *Practical secure aggregation for federated learning on user-held data*. In *NIPS Workshop on Private Multi-Party Machine Learning* (2016). URL <https://arxiv.org/abs/1611.04482>.
- [161] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan. *cpsgd: Communication-efficient and differentially-private distributed sgd*. *Advances in Neural Information Processing Systems* **31** (2018).
- [162] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong. *Personalized federated learning with differential privacy*. *IEEE Internet of Things Journal* **7**(10), 9530 (2020).
- [163] K. A. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. *Practical secure aggregation for federated learning on user-held data*. In *NIPS Workshop on Private Multi-Party Machine Learning* (2016). URL <https://arxiv.org/abs/1611.04482>.

- [164] S. Chhabra, Y. Solihin, R. Lal, and M. Hoekstra. *An analysis of secure processor architectures*. Transactions on computational science VII pp. 101–121 (2010).
- [165] T. Cucinotta, D. Cherubini, and E. Jul. *Confidential execution of cloud services*. In CLOSER, pp. 616–621 (2014).
- [166] A. Chhabra, K. Masalkovaitė, and P. Mohapatra. *An overview of fairness in clustering*. IEEE Access **9**, 130698 (2021).
- [167] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and S. Avestimehr. *Fairfed: Enabling group fairness in federated learning*. In 1st NeurIPS Workshop on New Frontiers in Federated Learning (2021). URL <https://arxiv.org/abs/1611.04482>.
- [168] L. Chu, L. Wang, Y. Dong, J. Pei, Z. Zhou, and Y. Zhang. *Fedfair: Training fair models in cross-silo federated learning*. arXiv preprint arXiv:2109.05662 (2021).
- [169] A. K. Menon and R. C. Williamson. *The cost of fairness in binary classification*. In Conference on Fairness, accountability and transparency, pp. 107–118 (PMLR, 2018).
- [170] M. Wick, J.-B. Tristan, et al. *Unlocking fairness: a trade-off revisited*. Advances in neural information processing systems **32** (2019).
- [171] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. *A survey on bias and fairness in machine learning*. ACM Computing Surveys (CSUR) **54**(6), 1 (2021).
- [172] S. Biswas and H. Rajan. *Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline*. In Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 981–993 (2021).
- [173] F. Kamiran and T. Calders. *Data preprocessing techniques for classification without discrimination*. Knowledge and information systems **33**(1), 1 (2012).
- [174] M. Wan, D. Zha, N. Liu, and N. Zou. *In-processing modeling techniques for machine learning fairness: A survey*. ACM Transactions on Knowledge Discovery from Data **17**(3), 1 (2023).
- [175] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. *Fairness without demographics in repeated loss minimization*. In International Conference on Machine Learning, pp. 1929–1938 (PMLR, 2018).
- [176] F. Petersen, D. Mukherjee, Y. Sun, and M. Yurochkin. *Post-processing for individual fairness*. Advances in Neural Information Processing Systems **34**, 25944 (2021).
- [177] A. Noriega-Campero, M. A. Bakker, B. Garcia-Bulle, and A. Pentland. *Active fairness in algorithmic decision making*. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 77–83 (2019).

- [178] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. *On the compatibility of privacy and fairness*. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 309–315 (2019).
- [179] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. *Geo-indistinguishability: Differential privacy for location-based systems*. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 901–914 (2013).
- [180] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon. *Federated learning: Strategies for improving communication efficiency*. In *NIPS Workshop on Private Multi-Party Machine Learning* (2016). URL <https://arxiv.org/abs/1610.05492>.
- [181] CMMS. *Centers for medicare and medicaid services* (2021). Accessed: 2022-09-21, URL <https://www.cms.gov/mmrr/News/mmrr-news-2013-03-hosp-chg-data.html>.
- [182] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar. *Leaf: A benchmark for federated settings*. Workshop on Federated Learning for Data Privacy and Confidentiality (2019).
- [183] R. Bassily, K. Nissim, U. Stemmer, and A. Guha Thakurta. *Practical locally private heavy hitters*. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., Red Hook, NY, USA, 2017). URL <https://proceedings.neurips.cc/paper/2017/file/3d779cae2d46cf6a8a99a35ba4167977-Paper.pdf>.
- [184] R. Bartlett, A. Morse, R. Stanton, and N. Wallace. *Consumer-lending discrimination in the fintech era*. *Journal of Financial Economics* **143**(1), 30 (2022).
- [185] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. *What can we learn privately?* *SIAM Journal on Computing* **40**(3), 793 (2011). <https://doi.org/10.1137/090756090>, URL <https://doi.org/10.1137/090756090>.
- [186] P. Kairouz, S. Oh, and P. Viswanath. *Extremal mechanisms for local differential privacy*. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds., *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Inc., 2014). URL https://proceedings.neurips.cc/paper_files/paper/2014/file/86df7dcfd896fcdf2674f757a2463eba-Paper.pdf.
- [187] J. Ullman. *Tight lower bounds for locally differentially private selection*. arXiv preprint arXiv:1802.02638 (2018).
- [188] R. Bassily, K. Nissim, U. Stemmer, and A. Guha Thakurta. *Practical locally private heavy hitters*. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017). URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3d779cae2d46cf6a8a99a35ba4167977-Paper.pdf.

- [189] F. Galli., S. Biswas., K. Jung., T. Cucinotta., and C. Palamidessi. *Group privacy for personalized federated learning*. In *Proceedings of the 9th International Conference on Information Systems Security and Privacy - ICISSP*, pp. 252–263. INSTICC (SciTePress, 2023).
- [190] J. Zhu and M. Blaschko. *R-gap: Recursive gradient attack on privacy* (2021). [2010.07733](https://arxiv.org/abs/2010.07733).
- [191] M. Lam, G.-Y. Wei, D. Brooks, V. J. Reddi, and M. Mitzenmacher. *Gradient disaggregation: Breaking privacy in federated learning by reconstructing the user participant matrix* (2021). [2106.06089](https://arxiv.org/abs/2106.06089).
- [192] C. E. Shannon. *A mathematical theory of communication*. The Bell System Technical Journal **27**(3), 379 (1948).
- [193] N. Fernandes. *Differential privacy for metric spaces : information-theoretic models for privacy and utility with new applications to metric domains*. Theses, Institut Polytechnique de Paris ; Macquarie University (Sydney, Australie) (2021). URL <https://theses.hal.science/tel-03344453>.
- [194] Liveen - blockchain-based social network platform that provides fair rewards for the users' contents. <https://www.liveen.com/>. (Accessed on 05/26/2021).
- [195] Datacoup - reclaim your personal data. <https://datacoup.com/>. (Accessed on 09/02/2021).
- [196] C. M. Bowen and J. Snok. *Comparative study of differentially private synthetic data algorithms from the nist pscr differential privacy synthetic data challenge*. arXiv preprint arXiv:1911.12704 (2019).
- [197] N. Volgushev, M. Schwarzkopf, B. Getchell, M. Varia, A. Lapets, and A. Bestavros. *Conclave: secure multi-party computation on big data*. In *Proceedings of the Fourteenth EuroSys Conference 2019*, pp. 1–18 (2019).
- [198] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti. *A survey on homomorphic encryption schemes: Theory and implementation*. ACM Computing Surveys (CSUR) **51**(4), 1 (2018).
- [199] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. *Privacy loss in apple's implementation of differential privacy on macos 10.12*. arXiv preprint arXiv:1709.02753 (2017).
- [200] J. Lee and C. Clifton. *How much is enough? choosing ϵ for differential privacy*. In *International Conference on Information Security*, pp. 325–340 (Springer, 2011).
- [201] Y. Chen, S. Chong, I. A. Kash, T. Moran, and S. Vadhan. *Truthful mechanisms for agents that value privacy*. ACM Transactions on Economics and Computation (TEAC) **4**(3), 1 (2016).
- [202] K. Ligett and A. Roth. *Take it or leave it: Running a survey when privacy comes at a cost*. In *International workshop on internet and network economics*, pp. 378–391 (Springer, 2012).

- [203] D. Xiao. *Is privacy compatible with truthfulness?* In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pp. 67–86 (2013).
- [204] K. Nissim, C. Orlandi, and R. Smorodinsky. *Privacy-aware mechanism design*. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 774–789 (2012).
- [205] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth. *Differential privacy: An economic method for choosing epsilon*. In *2014 IEEE 27th Computer Security Foundations Symposium*, pp. 398–410 (IEEE, 2014).
- [206] A. Ghosh and A. Roth. *Selling privacy at auction*. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 199–208 (2011).
- [207] P. Dandekar, N. Fawaz, and S. Ioannidis. *Privacy auctions for recommender systems*. *ACM Transactions on Economics and Computation (TEAC)* **2**(3), 1 (2014).
- [208] A. Roth. *Buying private data at auction: the sensitive surveyor’s problem*. *ACM SIGecom Exchanges* **11**(1), 1 (2012).
- [209] L. K. Fleischer and Y.-H. Lyu. *Approximately optimal auctions for selling privacy when costs are correlated with data*. In *Proceedings of the 13th ACM conference on electronic commerce*, pp. 568–585 (2012).
- [210] W. Li, C. Zhang, Z. Liu, and Y. Tanaka. *Incentive mechanism design for crowdsourcing-based indoor localization*. *IEEE Access* **6**, 54042 (2018).
- [211] R. Nget, Y. Cao, and M. Yoshikawa. *How to balance privacy and money through pricing mechanism in personal data market*. *arXiv preprint arXiv:1705.02982* (2017).
- [212] H. Oh, S. Park, G. M. Lee, H. Heo, and J. K. Choi. *Personal data trading scheme for data brokers in iot data marketplaces*. *IEEE Access* **7**, 40120 (2019).
- [213] C. Li, D. Y. Li, G. Miklau, and D. Suciu. *A theory of pricing private data*. *ACM Transactions on Database Systems (TODS)* **39**(4), 1 (2014).
- [214] C. Aperia and B. A. Huberman. *A market for unbiased private data: Paying individuals according to their privacy attitudes*. Available at SSRN 2046861 (2012).
- [215] K. Jung and S. Park. *Privacy bargaining with fairness: Privacy-price negotiation system for applying differential privacy in data market environments*. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1389–1394 (IEEE, 2019).
- [216] S. Krehbiel. *Choosing epsilon for privacy as a service*. *Proc. Priv. Enhancing Technol.* **2019**(1), 192 (2019).
- [217] T. Zhang and Q. Zhu. *On the differential private data market: Endogenous evolution, dynamic pricing, and incentive compatibility*. *arXiv preprint arXiv:2101.04357* (2021).

- [218] Z. Jorgensen, T. Yu, and G. Cormode. *Conservative or liberal? personalized differential privacy*. In *2015 IEEE 31st international conference on data engineering*, pp. 1023–1034 (IEEE, 2015).
- [219] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. *What can we learn privately?* SIAM Journal on Computing **40**(3), 793 (2011).
- [220] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. *Privacy loss in apple’s implementation of differential privacy on macos 10.12*. arXiv preprint arXiv:1709.02753 (2017).
- [221] J. Lee and C. Clifton. *How much is enough? choosing ϵ for differential privacy*. In *International Conference on Information Security*, pp. 325–340 (Springer, 2011).
- [222] J. Domingo-Ferrer and J. Soria-Comas. *From t -closeness to differential privacy and vice versa in data anonymization*. Knowledge-Based Systems **74**, 151 (2015).
- [223] N. Holohan, S. Antonatos, S. Braghin, and P. Mac Aonghusa. *(k, ϵ) -anonymity: k - anonymity with ϵ -differential privacy*. arXiv preprint arXiv:1710.01615 (2017).
- [224] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth. *Differential privacy: An economic method for choosing epsilon*. In *2014 IEEE 27th Computer Security Foundations Symposium*, pp. 398–410 (IEEE, 2014).
- [225] A. Ghosh and A. Roth. *Selling privacy at auction*. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 199–208 (2011).
- [226] A. Roth. *Buying private data at auction: the sensitive surveyor’s problem*. ACM SIGecom Exchanges **11**(1), 1 (2012).
- [227] L. K. Fleischer and Y.-H. Lyu. *Approximately optimal auctions for selling privacy when costs are correlated with data*. In *Proceedings of the 13th ACM conference on electronic commerce*, pp. 568–585 (2012).
- [228] R. Nget, Y. Cao, and M. Yoshikawa. *How to balance privacy and money through pricing mechanism in personal data market*. arXiv preprint arXiv:1705.02982 (2017).
- [229] C. Li, D. Y. Li, G. Miklau, and D. Suciu. *A theory of pricing private data*. ACM Transactions on Database Systems (TODS) **39**(4), 1 (2014).
- [230] T. Zhang and Q. Zhu. *On the differential private data market: Endogenous evolution, dynamic pricing, and incentive compatibility*. arXiv preprint arXiv:2101.04357 (2021).
- [231] K. Jung and S. Park. *Privacy bargaining with fairness: Privacy-price negotiation system for applying differential privacy in data market environments*. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1389–1394 (IEEE, 2019).
- [232] E. Winter. *The shapley value*. Handbook of game theory with economic applications **3**, 2025 (2002).

- [233] A. E. Roth. *The Shapley value: essays in honor of Lloyd S. Shapley* (Cambridge University Press, 1988).
- [234] E. ElSalamouny and C. Palamidessi. *Reconstruction of the distribution of sensitive data under free-will privacy* (2022). [2208.11268](#).
- [235] Y. Zhu. *How niantic is profiting off tracking where you go while playing “pokémon go”* (2016). URL <https://www.forbes.com/sites/yehongzhu/2016/07/29/how-niantic-is-profiting-off-tracking-where-you-go-while-playing-pokemon-go/>.
- [236] S. Dredge. *Tinder dating app was sharing more of users’ location data than they realised* (2014). URL <https://www.theguardian.com/technology/2014/feb/20/tinder-app-dating-data-location-sharing>.
- [237] K. Chatzikokolakis, C. Palamidessi, and M. Stronati. *A predictive differentially-private mechanism for mobility traces*. In E. De Cristofaro and S. J. Murdoch, eds., *Privacy Enhancing Technologies*, pp. 21–41 (Springer International Publishing, Cham, 2014).
- [238] D. Haussler. *Convolution kernels on discrete structures ucsc crl* (1999).
- [239] T. Hofmann, B. Schölkopf, and A. J. Smola. *Kernel methods in machine learning*. The Annals of Statistics **36**(3) (2008). URL <https://doi.org/10.1214%2F009053607000000677>.
- [240] I. J. Schoenberg. *Metric spaces and completely monotone functions*. Annals of Mathematics **39**(4), 811 (1938). URL <http://www.jstor.org/stable/1968466>.
- [241] S. Bernstein. *Sur les fonctions absolument monotones*. Acta Mathematica **52**(none), 1 (1929). URL <https://doi.org/10.1007/BF02592679>.
- [242] D. Widder. *The laplace transform, vol. 6 of*. Princeton Mathematical Series (1941).
- [243] H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics (Cambridge University Press, 2004).
- [244] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. *Emnist: Extending mnist to handwritten letters*. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926 (IEEE, 2017).
- [245] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. *An empirical investigation of catastrophic forgetting in gradient-based neural networks*. arXiv preprint arXiv:1312.6211 (2013).
- [246] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. *Overcoming catastrophic forgetting in neural networks*. Proceedings of the national academy of sciences **114**(13), 3521 (2017).
- [247] D. Lopez-Paz and M. Ranzato. *Gradient episodic memory for continual learning*. Advances in neural information processing systems **30** (2017).

-
- [248] Z. Zhang and M. Sabuncu. *Generalized cross entropy loss for training deep neural networks with noisy labels*. Advances in neural information processing systems **31** (2018).
- [249] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth. *On the lambert w function*. Advances in Computational mathematics **5**, 329 (1996).

Appendices



Proofs from Chapter 3

THEOREM 3.1. The privacy mechanism generated by BA produces an elastic location-privacy mechanism.

Proof. Let $x, y \in \mathcal{X}$ be any true and reported location, respectively. Letting \hat{C}_{BA} to be the limiting mechanism generated by BA, to show that \hat{C}_{BA} possesses an elastic distinguishability metric, we need to ensure that:

1. The probability of reporting y to obfuscate x given by \hat{C}_{BA} should be exponentially reducing w.r.t. the Euclidean distance between x and y , staying consistent with the essence of geo-ind (the property captured by (3.4)).
2. Under \hat{C}_{BA} , the probability of reporting y to obfuscate x should be taking into account the mass of reported points around y , i.e., the more geo-spatially isolated (from other reported points) y is in the space, the less likely it should be to report it, as, ideally, we would like to have x being reported as a location amidst a crowd of other reported locations (the property captured by (3.5)).

Let's simplify the notation and denote $\mathbb{P}[\hat{C}_{BA}(x) = y]$ as $\mathbb{P}_{BA}[y|x]$ and let $q(y)$ be the probability mass of the observed location y . Hence, for being an elastic location-privacy mechanism, \hat{C}_{BA} should satisfy (3.4) and (3.5), i.e., we must have:

$$\mathbb{P}_{BA}[y|x] \propto \exp\{-\beta d_E(x, y)\} \tag{A.1}$$

$$\mathbb{P}_{BA}[y|x] \propto q(y) \tag{A.2}$$

Therefore, in order to satisfy (A.1) and (A.2), it is sufficient to have:

$$\begin{aligned}\mathbb{P}_{\text{BA}}[y|x] &\propto \exp\{-\beta d_E(x, y) + \ln q(y)\} \\ \Rightarrow \mathbb{P}_{\text{BA}}[y|x] &\propto q(y) \exp\{-\beta d_E(x, y)\} \\ &= \frac{q(y) \exp\{\beta d_E(x, y)\}}{\sum_{z \in \mathcal{X}} q(z) \exp\{-\beta d_E(x, z)\}}\end{aligned}\quad (\text{A.3})$$

Now, it's sufficient to note that, if we interpret the mass of y as the probability of being reported by the mechanism, (A.3) is exactly the fixpoint of $\mathcal{G} \circ \mathcal{F}$, cf. Remarks 3.2.1 and 3.2.2. \square

THEOREM 3.2. For any $t \geq 1$, the mechanism generated by BA over \mathcal{X} at the t 'th iteration, seen as a stochastic matrix, is invertible.

Proof. For notational convenience, in this proof, we shall denote the Euclidean distance $d_E(\cdot)$ as $d(\cdot)$. For any $t \geq 1$, let $C^{(t)}$ be the channel generated at the t 'th iteration of BA. Hence, we have:

$$C_{x,y}^{(t)} = \frac{c_{t-1}(y) \exp\{-\beta d(x, y)\}}{\sum_{z \in \mathcal{Y}} c_t(z) \exp\{-\beta d(x, z)\}} \quad (\text{A.4})$$

Let $C' \in \mathcal{C}(\mathcal{X}, \mathcal{X})$ such that $C'_{x,y} = \exp\{-\beta d(x, y)\}$. Correspondingly, let us define $C''^{(t)}, C'''^{(t)} \in \mathcal{C}(\mathcal{X}, \mathcal{X})$ s.t. $C_{x,y}''^{(t)} = c_{t-1}(y) C'_{x,y}$ and $C_{x,y}'''^{(t)} = K_x C_{x,y}''^{(t)}$ where $K_x = (\sum_{z \in \mathcal{Y}} c_t(z) \exp\{-\beta d(x, z)\})^{-1}$. Therefore, we have $C'''^{(t)} = C^{(t)}$.

Exploiting the fact that scaling of rows and columns of matrices by real numbers (elementary operations on rows and columns) preserves their linear independence, we ensure that if C' is invertible, then so is $C''^{(t)}$ (elementary column operation on C') which, in turn, implies that $C'''^{(t)} = C^{(t)}$ is invertible (elementary row operation on $C''^{(t)}$). Therefore, in order to show $C^{(t)}$ is invertible, it is sufficient to prove that C' is invertible.

Note that $\exp\{-\beta d(x, y)^2\} = \exp\{-\beta \|x - y\|_2^2\}$ is the *Gaussian kernel* for any $\beta > 0$ and is positive definite [238, 239]. Furthermore, Schoenberg [240] observed that for any *completely monotone function* $g: \mathbb{R}_{\geq 0} \mapsto \mathbb{R}$, we can use *Hausdorff–Bernstein–Widder theorem* [241, 242] to deduce that *radial basis function (RBF) kernels* such as $\exp\{-\beta g(\|x - y\|_2^2)\}$ are also positive definite. Moreover, in addition to being positive definite, it was also shown that Gaussian kernels are *strictly positive definite* [239, 243].

Let $f: \mathbb{R}_{\geq 0} \mapsto \mathbb{R}$ be the *square-root function*, i.e., $f(x) = \sqrt{x}$ for all $x \in \mathbb{R}_{\geq 0}$. Therefore, observing that f is completely monotone and recalling that Gaussian kernels are strictly positive definite, i.e., $z^T \exp\{-\beta \|x_i - x_j\|_2^2\} z \geq 0$ for every $z \in \mathbb{R}^m$ with equality holding iff $z = \mathbf{0}$, we can use *Schoenberg theorem* [240] to show the strict positive definiteness of $\exp\{-\beta f(d(x, y)^2)\} = \exp\{-\beta d(x, y)\}$. Hence, noting that C' is the Gram matrix of the RBF kernel $\exp\{-\beta d(x, y)\}$, we must have $z^T C' z \geq 0$ for every $z \in \mathbb{R}^m$ with equality holding iff $z = \mathbf{0}$. This implies that C' is positive definite and, hence, invertible. Therefore, in turn, $C^{(t)}$ is invertible. \square

THEOREM 3.3. PRIVIC converges to the unique MLE of the true distribution.

Proof. For $1 \leq t \leq N$ and $1 \leq i \leq n$, in the t 'th round of PRIVIC, the i 'th sampled user locally sanitize their location $x_i^{(t)}$ with $\hat{C}^{(t)}$ and reports the noisy location $y_i^{(t)}$. Therefore, the *combined mechanism* (referred to as *output probability matrix* in [46]) for implementing GIBU is $\mathcal{G} = \begin{pmatrix} \hat{C}^{(1)} & \dots & \hat{C}^{(N)} \end{pmatrix}$ s.t.

$$\mathcal{G} \left(x, y_i^{(t)} \right) = \mathbb{P} \left[y_i^{(t)} \middle| x \right] = \hat{C}^{(t)} \left(x, y_i^{(t)} \right) \\ \forall x \in \mathcal{X}, i \in \{1, \dots, n\}.$$

By Theorem 3.2, $\hat{C}^{(t)}$ is invertible for every $t \geq 1$ and let $\hat{C}^{(t)-1}$ denote the corresponding inverse of $\hat{C}^{(t)}$. Therefore, defining \mathcal{G}' s.t $\mathcal{G}' = \frac{1}{N} \begin{pmatrix} \hat{C}^{(1)-1} & \dots & \hat{C}^{(N)-1} \end{pmatrix}^T$ ensures that $\mathcal{G} \cdot \mathcal{G}' = \mathbb{I}_m$ where $m = |\mathcal{X}|$. Therefore, \mathcal{G} is right-invertible.

Hence, combining the right-invertibility of \mathcal{G} with Theorem 3 (GIBU converges to MLEs) and Corollary 1 (right-invertibility of the combined channel of GIBU implies unique MLE) of [46], we can conclude that GIBU $\left(\left(\hat{C}^{(1)}, \mathbf{y}^{(1)} \right), \dots, \left(\hat{C}^{(N)}, \mathbf{y}^{(N)} \right) \right)$ estimates the unique MLE of the prior $\pi_{\mathcal{X}}$, implying that PRIVIC converges. \square

B

Tables from Chapter 3

Table B.1: EMD between the true and the estimated PMFs by PRIVIC on the Paris locations.

N	$\beta = 1$					$\beta = 0.5$				
	Round 1	Round 2	Round 3	Round 4	Round 5	Round 1	Round 2	Round 3	Round 4	Round 5
1	2.02262	2.02262	2.02262	2.02262	2.02262	2.02262	2.02262	2.02262	2.02262	2.02262
2	0.27104	0.27796	0.28247	0.27758	0.26276	0.57738	0.57994	0.55791	0.60717	0.56880
3	0.21916	0.23750	0.25035	0.25116	0.23241	0.51324	0.48295	0.47043	0.51285	0.45826
4	0.19156	0.21115	0.21726	0.20913	0.20408	0.43184	0.42398	0.41040	0.45230	0.41119
5	0.18241	0.19264	0.19570	0.19747	0.18728	0.39741	0.38771	0.37284	0.41176	0.36953
6	0.16526	0.18020	0.174578	0.17310	0.17268	0.37818	0.35482	0.36039	0.36375	0.38045
7	0.14643	0.18159	0.16092	0.17222	0.17139	0.35044	0.34383	0.34818	0.35760	0.36804
8	0.13860	0.17605	0.15938	0.17078	0.16192	0.34086	0.32983	0.35769	0.34430	0.34780
9	0.15047	0.16926	0.15153	0.17005	0.15266	0.33137	0.31749	0.34690	0.33840	0.34028
10	0.14734	0.14825	0.14585	0.15001	0.15459	0.3170	0.32975	0.32772	0.34529	0.32670
11	0.14227	0.14135	0.14326	0.13797	0.14507	0.31917	0.31851	0.32689	0.33667	0.32043
12	0.13818	0.14448	0.14703	0.13589	0.14142	0.32137	0.32451	0.31843	0.32556	0.34014
13	0.16111	0.13641	0.14893	0.14208	0.13808	0.31224	0.31219	0.31380	0.31515	0.31159
14	0.13894	0.13111	0.14094	0.14199	0.14192	0.30883	0.30496	0.30578	0.30726	0.30282
15	0.15106	0.14271	0.14601	0.13584	0.12413	0.29405	0.31198	0.30786	0.31167	0.30100

Table B.2: EMD between the true and the estimated PMFs by PRIVIC on the San Francisco locations.

N	$\beta = 1$					$\beta = 0.5$				
	Round 1	Round 2	Round 3	Round 4	Round 5	Round 1	Round 2	Round 3	Round 4	Round 5
1	7.37595	7.37595	7.37595	7.37595	7.37595	7.37595	7.37595	7.37595	7.37595	7.37595
2	0.37229	0.37038	0.36784	0.36949	0.36816	0.79621	0.80474	0.80219	0.80670	0.79940
3	0.29828	0.298370	0.30017	0.29784	0.29859	0.64931	0.66292	0.65362	0.65950	0.65285
4	0.26091	0.26029	0.26231	0.26180	0.26518	0.56896	0.57378	0.57338	0.57125	0.56618
5	0.23472	0.23419	0.23337	0.23740	0.23897	0.51672	0.51710	0.51735	0.51880	0.51138
6	0.21367	0.21432	0.21537	0.21777	0.21881	0.48194	0.48299	0.47992	0.48267	0.47791
7	0.19612	0.19761	0.19904	0.20120	0.20067	0.45531	0.45861	0.45122	0.45732	0.45450
8	0.18244	0.18412	0.18741	0.18724	0.18674	0.43588	0.43745	0.43558	0.43704	0.43584

C

Proofs from Chapter 4

THEOREM 4.1. [Compositionality Theorem for AGeoI] Let mechanisms \mathcal{K}_1 and \mathcal{K}_2 be $(\varepsilon_1, \delta_1)$ and $(\varepsilon_2, \delta_2)$ geo-indistinguishable, respectively. Then their composition is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$ -geo-indistinguishable. In other words, for every $S_1, S_2 \subseteq \mathcal{Y}$ and all $x_1, x'_1, x_2, x'_2 \in \mathcal{X}$:

$$\mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x_1, x_2)] \leq e^{\varepsilon_1 d(x_1, x'_1) + \varepsilon_2 d(x_2, x'_2)} \mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x'_1, x'_2)] + \left(\delta_1 e^{\varepsilon_1 d(x_1, x'_1)} + \delta_2 e^{\varepsilon_2 d(x_2, x'_2)} \right)$$

Proof. Let us simplify the notation and denote:

$$P_i = \mathbb{P}_{\mathcal{K}_i} [y_i \in S_i | x_i]$$

$$P'_i = \mathbb{P}_{\mathcal{K}_i} [y_i \in S_i | x'_i]$$

$$\tilde{\delta}_i = \delta_i e^{\varepsilon_i d(x_i, x'_i)}$$

for $i \in \{1, 2\}$. As mechanisms \mathcal{K}_1 and \mathcal{K}_2 are applied independently, we have:

$$\mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x_1, x_2)] = P_1 P_2 \tag{C.1}$$

$$\mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x'_1, x'_2)] = P'_1 P'_2 \tag{C.2}$$

Therefore, we obtain:

$$\mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x_1, x_2)] = P_1 P_2$$

$$\begin{aligned}
&\leq \left(\min \left(1 - \tilde{\delta}_1, e^{\varepsilon_1 d(x_1, x'_1)} P'_1 \right) + \tilde{\delta}_1 \right) \left(\min \left(1 - \tilde{\delta}_2, e^{\varepsilon_2 d(x_2, x'_2)} P'_2 \right) + \tilde{\delta}_2 \right) \\
&\leq m_1 m_2 + \tilde{\delta}_1 m_2 + m_1 \tilde{\delta}_2 + \tilde{\delta}_1 \tilde{\delta}_2 \\
&\left[\text{where } m_i = \min \left(1 - \tilde{\delta}_i, e^{\varepsilon_i d(x_i, x'_i)} P'_i \right) \right] \\
&\leq e^{\varepsilon_1 d(x_1, x'_1) + \varepsilon_2 d(x_2, x'_2)} P'_1 P'_2 \\
&+ \tilde{\delta}_1 - \tilde{\delta}_1 \tilde{\delta}_2 + \tilde{\delta}_2 - \tilde{\delta}_1 \tilde{\delta}_2 + \tilde{\delta}_1 \tilde{\delta}_2 \\
&\leq e^{\varepsilon_1 d(x_1, x'_1) + \varepsilon_2 d(x_2, x'_2)} \mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} \left[(y_1, y_2) \in S_1 \times S_2 | (x'_1, x'_2) \right] + \left(\delta_1 e^{\varepsilon_1 d(x_1, x'_1)} + \delta_2 e^{\varepsilon_2 d(x_2, x'_2)} \right)
\end{aligned}$$

□

LEMMA 4.2. For every $x_1, x_2 \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have $e^{-\varepsilon d(x_1, x_2)} \mathbb{P}_{\mathcal{L}} [y|x_1] - \mathbb{P}_{\mathcal{L}} [y|x_2] \leq 1$.

Proof.

$$\begin{aligned}
&e^{-\varepsilon d(x_1, x_2)} \mathbb{P} [y|x_1] - \mathbb{P} [y|x_2] \leq 1 \\
&\iff c \left(e^{-\varepsilon(d(x_1, x_2) + d(x_1, y))} - e^{-\varepsilon d(x_2, y)} \right) \leq 1
\end{aligned} \tag{C.3}$$

Now we observe that $d(x_1, x_2) + d(x_1, y) \geq d(x_2, y)$ due to the fact that d is a metric and it satisfies the triangle inequality. Immediately, we have $e^{-\varepsilon d(x_1, x_2) + d(x_1, y)} - e^{-\varepsilon d(x_2, y)} \leq 0$ for any $\varepsilon \in \mathbb{R}_{\geq 0}$. Therefore, as $c \geq 0$, (C.3) is trivially satisfied. □



Proofs from Chapter 5

PROPOSITION 5.1. Abusing the notation $\mathbb{P}[\cdot]$ to denote the PDF, let $\mathcal{L}_\varepsilon: \mathbb{R}^n \mapsto \mathbb{R}^n$ be the Laplace mechanism with distribution $\mathcal{L}_{x_0, \varepsilon}(x) = \mathbb{P}[\mathcal{L}_\varepsilon(x_0) = x] = K e^{-\varepsilon d(x, x_0)}$ with d being the Euclidean distance. If $\rho \sim \mathcal{L}_{x_0, \varepsilon}(x)$, then:

1. $\mathcal{L}_{x_0, \varepsilon}$ is ε - d -private and $K = \frac{\varepsilon^n \Gamma(\frac{n}{2})}{2\pi^{\frac{n}{2}} \Gamma(n)}$
2. $\|\rho\|_2 \sim \gamma_{\varepsilon, n}$ s.t. $\mathbb{P}[\|\rho\|_2 = r] = \frac{\varepsilon^n e^{-\varepsilon r} r^{n-1}}{\Gamma(n)}$
3. The i^{th} component of ρ has variance $\sigma_{\rho_i}^2 = \frac{n+1}{\varepsilon^2}$

where $\Gamma(n)$ is the Gamma function defined for positive reals as $\int_0^\infty t^{n-1} e^{-t} dt$ which reduces to the factorial function whenever $n \in \mathbb{N}$.

Proof. We provide proof of the three statements separately:

1. If $\mathcal{L}_{x_0, \varepsilon}(x) = K e^{-\varepsilon d(x, x_0)}$ is a probability density function of a point $x \in \mathbb{R}^n$ then K should be such that $\int_{\mathbb{R}^n} \mathcal{L}_{x_0}(x) dx = 1$. We note that it depends only on the distance x and x_0 and we can write $K e^{-\varepsilon d(x, x_0)}$ as $K e^{-\varepsilon r}$ where r is the radius of the ball in \mathbb{R}^n centered in x_0 . Without loss of generality, let us now take $x_0 = 0$. The probability density of the event $x \in \mathbb{S}_n(r) = \{x : \|x\|_2 = r\}$ is then $p(x \in \mathbb{S}_n(r)) = K e^{-\varepsilon r} \mathbb{S}_n(1) r^{n-1}$ where $\mathbb{S}_n(1)$ is the surface of the unitary ball in \mathbb{R}^n and $\mathbb{S}_n(r) = \mathbb{S}_n(1) r^{n-1}$ is the surface of a generic ball of radius r . Given that

$$\mathbb{S}_n(1) = \frac{2\pi^{n/2}}{\Gamma(\frac{n}{2})} \tag{D.1}$$

solving

$$\int_0^{+\infty} \mathbb{P}[x \in \mathbb{S}_n(r)] dr = \int_0^{+\infty} K e^{-\varepsilon r} S_n(1) r^{n-1} dr = K \frac{2\pi^{n/2} \Gamma(n)}{\varepsilon^n \Gamma(\frac{n}{2})} = 1 \quad (\text{D.2})$$

results in

$$K = \frac{\varepsilon^n \Gamma(\frac{n}{2})}{2\pi^{\frac{n}{2}} \Gamma(n)} \quad (\text{D.3})$$

where $\Gamma(\cdot)$ denotes the gamma function. By plugging $\mathcal{L}_{x_0, \varepsilon}(x) = K e^{-\varepsilon d(x, x_0)}$ in Equation 2.3:

$$K e^{-\varepsilon d(x, x_1)} \leq e^{\varepsilon d(x_1, x_2)} K e^{-\varepsilon d(x, x_2)} \quad (\text{D.4})$$

$$e^{\varepsilon(\|x-x_2\|_2 - \|x-x_1\|_2)} \leq e^{\varepsilon\|x_1-x_2\|} = e^{\varepsilon d(x_1, x_2)} \quad (\text{D.5})$$

which completes the poof of the first statement.

2. Without loss of generality, let us take $x_0 = 0$. Exploiting the radial symmetry of the Laplace distribution, we note that, in order to sample a point $\rho \sim \mathcal{L}_\varepsilon(x_0)$ in \mathbb{R}^n , it is possible to first sample the set of points distance r from x_0 and then sample uniformly from the resulting hypersphere. Accordingly, the p.d.f. of the L_2 -norm of ρ is the p.d.f. of the event $\rho \in \mathbb{S}_n(r) = \{\rho \in \mathbb{R}^n : \|\rho\|_2 = r\}$ which is then $\mathbb{P}[\rho \in \mathbb{S}_n(r)] = K e^{-\varepsilon r} S_n(1) r^{n-1}$, where $\mathbb{S}_n(r)$ is the surface of the sphere with radius r in \mathbb{R}^n . Hence, we can write

$$\|\rho\|_2 \sim \gamma_{\varepsilon, n} \text{ s.t. } \mathbb{P}[\|\rho\|_2 = r] = \frac{\varepsilon^n e^{-\varepsilon r} r^{n-1}}{\Gamma(n)} \quad (\text{D.6})$$

which completes the proof of the second statement.

3. With $\rho \sim \gamma_{\varepsilon, n}$ we have that, by construction,

$$\mathbb{E}[\rho^2] = \mathbb{E}\left[\sum_{i=1}^n \rho_i^2\right] = n \mathbb{E}[\rho_i^2] = n \sigma_{\rho_i}^2 \quad (\text{D.7})$$

With the last equality holding since $\mathcal{L}_{0, \varepsilon}$ is isotropic and centered in zero. Recalling that

$$\mathbb{E}[\rho^2] = \frac{d^2}{dt^2} M_\rho(t) \Big|_{t=0} \quad (\text{D.8})$$

with $M_\rho(t)$ the moment generating function of the gamma distribution $\gamma_{\varepsilon, n}$,

$$\begin{aligned} & \frac{d^2}{dt^2} \left(\left(1 - \frac{t}{\varepsilon}\right)^{-n} \right) \Big|_{t=0} = \\ &= \frac{n(n+1)}{\varepsilon^2} \left(1 - \frac{t}{\varepsilon}\right)^{-(n+2)} \Big|_{t=0} = \\ &= \frac{n(n+1)}{\varepsilon^2} \end{aligned}$$

which leads to $\sigma_{\rho_i}^2 = \frac{n+1}{\varepsilon^2}$, completing the proof of the third statement and of the Proposition. \square

PROPOSITION 5.2. Let $y = f(x, \theta)$ be the fitting function of a machine learning model parameterized by θ , and $(X, Y) = Z$ the dataset over which the RMSE loss function $F(Z, \theta)$ is to be minimized, with $x \in X$ and $y \in Y$. If $\rho \sim \mathcal{L}_{0,\varepsilon}$, the bound on the increase of the cost function does not depend on the direction of ρ , in first-order approximation, and:

$$\begin{aligned} \|F(Z, \theta + \rho)\|_2 - \|F(Z, \theta)\|_2 &\leq \\ \|J_f(X, \theta)\|_2 \|\rho\|_2 + o(\|J_f(X, \theta)\rho\|_2) \end{aligned} \quad (5.8)$$

Proof. The Root Mean Square Error loss function is defined as:

$$F = \sqrt{\frac{\sum_{i=1}^{|Z|} (f(x_i, \theta) - y_i)^2}{|Z|}} = \frac{\|f(X, \theta) - Y\|_2}{\sqrt{|Z|}} \quad (D.9)$$

If the model parameters θ are sanitized by the addition of a random vector $\rho \sim \mathcal{L}_{0,\varepsilon}$, we can evaluate how the cost function would change with respect to the non-sanitized parameters. Dropping the multiplicative constant we find:

$$\begin{aligned} &\|f(X, \theta + \rho) - Y\|_2 - \|f(X, \theta) - Y\|_2 \\ &\leq \|f(X, \theta + \rho) - Y - f(X, \theta) + Y\|_2 \\ &= \|f(X, \theta + \rho) - f(X, \theta)\|_2 \\ &= \|f(X, \theta) + J_f(X, \theta)\rho - f(X, \theta) + o(J_f(X, \theta)\rho)\|_2 \\ &= \|J_f(X, \theta)\rho + o(J_f(X, \theta)\rho)\|_2 \\ &\leq \|J_f(X, \theta)\|_2 \|\rho\|_2 + o(\|J_f(X, \theta)\rho\|_2) \end{aligned}$$

□

with $J_f(X, \theta)$ being the Jacobian of f with respect to X and $o(\cdot)$ being higher terms coming from the Taylor expansion. Thus we proved that the bound on the increase of the cost function does not depend on the direction of the additive noise, but on its norm, in first-order approximation.

THEOREM 5.3. Let \mathcal{K}_i be (ε_i) - d -private mechanism for $i \in \{1, 2\}$. Then their independent composition is $(\varepsilon_1 + \varepsilon_2)$ - d -private.

Proof. Let us simplify the notation and denote:

$$P_i = \mathbb{P}_{\mathcal{K}_i} [y_i \in S_i | x_i]$$

$$P'_i = \mathbb{P}_{\mathcal{K}_i} [y_i \in S_i | x'_i]$$

for $i \in \{1, 2\}$. As mechanisms \mathcal{K}_1 and \mathcal{K}_2 are applied independently, we have:

$$\begin{aligned}\mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x_1, x_2)] &= P_1 P_2 \\ \mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x'_1, x'_2)] &= P'_1 P'_2\end{aligned}$$

Therefore, we obtain:

$$\begin{aligned}\mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x_1, x_2)] &= P_1 P_2 \\ &\leq \left(e^{\varepsilon_1 d(x_1, x'_1)} P'_1 \right) \left(e^{\varepsilon_2 d(x_2, x'_2)} P'_2 \right) \\ &\leq e^{\varepsilon_1 d(x_1, x'_1) + \varepsilon_2 d(x_2, x'_2)} \mathbb{P}_{\mathcal{K}_1, \mathcal{K}_2} [(y_1, y_2) \in S_1 \times S_2 | (x'_1, x'_2)]\end{aligned}$$

□



Experimental settings in Chapter 5

All the following experiments are run on a local server running Ubuntu 20.04.3 LTS with an AMD EPYC 7282 16-Core processor, 1.5TB of RAM and NVIDIA A100 GPUs. Python and PyTorch are the main software tools adopted for simulating the federation of clients and their corresponding collaborative training.

Synthetic data

A total of 100 users holding 10 samples each, drawn from either one of the distributions, participate in a training of two initial hypotheses which are sampled from a Gaussian distribution centred in 0 and unit variance at iteration $t = 0$. A total of $U = 7$ users are asked to participate in the optimization at each round and train locally the hypothesis that fits their dataset better for $E = 1$ epochs each time. The noise multiplier is set to $\nu = 5$. Local step size $s = 0.1$ and batch size $B_s = 10$ complete the required inputs to the algorithm. To verify the training process, another set of users with the same characteristics are held out from training to perform validation and stop the federated optimization once there is no improvement in the loss function in Equation (D.9) for 6 consecutive rounds. Although first, the updates seem to be distributed all over the domain, in just a few rounds of training the process converges to values very close to the two optimal parameters. With the heuristic presented in Section 5.4.2 it is easy to find that whenever a user participates in an optimization round it incurs a privacy leakage of at most $n/\nu = 2/5 = 0.4$, in a differential private sense, with respect to points in its neighbourhood. Using the result in Theorem 5.3 clients can compute the overall privacy leakage of the optimization process, should they be required to participate multiple times. For any user, whether to participate or not in a training round can be decided right before releasing the updated parameters, in case that would increase the privacy leakage above a threshold value decided beforehand.

Hospital charge data

The dataset contains details about charges for the 100 most common inpatient services and the 30 most common outpatient services. It shows a great variety of charges applied by healthcare providers with details mostly related to the type of service and the location of the provider. Preprocessing of the dataset includes a number of procedures, the most important of which are described here:

- i) Selection of the 4 most widely treated conditions, which amount to simple pneumonia; kidney and urinary tract infections; heart failure and shock; esophagitis and digestive system disorders.
- ii) Transformation of ZIP codes into numerical coordinates in terms of longitude and latitude.
- iii) Setting as target the Average Total Payments, i.e. the cost of the service averaged among the times it was given by a certain provider.
- iv) As it is a standard procedure in the context of gradient-based optimization, dependent and independent variables are brought to be in the range of the *units* before being fed to the machine learning model. Note that this point takes the spot of the common feature normalization and standardization procedures, which we decided not to perform here to keep the setting as realistic as possible. In fact, both would require the knowledge of the empirical distribution of all the data. Although it is available in simulation, it would not be available in a real scenario, as each user would only have access to their dataset.

Given the preprocessing described above, the dataset results in 2947 clients, randomly split in train and validation subsets with 70 and 30 per cent of the total clients each. The goal is being able to predict the cost that a service would require given where it is performed in the country, and what kind of procedure it is. The model that was adopted in this context is a fully connected neural network (NN) of two layers, with a total of 11 parameters and Rectified Linear Unit (ReLU) activation function. Inputs to the model are an increasing index which uniquely defines the healthcare service, the longitude and latitude of the provider. Output of the model is the expected cost. Tests have been performed to minimize the RMSE loss on the clients selected for training (100 per round) and at each round the performance of the model is checked against a held-out set of validation clients, from where 200 are sampled every time. If 30 validation rounds are passed without improvement in the cost function, the optimization process is terminated. In order to decrease the variability of the results, a total of 10 runs have been performed with different seeds for every combination of number of hypotheses and noise multiplier.

FEMNIST image classification

The task consists in performing image classification on the FEMNIST [182] dataset, which is a standard benchmark dataset for federated learning, based on EMNIST [244] and with the data points grouped by user. It consists of a large number of images of handwritten digits,

lower and upper case letters of the Latin alphabet. As a preprocessing step, images of client c are rotated 90 degrees counter-clockwise depending on the realization of the random variable $\text{rot}_c \sim \text{Bernoulli}(0.5)$. This is a common practice in machine learning to simulate local datasets held by different clients being generated by different distributions [138, 245–247]. The chosen architecture is described in Table E.1 and yields a parameter vector $\theta \in \mathbb{R}^{n_0}$, $n_0 = 1206590$. Runs are performed with a maximum of 500 rounds of federated optimization, unless 5 consecutive validation rounds are conducted without improvements on the validation loss. The latter is evaluated on a held out set of clients, consisting of 10% of the total number. Validation is performed every 5 training rounds, thus the process terminates after 25 rounds without the model’s performance improvement. The optimization process aims to minimize either the RMSE loss or the Cross Entropy loss [248] between model’s predictions and the target class.

Layer	Properties
2D Convolution	kernel size: (2,2) stride: (1,1) nonlinearity: ReLU output features: 32
2D Convolution	kernel size: (2,2) stride: (1,1) nonlinearity: ReLU output features: 64
2D Max Pool	kernel size: (2,2) stride: (2,2) nonlinearity: ReLU
Fully Connected	nonlinearity: ReLU units: 128
Fully Connected	nonlinearity: ReLU units: 62

Table E.1: NN architecture adopted in the experiments of Section 5.5.4



Proofs from Chapter 6

LEMMA 6.1. Given a binary classification NN with no hidden layers, batch size 1 and a final layer containing bias terms, the gradient update using a cross-entropy loss function reveals the values of the input \mathbf{x} exactly.

Proof. Under the assumption that $|\mathbf{x}| \neq 0$, from the closed form above we can deduce the x_i exactly using the knowledge of the $p_k - y_k$ values. However, this assumes that the values $p_k - y_k$ are always non-zero. Observe that the softmax function can never be 0, and for 2 outputs can also never be 1. Thus, assuming that $y_k \in \{0, 1\}$ we have that $p_k - y_k$ must always be non-zero. And so \mathbf{x} is revealed exactly. \square

LEMMA 6.2. If \mathbf{x} and \mathbf{x}' are gradient clones, there is some $K \in \mathbb{R}$ such that $\mathbf{x}' = K\mathbf{x}$.

Proof. Let some \mathbf{x}, \mathbf{x}' be gradient clones. W.l.o.g. let $y(\mathbf{x}) = A$ and so $y_B(\mathbf{x}) = 0$. Then we have two possibilities: $y(\mathbf{x}') = A$ or $y(\mathbf{x}') = B$.

Case 1

$y(\mathbf{x}') = A$: Hence $y_B(\mathbf{x}') = 0$ and so from (6.11) we have that $\nabla \mathbf{w}_B(\mathbf{x}) = \nabla \mathbf{w}_B(\mathbf{x}')$, and so we deduce:

$$p_B(\mathbf{x})\mathbf{x} = p_B(\mathbf{x}')\mathbf{x}' \Rightarrow \tag{F.1}$$

$$\begin{aligned}
& \frac{e^{\sum_{i=1}^n x_i w_B(i) + b_B}}{e^{\sum_{i=1}^n x_i w_A(i) + b_A} + e^{\sum_{i=1}^n x_i w_B(i) + b_B}} x_i = \\
& \frac{e^{\sum_{i=1}^n x'_i w_B(i) + b_B}}{e^{\sum_{i=1}^n x'_i w_A(i) + b_A} + e^{\sum_{i=1}^n x'_i w_B(i) + b_B}} x'_i \Rightarrow \\
& \frac{e^{x \cdot w_B} x_i}{e^{x \cdot w_A + b_A} + e^{x \cdot w_B + b_B}} = \frac{e^{x' \cdot w_B} x'_i}{e^{x' \cdot w_A + b_A} + e^{x' \cdot w_B + b_B}} \quad (\text{F.2}) \\
& \forall i \in \{1, \dots, n\}
\end{aligned}$$

Setting

$$K = e^{w_B \cdot (x - x')} \frac{\left(e^{x' \cdot w_A + b_A} + e^{x'_i \cdot w_B + b_B} \right)}{\left(e^{x \cdot w_A + b_A} + e^{x \cdot w_B + b_B} \right)}$$

gives us our desired result.

Case 2

$y(x') = B$: Hence, $y_A(x') = 0$ and so from (6.10) we have that $\nabla w_A(x) = \nabla w_A(x')$ and so we deduce:

$$(p_A(x) - 1)x = p_A(x')x' \Rightarrow \quad (\text{F.3})$$

$$-p_B(x)x = p_A(x')x' \Rightarrow \quad (\text{F.4})$$

$$\begin{aligned}
& - \frac{e^{\sum_{i=1}^n x_i w_B(i) + b_B}}{e^{\sum_{i=1}^n x_i w_A(i) + b_A} + e^{\sum_{i=1}^n x_i w_B(i) + b_B}} x_i = \\
& \frac{e^{\sum_{i=1}^n x'_i w_A(i) + b_A}}{e^{\sum_{i=1}^n x'_i w_A(i) + b_A} + e^{\sum_{i=1}^n x'_i w_B(i) + b_B}} x'_i \quad (\text{F.5}) \\
& \forall i \in \{1, \dots, n\}
\end{aligned}$$

Setting

$$K = -e^{w_B \cdot x - w_A \cdot x' + (b_B - b_A)} \frac{\left(e^{x' \cdot w_A + b_A} + e^{x'_i \cdot w_B + b_B} \right)}{\left(e^{x \cdot w_A + b_A} + e^{x \cdot w_B + b_B} \right)}$$

gives us our desired result. \square

LEMMA 6.3. Setting $\gamma = x \cdot (w_A - w_B)$ and $\Delta = b_B - b_A$, x has a gradient clone iff there is some $K \in \mathbb{R}$ satisfying the equation

$$e^{K\gamma} - Ke^\gamma = (K - 1)e^\Delta. \quad (6.13)$$

The corresponding gradient clone of x would be x scaled by that K .

Proof. Let x' be a gradient clone of x . Using the fact that $x' = Kx$ for some $K \in \mathbb{R}$ (by Lemma 6.2), by comparing the signs of the components of $\nabla w(x)$ and $\nabla w(x')$ (Remark 6.4.2), and by Eqns. (6.10) and (6.11), we must have:

$$p_B(x)x = p_B(x')x'$$

$$\begin{aligned}
&\Rightarrow p_B(\mathbf{x})\mathbf{x} = p_B(\mathbf{x}')K\mathbf{x} \\
&\Rightarrow p_B(\mathbf{x}) = p_B(\mathbf{x}')K \text{ [as } |\mathbf{x}| \neq 0] \\
&\Rightarrow \frac{e^{\mathbf{x} \cdot \mathbf{w}_B} e^{b_B}}{e^{\mathbf{x} \cdot \mathbf{w}_A + b_A} + e^{\mathbf{x} \cdot \mathbf{w}_B + b_B}} = \frac{e^{\mathbf{x}' \cdot \mathbf{w}_B} e^{b_B} K}{e^{\mathbf{x}' \cdot \mathbf{w}_A + b_A} + e^{\mathbf{x}' \cdot \mathbf{w}_B + b_B}} \\
&\Rightarrow \frac{e^{\mathbf{x} \cdot \mathbf{w}_B}}{e^{\mathbf{x} \cdot \mathbf{w}_A + b_A} + e^{\mathbf{x} \cdot \mathbf{w}_B + b_B}} = \frac{e^{K\mathbf{x} \cdot \mathbf{w}_B} K}{e^{K\mathbf{x} \cdot \mathbf{w}_A + b_A} + e^{K\mathbf{x} \cdot \mathbf{w}_B + b_B}} \\
&\Rightarrow \frac{e^{K\mathbf{x} \cdot \mathbf{w}_A + b_A} + e^{K\mathbf{x} \cdot \mathbf{w}_B + b_B}}{e^{\mathbf{x} \cdot \mathbf{w}_A + b_A} + e^{\mathbf{x} \cdot \mathbf{w}_B + b_B}} = K e^{\mathbf{x} \cdot \mathbf{w}_B (K-1)} \\
&\Rightarrow \left(\frac{e^{K\mathbf{x} \cdot \mathbf{w}_B}}{e^{\mathbf{x} \cdot \mathbf{w}_B}} \right) \frac{e^{K(\mathbf{x} \cdot \mathbf{w}_A - \mathbf{x} \cdot \mathbf{w}_B) + b_A} + e^{b_B}}{e^{(\mathbf{x} \cdot \mathbf{w}_A - \mathbf{x} \cdot \mathbf{w}_B) + b_A} + e^{b_B}} \\
&= K e^{\mathbf{x} \cdot \mathbf{w}_B (K-1)} \\
&\Rightarrow e^{(K-1)\mathbf{x} \cdot \mathbf{w}_B} \left(\frac{e^{K\gamma} e^{b_A} + e^{b_B}}{e^\gamma e^{b_A} + e^{b_B}} \right) = K e^{\mathbf{x} \cdot \mathbf{w}_B (K-1)} \\
&\Rightarrow e^{K\gamma} e^{b_A} + e^{b_B} = K \left(e^\gamma e^{b_A} + e^{b_B} \right) \\
&\Rightarrow e^{K\gamma} + e^\Delta = K \left(e^\gamma + e^\Delta \right) \\
&\Rightarrow e^{K\gamma} - K e^\gamma = (K-1)e^\Delta
\end{aligned}$$

□

THEOREM 6.4. For any $K > 1$, if $\mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B) \leq \frac{\ln K}{K-1}$, then \mathbf{x} has no gradient clone scaled by that K .

Proof. Let $\gamma = \mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B)$ and $\Delta = b_B - b_A$. For \mathbf{x} to have a gradient clone scaled by K , by Lemma 6.3, we must satisfy (6.13). Moreover, we note that the RHS of (6.13) is positive for $K > 1$. In particular, we have:

$$\begin{aligned}
&e^{K\gamma} - K e^\gamma > 0 \\
&\iff e^{K\gamma} > e^{\gamma + \ln K} \\
&\iff K\gamma > \gamma + \ln K \\
&\iff \gamma > \frac{\ln K}{K-1}
\end{aligned}$$

But $\gamma \leq \frac{\ln K}{K-1}$ by assumption. □

THEOREM 6.5. \mathbf{x} cannot have a gradient clone \mathbf{x}' with $|\mathbf{x}'| > |\mathbf{x}|$ if there exists some $\lambda \in \mathbb{R}^+$ for which

$$\mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B) = 1 + \lambda + W_0 \left(\frac{e^\Delta}{e^{(1+\lambda)e^\Delta}} \right)$$

where $\Delta = b_B - b_A$ and W_0 is the principal branch of the Lambert W -function.

Proof. Let \mathbf{x}' be a gradient clone of \mathbf{x} with $|\mathbf{x}'| > |\mathbf{x}|$. By Lemma 6.2 and Remark 6.4.2, there must exist a $K > 1$ s.t. $\mathbf{x}' = K\mathbf{x}$ and let it be written as $K = 1 + T$ for some $T > 0$. Hence, setting

$\gamma = \mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B)$ and $\Delta = b_B - b_A$, by Lemma 6.3, we must have:

$$\begin{aligned} e^{K\gamma} - Ke^\gamma &= (K-1)e^\Delta \\ \Rightarrow e^{\gamma+\gamma T} - (1+T)e^\gamma &= Te^\Delta \\ \Rightarrow e^{\gamma T} &= 1 + T(1 + e^{\Delta-\gamma}) \end{aligned} \quad (\text{F.6})$$

As $e^{\gamma T} > 1 + \gamma T$, if we have $\gamma \geq 1 + e^{\Delta-\gamma}$, we would obtain $e^{\gamma T} > 1 + \gamma T \geq 1 + (1 + e^{\Delta-\gamma})T$, implying that (F.6) has no solution for any $T > 0$. Therefore, it is sufficient to have $\gamma \geq 1 + e^{\Delta-\gamma}$ to ensure that \mathbf{x} does not have any gradient clone \mathbf{x}' s.t. $|\mathbf{x}'| > |\mathbf{x}|$. But we observe:

$$\begin{aligned} \gamma &\geq 1 + e^{\Delta-\gamma} \\ \Leftrightarrow e^\gamma \gamma &\geq e^\gamma + e^\Delta \\ \Leftrightarrow e^\gamma(\gamma - 1) &\geq e^\Delta \Leftrightarrow e^{-\Delta}(\gamma - 1) \geq e^{-\gamma} \\ \Leftrightarrow e^{-\gamma} &= e^{-\Delta}(\gamma - 1) - \lambda \text{ for some } \lambda > 0 \\ \Leftrightarrow e^{-\gamma} &= e^{-\Delta}(\gamma - (1 + \lambda e^\Delta)) \end{aligned} \quad (\text{F.7})$$

By the theory of generalization of Lambert W -functions [249], we must have: $\gamma = 1 + \lambda + W\left(\frac{e^\Delta}{e^{(1+\lambda)e^\Delta}}\right)$, where $W(\cdot)$ is the Lambert W -function. But as $\frac{e^\Delta}{e^{(1+\lambda)e^\Delta}}$ is necessarily positive, $W\left(\frac{e^\Delta}{e^{(1+\lambda)e^\Delta}}\right)$ must lie on the principal branch of $W(\cdot)$ denoted by $W_0(\cdot)$. In other words, we must have:

$$\mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B) = 1 + \lambda + W_0\left(\frac{e^\Delta}{e^{(1+\lambda)e^\Delta}}\right)$$

□

THEOREM 6.6. \mathbf{x} has a gradient clone \mathbf{x}' iff all of the following hold:

- a) $\frac{e^\Delta}{e^\gamma + e^\Delta} - \frac{1}{\gamma} W\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right) \neq 1$
- b) $e^{-1}(e^\gamma + e^\Delta) < \gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}$
- c) $\mathbf{x}' = K\mathbf{x}$ with $K = \frac{e^\Delta}{e^\gamma + e^\Delta} - \frac{1}{\gamma} W\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right)$

where $\gamma = \mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B)$, $\Delta = b_B - b_A$, and $W(\cdot)$ is the Lambert W -function.

Proof. Let \mathbf{x}' be a gradient clone of \mathbf{x} . Then, by Lemma 6.2, \mathbf{x}' must be of the form $K\mathbf{x}$ for some $K > 0$ and, by Lemma 6.3, such a K must satisfy (6.13). In particular, we must have:

$$\begin{aligned} e^{K\gamma} - Ke^\gamma &= (K-1)e^\Delta \\ \Leftrightarrow e^{K\gamma} &= K(e^\gamma + e^\Delta) - e^\Delta \\ \Leftrightarrow e^{K\gamma} &= (e^\gamma + e^\Delta)\left(K - \frac{e^\Delta}{e^\gamma + e^\Delta}\right) \\ [\text{as } e^\gamma + e^\Delta &> 0] \end{aligned}$$

$$\Rightarrow K = \frac{e^\Delta}{e^\gamma + e^\Delta} - \frac{1}{\gamma} W\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right) \quad (\text{F.8})$$

[by the theory of generalization of Lambert W -function [249]]

a) By definition of a gradient clone as in Definition 6.4.1, $\mathbf{x}' \neq \mathbf{x}$. Hence, we must have $K \neq 1$ in (F.8).

b) $W\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right)$ and, correspondingly, (F.8) have no real solution if:

$$\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta} < -e^{-1} \Rightarrow e^{-1}(e^\gamma + e^\Delta) < \gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}$$

c) If conditions a and b hold, \mathbf{x}' is a gradient clone of \mathbf{x} iff $\mathbf{x}' = K\mathbf{x}$ where K is as in (F.8). □

COROLLARY 6.7. \mathbf{x} has at most two gradient clones in \mathbb{R}^n .

Proof. Let \mathbf{x} be any gradient clone of \mathbf{x} . By Theorem 6.6, $\mathbf{x}' = \left(\frac{e^\Delta}{e^\gamma + e^\Delta} - \frac{1}{\gamma} W\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right)\right) \mathbf{x}$ where $\gamma = \mathbf{x} \cdot (\mathbf{w}_A - \mathbf{w}_B)$, $\Delta = b_B - b_A$, and $W(\cdot)$ is the Lambert W -function. From the theory of Lambert W -functions [249], we recall that $W\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right)$

i) has *exactly one* real solution when $\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta} \in [0, \infty) \cup \{-e^{-1}\}$ which is given by its principal branch $W_0\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right)$.

ii) has *exactly two* real solutions when $\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta} \in (-e^{-1}, 0)$ which are given by $W_{-1}\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right)$.

iii) has *no* real solution when $\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta} < -e^{-1}$ as is addressed by condition c) of Theorem 6.6.

As the values of $\mathbf{w}_A, \mathbf{w}_B, b_A, b_B$ are all fixed, any gradient clone of \mathbf{x} has at most the same number of gradient clones as the number of real solutions of $W\left(\frac{-\gamma e^{\frac{\gamma e^\Delta}{e^\gamma + e^\Delta}}}{e^\gamma + e^\Delta}\right)$. □

LEMMA 6.8. If batches $B = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ and $B' = \{\mathbf{x}'_1, \dots, \mathbf{x}'_k\}$ are gradient clones, we must have $X\pi = X'\pi'$ where:

$$X = \begin{pmatrix} x_{11} & \dots & x_{k1} \\ \vdots & \vdots & \vdots \\ x_{1n} & \dots & x_{kn} \end{pmatrix}, \quad X' = \begin{pmatrix} x'_{11} & \dots & x'_{k1} \\ \vdots & \vdots & \vdots \\ x'_{1n} & \dots & x'_{kn} \end{pmatrix},$$

$$\pi = \begin{pmatrix} p_{11} - 1 & p_{21} & \cdots & p_{k1} & \cdots p_{m1} \\ p_{12} & p_{22} - 1 & \cdots & p_{k2} & \cdots p_{m2} \\ \vdots & \vdots & \ddots & \vdots & \\ p_{1k} & p_{2k} & \cdots & p_{kk} - 1 & \cdots p_{mk} \end{pmatrix},$$

$$\pi' = \begin{pmatrix} p'_{11} - 1 & p'_{21} & \cdots & p'_{k1} & \cdots p'_{m1} \\ p'_{12} & p'_{22} - 1 & \cdots & p'_{k2} & \cdots p'_{m2} \\ \vdots & \vdots & \ddots & \vdots & \\ p'_{1k} & p'_{2k} & \cdots & p'_{kk} - 1 & \cdots p'_{mk} \end{pmatrix},$$

and $p_{js} = \mathbb{P}[\mathbf{x}_s = L_j]$ (i.e., the output probability of the s^{th} input of batch X to be classified as label L_j) and, similarly, $p'_{js} = \mathbb{P}[\mathbf{x}'_s = L_j]$ for $s = 1, \dots, k$ and $j = 1, \dots, m$.

Proof. Generalising the derivation of the gradient updates w.r.t. the weights in the FC layer for binary classification (as in Eqns. (6.10), (6.11), and (6.12)) to a classification task involving m labels, for a specific training input $\mathbf{x} \in \mathbb{R}_{\geq 0}^n$ to the FC layer, we obtain:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_i} &= \sum_{j=1}^m \frac{\partial \mathcal{L}}{\partial P_j(\mathbf{x})} \frac{\partial P_j(\mathbf{x})}{\partial z_i} \frac{\partial z_i}{\partial w_i} \\ &= y_i(\mathbf{x})(P_i - 1)\mathbf{x} + \sum_{\substack{j=1 \\ j \neq i}}^m y_j(\mathbf{x})P_i\mathbf{x} \end{aligned} \quad (\text{F.9})$$

Therefore, if the true label of $\mathbf{x}_s = (x_{s1}, \dots, x_{sn}) \in \mathbb{R}_{\geq 0}^n$ was L_s for all $s \in \{1, \dots, k\}$, then for every $r \in \{1, \dots, k\}$ and $j \in \{1, \dots, n\}$:

$$(\nabla \mathbf{w}(x_{sj}))_r = \begin{cases} (P_s(\mathbf{x}) - 1)x_{sj} & \text{if } s = r \\ P_r(\mathbf{x})x_{sj} & \text{o.w.} \end{cases} \quad (\text{F.10})$$

Hence, for each $\mathbf{x}_s = (x_{s1}, \dots, x_{sn}) \in B$:

$$\begin{aligned} \nabla \mathbf{w}_B(\mathbf{x}_s) &= (\nabla \mathbf{w}(x_{s1}), \dots, \nabla \mathbf{w}(x_{sn}))^T = \\ &= \begin{pmatrix} P_1(\mathbf{x}_s)x_{s1} & \cdots & (P_s(\mathbf{x}_s) - 1)x_{s1} & \cdots & P_m(\mathbf{x}_s)x_{s1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_1(\mathbf{x}_s)x_{sn} & \cdots & (P_s(\mathbf{x}_s) - 1)x_{sn} & \cdots & P_m(\mathbf{x}_s)x_{sn} \end{pmatrix} \end{aligned}$$

Thus, the gradient update averaged over the batch members that is shared with the server is:

$$\frac{\sum_{i=1}^k \nabla \mathbf{w}_B(\mathbf{x}_i)}{k} = \frac{1}{k} W_B \quad (\text{F.11})$$

where $W_B \in \mathbb{R}^{n \times m}$ such that:

$$W_B(i, j) = \sum_{\substack{s=1 \\ s \neq j}}^k P_j(\mathbf{x}_s) x_{si} + (P_j(\mathbf{x}_j) - 1) x_{ji} \quad (\text{F.12})$$

Therefore, if batches $B = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ and $B' = \{\mathbf{x}'_1, \dots, \mathbf{x}'_k\}$ are gradient clones s.t. $y(\mathbf{x}_s) = y(\mathbf{x}'_s) = L_s$ for every $s \in \{1, \dots, k\}$, we must have $W_B = W_{B'}$

$$\iff \sum_{\substack{s=1 \\ s \neq j}}^k P_j(\mathbf{x}_s) x_{si} + (P_j(\mathbf{x}_j) - 1) x_{ji} = \sum_{\substack{s=1 \\ s \neq j}}^k P_j(\mathbf{x}'_s) x'_{si} + (P_j(\mathbf{x}'_j) - 1) x'_{ji} \quad (\text{F.13})$$

for every $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, k\}$.

Unifying the notations with the theorem statement (i.e., $P_j(\mathbf{x}_s) = p_{js}$ and $P_j(\mathbf{x}'_s) = p'_{js}$, as an alternative representation of (F.13), we conclude that if B and B' are gradient clones, we must have:

$$X\pi = X'\pi' \quad (\text{F.14})$$

where:

$$\begin{aligned} X &= \begin{pmatrix} x_{11} & \dots & x_{k1} \\ \vdots & & \vdots \\ x_{1n} & \dots & x_{kn} \end{pmatrix}, \quad X' = \begin{pmatrix} x'_{11} & \dots & x'_{k1} \\ \vdots & & \vdots \\ x'_{1n} & \dots & x'_{kn} \end{pmatrix}, \\ \pi &= \begin{pmatrix} p_{11} - 1 & p_{21} & \dots & p_{k1} & \dots & p_{m1} \\ p_{12} & p_{22} - 1 & \dots & p_{k2} & \dots & p_{m2} \\ \vdots & \vdots & \ddots & \vdots & & \\ p_{1k} & p_{2k} & \dots & p_{kk} - 1 & \dots & p_{mk} \end{pmatrix}, \\ \pi' &= \begin{pmatrix} p'_{11} - 1 & p'_{21} & \dots & p'_{k1} & \dots & p'_{m1} \\ p'_{12} & p'_{22} - 1 & \dots & p'_{k2} & \dots & p'_{m2} \\ \vdots & \vdots & \ddots & \vdots & & \\ p'_{1k} & p'_{2k} & \dots & p'_{kk} - 1 & \dots & p'_{mk} \end{pmatrix}. \end{aligned}$$

□

THEOREM 6.9. If the data in the batch that is used to train a certain round are *similar* and when the FL model *learns well*, the spatial arrangement of the members of the batch used for training is unique with respect to the shared gradients.

Proof. Noting that $\sum_{j=1}^m p_{js} = 1$, we have $\sigma_s - 1 = -(m-1)r_s$ for every $s \in \{1, \dots, k\}$. Therefore, we may rewrite (6.14) as:

$$\sum_{\substack{s=1 \\ s \neq j}}^m (x_{si} - x'_{si}) r_j + (x_{ji} - x'_{ji}) (-(m-1)r_j) = 0$$

$$\begin{aligned}
&\Rightarrow \sum_{\substack{s=1 \\ s \neq j}}^m (x_{si} - x'_{si})r_j = (x_{ji} - x'_{ji})(m-1)r_j \\
&\Rightarrow \sum_{s=1}^m (x_{si} - x'_{si})r_j = (x_{ji} - x'_{ji})mr_j \\
&\Rightarrow \frac{\sum_{s=1}^m (x_{si} - x'_{si})r_j}{m} = (x_{ji} - x'_{ji})r_j
\end{aligned} \tag{F.15}$$

Eqn. (F.15) essentially implies that:

$$(x_{1i} - x'_{1i})r_1 = (x_{2i} - x'_{2i})r_2 = \dots = (x_{ki} - x'_{ki})r_k \tag{F.16}$$

Moreover, realistically $\sigma_j < 1$ and, thus, $r_j > 0$ for every $j \in \{1, \dots, k\}$. Under this setting and as we are given $\mathbf{x}_{j_1} \cong \mathbf{x}_{j_2}$ for every $j_1, j_2 \in \{1, \dots, k\}$, Eqn. (F.16) would mean:

$$\begin{aligned}
&x_{1i} - x'_{1i} = x_{2i} - x'_{2i} = \dots = x_{ki} - x'_{ki} \\
&\Rightarrow x_{si} - x_{ti} = x'_{si} - x'_{ti} \forall s, t \in \{1, \dots, k\}, i \in \{1, \dots, n\}
\end{aligned} \tag{F.17}$$

From Eqn. (F.17) we can conclude that $d_p(\mathbf{x}_s, \mathbf{x}_t) = d_p(\mathbf{x}'_s, \mathbf{x}'_t) \forall s, t \in \{1, \dots, k\}$ where d_p is any ℓ^p norm. Therefore, in particular, with $p = 2$ we must have:

$$\|\mathbf{x}_s - \mathbf{x}_t\|_2 = \|\mathbf{x}'_s - \mathbf{x}'_t\|_2 \forall s, t \in \{1, \dots, k\} \tag{F.18}$$

Eqn. (F.18) illustrates that, under the aforementioned assumptions, if a potential batch B' is a gradient clone of the true batch B used for training, the spatial arrangement of B' must be the same as B' . \square

LEMMA 6.10. For batch size = 2, π (and, similarly, π') is right-invertible.

Proof. For $k = 2$, $\Pi = \begin{pmatrix} p_{11} - 1 & p_{21} & \dots & p_{m1} \\ p_{12} & p_{22} - 1 & \dots & p_{m2} \end{pmatrix}$ where $p_{js} = \mathbb{P}[\mathbf{x}_s = L_j]$ is the output (softmax) probability of the s^{th} input of batch X (with true label L_s) to be classified as label L_j . Let $\alpha, \beta \in \mathbb{R}$, not both 0, such that $\alpha\Pi[1] + \beta\Pi[2] = \mathbf{0}_{1 \times m}$.

Case 1: $\alpha\beta \geq 0$, i.e., α and β are of the same sign: W.l.o.g., let $\alpha \geq 0, \beta > 0$. Then $\alpha p_{31}, \dots, \alpha p_{m1} \geq 0$ and $\beta p_{32}, \dots, \beta p_{m2} > 0$. Hence, $\alpha\Pi[1]_j + \beta\Pi[2]_j > 0$ for every $j \in \{3, \dots, m\}$.

Case 2: α and β are of opposite signs: W.L.O.G., let $\alpha \geq 0, \beta < 0$. Then $\alpha p_{21} \geq 0$ and $\beta(p_{22} - 1) > 0$. Hence, $\alpha\Pi[1]_2 + \beta\Pi[2]_2 > 0$.

Both the cases make us arrive to a contradiction suggesting that no such α, β can exist. Hence, the rows of Π are linearly independent and, hence, Π is right-invertible. \square

THEOREM 6.12. $d_E(\mathbf{x}_i, \mathbf{x}'_i) = d_E(\delta_{1i}\mathbf{x}_1, -\delta_{2i}\mathbf{x}_2)$ for $i = 1, 2$, where $D = \mathbf{I}_{2 \times 2} + \delta$.

Proof. Let $D = \mathbf{I}_{2 \times 2} + \delta = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} = \begin{pmatrix} 1 + \delta_{11} & \delta_{12} \\ \delta_{21} & 1 + \delta_{22} \end{pmatrix}$ for some $\delta_{ij} \in \mathbb{R}$ for $i, j \in \{1, 2\}$.

Plugging in $D = \mathbb{I}_{2 \times 2} + \delta$ into Corollary 6.11, we have:

$$x'_{1i} = x_{1i} + \delta_{11}x_{1i} + \delta_{21}x_{2i} \quad (\text{F.19})$$

$$x'_{2i} = x_{2i} + \delta_{12}x_{1i} + \delta_{22}x_{2i} \quad (\text{F.20})$$

Let us start with the case of $i = 1$:

$$\begin{aligned} d_E(\mathbf{x}_1, \mathbf{x}'_1) &= \left(\sum_{i=1}^n |x_{1i} - x'_{1i}|^2 \right)^{\frac{1}{2}} \\ &= \left(\sum_{i=1}^n |\delta_{11}x_{1i} + \delta_{21}x_{2i}|^2 \right)^{\frac{1}{2}} \quad [\text{using (F.19)}] \\ &= d_E(\delta_{11}\mathbf{x}_1, -\delta_{21}\mathbf{x}_2) \end{aligned} \quad (\text{F.21})$$

The case of $i = 2$ proceeds exactly the same way. \square

LEMMA 6.13. Let \mathbf{z}_{k-1} be the vector of outputs from the penultimate layer $k - 1$ of a neural network. Assume a single non-zero value $z_{j,k-1}$ is revealed. Then every $z_{i,k-1}$ for $i \neq j$ can be deduced by an adversary with knowledge of the network and gradients.

Proof. From (6.19) we can deduce the value of $p_i(\mathbf{x}) - y_i(\mathbf{x})$ from the revealed $z_{j,k-1}$ and the corresponding gradient $\nabla w_{j,i,k-1}(\mathbf{x})$. Thus we can also deduce the values of the remaining $z_{i,k-1}$ using the partial derivatives $z_{i,k-1}(p_i(\mathbf{x}) - y_i(\mathbf{x}))$. \square

THEOREM 6.14. Let \mathbf{z}_{k-1} denote the vectors of outputs from penultimate layer of the NN, and let \mathbf{x} be the vector of inputs. Then an adversary with knowledge of \mathbf{z}_{k-1} learns the values in \mathbf{x} exactly.

Proof. The proof follows by induction. We first show that knowledge of the penultimate layer leaks the values in the previous layer. We have:

$$\begin{aligned} \nabla w_{i,j,k-1} &= z_{i,k-2} \frac{\partial \mathcal{L}}{\partial z_{i,k-1}} \\ &\quad [\text{from (6.23)}] \\ &= z_{i,k-2} \sum_m w_{i,m,k} \frac{\partial \mathcal{L}}{\partial z_{m,k}} \\ &\quad [\text{from (6.21)}] \\ &= z_{i,k-2} \sum_m w_{i,m,k} (p_m(\mathbf{x}) - y_m(\mathbf{x})) \\ &\quad [\text{from (6.22)}] \end{aligned}$$

Since $p_m(\mathbf{x}) - y_m(\mathbf{x})$ is leaked from \mathbf{z}_{k-1} and knowledge of the gradients (recall (6.19)), then we have that $z_{i,k-2}$ is deducible (since the $w_{i,m,k}$ and $\nabla w_{i,j,k-1}$ are known to the adversary). Thus the output values in layer $k - 2$ are leaked through knowledge of the values in layer $k - 1$.

The induction step gives us that from knowledge of any hidden layer (or more precisely, the derivatives in any hidden layer), we can infer the values and derivatives in the previous layer.

In particular, note that from any derivative $\frac{\partial \mathcal{L}}{\partial z_{i,h}}$ we can compute derivatives in the previous layer $\frac{\partial \mathcal{L}}{\partial z_{i,h-1}}$ (by (6.21)). And since we can compute those derivatives at the final layer (by (6.22)) then they are computable all the way through to the first hidden layer.

Secondly, knowledge of the derivatives at any layer h leaks the output values \mathbf{z}_{h-1} of the previous layer (by (6.23)), and when h is the first layer, leaks the values of \mathbf{x} (by (6.24)).

Finally, we address the issue caused by ReLU in that sometimes the values of the derivatives are 0 (eg. in (6.21)). Observe that from (6.21), that if the outputs from a layer are all zero'ed (by ReLU), then so are the derivatives $\frac{\partial \mathcal{L}}{\partial z_{i,h}}$ on the previous layers (follows by recursion). And therefore, by (6.23), the gradient updates are also zero, which means the network does not learn anything on any of the previous layers. We conclude that, for the network to learn, there must exist a node in each layer which is not zero'ed by the ReLU activation, and therefore by which the gradients can leak information about the input values \mathbf{x} .

Thus, we conclude by induction that knowledge of the penultimate layer leaks the values of \mathbf{x} exactly. \square



Proofs from Chapter 7

LEMMA 7.1. If u_i reports a price greater than her true price, i.e., $p_i \geq \pi_i$, then her utility will be less than the utility for the true price, i.e., $\rho(p_i) \leq \rho(\pi_i)$.

Proof. Suppose the i^{th} provider reports the privacy price as $p_i > \pi_i$. We want to show:

$$\begin{aligned} \rho(\pi_i) - \rho(p_i) &> 0 \\ \iff \mu(\pi_i) - \pi_i f(\pi_i) - \mu(p_i) + \pi_i f(p_i) &> 0 \\ \iff \mu(\pi_i) - \mu(p_i) - \pi_i(f(\pi_i) - f(p_i)) &> 0 \\ \iff \pi_i f(\pi_i) + \int_{\pi_i}^{\infty} f(z) dz - (p_i f(p_i) + \int_{p_i}^{\infty} f(z) dz) - \pi_i(f(\pi_i) - f(p_i)) &> 0 \\ \iff \pi_i f(p_i) - p_i f(p_i) + \int_{\pi_i}^{p_i} f(z) dz &> 0 \\ \iff \int_{\pi_i}^{p_i} f(z) dz &> f(p_i)(p_i - \pi_i) \end{aligned}$$

Note that f is a monotonically decreasing function. Let $g(p) = f(\pi_i) - f(p)$ for any $p \in \mathbb{R}$. Furthermore, $\int_{\pi_i}^{p_i} f(z) dz$ can be written as $f(p_i)(p_i - \pi_i) + \int_{\pi_i}^{p_i} g(z) dz$. Observe that f is monotonically decreasing means $g(\cdot)$ is monotonically decreasing, hence, $g(\pi_i) > g(p)$ for any $p > \pi_i$, $g(p_i) > 0$. Hence, $\int_{\pi_i}^{p_i} g(z) dz > 0 \Rightarrow \int_{\pi_i}^{p_i} f(z) dz > f(p_i)(p_i - \pi_i) \Rightarrow (2)$. \square \square

LEMMA 7.2. If u_i reports a price smaller than her true price, i.e., $p_i \leq \pi_i$, then her utility will be less than the utility for the true price, i.e., $\rho(p_i) \leq \rho(\pi_i)$.

Proof. Similar and symmetric argument to Lemma 7.1. \square \square

THEOREM 7.4. If $\tau(\cdot)$ is additive, then maximizing information and maximizing profit (desiderata (a) and (b)) are equivalent, in the sense that a ε -allocating function $f(\cdot, \cdot)$ that maximizes the one, maximizes also the other.

Proof. Let τ be a monotonically increasing and additive function representing the pay-off earned by the data consumer by processing the information obtained from the different data providers. We wish to maximize her profit, i.e., $\sum_{i=1}^n (\tau(\varepsilon_i) - \mu(p_i))$ for a fixed budget of expenses B , i.e., $\sum_{i=1}^n \mu(p_i) = B$. Therefore, we boil down to maximizing $\sum_{i=1}^n \tau(\varepsilon_i) = \tau(\sum_{i=1}^n \varepsilon_i)$ for $\sum_{i=1}^n \mu(p_i) = B$, where $\varepsilon_i = f(c, p_i)$. As τ is increasing, it attains maximum if and only if $\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n f(c, p_i)$ is maximum. Then just observe that $\sum_{i=1}^n f(c, p_i) = I(\mathbf{u})$ (cf. Definition 7.3.3). \square

THEOREM 7.6. If τ is additive, then there exists a c that gives an optimal **profit-maximizing** function $f(c, \cdot) \in \mathcal{F}$, for a fixed budget, and we can derive such c via the method of the Lagrangians.

Proof. The profit of the data consumer is $\sum_{i=1}^n (\tau(\varepsilon_i) - \mu(p_i))$, where $\varepsilon_i = f(c, p_i)$. Therefore, to maximize her profit for a fixed budget of expenses B (used to pay the incentives to the data providers), i.e., $\sum_{i=1}^n \mu(p_i) = B$, we just have to maximize $\sum_{i=1}^n \tau(\varepsilon_i) = \sum_{i=1}^n \tau(f(p_i))$, which by the assumption of additivity of τ , is equal to $\tau(\sum_{i=1}^n f(c, p_i))$.

Note that the definition of \mathcal{F} (cf. ((7.1))) ensures that all its functions are continuous on c . Since the sum of continuous functions is continuous, we have that also $\sum_{i=1}^n f(c, p_i)$ is continuous on c . Moreover, c ranges in a closed interval: $c \in [0, c_{\min}]$, where $c_{\min} = \min_{p \in \{p_1, \dots, p_n\}} \{c : f(c, p) = 0\}$. Thus, by *Extreme Value Theorem*, there exists a c which maximizes $\sum_{i=1}^n f(c, p_i)$, which, in turn, maximizes $\sum_{i=1}^n (\tau(\varepsilon_i) - \mu(p_i))$.

Furthermore, the condition of differentiability makes possible to apply the method of the lagrangians to find the maximum, by imposing that the partial derivatives are 0. The concavity condition implies that those points correspond to a (global) maximum. \square

THEOREM 7.7. There exists a c that gives an optimal **information-maximizing** function $f(c, \cdot) \in \mathcal{F}$, for a fixed budget, and we can derive such c via the method of the Lagrangians.

Proof. By the sequential compositionality of differential privacy, the total information obtained by the data consumer is $\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n f(c, p_i)$. The rest follows as the proof of Theorem 7.6. \square



Examples from Chapter 7

Example 7.3.3 Let $\mathcal{F} = \{\ln(e - cp) : c \in \mathbb{R}^+\}$. We want to maximize

$$F(c) = \sum_{i=1}^n f(c, p_i)$$

subject to the constraint

$$G(c) = \sum_{i=1}^n \mu(p_i) = \sum_{i=1}^n [p_i \ln(e - cp_i) + \int_{p_i}^{\frac{e-1}{c}} \ln(e - zp_i) dz] - B = 0,$$

for a fixed budget constant $B \in \mathbb{R}^+$. By the sequential compositionality of differential privacy, we have $\sum_{i=1}^n f(c, p_i) = \sum_{i=1}^n \ln(e - cp_i)$. Furthermore, $G(c) = \sum_{i=1}^n [p_i \ln(e - cp_i) + p_i \ln\left(\frac{e}{e - cp_i}\right) + \frac{e}{c} \ln(e - cp_i)] - K = 0$ where $K = B + \frac{n(e-1)}{c}$.

Using the method of Lagrange multipliers, we define the Lagrangian function $\mathcal{L}(c, \lambda) = F(c) - \lambda(G(c) - K) = \sum_{i=1}^n \ln(e - cp_i) - \lambda(\sum_{i=1}^n [p_i \ln(e - cp_i) + p_i \ln\left(\frac{e}{e - cp_i}\right) + \frac{e}{c} \ln(e - cp_i)] - K) = \sum_{i=1}^n \ln(e - cp_i) - \lambda(\sum_{i=1}^n [p_i + \frac{e}{c} \ln(e - cp_i)] - K)$. Thus we have

$$\mathcal{L}(c, \lambda) = \sum_{i=1}^n \ln(e - cp_i) - \lambda \sum_{i=1}^n p_i - \frac{\lambda e}{c} \sum_{i=1}^n \ln(e - cp_i) + \lambda K \quad (\text{H.1})$$

Hence, to find the optimal solution of c , we wish to solve the following:

$$\frac{\partial \mathcal{L}(c, \lambda)}{\partial \lambda} = 0 \quad (\text{H.2})$$

$$\frac{\partial \mathcal{L}(c, \lambda)}{\partial c} = 0 \quad (\text{H.3})$$

Solving equation (H.2), we get

$$\begin{aligned}
\frac{\partial \mathcal{L}(c, \lambda)}{\partial \lambda} &= 0 \\
\iff K - \sum_{i=1}^n [p_i + \frac{e}{c} \ln(e - cp_i)] &= 0 \\
\iff \sum_{i=1}^n [p_i + \frac{e}{c} \ln(e - cp_i)] &= K \\
\iff \ln \left(\prod_{i=1}^n e^{p_i} (e - cp_i)^{\frac{e}{c}} \right) &= K
\end{aligned} \tag{H.4}$$

Observe that:

$$\lim_{c \rightarrow 0} \ln \left(\prod_{i=1}^n e^{p_i} (e - cp_i)^{\frac{e}{c}} \right) = \infty \tag{H.5}$$

$$\lim_{c \rightarrow \infty} \ln \left(\prod_{i=1}^n e^{p_i} (e - cp_i)^{\frac{e}{c}} \right) = 0 \tag{H.6}$$

Combining (H.5), (H.6), the fact that $K = B + \frac{n(e-1)}{c} > 0$, and $\ln \left(\prod_{i=1}^n e^{p_i} (e - cp_i)^{\frac{e}{c}} \right)$ is continuous, by the Intermediate Value Theorem [206], we, thereby, can conclude that the curve $y = \ln \left(\prod_{i=1}^n e^{p_i} (e - cp_i)^{\frac{e}{c}} \right)$ intersects the curve $y = K$ at least once for $c \in (0, \infty)$, implying that we have at least one solution of (H.4) (\dagger_1).

Moreover, solving equation (H.3), we get

$$\begin{aligned}
\frac{\partial \mathcal{L}(c, \lambda)}{\partial c} &= 0 \\
\iff \frac{\partial (\sum_{i=1}^n \ln(e - cp_i)(1 - \frac{\lambda e}{c}) - \lambda \sum_{i=1}^n p_i + K)}{\partial c} &= 0 \\
\iff \frac{\partial (\sum_{i=1}^n \ln(e - cp_i)(1 - \frac{\lambda e}{c}))}{\partial c} &= 0 \\
\iff \sum_{i=1}^n \frac{-p_i}{e - cp_i} (1 - \frac{\lambda e}{c}) + \frac{\lambda e}{c^2} \sum_{i=1}^n \ln(e - cp_i) &= 0
\end{aligned} \tag{H.7}$$

Equation (H.7) is linear in λ , implying for every given c , we will find a λ satisfying (H.7) (\dagger_2).

Therefore, combining (\dagger_1) and (\dagger_2), we conclude that there is at least one optimal choice of $f(\cdot, \cdot)$ in \mathcal{F} that maximizes the information gathered by the data consumer, subject to the fixed budget.

Example 7.3.4 Let $\mathcal{F} = \{1 - cp : c \in \mathbb{R}^+\}$. We want to maximize

$$F(c) = \sum_{i=1}^n f(p_i) = \sum_{i=1}^n (1 - cp_i)$$

subject to

$$G(c) = \sum_{i=1}^n \mu(p_i) = \sum_{i=1}^n [p_i(1 - cp_i) + \int_{p_i}^{\frac{1}{c}} (1 - cz) dz] = B,$$

for a fixed budget $B \in \mathbb{R}^+$.

Observe that $G(c) = \frac{n}{2c} - \frac{c}{2} \sum_{i=1}^n p_i^2$. Using the method of Lagrange multipliers, we define the Lagrangian function $\mathcal{L}(c, \lambda) = F(c) - \lambda(G(c) - B) = -c \sum_{i=1}^n p_i - \lambda(\frac{n}{2c} - \frac{c}{2} \sum_{i=1}^n p_i^2 - B)$. Thus we have

$$\mathcal{L}(c, \lambda) = -c \sum_{i=1}^n p_i - \lambda(\frac{n}{2c} - \frac{c}{2} \sum_{i=1}^n p_i^2 - B) \quad (\text{H.8})$$

Hence, to find the optimal solution of c , we solve the following:

$$\frac{\partial \mathcal{L}(c, \lambda)}{\partial \lambda} = 0 \quad (\text{H.9})$$

$$\frac{\partial \mathcal{L}(c, \lambda)}{\partial c} = 0 \quad (\text{H.10})$$

Solving equation (H.9), we get

$$\begin{aligned} \frac{\partial \mathcal{L}(c, \lambda)}{\partial \lambda} &= 0 \\ \iff -(\frac{n}{2c} - \frac{c}{2} \sum_{i=1}^n p_i^2 - B) &= 0 \\ \iff \frac{n}{2c} - \frac{c}{2} \sum_{i=1}^n p_i^2 &= B \\ \iff c^2 \sum_{i=1}^n p_i^2 + 2Bc - n &= 0 \end{aligned} \quad (\text{H.11})$$

Note that the LHS of equation (H.11) is a quadratic equation in c , and as $4B^2 + 4n \sum_{i=1}^n p_i^2 > 0$, equation (H.11) has two distinct roots, or candidates for choosing the optimal c (\dagger'_1).

Moreover, solving equation (H.10), we get

$$\begin{aligned} \frac{\partial \mathcal{L}(c, \lambda)}{\partial c} &= 0 \\ \iff -\sum_{i=1}^n p_i + \frac{\lambda n}{2c^2} + \frac{\lambda}{2} \sum_{i=1}^n p_i^2 &= 0 \end{aligned} \quad (\text{H.12})$$

Note that for every given c , we will find a λ satisfying (H.12) (\dagger'_2).

Therefore, combining (\dagger'_1) and (\dagger'_2), we conclude that there is at least one optimal choice of $f(\cdot, \cdot)$ in \mathcal{F}_1 that maximizes the information gathered by the data consumer, subject to the fixed budget.



Proofs from Chapter 8

THEOREM 8.1. If the privacy valuation function used in order to impose the *penalty scheme* to any member p of a federation F is $f(m) = K_1(e^{K_2 m} - 1)$, the Shapley valuation function ψ chosen must satisfy

$$\frac{\ln\left(\frac{\varepsilon_p^T}{K_1} + 1\right)}{K_2} < \psi\left(\varepsilon_p^T, \frac{\ln\left(\frac{w^* \varepsilon_p^T}{K_1} + K\right)}{K_2}\right),$$

where $K = \frac{1}{K_1} \sum_{\substack{p' \neq p \\ p' \in F}} d(m)_{p'} \varepsilon_{p'}^T + 1$ and w^* is the scaling parameter as given by (8.2).

Proof. Using the privacy valuation function $f(m) = K_1(e^{K_2 m} - 1)$, we have $f^{-1}(\varepsilon) = \frac{\ln\left(\frac{\varepsilon}{K_1} + 1\right)}{K_2}$. Let p be an arbitrary member of F with a maximum privacy threshold ε_p^T . Therefore, in order to impose a penalty scheme on p , it needs to be ensured that

$$\begin{aligned} \frac{\ln\left(\frac{\varepsilon_p^T}{K_1} + 1\right)}{K_2} < \psi(v, M) &\Rightarrow \frac{\ln\left(\frac{\varepsilon_p^T}{K_1} + 1\right)}{K_2} < \psi(v, f^{-1}(\varepsilon_F^P)) \\ [\because w^* \in [0, 1] \text{ is the scaling parameter chosen by } D \text{ and } \varepsilon_F^P = w^* \varepsilon_F^T] \\ \Rightarrow \frac{\ln\left(\frac{\varepsilon_p^T}{K_1} + 1\right)}{K_2} < \psi\left(v, \frac{\ln\left(\frac{\varepsilon_F^P}{K_1} + 1\right)}{K_2}\right) &\Rightarrow \frac{\ln\left(\frac{\varepsilon_p^T}{K_1} + 1\right)}{K_2} < \psi\left(v, \frac{\ln\left(\frac{C_0 + w^* \varepsilon_p^T}{K_1} + 1\right)}{K_2}\right) \end{aligned}$$

$$\begin{aligned}
& \left[\text{where } C_0 = \sum_{p' \neq p \in F} d'_p \varepsilon_{p'}^T \text{ is a constant} \right] \\
& \Rightarrow \frac{\ln\left(\frac{\varepsilon_p^T}{K_1} + 1\right)}{K_2} < \psi\left(v, \frac{\ln\left(\frac{C_0 + w^* \varepsilon_p^T}{K_1} + 1\right)}{K_2}\right) \Rightarrow \frac{\ln\left(\frac{\varepsilon_p^T}{K_1} + 1\right)}{K_2} < \psi\left(v, \frac{\ln\left(\frac{w^* \varepsilon_p^T}{K_1} + K\right)}{K_2}\right) \quad (\text{I.1}) \\
& \left[\text{where } K = \frac{C_0}{K_1} + 1 \right]
\end{aligned}$$

□

Titre : Comprendre et optimiser le compromis entre vie privée et utilité d'un point de vue fondamental

Mots clés : confidentialité différentielle, confidentialité de localisation, apprentissage automatique, optimisation de la confidentialité et de l'utilité

Résumé : Avec les récentes avancées technologiques, les menaces de violation de la vie privée concernant les données personnelles des individus augmentent plus que jamais. Si la protection de la confidentialité des informations sensibles devient plus importante que jamais, il est également crucial de préserver l'utilité des données dans la société contemporaine basée sur l'information. Differential Privacy (DP) est considérée comme l'étalon-or des garanties formelles de protection de la vie privée. Son applicabilité étendue et sa technique de mise en œuvre simple ont conduit à une croissance rapide de sa popularité et de l'intérêt pour l'étude et l'application de la confidentialité différentielle à une variété de domaines, tant dans les universités que dans l'industrie. Au fil du temps, la communauté a exploré diverses variantes de la DP pour répondre aux préoccupations en matière de protection de la vie privée dans différents contextes et dans le cadre de divers modèles de menace.

Malgré l'acceptation prolifique du DP, l'interprétation

de son interaction avec l'utilité des données reste nébuleuse, ce qui accroît le besoin de réponses à des questions rudimentaires telles que la manière dont l'ajout de bruit DP affecte l'utilité des données partagées (par exemple, la qualité de service des propriétaires de données, l'utilité statistique des fournisseurs de services, la précision de l'analyse et de la formation de modèles effectuées, etc.) et s'il existe un mécanisme DP optimal en ce qui concerne l'utilité des données dans différents domaines et contextes. L'objectif de cette thèse était de répondre à ces questions et, en particulier, d'établir une base théorique pour analyser de manière exhaustive le compromis entre la protection de la vie privée et l'utilité des données sensibles selon diverses perspectives et dans le contexte de différents cas d'utilisation. Outre la dissection de la bataille séculaire entre la vie privée et l'utilité, cette thèse développe également des mécanismes de préservation de la vie privée afin d'optimiser la perte d'utilité avec des garanties formelles de respect de la vie privée dans divers domaines d'applicabilité.

Title : Understanding and optimizing the trade-off between privacy and utility from a foundational perspective

Keywords : differential privacy, location privacy, machine learning, privacy and utility optimisation

Abstract : With recent advancements in technology, the threats of privacy violations of individuals' personal data are surging like never before. While protecting the privacy of sensitive information is becoming more important than ever before, it is also crucial to uphold the utility of the data in the contemporary information-based society. Differential privacy (DP) is considered to be the gold standard of formal privacy guarantees. Its widespread applicability and uncomplicated implementation technique have led to a rapid growth in its popularity and interest in studying and applying DP to a variety of domains in academia and industry alike. Over time, the community has explored various variants of DP addressing privacy concerns in different contexts and under a variety of threat models.

Despite the prolific acceptance of DP, it is still nebulous to interpret how it interacts with the utility of data,

escalating the need for answers to rudimentary questions like how adding DP noise affects the utility of the shared data (e.g., the quality of service of the data owners, the statistical utility of the service providers, the accuracy of the analysis and model training performed, etc.) and does there exist some optimal DP mechanism with respect to the usefulness of data in different realms and contexts. The objective of this thesis has been to address these questions and, in particular, establish a theoretical foundation to comprehensively analyze the trade-off between privacy and the utility of sensitive data from a variety of perspectives and in the context of different use cases. Aside from dissecting the age-old battle between privacy and utility, this thesis also develops privacy-preserving mechanisms to proceed in the direction of optimizing utility loss with formal privacy guarantees in diverse domains of applicability.